EAS Publications Series, Volume 59, 2013

New Concepts in Imaging: Optical and Statistical Models

Nice, France, June 18, 2012
Fréjus, France, June 20–22, 2012

Edited by: *D. Mary, C. Theys and C. Aime*

17 avenue du Hoggar, PA de Courtabœuf, B.P. 112, 91944 Les Ulis cedex A, France

Sponsors

Centre National de la Recherche Scientifique (École thématique)
Université de Nice Sophia Antipolis
Laboratoire Lagrange, UMR 7293
Collège de France

Local Committee

Céline Theys (UMR 7293, Université de Nice Sophia Antipolis)
David Mary (UMR 7293, Université de Nice Sophia Antipolis)
Claude Aime (UMR 7293, Université de Nice Sophia Antipolis)
Antoine Labeyrie (Collège de France)
Farrokh Vakili (Observatoire de la Côte d'Azur)
Thierry Lanz (UMR 7293, Université de Nice Sophia Antipolis)
Caroline Daire (UMR 7293, Université de Nice Sophia Antipolis)
Catherine Blanc (UMR 7293, Université de Nice Sophia Antipolis)
Rose Pinto (UMR 7293, Université de Nice Sophia Antipolis)
Isabelle Giaume (Observatoire de la Côte d'Azur)

Cover Figure

Seaside at morning.

The cover image is a watercolor evoking the seaside of "Côte d'Azur", near Fréjus, the place where the school took place, in June 2012. The picture shows typical red rocks of the Esterel mounts, with "Cap Drammont" and "Ile d'or", in the distance. The watercolor is by Yves Rabbia, astronomer at OCA and sometimes amateur painter.

Indexed in: ADS, Current Contents Proceedings – Engineering & Physical Sciences, ISTP®/ISI Proceedings, ISTP/ISI CDROM Proceedings.

ISBN 978-2-7598-0958-5 EDP Sciences Les Ulis
ISSN 1633-4760
e-ISSN 1638-1963

Printed in UK


Photometry with IRDIS instrument
Divergences de Csisz
Toujours convexes

List of Participants

Aime Claude, Laboratoire Lagrange, claude.aime@unice.fr
Anterrieu Eric, LATT, Anterrieu@irap.omp.eu
Aristidi Eric, Laboratoire Lagrange, aristidi@unice.fr

Bertero Mario, Dipartimento di Informatica e Scienze dell' Informazione (DISI), bertero@disi.unige.it
Bijaoui Albert, Laboratoire Lagrange, Albert.Bijaoui@oca.eu
Boccacci Patrizia, Dipartimento di Informatica e Scienze dell' Informazione (DISI), boccacci@disi.unige.it
Bonnefois Aurélie, ONERA, aurelie.bonnefois@onera.fr
Bouquillon Sébastien, Observatoire de Paris-Meudon, sebastien.bouquillon@obspm.fr
Bourguignon Sébastien, IRCCYN, Sebastien.Bourguignon@irccyn.ec-nantes.fr
Bremer Michael, IRAM, bremer@iram.fr
Brie David, CRAN, david.brie@cran.uhp-nancy.fr

Carbillet Marcel, Laboratoire Lagrange, marcel.carbillet@unice.fr
Carfantan Hervé, IRAP, Herve.Carfantan@irap.omp.eu
Chen Jie, Laboratoire Lagrange, chen@unice.fr
Cruzalebes Pierre, OCA, pierre.cruzalebes@oca.eu

Dabbech Arwa, OCA, dabbech@oca.eu
Denis Loic, Laboratoire Hubert Curien, loic.denis@univ-st-etienne.fr
Dobigeon Nicolas, Laboratoire IRIT, nicolas.dobigeon@enseeiht.fr

Ferrari André, Laboratoire Lagrange, ferrari@unice.fr
Folcher Jean-Pierre, Laboratoire Lagrange, Jean-Pierre.folcher@unice.fr

Hadjara Massinissa, OCA, hadjara@oca.eu

Kluska Jacques, IPAG, jacques.kluska@obs.ujf-grenoble.fr

Labeyrie Antoine, Collège de France, antoine.labeyrie@oca.eu
Lantéri Henri, Laboratoire Lagrange, Henri.Lanteri@unice.fr
Liyong Liu, Laboratoire Lagrange, liuly@nao.cas.cn

Mary David, Laboratoire Lagrange, david.mary@unice.fr
Mourard Denis, OCA, denis.mourard@oca.eu

Nguyen Hoang Nguyen, Laboratoire Lagrange, nguyenh@oca.eu

Paris Silvia, Laboratoire Lagrange, silvia.paris@unice.fr

Rabbia Yves, Laboratoire Lagrange, rabbia@oca.eu
Raja Suleiman Raja Fazliza, Laboratoire Lagrange, raja.fazliza@unice.fr

Réfrégier Philippe, École centrale Marseille, philippe.refregier@fresnel.fr
Richard Cédric, Laboratoire Lagrange, cedric.richard@unice.fr
Roche Muriel, École centrale Marseille, muriel.roche@fresnel.fr

Schmider François-Xavier, Laboratoire Lagrange, schmider@oca.eu
Soldo Yan, laboratoire CESBIO, yan.soldo@cesbio.cnes.fr
Soulez Ferréol, CRAL, ferreol.soulez@univ-lyon1.fr

Theys Céline, Laboratoire Lagrange, celine.theys@unice.fr
Thiébaut Éric, CRAL, thiebaut@obs.univ-lyon1.fr

Zurlo Alice, LAM, alice.zurlo@oamp.fr

Contents

Statistical Models in Signal and Image Processing

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 1–2*

Foreword

This book is a collection of 19 articles which reflect the courses given at the Collège de France/Summer school "Reconstruction d'images – Applications astrophysiques" held in Nice and Fréjus, France, from June 18 to 22, 2012.

The articles presented in this volume share a common point: they all address emerging concepts and methods that are useful in the complex process of improving our knowledge of the celestial objects, including Earth.

In the spirit of a school, several articles underline using historical elements how essential have been instruments of high angular resolution, mathematical description of the observations, transmission of knowledge and reliance on long term research projects to our current representation of the Universe. Many articles can be read as tutorials of the specific research field they address.

The book contains three parts.

The first part is titled **"Physical bases and new challenges in high resolution imaging"**. In these articles, the strategy followed for attacking such challenges relies on a careful description of the electromagnetic waves emitted by the celestial sources, and of their perturbations. This part draws a picture of some of the high angular resolution instruments of the near to far future, and of the issues to overcome to make this picture real. It deals with hypertelescopes, optical interferometry, adaptive optics, wavefront coding, and with polychromatic astrophysical models.

The point of view of the articles of the second part, titled **"Physical models and data processing"**, is twofold. Of concern to these articles are not only the data description using physical modeling of electromagnetic waves, but also the resulting data processing. These articles address issues such as sampling, information modeling and restoration in radio and optical interferometry, including hypertelescopes.

The third part is titled **"Statistical models in signal and image processing"**. These contributions cover past and recent developments in multiresolution analysis, Bayesian modeling, sparsity and convex optimization. The last three papers deal specifically with hyperspectral data of Earth and of the deep Universe, images recorded at hundreds of wavelengths resulting in massive data. This last part illustrates the benefits brought by a careful data processing, and comes perhaps in contrast to the conventional wisdom which claims that too much information kills information.

DOI: 10.1051/eas/1359000

While the volume is divided in three parts to clarify which topics are covered and where, the issues addressed in the first and third parts are in reality connected by the observation instruments. These connections are mentioned at various places, and especially in the articles of the second part, which is a hyphen. As it goes, the alert reader will notice many more such connections.

Obviously, the successful realization of more powerful observation technologies and the best extraction of the astrophysical information encapsulated in their data involve the joint expertise of several research communities. The various articles collected in this book may contribute to such a synergy.

Physical Bases and New Challenges in High Resolution Imaging

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 5–23*

HYPERTELESCOPES: THE CHALLENGE OF DIRECT IMAGING AT HIGH RESOLUTION

A. Labeyrie[1]

Abstract. Sparse optical interferometric arrays of many apertures can produce direct images in the densified-pupil mode, also called "hypertelescope" mode. Pending the introduction of adaptive optics for cophasing, indirect images can also be reconstructed with speckle imaging techniques. But adaptive phasing is preferable, when a sufficiently bright guide star is available. Several wave sensing techniques, byproducts of those used on monolithic telescopes for some of them, are potentially usable. For cophased direct images of very faint sources in the absence of a natural guide star, a modified form of the Laser Guide Star techniques demonstrated on conventional and segmented telescopes is described. Preliminary testing in laboratory suggests further investigation. Recorded images, assumed co-phased, are also improvable post-detection with optical aperture-synthesis techniques such as Earth rotation synthesis, where data from successive exposures are combined incoherently. Nevertheless, the gain becomes modest if hundreds of sub-apertures are used. Image deconvolution techniques are also applicable, if suitably modified as demonstrated by Aime *et al.* (2012), and Mary (2012). Their modified deconvolution algorithms can extend the Direct Imaging Field (also called Clean Field) of hypertelescopes. More sub-apertures at given collecting area, implying that their size is reduced, improve the direct-imaging performance. The predictable trend thus favors systems combining hundreds of sub-apertures of modest size, if workable designs can be evolved. One such design, the "Ubaye Hypertelescope" entering the initial testing phase in the southern Alps, has a fixed spherical meta-mirror with a 57 m effective aperture, expandable to 200 m. Preliminary results suggest that larger versions, whether spherical or active paraboloidal, can reach a kilometric aperture size at terrestrial sites having a suitable concave topography. In space, hypertelescope meta-apertures spanning up to 100 000 km are in principle feasible in the form of a flotilla of mirrors, driven by micro-thrusters or by the radiation pressure of the Sun or lasers.

[1] Collège de France and Observatoire de la Côte d'Azur, 06460 Caussols, France

DOI: 10.1051/eas/1359001

1 Introduction

In the way of improved astronomical observation, the time has come for many-aperture optical interferometers providing direct high-resolution images. Most efficient perhaps will be the "hypertelescope" approach, using a sparse array of many sub-apertures and a densified pupil (Labeyrie 1996; Lardiere *et al.* 2007; Aime *et al.* 2012). Various opto-mechanical architectures and design concepts have been considered for versions on Earth and in space. Following the testing of a 3-mirror prototype with 9 m spacings at Haute-Provence (Le Coroller 2012), a larger instrument installed in the Ubaye range of the southern Alps has reached the testing stage (Labeyrie 2012). I describe some of the challenges raised by the direct and indirect production of hypertelescopic images, their phasing with adaptive optics and the observability of faint sources with a modified Laser Guide Star.

2 The basic optics of hypertelescopes

Among the various forms of optical interferometry which are considered for enhancing the resolution of astronomical observations, there had been some debate on the respective merits of pupil-plane *vs.* image plane arrangements. And on uni-axial *vs.* multi-axial devices. The hypertelescope concept involves a multi-axial beam combiner and densifier which can work either in the pupil or image planes, as well as through optical fibers where the notions of pupil and image vanish, as illustrated by Mourard *et al.* (2012) and Tarmoul *et al.* (2009). Pupil densification was already used by pioneer A.A. Michelson in his 20 feet interferometer at Mt Wilson, and the resulting intensification of the two-aperture fringe pattern on his retina, about 1600x at the full baseline setting, probably contributed to his observing success.

Since described in Labeyrie 1996, the hypertelescope principle has been discussed by different authors (Tarmoul 2009; Lardiere *et al.* 2007; Bouyeron *et al.* 2012; Patru *et al.* 2011; Aime *et al.* 2012). Various optical architectures can be adopted, one of which is a N-aperture Fizeau interferometer equipped with a pupil densifier, typically a small or even micro-optical accessory which can fit near the focal camera. Its effect is to shrink the diffractive envelope of the combined image and thus concentrate light into the central part of the interference function, thereby intensifying the image without affecting its pixel sampling. This is achieved at the expanse of the Direct Imaging Field of view, also called "Clean Field", which becomes shrunk down to λ/s, in terms of angular sky coverage, if s is the minimal spacing of the sub-apertures. However, it now appears that suitably modified deconvolution techniques can retrieve in part the missing field (Mary, this volume).

At given collecting area, many small sub-pupils improve the imaging performance, with respect to fewer larger ones, both for direct imaging and for image reconstruction by aperture synthesis with a varying or rotating aperture pattern. The simple way of producing direct images which uses a Fizeau-type beam

combiner, without pupil densification, does not exploit efficiently the exposures since much light is diffracted outside of the central interference peak in the spread function. The recorded pattern can be intensified by densifying the pupil, and considerably so if the aperture is highly diluted. In the absence of adaptive phasing, and in the use of "speckle interferometry" or "speckle imaging" for reconstructing the image as discussed in Section 2.2, such densification improves the signal/noise ratio and therefore the limiting magnitude. With adaptive phasing, generating a dominant peak in the N-aperture interference function, pupil densification is obviously beneficial for thresholded detectors, such as Michelson's retina and some infra-red cameras, since it can bring the level of interference peaks above the threshold.

Among the possible opto-mechanical design concepts for hypertelescopes, there are: a) arrays of telescopes having coudé foci, whether mirror-based or fibered, with optical delay lines feeding a beam-combiner; b) simplified designs resembling a giant sparse telescope, *i.e.* similar to the Arecibo radio-telescope although utilizing a sparse primary mirror, spherical and static, and not requiring delay lines since the moving focal combiner maintains the balance of all optical path lengths; c) Active versions of the latter ensuring a paraboloidal figure for the primary meta-mirror (Arnold *et al.* 1996). Terrestrial versions of types b and c, called Carlina, require a concave site for hosting the sparse mirror, and the absence of delay lines favors the use of numerous apertures, hundreds to thousands, having the potential of producing information-rich direct images. In space, versions as large as 100 000 km are expected to become feasible in the form of a mirror flotilla. Laser-trapped versions are proposed.

2.1 Imaging performance

At radio wavelengths, interferometric arrays of antennas have achieved increasing success by using tens, hundreds, and thousands of elements. More than 10 000 are now considered in some projects. A similar trend is also expected at optical wavelengths, following the science gains recently demonstrated with multi-telescope interferometers combining up to 6 beams (VLTI, CHARA...). These begin to produce reconstructed images, using aperture synthesis techniques analogous to those developed for radio interferometry. A basic difference, however, between the radio and optical versions of aperture synthesis currently utilized results from the incoherence of the optical data recorded sequentially at different times, with different aperture patterns. These optical exposures cannot be combined coherently for image reconstruction, since the phase distribution is missing. The incoherent combination gives a reconstructed spread function which is distorted and has a degraded dynamic range, compared to that provided by more sub-apertures, providing the same collecting area but simultaneously exploited for direct imaging. At radio wavelengths instead, the heterodyne detection can provide exposures shorter than the coherence time, and their complex amplitudes mapped at different times, with different aperture patterns, can be added for a coherent form of aperture synthesis providing a true synthetized aperture and the corresponding image.

Conceivably, exposures as short as a pico- or femtosecond may become feasible at the combined focus of an optical interferometer, with an heterodyne beat to visualize the phase and thus obtain the complex amplitude map (Riaud 2012). Indeed, any incoherent source becomes spatially coherent if suitably filtered and observed with a sufficiently short exposure, shorter than the coherence time defined by the filter's bandwidth. Changing the aperture pattern and repeating the procedure would then provide a summed map of complex amplitudes, identical to that provided by the global aperture thus synthesized. And the intensity map which can be calculated is the object's image, convolved with the intensity spread function.

Although typically achieved at radio to sub-mm wavelengths, this has not yet been possible at wavelengths shorter than 10 microns. A reason is the very small number of photons per spatio-temporal field mode, at visible and near-infrared wavelengths, from usual astrophysical sources. If the collected starlight can be dispersed into a large number of narrow wavelength channels, each receiving a properly tuned heterodyne "clock" beam, and equipped with an ultra-fast detector, it could be attempted to analyze the data statistically to overcome the low photon count per mode. Whether this would theoretically improve the degraded form of aperture-synthesis process heretofore used at optical wavelengths remains to be explored. The practical implementation appears difficult.

It is therefore of interest to build efficient forms of many-aperture interferometers, forming a sparse meta-aperture much larger than feasible with a monolithic mirror or an "Extremely Large Telescope" (ELT) mosac mirror, the size of which is limited by the pointable supporting mount. On Earth, the size of such meta-apertures may likely reach 1000 or 1200 m if built somewhat like the Arecibo radio-telescope within a natural depression. Larger versions, spanning perhaps 10 km, can also be considered, but in the absence of sufficiently wide and deep depressions, long delay lines would be needed and their high cost may constrain the sub-aperture count. Also, the implementation of laser guide star systems, needed for cophased direct imaging on faint sources, may be more difficult with such delay lines and complex coudé trains.

In space, the technical challenge is very different, and baselines spanning 100 000 km appear feasible at some stage (Labeyrie 2009). First-generation proposals for hypertelescope flotillas of mirrors have been submitted to the space agencies NASA and ESA (Labeyrie 2009). A low-cost version involving thousands of inch-sized mirrors, accurately controlled by a pair of laser beams forming a "standing wave trap", has also been conceived and subjected to laboratory tests in high vacuum (Labeyrie *et al.* 2010).

2.2 Pending phasing: The speckle imaging mode of hypertelescopes

Pending adaptive phasing, early science can be performed with a hypertelescope using speckle interferometry and speckle imaging with triple correlations (Lohmann *et al.* 1983; Tcherniavski *et al.* 2011). This has been simulated for the hypertelescope case by Surya *et al.* (2012) who obtained encouraging results. In monolithic

telescopes, higher limiting magnitudes, beyond $m_v = 18$, have been achieved with speckle interferometry than with adaptive optics using a natural guide star, the magnitude of which rarely exceeds $m_v = 13$. Pierre Riaud (private communication) suggests that the limiting magnitudes are in fact the same in both modes, a point which deserves verification. If speckle imaging proves more sensitive in hypertelescopes, it may remain of interest, even after adaptive cophasing becomes installed, on rather faint and simple objects not located within the isoplanatic patch of a brighter star. But Laser Guide Star systems, if they can be implemented on hypertelescopes for cophasing on faint sources (section below), should become preferable in all cases.

Does pupil densification improve the speckle interferometric signal/noise ratio and limiting magnitude? In monolithic telescopes, the photon-limited signal-to-noise ratio in speckle interferometry is classically known to be $S/N = 1/2$ $(1 + k)^{-1} N_{phs} N_p^{1/2} N_s^{1/2}$ (Dainty 1974), if N_{phs} is the number of star photons per speckle, k the number of photons from the sky background and N_p the number of recorded short exposures. $N_s = (D/r_0)^2$ is the number of speckles per exposure, if D is the aperture diameter and r_0 Fried's radius describing the size of phase cells. In a Fizeau interferometer having a meta-aperture of size D_m, containing N_a sub-apertures of size d, here assumed to be distributed non-redundantly and to match the size r_0 of seeing cells, the speckle count within the image's diffractive envelope is now $N_{sa} = (D_m/r_0)^2 = (D/d)^2 = N_s(D_m/D)^2$. The speckle count thus increases quadratically with the array size D_m. Densifying the pupil by a factor γ_d shrinks the speckle envelope in the same ratio, and therefore also decreases the number of speckles N_s as $1/\gamma_d^2$. Energy being conserved, the number of photons per speckle correspondingly increases, and eventually, at full densification, reaches that of a monolithic aperture having the same collecting area. According to Dainty's expression, pupil densification thus increases the signal/noise ratio, and matches that of a monolithic aperture having equivalent collecting area if also operated in the speckle interferometry mode. But it must be verified whether Dainty's expression remains applicable in the hypertelescope situation. In practice, pupil densification also relaxes the monochromaticity requirement, down to the monolithic value, thus also further enhancing the photon count per speckle N_{phs} and therefore the S/N ratio, unless perhaps if many narrow wavelength channels can be simultaneously exploited in parallel by a photon-counting camera. The effect of densification on the signal/noise ratio of "speckle imaging" reconstructions with triple correlations of recorded images also deserves to be explored.

2.3 Adaptive phasing

Adaptive phasing is highly desirable when a guide star, whether natural or artificial, is available near the observed source. Commercial deformable mirrors such as Boston Micromachines' MEMs with tip-tilt-piston facets appear suitable and may be installed at the exit of the pupil densifier. The usual types of wave sensor, such as the Shack-Hartmann or curvature sensor, serving in conventional telescopes, however, are not suitable since the measurements of local slope or curvature errors

in the wavefront assume its continuity to reconstruct it. Other methods which appear suitable are:

a) Hierarchical phasing (Pedretti & Labeyrie 1999);
b) A modified version of the Shack-Hartmann method, with triplets of adjacent sub-apertures feeding each lenslet, with overlap, to provide polychromatic interference honeycombs from which phase maps can be derived (Cuevas 2007);
c) The dispersed-speckle method (Borkowski & Labeyrie 2004; Martinache 2004), specifically developed for hypertelescopes;
d) The chromatic phase diversity method (Mourard *et al.* 2012);
e) The modified phase diversity method of Bouyeron *et al.* (2012) using a genetic algorithm.

Among these methods, b) is analogous to Shack-Hartmann and curvature sensing in the sense that it reconstructs the global map of piston errors from local slopes measured among clusters of adjacent sub- apertures. A difference, however, is that the local slope signal is derived from the position of polychromatic honeycomb-like interference patterns. The guide star should not be much resolved by the clusters of subapertures, but can be resolved by the global aperture. Methods a, c, d and e exploit interference speckles, which contain contributions from all baselines, short and long. They are therefore affected if the star is resolved by the latter.

2.4 Principle of a "Hypertelescope Laser Guide Star" (H-LGS) system

For observing faint sources, providing less than a few photons per seeing cell and spectral channel in exposures shorter than the turbulence lifetime, Laser Guide Stars have been successfully used with adaptive optics on monolithic or mosaic telescopes, and also provide the best hope of cophasing terrestrial hypertelescopes. If somewhat modified, as briefly suggested in Labeyrie (2008), the Laser Guide Star systems developed for monolithic apertures may also become suitable for the sparse apertures of hypertelescopes. As shown in Figure 1, Young's fringes can be projected by a sodium laser toward the sodium layer at 92 km altitude, using three or more apertures distributed like those of the hypertelescope, or actually belonging to its mirror array. Back-scattered light returning through the same apertures carries information on the cophasing errors.

The intensity spread functions of a multiple aperture for the up-going beam reaching the sodium layer and the down-going light back-scattered through the same multiple aperture are identical and similarly oriented if seen from below, while the down-going geometrical imaging of the spatially incoherent apparent Fizeau pattern in the sodium layer is inverted (Fig. 1). The double-pass pattern recorded by the camera is therefore a convolution of the single-pass Fizeau intensity pattern $I_h(x, y)$, projected at altitude h_s within the sodium layer, with an inverted copy of itself $I_h(-x, -y)$.

$$I(x, y) = I_h(x, y) \otimes I_h(-x-, y). \tag{2.1}$$

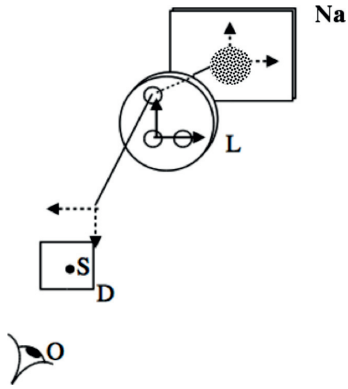


Fig. 1. Principle of H-LGS and coordinate symmetry. A sodium laser source S is projected by lens L, carrying a multi-aperture mask, toward the sodium layer Na, where interference fringes are formed within the diffraction envelope. The resonantly scattered light is made incoherent by the fast motion of sodium atoms, and part of it propagates back through the sub-apertures toward S, where a camera (not shown) records the double-pass interference pattern. Arrows indicate the coordinate symmetry, affected by the inversion of the lens image.

Phase information can be extracted from the recorded fringe exposure by calculating its two-dimensional Fourier transform, which is the product of the pupil's complex autocorrelation with the inverted copy of it, which is also its complex conjugate. The product is therefore a real function, and its modulus is the square of the pupil's single-pass autocorrelation function modulus:

$$I(u,v) = AC[P_h(u,v)]AC[P_h(-u,-v)] \quad (2.2)$$
$$= AC[P_h(u,v)]\overline{AC}[P_h(u,v)] \quad (2.3)$$
$$= |AC[P_h(u,v)]|^2 \quad (2.4)$$

where $P_h(u,v)$ describes the complex pupil carrying the single-pass phase error, and AC is the auto-correlation. If there are redundant baselines in the cluster of sub-apertures, each corresponding auto-correlation peak contains a sum of complex terms, and their modulus is consequently sensitive to any phase difference among these terms, thus allowing their measurement.

This method differs from those developed by Bonaccini (2004) and by Rabien *et al.* (2006) for reducing the LGS cone effect in a large telescope. The H-LGS principle is more related to the analysis of Mavroidis (1991), showing in a different context that retro-propagation through the same cluster of sub-apertures and seeing preserves phase information.

The typical fringe period in the sodium layer is of the order of 6 mm, within a diffractive envelope spanning 500 mm if the emitting sub-apertures are 120 mm wide, assumed smaller than Fried's r0 value. If observed with temporal resolution

longer than microseconds, the light resonantly backscattered by the sodium atoms appears to be spatially incoherent, owing to the Doppler spread induced by their fast Brownian motion. The pattern thus projected onto the sodium layer, fringed and serving as the artificial guide star, jitters in response to atmospheric phase shifts in the up-going beams. Nevertheless wave sensing in the returning beams remains possible with methods such as a) and b) mentioned in Section 2.3. Indeed, the incoherent fringed pattern, as seen in the sky through any cluster of adjacent sub-apertures in the hypertelescope array, gives a convolved image which retains appreciable fringe contrast, particularly if the aperture pattern is periodic or if the same cluster serves for the up- and down-going laser light. Also, the cluster should have more than three sub-apertures, so that at least one phase error can be calculated. The smallest type of cluster meeting these conditions is a quadruplet of sub-apertures, and they should preferably be adjacent to minimize anisoplanaticity effects.

If both wavelengths of the sodium doublet are emitted by the laser, and if the piston errors are pre-adjusted to be less than the doublet's coherence length, about 100 microns (instead of 100 mm for the Doppler-shifted back-scattered light from a single line), then piston errors can be derived and mapped for adaptive correction.

However, this is affected by the "cone effect", *i.e.* the propagation mismatch of light rays from the artificial star and the observed natural source, becoming particularly strong with a kilometric meta-aperture. To avoid it, each sub-aperture should capture back-scattered laser light which includes rays co-propagating with those from the observed source, thus crossing the same atmospheric turbules which affect their phase. The condition can be met by using in parallel many such artificial guide stars, arrayed within the sodium layer as a projection of the hypertelescope's aperture pattern toward the observed celestial source. If Fried's radius r_0 and the sub-aperture size d are such that $r_0 = d = 25$ cm, then the size of the projected Airy spots in the sodium layer matches that of the sub-apertures. Starlight crossing the sodium layer through these laser spots enters the corresponding sub-apertures, together with some of the back-scattered laser rays which follow the same path through the atmosphere and are thus affected by the same turbulence. In addition, since each laser-illuminated spot in the sodium layer must contain interference fringes, it must also receive coherent laser light from a few other sub-apertures.

This is achievable if their common laser source is a point source located some distance above the science camera, suspended in the focal plane of the meta-mirror, so that it be imaged by it at the sodium layer. As sketched in Figure 2, the arrangement is parallelized by installing an array of such laser sources, each illuminating a cluster of sub-apertures which projects a separate laser spot into the sodium layer.

For connecting the local phases thus obtained and deriving a global map of the starlight phases at all sub-apertures, the clusters have to contain at least four sub-apertures and to be partially overlapping. The map calculation has to assume that the atmosphere's isoplanatic angle i_0 is larger than the apparent spacing s/h_s

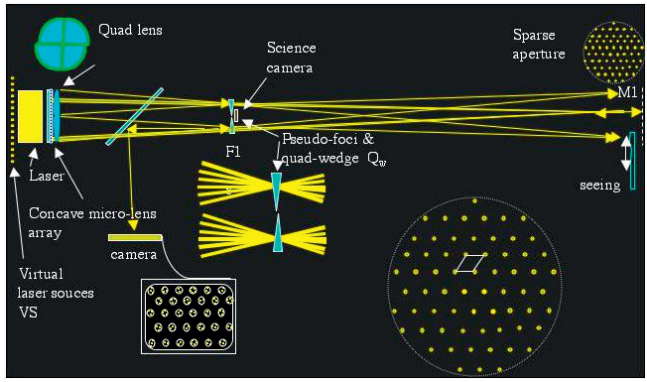


Fig. 2. Example of Hypertelescope Laser Guide Star system: segment M1i of the meta-mirror M1 is focusing onto the sodium layer (not shown, as it would be located far at left on a scale drawing) the image S'i of a point source Si of laser light, located slightly above F1, the focal image of the observed star. The exact transverse position of segment M1i is such that the reflected laser ray is directed toward the observed star, thus following in the opposite direction the propagation path of its ray reaching the segment. The same laser source Si also illuminates at least three additional M1 segments, which focus the corresponding beams at the same S'i position, thus generating interference fringes within the diffractive envelope in the sodium layer. The arrangement is parallelized with additional laser sources, projecting into the sodium layer an array of fringed spots. The array pattern is identical to M1's sub-aperture pattern, and the clusters of M1 elements focusing each S'i spot are partially overlapping so that the measured fringe phases be propagated for building a global map of phase errors affecting the observed source's image. The multiple laser source, where each laser illuminates a single cluster of M1 segments, is here implemented with a single laser illuminating a concave micro-lens array, next to a group of four or more adjacent lenses and facing wedges.

of the laser spots in the sodium layer, if s is the sub-aperture spacing, matching the laser spot spacing, and h_s the altitude of the sodium layer. The spacing matching the condition is then $s = h_s i_0$, typically amounting to 10 m with $20''$ isoplanatism, and then implying $N_a = 10\,000$ sub-apertures within a kilometric meta-aperture. With such large numbers, the needed laser power may become a practical challenge since a few tens of watts are needed per sub-aperture.

I made a simple laboratory experiment, sketched in Figure 3, to verify the phase sensing scheme with a single cluster of sub-apertures. A laser source is focused toward a rotating reflective diffuser, simulating the sodium layer, through a multiple aperture. The pattern of backscattered light returning through the same multiple aperture and lens is recorded by a camera, virtually located in the plane of the laser source, but separated by a beam splitter. Rotating the diffuser simulates the motion of sodium atoms in the atmospheric sodium layer, thus smoothing the speckled backscattered light reaching the aperture mask and causing the apparent

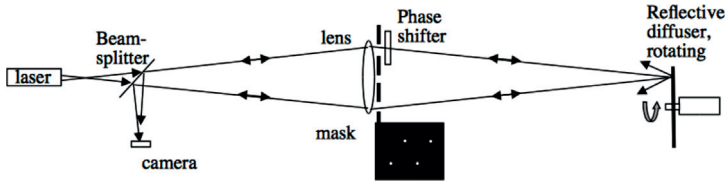


Fig. 3. Laboratory simulator of the H-LGS scheme. Laser light is focused toward a pinhole, itself relayed by a lens toward a reflective diffuser, which can be rotated to simulate the motion of sodium atoms in the atmospheric sodium layer. A multi-aperture mask inserted near the lens simulates an array of hypertelescope sub-apertures, and a cluster such as the shown rhombus can be selected with a second mask. A bumpy glass plate near the sub-apertures produces phase errors. The back-scattered light returning from the diffuser and through the same mask is recorded by a camera, through a beam-splitter.

fringe pattern on the rotating diffuser to become spatially incoherent, which reduces somewhat the fringe contrast in the recorded long exposures. Phase errors, introduced on sub-apertures by a distorted glass plate, are measured by Fourier analyzing the recorded fringe patterns.

Figure 4 shows the recorded exposures. Appreciable contrast is retained in the fringe patterns when the diffuser is rotated. The patterns are seen to be influenced by the phase error, as it is also apparent in their calculated Fourier transforms by comparing the intensities of the four median peaks, each involving a pair of redundant baselines.

2.5 Phase determination with overlapping clusters of four subapertures

The laboratory simulation has confirmed that the recorded fringe pattern is influenced by phase errors, using either a rhombus-shaped cluster of four adjacent sub-apertures among a triangular array, or a centered-hexagon cluster of 6+1 sub-apertures. Further testing in the laboratory is undertaken by Paul D. Nuñez for a more realistic simulation and hardware development, toward some real testing envisaged with the "Ubaye Hypertelescope" prototype and the development of a working system. Its optical scheme, providing from a single laser source the required multiplicity of laser beams, is sketched in (Fig. 2). The design can be downsized for laboratory simulations.

The laboratory simulation is complemented by numerical simulations and data analysis that will quantify the accuracy of the phase determination accross the array in the presence of non-uniform illumination in the sub-apertures, photon noise, etc. Paul D. Nuñez has used the formalism described in the preceeding section to perform numerical simulations of similar data to that presented in Figure 4. An example is presented for illustrative purposes in Figure 5. The top and bottom set of images differ from each other by a phase difference introduced in one of the four sub-apertures.

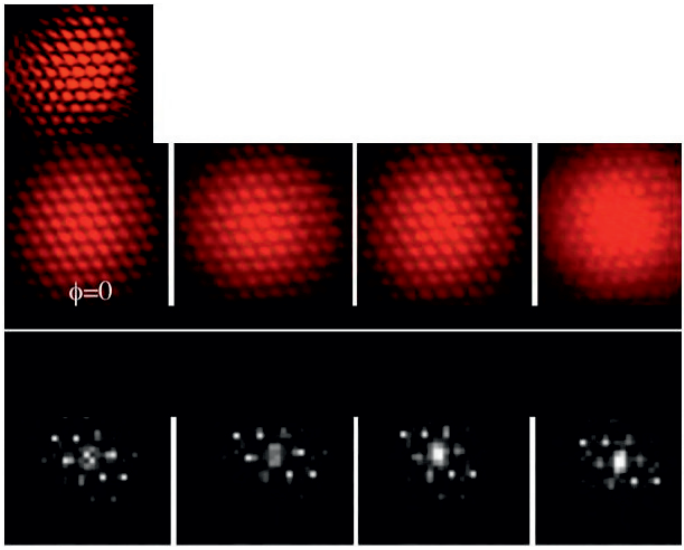


Fig. 4. Exposures obtained with the rhombus cluster of four sub-apertures and different phases errors. The top left exposure involves a static diffuser and no phase error. The middle row involves the rotating diffuser and shows the effect of various phase errors on the long-exposure pattern. The bottom row shows the corresponding Fourier transforms, where the median peaks generated by each redundant pair of baselines have their intensities influenced by the phase errors.

2.5.1 Deriving the phase errors

With a rhombus cluster of sub-apertures, three of the phase values are known or can be taken as zero if image motion is accepted, and the fourth has the unknown value ϕ_4. In the complex autocorrelation of the pupil, the peak related to baselines 2-3 and 1-4, for example, has a value $1 + e^{i\phi_4}$, and its square modulus $I_{14} = (1 + e^{i\phi_4})(1 + e^{-i\phi_4}) = 2(1 + \cos(\phi_4))$ is the modulus of the corresponding peak observed in the calculated Fourier transform of the camera image. Its value varies from 0 to 4. The central peak's modulus is $I_0 = 4^2 = 16$. The unknown phase ϕ_4 is thus determined as: $\phi_4 = \pm \arccos(I_{14}/2 - 1) = \pm \arccos(8I_{14}/I_0 - 1)$.

The sign ambiguity is resolvable by trial and error, in one or two steps of piston correction, or by a phase diversity method, using a second camera exposure with a known amount of defocus, achieved by range-gating laser pulses to select back-scattered light from a different sub-layer of the sodium layer. Pierre Riaud suggests to record a second image separated with a beam splitter, and where a known phase shift is added. For increased sensitivity, many such layers can be simultaneously observed in separate temporal channels. More than four apertures can presumably be used in a cluster, although the phase extraction is more elaborate with more redundancy, *i.e.* if more than two baselines contribute to each Fourier peak, as it is the case for a centered hexagon.

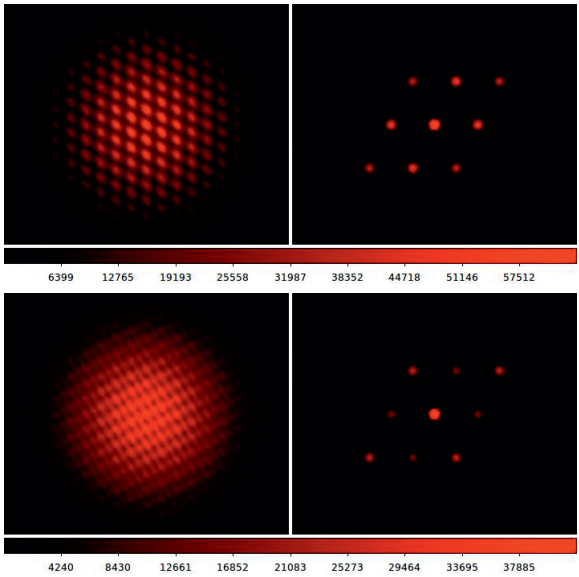


Fig. 5. Numerical simulation, made by Paul D. Nuñez, of fringes (*left*) and their Fourier transform (*right*, and in square root scale for better visibility) as would be obtained with a laser guide star. These correspond to four sub-apertures arranged in a rhombus pattern. The two sets of images differ by introducing a phase delay of $PI/2$ in one of the sub-apertures.

2.6 Deconvolution of phased images

Image deconvolution has been fruitfully used in different situations of astronomical imaging, particularly with radio-interferometers and other situations where the spread function has numerous and strong sidelobes, as is the case with interferometers containing few apertures, or when dealing with highly contrasted sources, such as a star and its planets. Aime *et al.* (2012) and Mary (this volume) have began exploring the case of hypertelescope images. Because of the pseudo-convolution which accounts for image formation in this case, they had to modify the established deconvolution methods and obtained encouraging results. An unexpected gain has been the retrieval of sources located outside of the "Direct Imaging Field", also called "Clean Field". Further work toward characterizing the potential of these methods will be beneficial for assessing the science program of large hypertelescopes (section below). Like in the case of speckle imaging (section above), the effect of the pupil densification factor on the signal/noise ratio, using detectors which are photon-limited or not, should also be analyzed.

Another aspect that can be studied with deconvolution is the dynamic range. The array configuration has an effect on the dynamic range of the direct image (Lardiére *et al.* 2007), and some configurations may be better than others in this respect. However, it is not known if this will remain true in the deconvolved images. This is one of the issues that will start to be addressed in an upcoming

paper that deconvolves simulated images of Betelgeuse (Patru *et al.* 2013 in prep.). Paul D. Nuñez, in collaboration with David Mary, Fabien Patru and Claude Aime are making a more systematic study of these questions.

3 Construction of the "Ubaye Hypertelescope" prototype

Following more than a decade of analysis and simulation by various authors, together with laboratory testing, the direct-imaging performance and sensitivity gain foreseen for hypertelescopes has prompted our group to build two prototype versions of a Carlina hypertelescope: 1) a first prototype at Haute-Provence observatory, utilizing a tethered balloon to carry the focal optics and camera 35 m above a triplet of mirrors spaced 9 m apart (LeCoroller *et al.* 2012); 2) and then the larger "Ubaye hypertelescope" prototype utilizing a suspending cable, stretched across a deep valley in the Ubaye mountains of the southern Alps to carry a focal gondola 100 m above a 57 m meta-aperture, expandable to 200 m with potentially 100 or more mirror elements (Labeyrie *et al.* 2012).

3.1 Opto-mechanical concept and design

Among the various possible architectures for Earth-based hypertelescopes, flat or concave, the Arecibo-like concept involves a large fixed and spherical concave meta-mirror focusing images toward one or more gondolas driven along the focal surface. It does not require delay lines and thus favors the use of numerous sub-apertures, in the form of mirror segments sparsely arrayed across the large "meta-mirror". For a given collecting area, a large number of small apertures are preferred to fewer large ones, in accordance with the theoretical results indicating that the direct-imaging performance improves markedly with the number of sub-apertures. Such opto-mechanical designs are called "Carlina", name of a large alpine thistle flower which is stem-less and contains hundreds of smaller flowers within its "meta-flower".

As previously described (Labeyrie *et al.* 2012), the optical train of the "Ubaye hypertelescope" has a fixed, sparse and spherical M1 primary mirror. It also has a suspended focal gondola, accurately driven by six oblique tethers to track the star's image. It contains a spherical M2 mirror at the best focus within the caustic surface, a strongly aspheric M3 mirror located in the pupil plane, and additional lenses for densifying the pupil. An optional flat coudé folding mirror addresses a small collecting telescope located in the North-facing slope at the polar projection of M1's curvature center C1.

Following the optical design study by André Rondi, further developed by Rijuparna Chakraborthy, the concept has been also analyzed by Enmark *et al.* (2011) who included a detailed model of the gondola's drive, including the effect of wind disturbances. A price to be paid for the optical simplification of the Carlina design, with respect to hypertelescopes using delay lines, is the pupil drift occurring within the gondola as it tracks the star, while keeping its optical axis aligned with it. This can be accommodated by moving an optical element within the gondola, whether passively with a pendulum drive such as adopted for the

Haute-Provence prototype, or actively with for example a motorized version of it. Another possibility is the use of a photo-activated pupil densifier driven by laser beams received from each M1 mirror. The latter putative type of integrated "pupil-tracking densifier" can use photo-active materials such as a thermoplastic/photoconductive film stack (Crandall *et al.* 1985) where the incident array of projected laser spots reversibly imprints micro-lenses, enabling thus to follow the drifting motion of the laser spot array. Variants with photo-refractive films may also be considered. Pupil densifier types using a single micro-lens array, as described in Pedretti *et al.* (1999), are particularly suited for incorporating such devices.

3.2 Experience gained with "Ubaye Hypertelescope"

As described in Labeyrie *et al.* (2012), the construction of the "Ubaye Hypertelescope" prototype since 2011 in the mountain valley at 2100–2300 m altitudes has raised unusual challenges, in comparison with those experienced when building interferometers of the previous generation. Both the optical and mechanical concept were simplified, but unexpected issues were met, such as the manipulation of long cables and their protection from avalanches and tree branches, the invasion by sheep flocks tangling their legs in wires, and the colonization of mirror supports by nesting wasps. Also, part of the site being within the Parc National du Mercantour, it has been important to protect birds from collisions with the cables (particularly the Black Grouse Lyurus tetrix and the large endangered Bearded Vulture Gypaetus barbatus). Some team members expressed fears of attacks by wolves which are present at the site.

For an optimal insertion of the M1 locus in the terrain topography, its mapping has been done with GPS/RTK techniques approaching decimetric accuracy by Loic Evrard and Marion Gaudon (Institut Geographique National), and pursued by Martine Roussel and Jerome Maillot, who also used a laser theodolite. Remi Prudhomme has designed and assembled the driving electronics and the extended wifi link between its elements, located together with the winches a few hundred meters apart. Denis Mourard has tested the system and written the high-level control code based on the spherical trigonometric model. He has been able to observe the focal gondola's stellar tracking motion with the coudé telescope and verify its accuracy, expected to reach 1mm during typical good observing nights where wind velocity is low.

We have installed a pair of mirrors on stiff tripods, 16 m aprart, for assessing the system with fringes on a bright star. This has not yet been achieved, but no major problem has been identified. We have developed techniques for the optical alignment, including the acquisition of the coudé beam toward the 20 cm collecting telescope and the gondola's drifting attitude which must be controlled in addition to its drifting position.

With the heavy snow and difficult accessibility in wintertime, it is expected that remote observing will become possible at a later stage, without any human attendance.

3.3 Spherical or paraboloidal meta-mirror?

The LOVLI concept for a Moon-based interferometric array proposed by Arnold *et al.* (1996) involved a sparse and active paraboloidal meta-mirror, the segments of which were controlled by actuators to keep the paraboloid axis pointed toward the source being tracked. Also, a larger variant of the Arecibo radio-telescope, the 500 m FAST radio-telescope becoming built in China, has an active paraboloidal mirror. Its shape is dynamically adjusted by actuators so as to remain paraboloidal while its axis direction tracks the observed source.

A paraboloidal design comparable to the LOVLI is now studied by Yves Bresson as a possible future upgrade for Ubaye Hypertelescope. His initial ray-tracing analysis with Zemax code indeed suggests that the spherical meta-mirror, once equipped with a few actuators on each segment, can be re-shaped into a paraboloid by computer commands in a matter of minutes and reversibly. The shifting also requires removing the focal corrector of spherical aberration needed in the spherical case.

The old controversy on the merits of paraboloidal and spherical telescopes, correctible by Schmidt plates or smaller elements near the focal plane, suggests that each mode has specific advantages and drawbacks. For example, the focal optics is simpler in the paraboloidal case, especially if there is no coma corrector. But, with a spherical M1, its static geometry and the multi-gondola option for simultaneous observations in widely separated fields may impact the cost and science output. Among the comparison elements deserving further study are: 1) the cost of active M1 segments, *vs.* that of spherical aberration correctors in multiple gondolas, and: 2) the field-of-view coverage achievable with coma correctors in the single focal gondola of a parabolodal meta-mirror, *vs.* that achievable in each gondola exploiting a spherical M1. In both cases, coma correction can be extended with local correctors within each field channel covering the diffraction lobe of the sub-aperture.

3.4 Science capabilities

In addition to stellar physics, with spatio-spectral direct imaging on resolved stars and their environment, a challenging related possibility is the production of transit images when an orbiting exo-planet crosses the star's apparent disk. Such displays should be reasonably contrasted if there are many apertures and if the planet is resolved or nearly so. They are potentially easier to observe than the presence of the same planets when not transiting. In the latter case, advanced coronagraphic techniques are needed to evidence the comparatively faint planet surface. In the limit case of a planet just entering transit, or emerging from it, bright refractive arcs are also likely observable, such as seen during the recent Venus transit across the Sun. This should provide valuable opportunities to probe spectroscopically the exoplanet's atmosphere thus sampled by the star's light at grazing incidence, as needed for searching bio-signature molecules (Leger *et al.* 2011).

But the large gain in limiting magnitude expected with hypertelescopes, especially when equipped with a Laser Guide Star, potentially brings a diversity of

galactic and extra-galactic sources within observing reach: Active Galactic Nuclei and their fast-orbiting central stars giving information on the mass of a central dark hole, jet structures of galaxies or QSOs, and gravitational lenses.

4 Feasibility of 1000–1200 m Carlina hypertelescopes on Earth

The preliminary experience gathered while building Ubaye Hypertelescope already suggests that much larger versions can be built in larger and deeper Andean or Himalayan valleys having a suitable topography. The feasibility and the design options for a 1200 m "Extremely Large Hypertelescope" (ELHyT) have been discussed by Labeyrie *et al.* (2012). An H-LGS system such as discussed in Section 2.4 above is essential for fully exploiting the science potential of an ELHyT toward the faintest limiting magnitudes for observing cosmological sources. These limiting magnitudes can match in principle those accessible to an ELT of similar collecting area, and similarly equipped with a Laser Guide Star (Boccaletti *et al.* 2000). Together with the 100 microarc-second resolution attainable at visible wavelengths, a 30x gain with respect to a 40 m ELT, this announces major science inroads toward cosmology on the faintest known galaxies at the edge of the observable Universe.

The prospect deserves a comparative study of ELT and ELHyT technologies in terms of science, readiness and cost efficiency, which the funding institutions should initiate as part of their decision-making process. In the ELHyT case, features which may impact the compared cost are: 1) the absence of a pointable mount; 2) absence of a dome; 3) the smaller mirror elements, preferable for improved imaging performance at given collecting area, allowing a reduction of the glass thickness; 4) the resulting use of lower-grade glass such as Pyrex, becoming adequate for the smaller mirrors; 5) the high-yield mirror figuring techniques then also becoming available; 6) the risk reduction achievable by deploying and testing a few mirror elements with a focal gondola and its tracking system; 7) the progressive construction and early science achievable, as demonstrated with the ALMA; 8) the availability of the numerical model of gondola control and dynamic behaviour, developed by Enmark *et al.* (2011); 9) the existence of an access road at some of the mountain sites considered, such as Spiti valley (India, at 4000 m altitude).

5 Hypertelescopes in space

Large flotillas of numerous small mirrors may become feasible in space, using micro-thrusters for control, as demonstrated by the PRISMA test of the Swedish National Space Board with a pair of agile micro- satellites, controlled with centimeter accuracy. A version controlled by solar radiation pressure has been proposed to the European Space Agency (Labeyrie *et al.* 2009), and mentioned in the U.S. Decadal Survey as suitable for observing faint extra-galactic sources such as AGNs (Kraemer *et al.* 2010). Another version using inch-sized mirrors, each accurately controlled by radiation pressure within standing waves from a pair of counter-propagating laser beams, is described by Labeyrie *et al.* (2010).

Such hypertelescope flotillas have been studied in preliminary detail for versions at the kilometer scale, the 100 km scale needed to resolve exo-planetary detail, and the 100 000 km scale needed to resolve details of the Crab Pulsar. The latter version appears workable in terms of orbit control at the Earth-Sun Lagrangian L2 point, as analyzed by Infeld (2006), but needs rather large mirror elements, typically 6 or 8 m, to efficiently collect their diffracted light at the combined focus. Such large mirrors may themselves be feasible in "laser-trapped" form (Labeyrie 1979) with a thin membrane, possibly made of diamond or graphene.

The prospect for such large dilute astronomical instruments emitting many laser beams raises the question of whether some advanced exo-civilizations, if they exist, may have built them before being capable of building Dyson spheres (2011). If so, the peculiar characteristics of their laser light emission, with periodically swept wavelength and potentially detectable with ELTs or hypertelescopes, may provide better signatures of extra-terrestrial intelligence than the infra-red excess heretofore searched by SETI programs for detecting Dyson spheres on stars.

6 Conclusions and future work

In the way of higher angular resolution at optical wavelengths, the potential of direct imaging with hypertelescopes, and induced sensitivity gain, has been confirmed by computer simulations, laboratory experiments and limited sky observations with miniature versions. The technical feasibility of large Earth-based versions, currently tested at Haute Provence and in the Ubaye range, becomes confirmed for hectometric aperture sizes, and kilometric sizes are now considered. Although images can be reconstructed with speckle interferometry techniques, the direct images obtainable with adaptive cophasing are a major goal.

The attainment of high limiting magnitudes appears feasible with modified forms of Laser Guide Star systems under development, thus greatly extending the scope of high angular resolution observing in astronomy. Also, image deconvolution techniques becoming developed for the hypertelescope case give encouraging results. These prospects suggest that a major breakthrough is possible in the way of high-resolution observing, extending to extra-galactic and cosmological sources. It appears to justify efforts toward building progressively larger instruments, especially considering the feasibility of expandable arrays not requiring the kind of major initial investment needed for the dome and mount of an ELT. In space, considerably larger apertures should become feasible in the form of mirror flotillas, eventually spanning up to perhaps 100 000 km. Such instruments would provide enough resolution to resolve the central body, believed to be a 20 km neutron star, of the Crab pulsar.

References

Aime, C., Lanteri, H., Diet, M., & Carlotti, A., 2012, A&A, 543A, 42A

Arnold, L., Labeyrie, A., Mourard, D., 1996, Adv. Space Res., 18, 49

Boccaletti, *et al.*, 2000, Icarus, 145, 636

Bonaccini Calia, D., Myers, R.M., Zappa, F., *et al.*, 2004, SPIE, 5490, 1315

Borkowski, V., & Labeyrie, A., 2004, EAS Publications Series, 12, 287

Buscher, D.F., Love, G.D., & Myers, R.M., 2002, Opt. Lett., 27, 149

Bouyeron, L., Delage, L., Grossard, L., & Reynaud, F., 2012, A&A, 545, A18

Chapa, O., Cuevas, S., Sánchez, B., *et al.*, 2007, Rev. Mex. Astron. Astrofis. Conf. Ser., 28, 82

Crandall, R.S., *et al.*, 1985, "Reversible optical storage medium and a method for recording information therein" US patent, `http://www.google.com/patents/US4320489`

Dainty, J.C., 1974, MNRAS, 169, 631

Dyson, F., 2011, see `http://en.wikipedia.org/wiki/Dyson_sphere`

Infeld, S.I., 2006, "Optimization of Mission Design for Constrained Libration Point Space Missions" Ph.D. Stanford, `http://www.stanford.edu/group/SOL/dissertations/samantha-thesis.pdf`

Kraemer, S., Windhorst, R., Carpenter, K.G., *et al.*, 2010, in "Astro2010: The Astronomy and Astrophysics Decadal Survey, Science White Papers, 162"

Le Coroller, H., Dejonghe, J., Arpesella, C., Vernet, D., & Labeyrie, A., 2004, A&A, 426, 721

Enmark, A., Andersen, T., Owner-Petersen, M., Chakraborty, R., & Labeyrie, A., 2011, Integrated model of the Carlina Telescope", in "Integrated Modeling of Complex Optomechanical Systems", ed. Andersen, Torben, Enmark & Anita, Proceedings of the SPIE, Vol. 8336, 83360J-83360J-14

Lardière, O., Martinache, F., & Patru, F., 2007, MNRAS, 375, 977

Labeyrie, A., 1996, A&A, 118, 517

Labeyrie, A., Le Coroller, H., & Dejonghe, J., 2008, SPIE, 7013

Labeyrie, A., 2008, Proceedings of the SPIE, Vol. 6986, 69860C-69860C-12

Labeyrie, A., *et al.*, 2009, Exper. Astron. 23, 463

Labeyrie, A., *et al.*, 2010, "Resolved Imaging of Extra-Solar Photosynthesis Patches with a "Laser Driven Hypertelescope Flotilla", in "Pathways Towards Habitable Planets", proceedings of a workshop held 14 to 18 September 2009 in Barcelona, Spain, ed., Vincent Coudé du Foresto, Dawn M. Gelino & Ignasi Ribas (San Francisco: Astronomical Society of the Pacific), 239

Labeyrie, A., *et al.*, 2012, Optical and Infrared Interferometry III. Proceedings of the SPIE, Vol. 8445, id. 844512-844512-9

Labeyrie, A., *et al.*, 2012, Optical and Infrared Interferometry III. Proceedings of the SPIE, Vol. 8445, id. 844511-844511-9

Labeyrie, A., 1979, A&A, 77, L1

Labeyrie, A., Guillon, M., & Fournier, J.M., 2012, "Optics of Laser Trapped Mirrors for large telescopes and hypertelescopes in space", SPIE conf.

Le Coroller, H., Dejonghe, J., Regal, X., *et al.*, 2012, A&A, 539, A59

Leger, *et al.*, 2011, Astrobiology, 11, 4

Lohmann, A.W., Weigelt, G., & Wirnitzer, B., 1983, Appl. Opt., 22, 4028

Martinache, F., 2004, J. Opt. A: Pure Appl. Opt., 6, 216

Martinache, F., 2012, A&A, 286, 365

Mary, *et al.*, 2013, EAS Publications Series, 59, 213

Mavroidis, T., Solomon, C.J., & Dainty, J.C., 1991, J. Opt. Soc. Am. A, 8, 1003

Mourard, D., *et al.*, 2012, Optical and Infrared Interferometry III. Proceedings of the SPIE, Vol. 8445, id. 84451M-84451M-10

Patru, F., Tarmoul, N., Mourard, D., & Lardière, O., 2009, MNRAS, 395, 2363

Patru, F., Chiavassa, A., Mourard, D., & Tarmoul, N., Direct imaging with a hypertelescope of red supergiant stellar surfaces [`eprint arXiv:1108.2320`]

Pedretti, E., & Labeyrie, A., 1999, A&AS, 137, 543

Rabien, S., Eisenhauer, F., Genzel, R., Davies, R. I., & Ott, T., 2006, A&A, 450, 1, 2006, 415

Riaud, P., 2012, Eur. Phys. J. D, 66, 8

Surya, A., 2012, in preparation

Tcherniavski, I., 2011, Optical Engineering, 50, 3

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 25–36*

OPTICAL LONG BASELINE INTERFEROMETRY: EXAMPLES FROM VEGA/CHARA

D. Mourard[1]

Abstract. In this paper I review some of the fundamental aspects of optical long baseline interferometry. I present the principles of image formation, the main difficulties and the ways that have been opened for high angular resolution imaging. I also review some of the recent aspects of the science program developed on the VEGA/CHARA interferometer.

1 Introduction

Astrophysics is based on observations and physical analysis. From the point of view of observations, this science has mainly been developed through the progresses in the techniques of imaging, astrometry, photometry, spectroscopy and polarimetry. However, through these techniques, objects are almost always considered as point-like source and no information is obtained on their brightness distribution. This is of course due to the diffraction principle, the limited size of the collecting optics used in telescopes and the very small apparent angular sizes of these objects.

In 1974, A. Labeyrie succeeded for the first time to obtain interference fringes on a stellar source with two separate telescopes. This achievement opened the road for the modern development of optical interferometry and allowed to give access to astrophysics at very high angular resolution. Today, the situation is dominated by a few facilities: mainly the VLTI (Glindemann *et al.* 2004), KECK (Colavita *et al.* 2006) and the CHARA array (Ten Brummelaar *et al.* 2005), allowing combination of 4 to 6 telescopes from the visible to the thermal infrared domain. With almost 50 scientific papers per year, the progression of the astrophysical impact of long baseline optical interferometry is almost following, with a time shift of 30 years, the development of radio interferometry.

The main scientific domains of modern optical long baseline interferometry are the study of brightness distribution of compact objects such as stellar surfaces,

[1] Laboratoire Lagrange, UMR 7293, UNS-CNRS-OCA, Boulevard de l'Observatoire, BP. 4229, 06304 Nice Cedex 4, France

DOI: 10.1051/eas/1359002

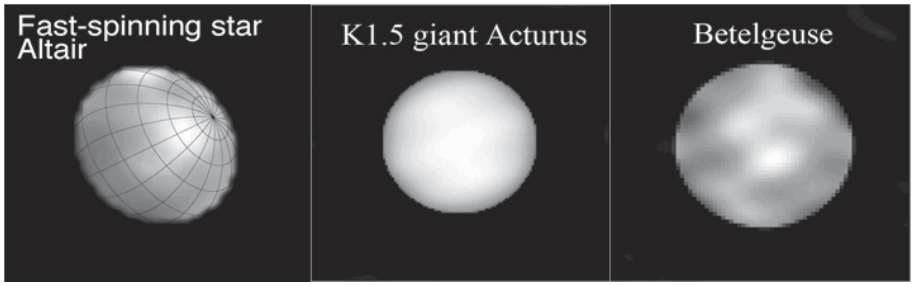


Fig. 1. Recent examples of stellar surface images obtained by optical long baseline interferometers. *Left*: Monnier *et al.* (2007). *Middle*: Lacour *et al.* (2008). *Right*: Haubois *et al.* (2008).

cores of young stellar object environments or central regions of active galactic nuclei. These studies require high resolution in the space, time and spectral domains for a correct understanding of the physical processes in action in these objects. As an imaging technique, optical long baseline interferometry performance is highly related to the properties of the field of view and of the transfer function. Recent advances by different groups in the world have lead to the first images of stellar surfaces (see Fig. 1).

Although these first images show remarkable progresses in that field, it is clear however that more technical work is needed to improve the impact of long baseline interferometry. The main issue is certainly the need for angular resolution that requires long baseline ($B > 100$ m) and short wavelengths ($\lambda < 1$ μm) to reach resolution lower than the millisecond of degree needed to resolve details on the surface of stars. Also a much higher dynamic range in the images will be necessary which corresponds in fact to a better sensitivity and an improved signal to noise ratio in the raw measurements. This last point is of course related to the improvement of the limiting magnitude of the technique which is absolutely mandatory for large programs in the extragalactic domain.

In Section 2, I review some of the general principles of optical long baseline interferometry. In Section 3, I will show that optical interferometry is mainly an imaging technique and will detail the most important aspects of this point of view. I present in Section 4, the main limitations encountered and the way optical long baseline interferometry is currently implemented as an observing technique. After a rapid presentation of the CHARA Array and the VEGA instrument in Section 5, I will present recent results obtained by the VEGA group in Section 6.

2 Principles of optical interferometry

This section does not intend to present a complete and rigorous demonstration of the principles of optical interferometry. This is of course out of the scope of this paper and the reader could refer to the excellent book of Goodman (2000) as well

as to many other reviews. The idea is to present by different point of views the principle of the physical properties of long baseline interferometry.

2.1 Coherence of the stellar wave

If we consider a star located at infinity and presenting an angular diameter θ, this object defines a solid angle Ω defined by:

$$\Omega = \pi \left(\frac{\theta}{2}\right)^2. \tag{2.1}$$

We consider a screen of radius r receiving the stellar wave. This screen has a surface $S = \pi r^2$. This defines a beam etendue ϵ that can be written as:

$$\epsilon = S\Omega = \pi^2 r^2 \left(\frac{\theta}{2}\right)^2. \tag{2.2}$$

The principle of coherence, as defined by Goodman in his book, indicates that we can consider the wave as coherent if $\epsilon < \lambda^2$. This defines a so-called radius of coherence r_c:

$$r_c = \frac{\lambda}{\pi(\frac{\theta}{2})}. \tag{2.3}$$

One can note that in the case of a star with an angular diameter $\theta = 10$ mas and at a wavelength $\lambda = 1$ μm, this leads to a value of $r_c \simeq 13$ m. We thus understand that it exists a relation between the coherence of the wave and the angular diameter of the star. The coherence of the electromagnetic wave ψ could be determined by the computation of the complex degree of mutual coherence (Γ_{12}) between two points of the collecting screen separated by a distance B.

$$\Gamma_{12} = \frac{|\psi_1 {\psi_2}^*|}{\sqrt{|\psi_1|^2 |\psi_2|^2}}. \tag{2.4}$$

By using the Van-Cittert Zernike theorem and the notation $\tilde{\mathrm{O}}$ for the Fourier Transform of the star brightness distribution, we can write the following relation:

$$\Gamma_{12} = \frac{\tilde{O}(\frac{B}{\lambda})}{\tilde{O}(0)}. \tag{2.5}$$

Considering the star as a uniform disk, we finally obtain:

$$\Gamma_{12} = \left| \frac{2J_1(\pi B\theta/\lambda)}{\pi B\theta/\lambda)} \right|. \tag{2.6}$$

The definition of coherence by Goodman corresponds to the case where $\Gamma_{12} = 0.5$ which corresponds to $\pi B\theta/\lambda = 2$ and thus to $B = r_c = \frac{\lambda}{\pi(\frac{\theta}{2})}$, which is an other way of defining the coherence ($\epsilon < \lambda^2$).

This simple calculation shows that the coherence of the electromagnetic wave of stellar sources could be measured through a spatial sampling if one can access to very long baselines (B larger than 100 m typically). In this paper we only consider the case of direct interferometry in the optical domain, which means that we use detectors sensitive to the intensity of the electromagnetic wave and that we record the intensity resulting from the coherent addition of the complex waves. Coming back to the simple case of an instrumental setup dedicated to the measurement of the complex degree of mutual coherence, a practical implementation of this experiment is thus to consider the coherent addition of the two complex waves collected at point 1 and 2 with a phase shift on the second one dedicated to the necessary adjustment of the optical path between the two wave collectors. Thus we obtain the total intensity I as:

$$I = |\psi_1 + \psi_2 e^{i\phi}|, \tag{2.7}$$

$$I = {\psi_1}^2 + {\psi_2}^2 + 2\psi_1{\psi_2}^* cos(\phi). \tag{2.8}$$

Denoting I_i the intensity of the wave at point i, we finally obtain:

$$I = (I_1 + I_2)\left(1 + \frac{2\sqrt{I_1 I_2}}{I_1 + I_2} * \frac{\psi_1{\psi_2}^*}{\sqrt{|\psi_1|^2|\psi_2|^2}} * cos(\phi)\right). \tag{2.9}$$

The term with the cosinus function represents, if one introduces variations of ϕ either by temporal or spatial sampling, a modulation in the measured intensity, which is also called interference fringes. The amplitude of the modulation is defined by the factor in front of the cosinus. It contains two parts: the photometric one $\left(\frac{2\sqrt{I_1 I_2}}{I_1+I_2}\right)$ and the coherence one $\left(\frac{\psi_1{\psi_2}^*}{\sqrt{|\psi_1|^2|\psi_2|^2}}\right)$ where we recognizes Γ_{12} the complex degree of mutual coherence of the two collected waves.

As a conclusion of this section, we see that we have indeed a way to measure complex degrees of mutual coherence of stellar waves allowing us to sample the Fourier transform of brightness distributions at very high spatial frequencies. We will see in Section 4 how this method is now implemented in reality but before coming into the instrumental part of this technique an other point of view is also very important for a correct understanding of this observing technique.

3 Interferometry and images

Astronomers have developed optical interferometry in order to improve the resolving power of the telescopes. Indeed image formation in a telescope is a standard diffraction problem and it is known for a long time that an image is obtained as the convolution of the brightness distribution of the source by the point spread function of the optical device. When this convolution relation is translated into the Fourier domain, it shows that the spatial frequency spectrum of an image is the spatial frequency spectrum of the object filtered by the optical transfer function

(OTF) of the optical device. Thanks to the diffraction principle, it could be easily shown that the modulus of the optical transfer function, called the modulation transfer function (MTF), is obtained as the autocorrelation of the pupil function, defining the entrance plane of the optical device.

In the case of a monolithic telescope of diameter D, the OTF acts as a low pass filter transmitting the spatial frequencies of the object brightness distribution up to D/λ. This corresponds to what is usually called the diffraction limit λ/D of the telescope. We do not consider here the perturbations induced by the atmosphere and we just consider the ideal case of a perfect optical instrument.

In the case of an interferometer with two telescopes of diameter D and separated by a vector $\overrightarrow{B}$, the support of the OTF (also called the (u,v) plane), is made of a low frequency peak of extent $\pm D/\lambda$ and two high frequency peaks of extent $\pm D/\lambda$ and located at $\pm\overrightarrow{B}/\lambda$. The interferometer acts thus as a high frequency band pass filter, allowing to reach information at a resolution of $\lambda/|\overrightarrow{B}|$.

In the general case, the (u,v) plane (support of the OTF) is a function of the input baselines, of the latitude of the observatory, of the target coordinates, of the wavelength and of the time (because of the earth rotation). The (u,v) plane coverage defines the sampling of Fourier transform of the object brightness distribution.

The properties of the image obtained directly at the focus of an interferometer clearly depend on the (u,v) plane coverage but it can also be shown (Labeyrie 1996) that the beam combination scheme plays also an important role in that domain. I refer the reader to the important papers published in that domain (see Labeyrie *et al.*, these proceedings). As an illustration we present in Figure 2, some examples of (u,v) plane coverage and point spread function for different kind of optical interferometers.

Currently, no interferometer is working in a direct imaging scheme except maybe the Large Binocular Telescope. The limitations of coherence for ground based projects in the optical domain are clearly difficult to overcome. Progresses are being made in that direction but for the moment, imaging at high angular resolution, is not working directly at the focus of the interferometer. Instead, astronomers are using the (u,v) plane coverage to sample the Fourier transform of the brightness distribution and then to reconstruct images. This method has made great progresses in the recent years as shown in Figure 1. The quality of the reconstructed images highly depends of the (u,v) plane coverage and of the a priori information (regularization constraints) introduced in the reconstruction algorithm. I do not intend to describe this method in the present paper and I refer the reader to the chapters written in these proceedings by E. Thiébaut, D. Mary, C. Aime.

I will conclude this section by giving some general considerations about image reconstruction with an interferometer. First of all, an interferometer made of N telescopes produces $N(N-1)/2$ baselines and thus samples $N(N-1)/2$ frequencies in the Fourier transform of the brightness distribution of the object. We thus have a problem with $N(N-1)$ unknowns.

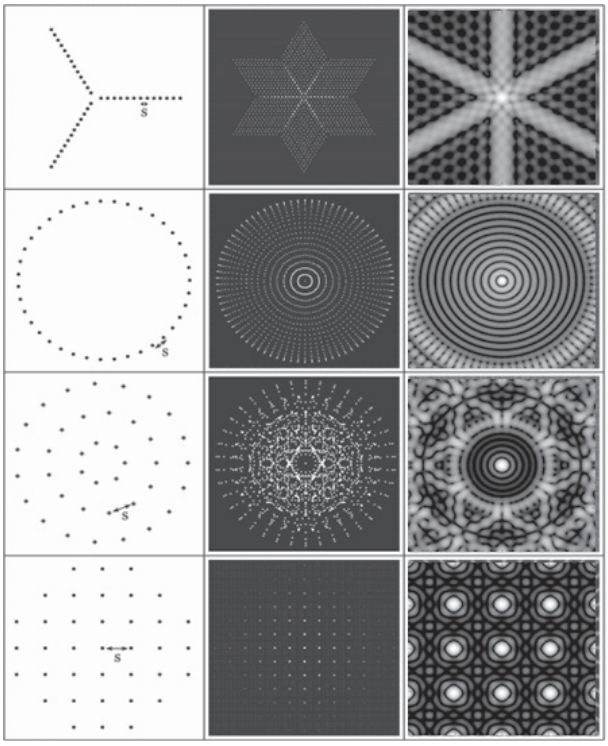


Fig. 2. Examples of (u,v) plane coverage (*middle column*) and of the corresponding point spread function (*right column*) for four different input pupil (*left column*) configurations.

We have already indicated that the limitations of ground based interferometers are dominated by the phase effects introduced by the atmospheric turbulence. If it is easy to measure the modulus of the Fourier transform over the $N(N-1)/2$ points, the phase measurements are highly corrupted by the turbulence. As in radio interferometry, astronomers overcome this difficulty by computing closure phase measurements over triplets of apertures. It can be shown easily that the atmospheric phase terms are removed in the sum of the phase of three interference fringes over any triplet of telescopes. Thus closure phase measurements give us access to $(N-1)(N-2)/2$ additional measurements. With this in hand, we understand that the problem is not well constrained because the number of unknowns is always larger than the number of measurements. A representation of these numbers is presented on Figure 3.

4 Reality of optical interferometry

The current implementation of optical interferometry involves a limited number of apertures. The VLTI is able to recombine four telescopes at the same time with the

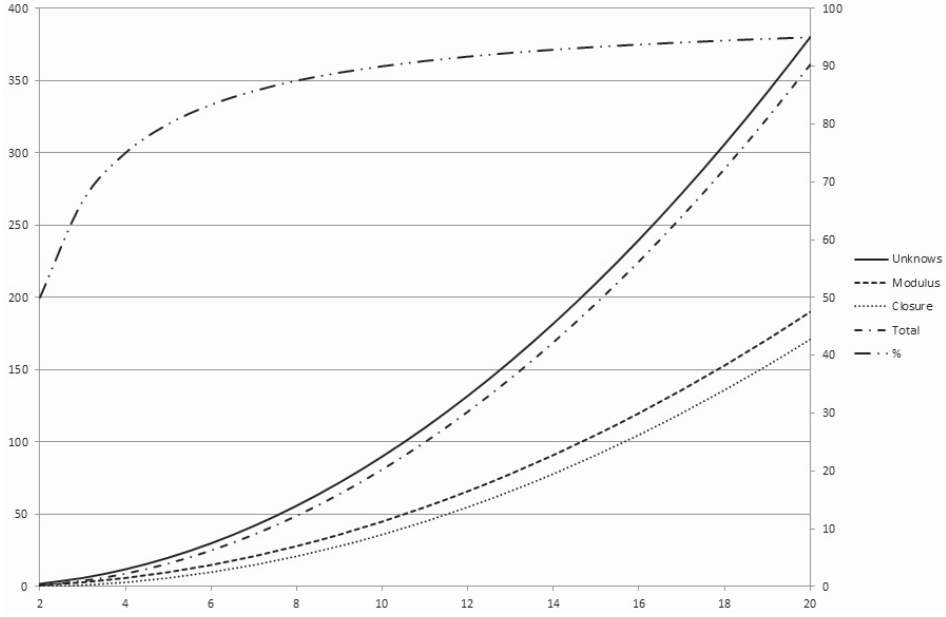


Fig. 3. Number of unknowns, of modulus and closure phase measurements as a function of the number of telescopes. The dark curve represents the percentage of information measured as a function of the number of telescopes.

PIONIER instrument (Le Bouquin *et al.* 2011) whereas CHARA can recombine up to 6 telescopes simultaneously with the MIRC instrument (Monnier *et al.* 2008). While this clearly corresponds to a great advance, it appears that imaging with optical interferometry is still very limited for the moment. Thus astronomers are mainly using the interferometric raw measurements (visibility, closure phase, complex differential visibility) to constrain the geometrical distribution of the emitting sources. An important effort has been devoted in the last years in the development of model fitting tools or of image reconstruction algorithms.

But before dealing with the calibrated products of the interferometer, many actions have to been done for a correct operation of the interferometer. One of the main difficulty concerns the overall reliability of the array. Indeed, an array is made of an important number of subsystems and the observing efficiency highly depends on the reliability of each element. This concerns the telescope (pointing, tracking, tip/tilt corrections, adaptive optics efficiency), the beam transportation and the optical path length equalization (vacuum pipes, optical fibers, optical quality, delay line, laser metrology, dumping of vibrations) and finally the beam combination and the signal sampling (spatial or temporal modulation, detection).

But on the ground and at optical wavelengths, the main difficulties for optical interferometry are in the domain of the perturbations induced by the atmospheric turbulence. The signal to noise ratio of a coherence measurement as described

before depends of course on the coherence volume. This volume has in fact many dimensions. Spatial firstly and here the atmospheric turbulence highly reduces this dimension to the so-called Fried parameter (r_0). Depending on the atmospheric conditions, the value of r_0 is in the range from 5 to 15–20 cm, well below the diameter of the individual collectors. Temporal secondly and the reduction is here drastic with coherence time t_0 between 2 and 15 ms typically. Finally the spectral dimension should also be considered because the atmospheric turbulence has a spectral coherence characteristic that limit the useful spectral bandwidth to about 20 to 30 nm at visible wavelengths (Berio *et al.* 1997). Important efforts are devoted to improve this volume of coherence through the implementation of dilute adaptive optics systems: first adaptive optics to correct each aperture and then cophasing systems allowing to control the phase between the different sub apertures. With these conditions, long exposures will be possible and thus fainter magnitude and/or higher quality will be reached.

5 The CHARA array and the VEGA spectro-interferometer

The Center for High Angular resolution (CHARA) of the Georgia State University operates an optical interferometric array located at the Mount Wilson Observatory that consists of six one meter telescopes placed in pairs along the arms of a Y-shaped array. It yields 15 baselines ranging from 34 to 331 m. Operating in the near-infrared with the instruments CLASSIC (Ten Brummelaar *et al.* 2005), CLIMB (Sturmann *et al.* 2010), FLUOR (Coude du Foresto *et al.* 2003), and MIRC (Monnier *et al.* 2008), and in the visible with PAVO (Ireland *et al.* 2008) and VEGA (Mourard *et al.* 2009, 2011), the CHARA array allows a maximum angular resolution of 1.3 and 0.3 millisecond of arc in the K and V band, respectively.

The VEGA spectrograph is designed to sample the visible band from 0.45 to 0.85 μm. It is equipped with two photon counting detectors looking at two different spectral bands. The main characteristics are summarized in Table 1. The optical design allows simultaneous recording of data, in medium spectral resolution, of the spectral region around $H\alpha$ with the red detector and around $H\beta$ with the blue detector. Observing in the blue requires good seeing conditions but increases by 30% the limit of spatial resolution of the instrument with respect to its operation around 700 nm.

Table 1. Spectral resolution (R) and bandwidth ($\Delta\lambda$) of the VEGA spectrograph, as well as the spectral separation between the two detectors.

Grating	**R**	$\Delta\lambda$ **(Blue)**	$\Delta\lambda$ **(Red)**	$\lambda_R - \lambda_B$
R1: 1800gr/mm	30 000	5 nm	8 nm	25 nm
R2: 300gr/mm	5000	30 nm	45 nm	170 nm

The limiting magnitudes of VEGA/CHARA are presented in Table 2. They of course highly depend on the actual seeing conditions and on the intrinsic target visibility.

Table 2. Estimation of typical limiting magnitude as a function of the different spectral resolution modes. These values are presented for the median value of the Fried parameter r_0 at Mount Wilson *i.e.* 8 cm. We also indicate the best performances assuming an r0 of 15 cm.

Resolution	R	Typical Lim. Magnitude	Best perf.
Medium	6000	6.5	7.5
High	30 000	4.2	5.5

VEGA is in routine operation at Mount Wilson and benefits from about 60 nights per year. Many observations are now done remotely from Nice Observatory. Our group has recently improved the photon counting detectors. A new image intensifier has been installed with better quantum efficiency (approximately a factor 1.5 better) in the red part of the spectrum and the Dalsa sensor (Blazit *et al.* 2008) behind the two image intensifiers has been replaced by a Gazelle sensor from the Point Grey company. This new sensor allows a faster frame rate (10 ms) and a much lower dead time during two frames (1 ms instead of 2 ms). The duty cycle of the sensor is now of the order of 90% instead of 60% with the old camera. An improvement of 1.5 magnitude has thus been recently demonstrated as well as a much better detector cosmetics important for spectrum measurements.

We are also considering a future evolution of VEGA in order to correctly benefit from the future installation of adaptive optics on the CHARA telescopes. The high Strehl ratio that will be allowed thanks to these new devices will highly increase the signal to noise of our measurements. However it will also concentrate the flux in a small part of the detector and thus will lead to an increase of the saturation effect with the current generation of photon counting detector. We are thus considering using analogical detector such as EMCCD or OCAM2 (Feautrier *et al.* 2011) that allows a very high frame rate (up to 1500 fps) and a very low readout noise ($0.13e^-$/pix/frame). Coupling this kind of detector with a beam combiner using spatial filtering and high efficiency optical devices (P. Bério, in preparation) will permit to enhance the scientific domain of VEGA/CHARA in the future.

6 Recent results from VEGA/CHARA

The most remarkable properties of VEGA/CHARA are first the access to unprecedented angular resolution thanks to the 300 meters baseline and the short wavelengths and second the access to high angular resolution measurements at very high spectral resolution (up to 30 000).

The medium (6000) and high (30 000) spectral resolutions are well suited to perform kinematic analysis of the interferometric signal, providing resolution of 60 and 10 $\mathrm{km\,s^{-1}}$ respectively. These spectral resolutions are best dedicated to the extraction of differential spectral information. Radiative winds and fast rotating

photospheres of hot stars can be probed efficiently with the medium spectral resolution. Some recent examples of such studies could be found for the Be stars 48 Per and ϕ Per (Delaa *et al.* 2011) where the authors characterize the rotating disks in term of extension, ellipticity and kinematical field. In Meilland *et al.* (2011), the authors use the combination of VEGA and VLTI/AMBER (Petrov *et al.* 2007) data to constrain both the orbital elements of the famous Be binary δ Sco and the disk's parameters. The interactive binaries β Lyrae and υ Sagitarri (Bonneau *et al.* 2011) have also been studied in the same way. Perraut *et al.* (2010) succeeded also for the first time to spectrally and spatially resolve the $H\alpha$ emitting region of the prototype of the young stellar objects AB Aurigae. High spectral and angular resolutions bring also complementary views on old and famous problems such as the mysterious eclipsing system ϵ Aurigae (Mourard *et al.* 2012) or on the chromosphere of K giants (Berio *et al.* 2011).

The medium resolution is also well suited to absolute visibility studies and are also well adapted for the study of binaries or multiple systems. In that field the main goal is the study of fundamental stellar parameters through angular diameters measurements and analysis through classical stellar modeling and/or confrontation with other observing techniques such as spectroscopy and asteroseismology. Recent results of such programs concern the study of the ro Ap star γ Equulei (Perraut *et al.* 2011) or the famous CoRoT targets HD49933 (Bigot *et al.* 2011) and more recently the study of four exoplanet hosts stars (Ligi *et al.* 2012). These exploratory programs are now coordinated as large programs where many tens of objects are being studied in order to have a good analysis of the stellar properties in different part of the Herztsprung-Russel diagram.

Another interesting possibility is the presence of a polarimeter that could be inserted into the beam. This gives new insight into many physical processes. Many science sources are linearly polarized, in particular at a small angular scale, and the interferometric polarized signal is a powerful probe of circumstellar scattering environments that contain ionized gas or dust (Chesneau *et al.* 2003; Elias *et al.* 2008; Ireland *et al.* 2005) and of magnetic properties (Rousselet-Perraut *et al.* 2000, 2004). This possibility has not yet been really exploited on the VEGA/CHARA interferometer but it could bring interesting new programs.

7 Conclusion

With this lecture and this paper, my intention has been first to describe the way optical interferometry should be understood from a physical point of view and second to show the recent advances in that field in terms of astrophysical programs and in terms of observing possibilities. The dream of the groups working in optical interferometry is clearly to push towards a large facility with remarkable capabilities highly complementary to what will bring, in the future, the Extremely Large Telescopes or the large radio arrays.

In complement to the science addressed by the large radio arrays and the Extremely Large telescopes, we consider that optical interferometry can bring important answers, firstly on the possibility of fighting off the expected confusion

limit of ELT and secondly for the direct imaging with spatial, temporal and spectral resolution of compact sources such as the inner part of young stellar objects where planets are formed or the inner parsec around active galactic nuclei. In all cases, the quality of the synthetic point spread function will be fundamental both for the sensitivity and for the resolving power. The control of such imaging machines for nulling or phase-controlled coronagraphy is also of utmost importance for the detection and characterization of planets in the habitable zone. In this latter case, the effort is more in the control of the dynamic in the image than in the angular resolution. Debates around the future concepts have almost concluded around three main classes of future optical arrays: 1) a VLTI-like interferometer with a very small number of ELT-like telescopes on a compact array, 2) kilometric baselines with a small number of 8-m class telescopes and 3) a dense array of a large number of small telescopes over possibly kilometric baselines. If the conceptual design of the two first classes of array could certainly rest on the current concepts of classical telescopes + delay lines, it is clear that expanding the number of apertures to 50, 100 or even more individual apertures encounters a real limitation for the implementation. This represents a major difference to the situation of radio interferometry and many conceptual and prototyping efforts are now engaged in that direction.

References

Berio, P., Mourard, D., Vakili, F., *et al.*, 1997, JOSA-A, 14
Berio, P., Merle, T., Thevenin, F., *et al.*, 2011, A&A, 535
Bigot, L., Mourard, D., Berio, P., *et al.*, 2011, A&A, 534
Blazit, A., Rondeau, X., Thiebaut, E., *et al.*, 2008, Appl. Opt., 47
Bonneau, D., Chesneau, O., Mourard, D., *et al.*, 2011, A&A, 532
Chesneau, O., Wolf, S., & Domiciano de Souza, A., 2003, A&A, 410
Colavita, M., Serabyn, G., Wizinowich, P., & Akeson, R., 2006, Proc. SPIE, 6268
Coude du Foresto, V., Borde, P., Merand, A., *et al.*, 2003, SPIE Conf. Proc., 4838
Delaa, O., Stee, P., Meilland, A., *et al.*, 2011, A&A, 529
Elias, N., Schmitt, H., Jorgensen, A., *et al.*, 2008 [arXiv:0811.3139]
Feautrier, P., Gach, J.L., Balard, P., *et al.*, 2011, PASP, 123
Glindemann, A., *et al.*, 2004, Spie Proc., 5491
Goodman, J.W., 2000, "Statistical Optics" (Wiley)
Haubois, *et al.*, 2008, A&A
Ireland, M., Tuthill, P., Davis, J., & Tango, W., 2005, MNRAS, 361
Ireland, M., ten Brummelaar, T., Tuthill, P.G., *et al.*, 2008, SPIE Conf. Proc., 7013
Labeyrie, A., 1975, ApJ, 196
Labeyrie, A., 1996, A&AS, 118
Lacour, *et al.*, 2008, A&A
Le Bouquin, J.B., Berger, J.P., Lazareff, B., *et al.*, 2011, A&A, 535
Ligi, R., Mourard, D., Lagrange, A.M., *et al.*, 2012, A&A, 545

Meilland, A., Delaa, O., Stee, P., *et al.*, 2011, A&A, 532

Monnier, *et al.*, 2007, Science

Monnier, J., Zhao, M., Pedretti, E., *et al.*, 2008, SPIE Conf. Proc., 7013

Mourard, D., Clausse, J.M., Marcotto, A., *et al.*, 2009, A&A, 508

Mourard, D., Berio, P., Perraut, K., *et al.*, 2011, A&A, 531

Mourard, D., Harmanec, P., Stencel, R., *et al.*, 2012, A&A, 544

Perraut, K, Benisty, M., Mourard, D., *et al.*, 2010, A&A, 516

Perraut, K., Brandao, I., Mourard, D., *et al.*, 2011, A&A, 526

Petrov, R., Malbet, F., Weigelt, G., *et al.*, 2007, A&A, 464

Rousselet-Perraut, K., Chesneau, O., Berio, P., & Vakili, F., 2000, A&A, 354

Rousselet-Perraut, K., Stehle, C., Lanz, T., *et al.*, 2004, A&A, 422

Sturmann, J., ten Brummelaar, T., Sturmann, L., & Mc Alister, H.A., 2010, SPIE Conf. Proc., 7734

ten Brummelaar, T., McAlister, H.A., Ridgway, S., *et al.*, 2005, ApJ, 628

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 37–58*

THE FRESNEL DIFFRACTION: A STORY OF LIGHT AND DARKNESS

C. Aime[1], É. Aristidi[1] and Y. Rabbia[1]

Abstract. In a first part of the paper we give a simple introduction to the free space propagation of light at the level of a Master degree in Physics. The presentation promotes linear filtering aspects at the expense of fundamental physics. Following the Huygens-Fresnel approach, the propagation of the wave writes as a convolution relationship, the impulse response being a quadratic phase factor. We give the corresponding filter in the Fourier plane. As an illustration, we describe the propagation of wave with a spatial sinusoidal amplitude, introduce lenses as quadratic phase transmissions, discuss their Fourier transform properties and give some properties of Soret screens. Classical diffractions of rectangular diaphragms are also given here. In a second part of the paper, the presentation turns into the use of external occulters in coronagraphy for the detection of exoplanets and the study of the solar corona. Making use of Lommel series expansions, we obtain the analytical expression for the diffraction of a circular opaque screen, giving thereby the complete formalism for the Arago-Poisson spot. We include there shaped occulters. The paper ends up with a brief application to incoherent imaging in astronomy.

1 Historical introduction

The question whether the light is a wave or a particle goes back to the seventeenth century during which the mechanical corpuscular theory of Newton took precedence over the wave theory of Huygens. Newton's particle theory, which explained most of the observations at that time, stood as the undisputed model for more than a century. This is not surprising since it was not easy to observe natural phenomena resulting from the wave nature of light. At that time light sources like the Sun or a candle light were used. They are incoherent extended sources while a coherent source is needed to see interference phenomenas, unquestionable signatures of the wave nature of light.

[1] Laboratoire Lagrange, Université de Nice Sophia-Antipolis, Centre National de la Recherche Scientifique, Observatoire de la Côte d'Azur, Parc Valrose, 06108 Nice, France

DOI: 10.1051/eas/1359003

The starting point of the wave theory is undoubtedly the historical double-slits experiment of Young in 1801. The two slits were illuminated by a single slit exposed to sunlight, thin enough to produce the necessary spatially coherent light. Young could observe for the first time the interference fringes in the overlap of the light beams diffracted by the double slits, demonstrating thereby the wave nature of light against Newton's particle theory. Indeed, the darkness in the fringes cannot be explained by the sum of two particles but easily interpreted by vibrations out of phase. This argument was very strong. Nevertheless Einstein had to struggle in his turn to have the concept of photons accepted by the scientific community a century later. In astronomy, we can use the simplified semiclassical theory of photodetection, in which the light propagates as a wave and is detected as a particle (Goodman 1985).

In Young 's time, Newton's prestige was so important that the wave nature of light was not at all widely accepted by the scientific community. About fifteen years later, Fresnel worked on the same problematics, at the beginning without being aware of Young's work. Starting from the Huygens approach, Fresnel proposed a mathematical model for the propagation of light. He competed in a contest proposed by the Academy of Sciences. The subject was of a quest for a mathematical description of diffraction phenomena appearing in the shade of opaque screens. Poisson, a jury's member, argued that according to Fresnel's theory, a bright spot should appear at the center of a circular object's shadow, the intensity of the spot being equal to that of the undisturbed wavefront. The experiment was soon realized by Arago, another jury's member, who indeed brilliantly confirmed Fresnel's theory. This bright spot is now called after Poisson, Arago or Fresnel.

Arago's milestone experience was reproduced during the CNRS school of June 2012, using learning material for students in Physics at the University of Nice Sophia – Antipolis. The results are given in Figure 1. A laser and a beam expander were used to produce a coherent plane wave. The occulter was a transparent slide with an opaque disk of diameter 1.5 mm, while Arago used a metallic disk of diameter 2 mm glued on a glass plate. Images obtained in planes at a distance of $z = 150$, 280 and 320 mm away from the screen of observation are given in the figure. The Arago bright spot clearly appears and remains present whatever the distance z is. The concentric circular rings are not as neat as expected, because of the poor quality of the plate and of the occulting disk, a difficulty already noted by Fresnel (see for example in de Senarmont *et al.* 1866). We give the mathematical expression for the Fresnel diffraction in the last section of this paper.

The presentation we propose here is a short introduction to the relations of free space propagation of light, or Fresnel diffraction. It does not aim to be a formal course or a tutorial in optics, and remains in the theme of the school, for an interdisciplinary audience of astronomers and signal processing scientists. We restrict our presentation to the scalar theory of diffraction in the case of paraxial optics, thus leaving aside much of the work of Fresnel on polarization. We show that the propagation of light can be simply presented with the formalism of linear filtering. The reader who wishes a more academic presentation can refer to books of Goodman (2005) and Born & Wolf (2006).

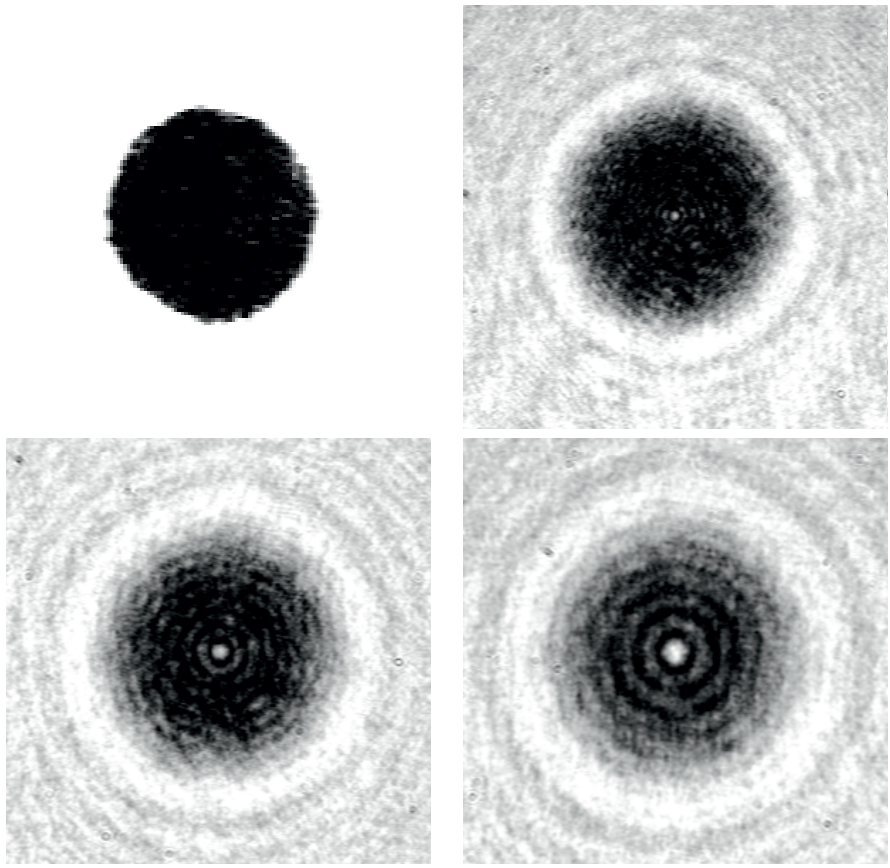


Fig. 1. Reproduction of Arago's experience performed during the CNRS school. *Top left*: the occulter (diameter: 1.5 mm), and its Fresnel diffraction figures at the distances of 150 mm (*top right*), 280 mm (*bottom left*) and 320 mm (*bottom right*) from the screen of observation.

The paper is organized as follows. We establish the basic relations for the free space propagation in Section 2. An illustration for the propagation of a sinusoidal pattern is given in Section 3. Fourier properties of lenses are described in Section 4. Section 5 is devoted to the study of shadows produced by external occulters with application to coronagraphy. Section 6 gives a brief application to incoherent imaging in astronomy.

2 Basic relations for free space propagation, a simplified approach

We consider a point source S emitting a monochromatic wave of period T, and denote $A_S(t) = A\exp(-2i\pi t/T)$ its complex amplitude. In a very simplified model

where the light propagates along a ray at the velocity $v = c/n$ (n is the refractive index), the vibration at a point P located at a distance s from the source is:

$$A_P(s,t) = A\exp\left(-2i\pi\frac{(t-s/v)}{T}\right) = A\exp\left(-2i\pi\frac{t}{T}\right)\exp\left(2i\pi\frac{ns}{\lambda}\right) \qquad (2.1)$$

where $\lambda = cT$ is the wavelength of the light, the quantity ns is the optical path length introduced by Fermat, a contemporary of Huygens. The time dependent factor $\exp(-2i\pi t/T)$, common to all amplitudes, is omitted later on in the presentation. For the sake of simplicity, we moreover assume a propagation in the vacuum with $n = 1$.

In the Huygens-Fresnel model, the propagation occurs in a different way. First of all, instead of rays, wavelets and wavefronts are considered. A simple model for a wavefront is the locus of points having the same phase, *i.e.* where all rays originating from a coherent source have arrived at a given time. A spherical wavefront becomes a plane wavefront for a far away point source. According to Malus theorem, wavefronts and rays are orthogonal. The Huygens-Fresnel principle states that each point of any wavefront irradiates an elementary spherical wavelet, and that these secondary waves combine together to form the wavefront at any subsequent time.

We assume that all waves propagate in the z positive direction in a $\{x, y, z\}$ coordinate system. Their complex amplitudes are described in parallel transverse planes $\{x, y\}$, for different z values. If we denote $A_0(x, y)$ the complex amplitude of a wave in the plane $z = 0$, its expression $A_z(x, y)$ at the distance z may be obtained by one of the following equivalent equations:

$$\begin{aligned} A_z(x,y) &= A_0(x,y) * \frac{1}{i\lambda z}\exp\frac{i\pi(x^2+y^2)}{\lambda z} \\ A_z(x,y) &= \Im^{-1}\left[\hat{A}_0(u,v)\exp(-i\pi\lambda z(u^2+v^2))\right] \end{aligned} \qquad (2.2)$$

where λ is the wavelength of the light, the symbol $*$ stands for the 2D convolution. $\hat{A}_0(u, v)$ is the 2D Fourier transform of $A_0(x, y)$ for the conjugate variables (u, v) (spatial frequencies), defined as

$$\hat{A}_0(u,v) = \iint A_0(x,y)\exp(-2i\pi(ux+vy))dxdy. \qquad (2.3)$$

The symbol $\Im^{-1}$ denotes the two dimensional inverse Fourier transform. It is interesting to note the quantity $\sqrt{\lambda z}$, playing the role of a size factor in Equation (2.2), as we explain in the next section in which these relationships are established and their consequences analyzed.

2.1 The fundamental relation of convolution for complex amplitudes

The model proposed by Huygens appears as a forerunner of the convolution in Physics. In the plane z, the amplitude $A_z(x, y)$ is the result of the addition of

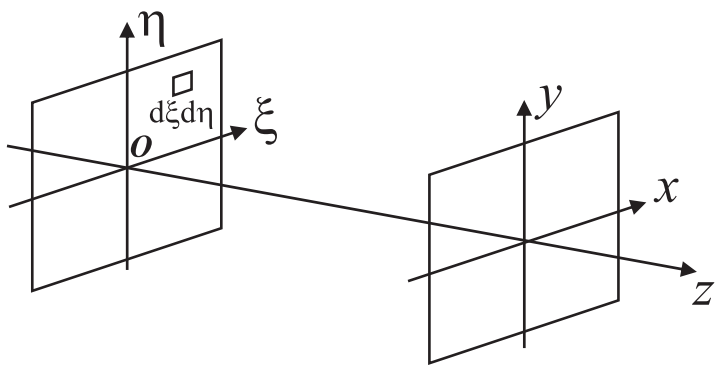


Fig. 2. Notations for the free space propagation between the plane at $z = 0$ of transverse coordinates ξ and η and the plane at the distance z of transverse coordinates x and y.

the elementary wavelets coming from all points of the plane located at $z = 0$. To build the relations between the two waves in the planes $z = 0$ and z, we will have to consider the coordinates of the points of these planes. To make the notations simpler, we substitute ξ and η to x and y in the plane $z = 0$.

Let us first consider the point-like source at the origin O of coordinates $\xi = \eta = 0$, and the surrounding elementary wavefront of surface $\sigma = d\xi\, d\eta$. The wavelet emitted by O is an elementary spherical wave. After a propagation over the distance s, the amplitude of this spherical wave can be written as $(\alpha/s)\ \sigma A_0(0,0)\times \exp(2i\pi s/\lambda)$, where α is a coefficient to be determined. The factor $1/s$ is required to conserve the energy.

Now we start deriving the usual simplified expression for this elementary wavelet in the plane $\{x, y\}$ at the distance z from O. Under the assumption of paraxial optics, *i.e.* x and $y \ll z$, the distance s is approximated by

$$s = (x^2 + y^2 + z^2)^{1/2} \simeq z + (x^2 + y^2)/2z. \tag{2.4}$$

The elementary wavelet emitted from a small region $\sigma = d\xi d\eta$ around O (see Fig. 2) and received in the plane z can be written:

$$A_0(0,0)\,\sigma \times \frac{\alpha}{s} \exp\left(2i\pi\frac{s}{\lambda}\right) \simeq A_0(0,0)\, d\xi d\eta \times \exp\left(2i\pi\frac{z}{\lambda}\right) \frac{\alpha}{z} \exp\left(i\pi\frac{x^2+y^2}{\lambda z}\right). \tag{2.5}$$

The approximation $1/s \simeq 1/z$ can be used, when s works as a factor for the whole amplitude, since this latter is not sensitive to a small variation of s. On the contrary the two terms in the Taylor expansion of Equation (2.4) must be kept in the argument of the complex exponential, since it expresses a phase and is very sensitive to a small variation of s. For example a variation of s as faint as λ induces a phase variation of 2π.

For a point source P at the position (ξ, η), the response is:

$$dA_z(x,y) = A_0(\xi,\eta)\, \exp\left(2i\pi\frac{z}{\lambda}\right)\, \frac{\alpha}{z}\, \exp\left(i\pi\frac{(x-\xi)^2+(y-\eta)^2}{\lambda z}\right) d\xi d\eta. \quad (2.6)$$

According to the Huygens-Fresnel principle, we sum the wave amplitudes for all point like sources coming from the plane at $z = 0$ to obtain the amplitude in z:

$$\begin{aligned} A_z(x,y) =& \exp\left(2i\pi\frac{z}{\lambda}\right) \iint A_0(\xi,\eta)\, \frac{\alpha}{z}\, \exp\left(i\pi\frac{(x-\xi)^2+(y-\eta)^2}{\lambda z}\right) d\xi d\eta \\ =& \exp\left(2i\pi\frac{z}{\lambda}\right)\, A_0(x,y) * \frac{\alpha}{z} \exp\left(i\pi\frac{x^2+y^2}{\lambda z}\right). \end{aligned} \quad (2.7)$$

Equation (2.7) results in the convolution of the amplitude at $z = 0$ with the amplitude of a spherical wave. The factor $\exp(2i\pi z/\lambda)$ corresponds to the phase shift induced by the propagation over the distance z, and will be in general omitted as not being a function of x and y. The coefficient α is given by the complete theory of diffraction. We can derive it considering the propagation of a plane wave of unit amplitude $A = 1$. Whatever the distance z we must recover a plane wave. So we have:

$$1 * \frac{\alpha}{z} \exp\left(i\pi\frac{x^2+y^2}{\lambda z}\right) = 1 \quad (2.8)$$

which leads to the value $\alpha = (i\lambda)^{-1}$, as the result of the Fresnel integral. The final expression is then:

$$A_z(x,y) = A_0(x,y) * \frac{1}{i\lambda z} \exp\left(i\pi\frac{x^2+y^2}{\lambda z}\right) = A_0(x,y) * D_z(x,y). \quad (2.9)$$

The function $D_z(x,y)$ behaves as the point-spread function (PSF) for the amplitudes. It is separable in x and y:

$$D_z(x,y) = D_z^0(x) D_z^0(y) = \frac{1}{\sqrt{i\lambda z}} \exp i\pi\frac{x^2}{\lambda z} \cdot \frac{1}{\sqrt{i\lambda z}} \exp i\pi\frac{y^2}{\lambda z} \quad (2.10)$$

where $D_z^0(x)$ is normalized in the sense that $\int D_z^0(x)dx = 1$. It is important to note that $D_z(x,y)$ is a complex function, essentially a quadratic phase factor, but for the normalizing value $i\lambda z$.

2.1.1 The Fresnel transform

Another form for the equation of free space propagation of the light can be obtained by developing Equation (2.9) as follows

$$\begin{aligned} A_z(x,y) =& \frac{1}{i\lambda z} \exp\left(i\pi\frac{x^2+y^2}{\lambda z}\right) \times \\ & \iint \left\{ A_0(\xi,\eta)\, \exp\left(i\pi\frac{\xi^2+\eta^2}{\lambda z}\right)\right\} \exp\left(-2i\pi\left(\xi\frac{x}{\lambda z}+\eta\frac{y}{\lambda z}\right)\right)\, d\xi\, d\eta. \end{aligned} \quad (2.11)$$

The integral clearly describes the Fourier transform of the function between brackets for the spatial frequencies $x/\lambda z$ and $y/\lambda z$. It is usually noted as:

$$A_z(x,y) = \frac{1}{i\lambda z} \exp\left(i\pi \frac{x^2+y^2}{\lambda z}\right) \mathcal{F}_z\left[A_0(x,y) \exp\left(i\pi \frac{x^2+y^2}{\lambda z}\right)\right]. \qquad (2.12)$$

Following Nazarathy & Shamir (1980), it is worth noting that the symbol $\mathcal{F}_z$ can be interpreted as an operator that applies on the function itself, keeping the original variables x and y, followed by a scaling that transform x and y into $x/\lambda z$ and $y/\lambda z$. Although it may be of interest, the operator approach implies the establishment of a complete algebra, and does not present, at least for the authors of this note, a decisive advantage for most problems encountered in optics.

The Fresnel transform and the convolution relationship are strictly equivalent, but when multiple propagations are considered, it is often advisable to write the convolution first, and then apply the Fresnel transform to put in evidence the Fourier transform of a product of convolution.

2.2 Filtering in the Fourier space

The convolution relationship in the direct plane corresponds to a linear filtering in the Fourier plane. If we denote u and v the spatial frequencies associated with x and y, the Fourier transform of Equation (2.9) becomes

$$\hat{A}_z(u,v) = \hat{A}_0(u,v).\hat{D}_z(u,v) \qquad (2.13)$$

where:

$$\hat{D}_z(u,v) = \exp(-i\pi\lambda z(u^2+v^2)) \qquad (2.14)$$

is the amplitude transfer function for the free space propagation over the distance z. Each spatial frequency is affected by a phase factor proportional to the square modulus of the frequency. For the sake of simplicity, we will still denote this function a modulation transfer function (MTF), although it is quite different from the usual Hermitian MTFs encountered in incoherent imagery.

The use of the spatial filtering is particularly useful for a numerical computation of the Fresnel diffraction. Starting with a discrete version of $A_0(x,y)$, we compute its 2D Fast Fourier Transform (FFT) $\hat{A}_0(u,v)$, multiply it by $\hat{D}_z(u,v)$ and take the inverse 2D FFT to recover $A_z(x,y)$. We used this approach to derive the Fresnel diffraction of the petaled occulter given in the last section of this paper.

Before ending this section, we can check that the approximations used there do not alter basic physical properties of the wave propagation. To obtain the coefficient α, we have used the fact that a plane wave remains a plane wave along the propagation. The reader will also verify that a spherical wave remains also a spherical wave along the propagation. This is easily done using the filtering in the Fourier space. A last verification is the conservation of energy, *i.e.* the fact that the flux of the intensity does not depend on z. That derives from the fact that the MTF is a pure phase filter and is easily verified making use of Parseval theorem.

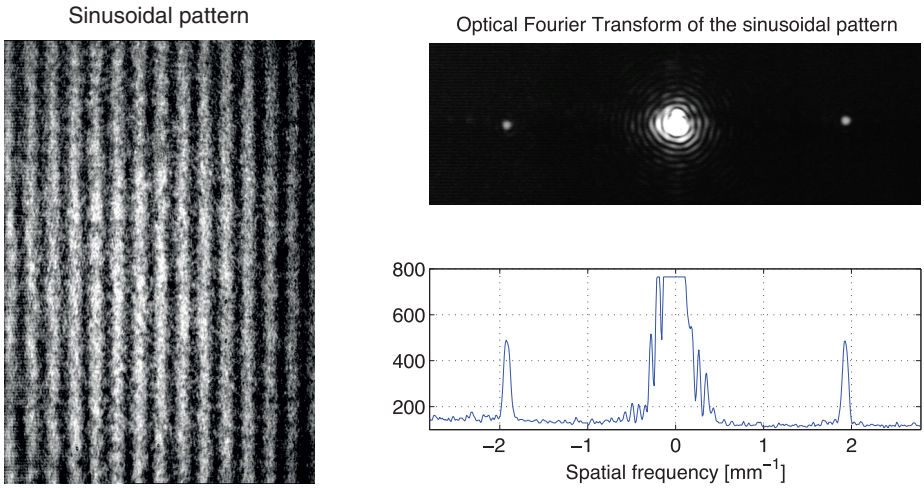


Fig. 3. *Left*: image of a two-dimensional sinusoidal pattern (Eq. (3.1)). The spatial period is $1/m = 0.52$ mm. *Top right*: optical Fourier tranform of the sinusoidal pattern in the (u, v) plane. *Bottom right*: plot of the intensity of the optical Fourier transform as a function of the spatial frequency u for $v = 0$.

3 Fresnel diffraction from a sinusoidal transmission

The particular spatial filtering properties of the Fresnel diffraction can be illustrated observing how a spatial frequency is modified in the free space propagation. The experiment was presented at the CNRS school observing the diffraction of a plate of transmission in amplitude of the form:

$$f_1(x, y) = (1 - \epsilon) + \epsilon \cos(2\pi(m_x x + m_y y)). \tag{3.1}$$

The plate is a slide of a set of fringes. This transmission in fact bears three elementary spatial frequencies at the positions $\{u, v\}$ respectively equal to $\{0, 0\}$, $\{m_x, m_y\}$ and $\{-m_x, -m_y\}$. For the simplicity of notations we assume in the following that the fringes are rotated so as to make $m_y = 0$ and $m_x = m$, and we assume $(1 - \epsilon) \sim 1$. The fringes and their optical Fourier transform are shown in Figure 3. We describes further in the paper how the operation of Fourier transform can be made optically.

As one increases the distance z, the fringes in the images almost disappear and appear again periodically with the same original contrast. A careful observation makes it possible to observe an inversion of the fringes in two successive appearances. This phenomenon is a consequence of the filtering by $\hat{D}_z(u, v)$. The frequencies at $u = \pm\, m$ are affected by the same phase factor $\exp -(i\pi\lambda z m^2)$, while the zero frequency is unchanged. At a distance z behind the screen the complex amplitude therefore expresses as

$$U_z(x, y) \sim 1 + \epsilon \cos(2\pi m x)\ \exp(-i\pi\lambda z m^2). \tag{3.2}$$

When $\lambda z m^2$ is equal to an integer number k, the amplitude is purely real and equal to $1 \pm \epsilon \cos(2\pi m x)$. When $\lambda z m^2 = 1/2 + k$, the amplitude modulation is an imaginary term, and $U_z \simeq 1 \pm i\epsilon \cos(2\pi m x)$. For ϵ very small, the wavefront is then almost a pure phase factor of uniform amplitude. It can be represented as an undulated wavefront, with advances and delays of the optical path compared to the plane wave. The wave propagates towards the z direction, continuously transforming itself from amplitude to phase modulations, as illustrated in Figure 4.

The observed intensity is

$$I_z(x, y) \sim 1 + 2\epsilon \cos(2\pi m x)\ \cos(\pi \lambda z m^2). \tag{3.3}$$

The fringes almost disappear for $z = (k + 1/2)/(\lambda m^2)$, with k integer. They are visible with a contrast maximum for $z = k/(\lambda m^2)$, and the image is inverted between two successive values of k.

At the CNRS school we have also shown the Fresnel diffraction of a Ronchi pattern, a two dimensional square wave $f_R(x, y)$ made of alternate opaque and transparent parallel stripes of equal width, as illustrated in Figure 5. Making use of the Fourier series decomposition, we can write the square wave as a simple addition of sinusoidal terms of the form:

$$f_R(x, y) = \frac{1}{2} + \frac{2}{\pi} \sum_{n=0}^{\infty} \frac{1}{2n+1} \sin(2\pi(2n+1)mx). \tag{3.4}$$

The complex amplitude $U_z(x, y)$ at a distance z behind the Ronchi pattern is simply obtained by the sum of the sine terms modified by the transfer function. We have:

$$U_z(x, y) = \frac{1}{2} + \frac{2}{\pi} \sum_{n=0}^{\infty} \frac{1}{2n+1} \sin(2\pi(2n+1)mx)\ \exp(-i\pi\lambda z(2n+1)^2 m^2). \tag{3.5}$$

As the wave propagates, each sinusoidal component experiences a phase modulation depending on its spatial frequency. One obtains an image identical to the Ronchi pattern when all the spatial frequencies in $U_z(x, y)$ are phase-shifted by a multiple of 2π. The occurrences of identical images are obtained for $\lambda z m^2 = 2k$, as for the single sine term. This property is known as the Talbot effect.

4 Focusing screens and Fourier transform properties of lenses

The Fresnel transform makes easy to introduce the converging lens and its properties relative to the Fourier transform. To make the notations simpler, we assume that the wavefront $A_0(x, y)$ is simply of the form $A \times f(x, y)$, where A stands for an incident plane wave and $f(x, y)$ is the transmission of a screen. Let us consider that we can manufacture a phase screen with the following transmission:

$$L_\phi(x, y) = \exp\left(-i\frac{\pi(x^2 + y^2)}{\lambda\phi}\right) \tag{4.1}$$

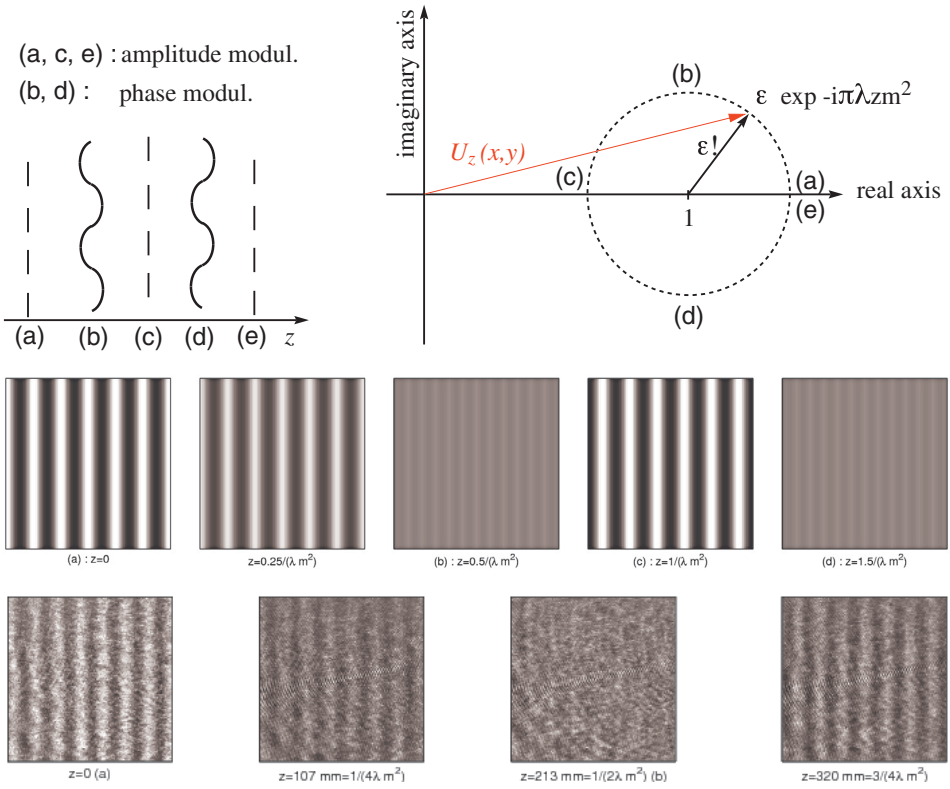


Fig. 4. Illustration of the Fresnel diffraction of the sinusoidal mask (Eq. (3.1)) using Equation (3.2). *Top left*: z positions (a), (c), (e) where the amplitudes become again identical to the mask. Corresponding values for the distances are $z_a = \frac{k}{\lambda m^2}$, $z_c = \frac{k+1}{\lambda m^2}$, $z_e = \frac{k+2}{\lambda m^2}$ (k integer). Positions (b) and (d) correspond to almost pure phase modulation (uniform intensity): $z_b = \frac{k+1/2}{\lambda m^2}$, $z_d = \frac{k+3/2}{\lambda m^2}$. *Top right*: illustration of Equation (3.2) in the complex plane. *Middle row*: simulated images as seen at distances $z = z_a$, $z = \frac{1/4}{\lambda m^2}$, $z = z_b$, $z = z_c$ and $z = z_d$. Notice the contrast inversion between positions z_a and z_c. *Bottom row*: experimental images obtained with a sinusoïdal grid of frequency $m = 1/0.52$ mm^{-1}. From *left* to *right*: positions $z = 0$, $z = \frac{1/4}{\lambda m^2}$, $z = z_b$ and $z = \frac{3/4}{\lambda m^2}$. Here again, the contrast inversion between the first and last images is visible.

that we affix to $f(x, y)$. At the distance $z = \phi$, Equation (2.12) shows that the amplitude becomes $\exp(i\pi\frac{x^2+y^2}{\lambda\phi})\hat{f}(\frac{x}{\lambda\phi}, \frac{y}{\lambda\phi})$, and the intensity appears here as a scaled Fourier transform of $f(x, y)$.

In the absence of screen (or $f(x, y) = 1$), the diffracted amplitude is proportional to a Dirac function $\delta(x, y)$, which explicits the focusing effect of a perfect lens on the axis. Such a phase screen is a converging lens (a thin piece of glass formed between a plane and a sphere gives the desired transmission), or a parabolic mirror of focal length ϕ.

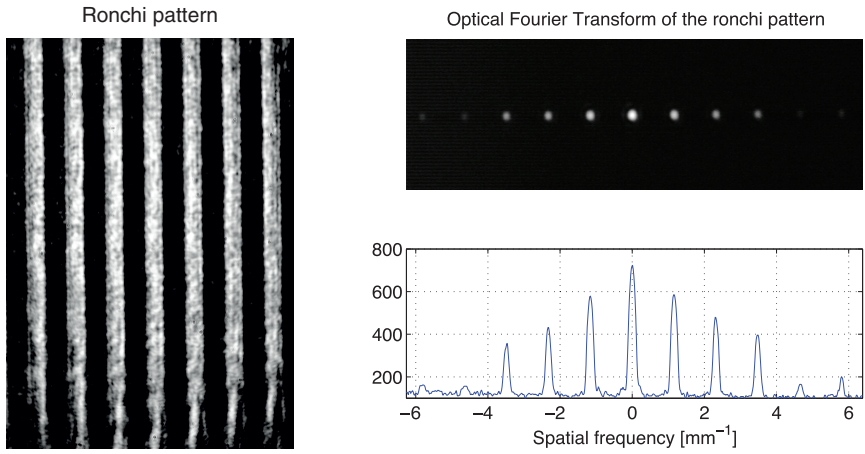


Fig. 5. *Left*: image of a Ronchi pattern. The period if $1/m = 0.86$ mm (see Eq. (3.4)). *Top right*: optical Fourier tranform of the sinusoidal pattern in the (u, v) plane. *Bottom right*: intensity of the optical Fourier transform as a function of the spatial frequency u for $v = 0$. Note that the even harmonics of the frequency m are present, while they should not with a perfect Ronchi pattern with identical widths of white and black strips.

The phase factor $\exp(i\pi\frac{x^2+y^2}{\lambda\phi})$ which remains in this focal plane corresponds to a diverging lens $L_{-\phi}(x, y)$. It can be cancelled adding here a converging lens of focal length ϕ. So, a system made of two identical converging lenses of focal length ϕ separated by a distance ϕ optically performs the exact Fourier transform of the transmission in amplitude of a screen. Such a device is called an optical Fourier transform system. This property becomes obvious if we re-write the Fresnel transform of Equation (2.12) making explicit the expression corresponding to diverging lenses:

$$A_z(x, y) = \frac{1}{i\lambda z} L_{-z}(x, y)\, \mathcal{F}_z[A_0(x, y)L_{-z}(x, y)]. \tag{4.2}$$

It is clear here that for the optical Fourier transform system the two converging lenses cancel the diverging terms of propagation. Another similar Fourier transform device can be obtained with a single converging lens of focal length ϕ, setting the transmission $f(x, y)$ in front of it at the distance ϕ and observing in its focal plane. Such systems have been used to perform image processing, as described by Françon (1979).

It is important to note that phase factors disappear also when the quantity of interest is the intensity, as for example in incoherent imagery (see Sect. 6). Optical Fourier transforms were actually used in the past to analyse speckle patterns at the focus of large telescopes (Labeyrie 1970).

4.1 Focusing screens with a real transmission

It is possible to make screens of real transmission (between 0 and 1) acting as converging lenses. For that, the transmission of the screen must contain a term similar to $L_\phi(x,y)$. To make its transmission real, we can add the transmission of a diverging lens $L_{-\phi}(x,y)$. Doing so we get a cosine term. It is then necessary to add a constant term and use the right coefficients to make the transmission of the screen between 0 and 1. We result in:

$$s_\phi(x,y) = \frac{1}{2} + \frac{1}{4}\{L_\phi(x,y) + L_{-\phi}(x,y)\} = \frac{1}{2} + \frac{1}{2}\cos\left(\frac{\pi(x^2+y^2)}{\lambda\phi}\right). \qquad (4.3)$$

Such a screen acts as a converging lens of focal length ϕ, but with a poor transmission (1/4 in amplitude, 1/16 in intensity). It will also act as a diverging lens and as a simple screen of uniform transmission. A different combination of lenses leads to a transmission with a sine term.

The variable transmission of such screens is very difficult to manufacture with precision. It is easier to make a screen of binary transmission (1 or 0). This can be done for example by the following transmission:

$$\begin{aligned} S_\phi(x,y) &= \frac{1}{2} + \frac{2}{\pi}\sum_{n=0}^{\infty}\frac{1}{2n+1}\sin\left(\pi(2n+1)\,\frac{x^2+y^2}{\lambda\phi}\right) = \frac{1}{2} + \\ &\frac{1}{i\pi}\sum_{n=0}^{\infty}\frac{1}{2n+1}\left\{\exp\left(i\pi(2n+1)\,\frac{x^2+y^2}{\lambda\phi}\right) - \exp\left(-i\pi(2n+1)\,\frac{x^2+y^2}{\lambda\phi}\right)\right\} \\ &= \frac{1}{2} + \frac{i}{\pi}\sum_{n=0}^{\infty}\frac{1}{2n+1}\{L_{\phi/(2n+1)}(x,y) - (L_{-\phi/(2n+1)}(x,y)\}. \end{aligned} \qquad (4.4)$$

The transmission of such a screen is given in Figure 6 (top left). Its efficiency to focus in the plane $z = \phi$ is slightly improved (from 1/4 to $1/\pi$) at the expense of an infinite number of converging and diverging lenses (of focal lengths $\phi/(2n+1)$). A few of these ghost focal planes are shown in Figure 6 (experimental results).

These systems may found interesting applications at wavelengths for which it is difficult to manufacture classical lenses or mirrors. It is interesting to note that screens based on this principle have been proposed also for astronomical applications in the visible domain by Koechlin *et al.* (2009).

5 Fresnel diffraction and shadowing in astronomy: Application to coronagraphy

5.1 Fresnel diffraction with complementary screens

Let us consider two complementary screens of the form $t(x,y)$ and $1 - t(x,y)$. The amplitude diffracted by the complementary screen is 1 minus the diffracted

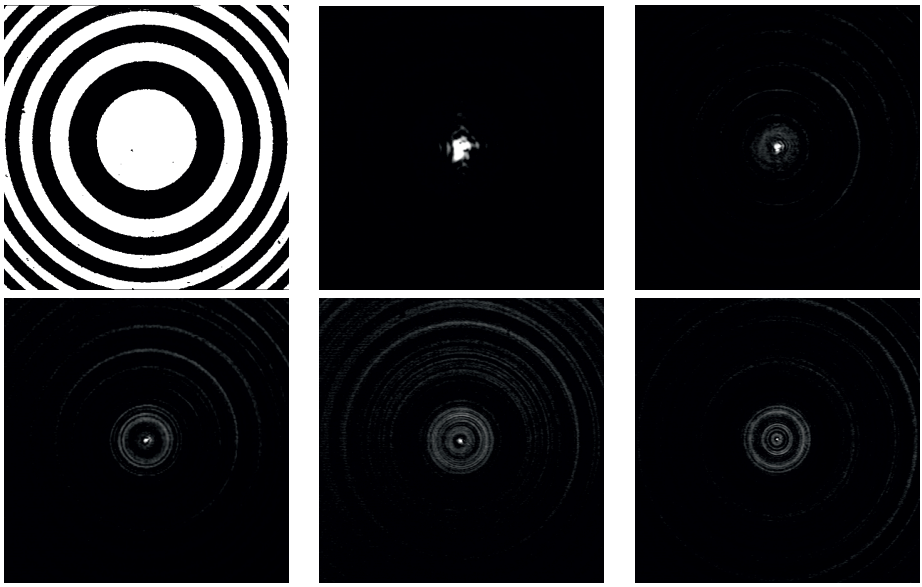


Fig. 6. *Top row*, from *left* to *right*: image of a Soret screen, intensity distribution in the plane $z = \phi$ and $z = \phi/3$ (corresponding to the focal planes of the lenses $L_{\phi/(2n+1)}$ for n=0,1 in Eq. (4.4)). *Bottom row*: intensity distribution at $z = \phi/5$, $z = \phi/7$ and $z = \phi/9$ (terms n=2,3,4 of the sum). The intensity of the central spot decreases with n as predicted.

amplitude from $t(x, y)$. Indeed, at a distance z, we have for an incident plane wave of unit amplitude:

$$(1 - t(x, y)) * D_z(x, y) = 1 - t(x, y) * D_z(x, y) \tag{5.1}$$

a property which is sometimes confused with Babinet's principle in the literature (see Cash 2011, for example).

5.2 Diffraction with rectangular apertures

The diffraction of rectangular diaphragms (infinite edge, slit, square or rectangle) can be easily computed making use of the separability in x and y of these functions and the corresponding properties of the convolution. Indeed, if the transmission $t(x, y)$ can be written as $t_x(x) \times t_y(y)$, then:

$$D_z(x, y) * t(x, y) = D_z^0(x) * t_x(x) \times D_z^0(y) * t_y(y). \tag{5.2}$$

In these cases, many problems find a solution using the Fresnel integrals $C(x)$ and $S(x)$, that can be defined as:

$$F(x) = C(x) + iS(x) = \int_0^x \exp\left(i\frac{t^2}{2}\right) dt. \tag{5.3}$$

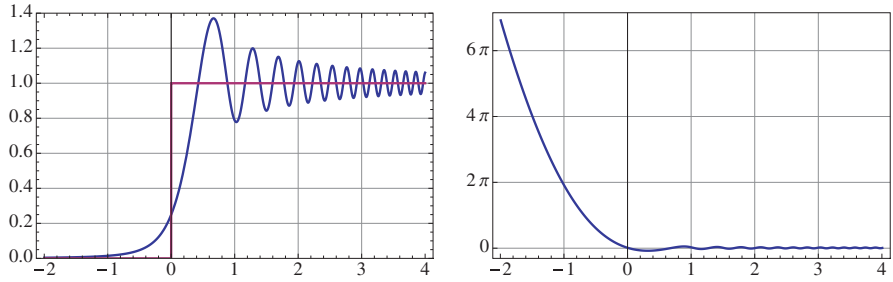


Fig. 7. *Left*: normalized intensity and *right*: phase (unwrapped) of the Fresnel diffraction of an infinite edge $H(x)$, outlined in the left figure. The observing plane is at 1 m from the screen, the wavelength is 0.6 μm. The x-axis is in mm.

The complex amplitude diffracted by an edge is obtained computing the convolution of $D_z(x, y)$ with the Heaviside function $H(x)$ for all x and y (we may denote its transmission as $t(x, y) = H(x) \times 1(y)$ for clarity). We have:

$$A_H(x, y) = D_z(x, y) * H(x) = D_z^0(x) * H(x) = \frac{1}{2} + \frac{1}{\sqrt{2i}} F\left(x\sqrt{\frac{2}{z\lambda}}\right). \qquad (5.4)$$

The intensity and the phase of the wave are given in Figure 7. The intensity is very often represented in Fresnel diffraction, but this is not the case for the phase. The rapid increase of phase in the geometrical dark zone may be heuristically interpreted as a tilted wavefront, the light coming there originating from the bright zone.

Similarly, the free space propagation of the light for a slit of width L can be directly derived from the above relation assuming that the transmission is $t(x, y) = H(x + L/2) - H(x - L/2)$. We have:

$$A_{L+}(x, y) = \frac{1}{\sqrt{2i}} \left\{ F\left(\frac{2x + L}{\sqrt{2\lambda z}}\right) - F\left(\frac{2x - L}{\sqrt{2\lambda z}}\right) \right\} \qquad (5.5)$$

where the subscript $L+$ stands for a clear slit of width L. Graphical representations of the corresponding intensity and phase are displayed in Figure 8.

As already written, the Fresnel diffracted amplitude from the complementary screen can be obtained as 1 minus the diffracted amplitude from the slit. It can be also be written as the sum of the diffraction of two bright edges $H(x - L/2)$ and $H(-x - L/2)$. Then it is clear that ripples visible in the shadow of the slit are due to phase terms produced by the edges. We have:

$$\begin{aligned} A_{L-}(x, y) &= 1 - \frac{1}{\sqrt{2i}} \left\{ F\left(\frac{2x + L}{\sqrt{2\lambda z}}\right) + F\left(\frac{2x - L}{\sqrt{2\lambda z}}\right) \right\} \\ &= \frac{1}{\sqrt{2i}} F\left(\frac{-2x - L}{\sqrt{2\lambda z}}\right) + \frac{1}{\sqrt{2i}} F\left(\frac{2x - L}{\sqrt{2\lambda z}}\right) \end{aligned} \qquad (5.6)$$

where $L-$ stands for an opaque strip of width L.

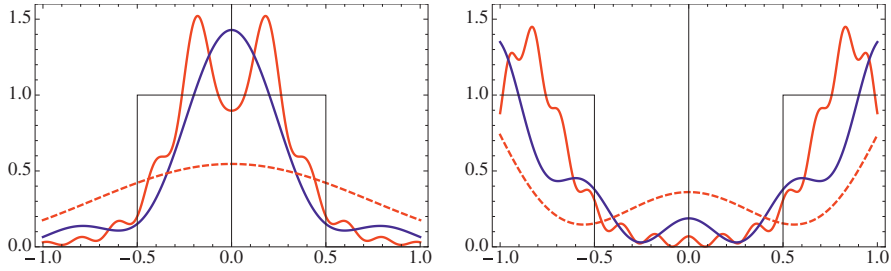


Fig. 8. Fresnel diffractions (in intensity) of a transmitting slit (*left*) and an opaque slit (*right*) of width 1 mm (slits are outlined in the figures). Observing planes are at 0.3 m (red), 1 m (blue) and 3 m (dashed) from the screen, the wavelength is 0.6 μm.

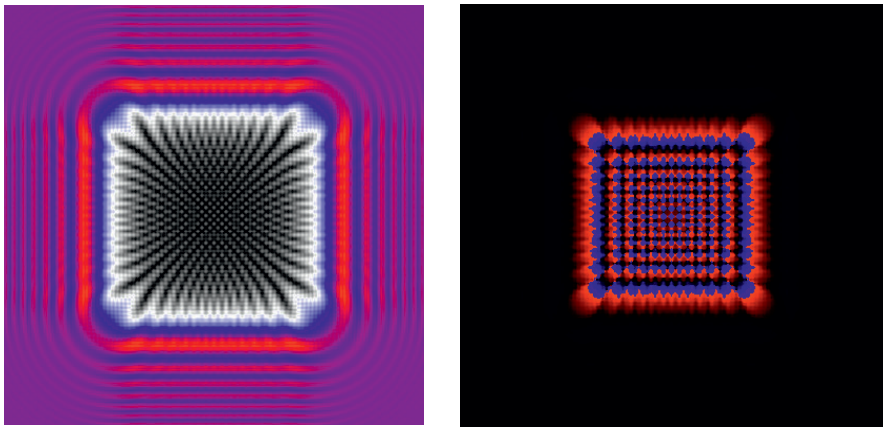


Fig. 9. Fresnel diffraction (*left*: amplitude, *right*: phase) of a square occulter of side 50 m at 80 000 km, with $\lambda = 0.6$ μm. The region represented in the figures is 100 m $\times$ 100 m. The color scale for the amplitude (black, white, blue, red) is chosen so as to highlight the structures in the dark zone of the screen. The color scale for the phase is blue for $-\pi$, black for 0 and red for $+\pi$ (the phase is not unwrapped here).

A transmitting square (or rectangle) aperture can be written as the product of two orthogonal slits. Therefore the Fresnel diffraction of the open square $A_{L^2+}(x, y)$ is the combination of two Fresnel diffractions in x and y. This property of separability is no longer verified for the diffraction of the opaque square $A_{L^2-}(x, y)$, which transmission must be written as 1 minus the transmission of the open square. We have:

$$\begin{aligned} A_{L^2+}(x, y) &= A_{Lx+}(x, y) \times A_{Ly+}(x, y) \\ A_{L^2-}(x, y) &= 1 - A_{L^2+}(x, y). \end{aligned} \tag{5.7}$$

We give in Figure 9 an example of the amplitude and phase of the wave in the shadow of a square occulter of 50×50 meters at a distance of 80 000 km, and that

could possibly be used for exoplanet detection. These parameters are compatible with the observation of a planet at about 0.1 arcsec from the star (a Solar – Earth system at 10 parsec) with a 4-m telescope. It is however interesting to note the strong phase perturbation in the center of the shadow, while it is almost zero outside. In such an experiment, the telescope is set in the center of the pattern to block the direct starlight, and the planet is observable beyond the angular dark zone of the occulter. The level of intensity in the central zone is of the order of 10^{-4} of that of the direct light. This is still too bright to perform direct detection of exoplanets: the required value is 10^{-6} or less. To make the shadow darker it would be necessary to increase the size of the occulter and the distance between the occulter and the telescope.

5.3 Fresnel diffraction with a circular occulter: The Arago-Poisson spot

The transmission of a circular occulter of diameter D can be written as $1-\Pi(r/D)$, where $r = \sqrt{x^2+y^2}$, and $\Pi(r)$ is the rectangle function of transmission 1 for $|r| < 1/2$ and 0 elsewhere. Since the occulter is a radial function, its Fresnel diffraction is also a radial function that can be written as:

$$A_D(r) = 1 - \frac{1}{i\lambda z}\exp\left(i\pi\frac{r^2}{\lambda z}\right)\int_0^{D/2} 2\pi\xi\,\exp\left(i\pi\frac{\xi^2}{\lambda z}\right)\,J_0\left(2\pi\frac{\xi r}{\lambda z}\right)d\xi \qquad (5.8)$$

where $J_0(r)$ is the Bessel function of the first kind. Here again, the Fresnel diffraction from the occulter writes as 1 minus the Fresnel diffraction of the hole. At the center of the shadow we have $A_D(0) = \exp[i\pi D^2/(4\lambda z)]$ and we recover the value of 1 for the intensity.

Obtaining the complete expression of the wave for any r value is somewhat tricky. The integral of Equation (5.8) is a Hankel transform that does not have a simple analytic solution. A similar problem (the wave amplitude near the focus of a lens) has been solved by Lommel, as described by Born & Wolf (2006). It is possible to transpose their approach to obtain the Fresnel diffraction from a circular occulter.

After a lot of calculations, we obtain the result in the form of alternating Lommel series for the real and imaginary parts of the amplitude. The result can be represented in a concise form as:

$$\begin{aligned}
&&\Psi(r) = \\
r < D/2: &\quad A\exp\left(i\frac{\pi r^2}{\lambda z}\right)\exp\left(i\frac{\pi D^2}{4\lambda z}\right)\times\sum_{k=0}^{\infty}(-i)^k\left(\frac{2r}{D}\right)^k J_k\left(\frac{\pi D r}{\lambda z}\right)\\
r = D/2: &\quad \frac{A}{2}\left[1+\exp\left(i\frac{\pi D^2}{2\lambda z}\right)J_0\left(\frac{\pi D^2}{2\lambda z}\right)\right]\\
r > D/2: &\quad A - A\exp\left(i\frac{\pi r^2}{\lambda z}\right)\exp\left(i\frac{\pi D^2}{4\lambda z}\right)\times\sum_{k=1}^{\infty}(-i)^k\left(\frac{D}{2r}\right)^k J_k\left(\frac{\pi D r}{\lambda z}\right).
\end{aligned} \qquad (5.9)$$

Two expressions are needed to ensure the convergence of the sum depending on the value of $2r/D$ compared to 1. The convergence is fast except for the transition zone around $r \sim D/2$, and luckily there is a simple analytical form there. An upper bound of the series limited to n terms is given by the absolute value of the $n+1$ term, according to Leibniz' estimate.

An illustration of this formula is given in Figure 10 for an occulter of diameter 50 m, observed at a distance of 80 000 km, at $\lambda = 0.55$ μm, which corresponds to data for the exoplanet case. For this figure, we computed the series for 100 terms, which can be rapidly done using *Mathematica* (Wolfram 2012) and gives a sufficient precision everywhere. The Arago spot is clearly visible at the center of the diffraction zone. For $r \ll D$, the amplitude is fairly described by the only non-zero term of the Lommel series that is the Bessel function $J_0(\pi r D/(\lambda z))$. Its diameter is approximately $1.53\lambda z/D$.

As mentioned in the introduction, Arago's experience was reproduced during the CNRS school of June 2012. Fresnel diffraction patterns (intensity) of a small occulter, reproduced in Figure 1 show the Arago spot at their center.

Thus a circular screen is not a good occulter.

For the detection of exoplanets, several projects envisage petaled occulters (Arenberg *et al.* 2007; Cash 2011) and we give an illustration of the performances in Figure 11. The analytic study of circular occulters remains however of interest for solar applications. Indeed, because of the extended nature of the solar disk, it seems difficult to use shaped occulters there, even if serrated edge occulters have been envisaged for that application (Koutchmy 1988).

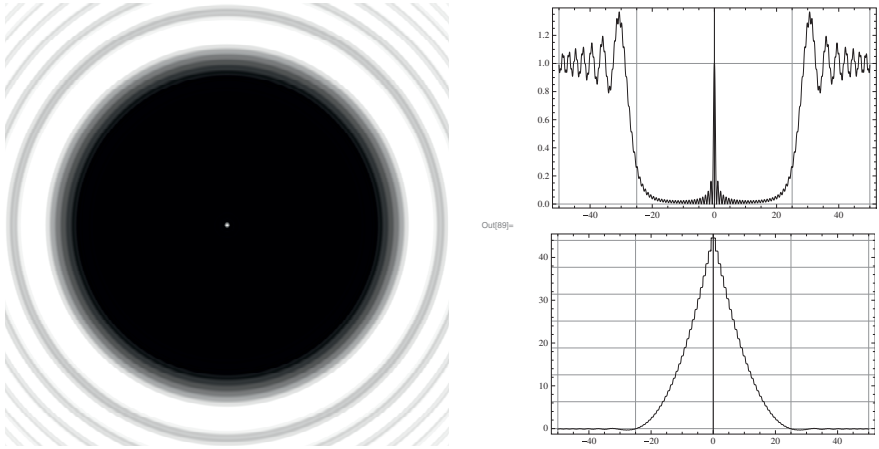


Fig. 10. Fresnel diffraction of an occulter of diameter 50 m, observed at a distance of 80 000 km, at $\lambda = 0.55$ μm. *Left*: 2D intensity, *top right*: central cut of the intensity, *bottom right*: central cut of the unwrapped phase. Notice the strong Arago spot at the center of the shadow and the important phase variation.

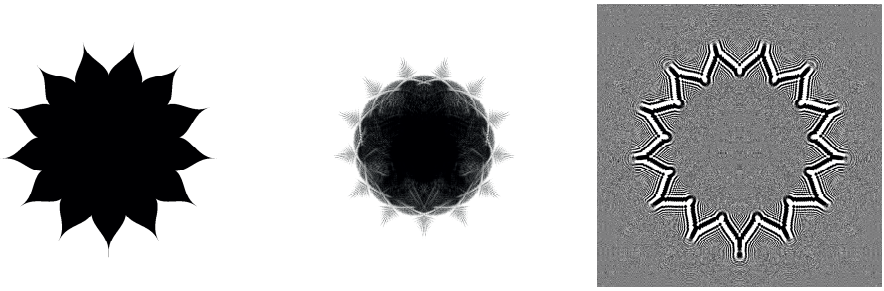


Fig. 11. Fresnel diffraction of an occulter with a petal shape. From *left* to *right*: the occulter, the intensity (×10) in the shadow and the phase. The parameters are the same as in Figure 10 $D = 50$ m, $z = 80\,000$ km, $\lambda = 0.55$ μm. The shadow at the center of the screen is much darker (no Arago spot) and the phase variation is weak there.

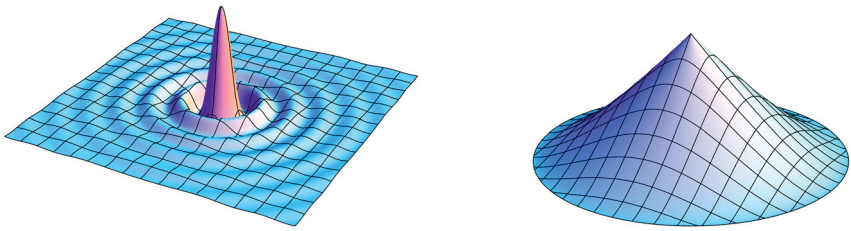


Fig. 12. Numerical 3D representation of the PSF (left), here an Airy function, and the corresponding OTF (right) of a perfect telescope with a circular entrance aperture.

6 Application to incoherent imaging in astronomy

The formation of an image at the focus of a telescope in astronomy can be divided into two steps, one corresponding to a coherent process leading to the point spread function (PSF) and the other corresponding to a sum of intensities, *e.g.* an incoherent process. Equation (4.2) makes it possible to write the PSF observed in the focal plane of the telescope as a function of the spatial (x, y) or angular $(\alpha = x/\phi, \beta = y/\phi)$ coordinates. For an on-axis point-source of unit intensity, we have:

$$\begin{aligned} R_\phi(x,y) &= \frac{1}{S\lambda^2\phi^2} \left| \hat{P}\left(\frac{x}{\lambda\phi}, \frac{y}{\lambda\phi}\right)\right|^2 \\ R(\alpha,\beta) &= \frac{1}{S\lambda^2} \left| \hat{P}\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}\right)\right|^2 \end{aligned} \tag{6.1}$$

where ϕ is the telescope focal length and $P(x, y)$ is the function that defines the telescope transmission. Aberrations or other phase defaults due to atmospheric turbulence can be included in the term $P(x, y)$. The division by the surface area

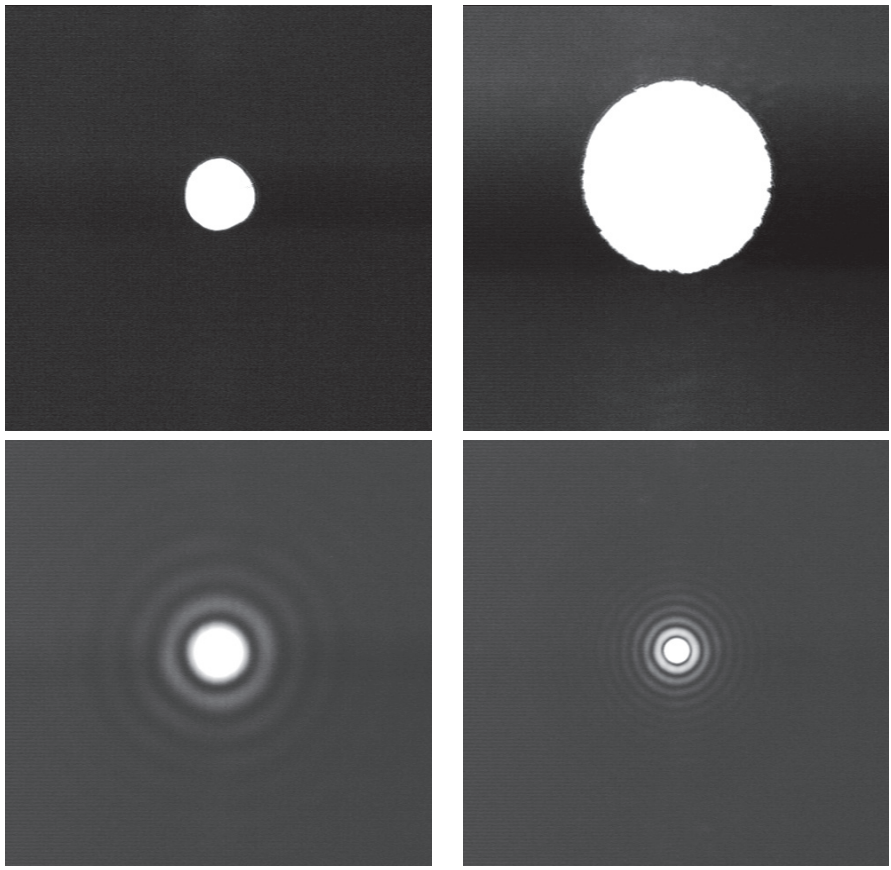


Fig. 13. Example of PSFs shown during the CNRS school using a simple optical setup. *Top*, circular apertures, *bottom*, corresponding PSFs. Note the inverse relationship between the size of the PSF and the aperture diameter.

S of the telescope allows the calculation of a normalized PSF. The normalizing coefficients ensure the energy conservation of the form:

$$\iint R_\phi(x,y)\,dxdy = \iint R(\alpha,\beta)\,d\alpha d\beta = \frac{1}{S}\iint |P(\xi,\eta)|^2\,d\xi d\eta = 1. \qquad (6.2)$$

We have made use of Parseval theorem to write the last equality. For a telescope of variable transmission, see Aime (2005).

It is convenient to consider angular coordinates independent of the focal length of the instrument. Each point of the object forms its own response in intensity shifted at the position corresponding to its angular location. This leads to a convolution relationship. The focal plane image is reversed compared to the object sense. By orienting the axes in the focal plane in the same direction as in the sky,

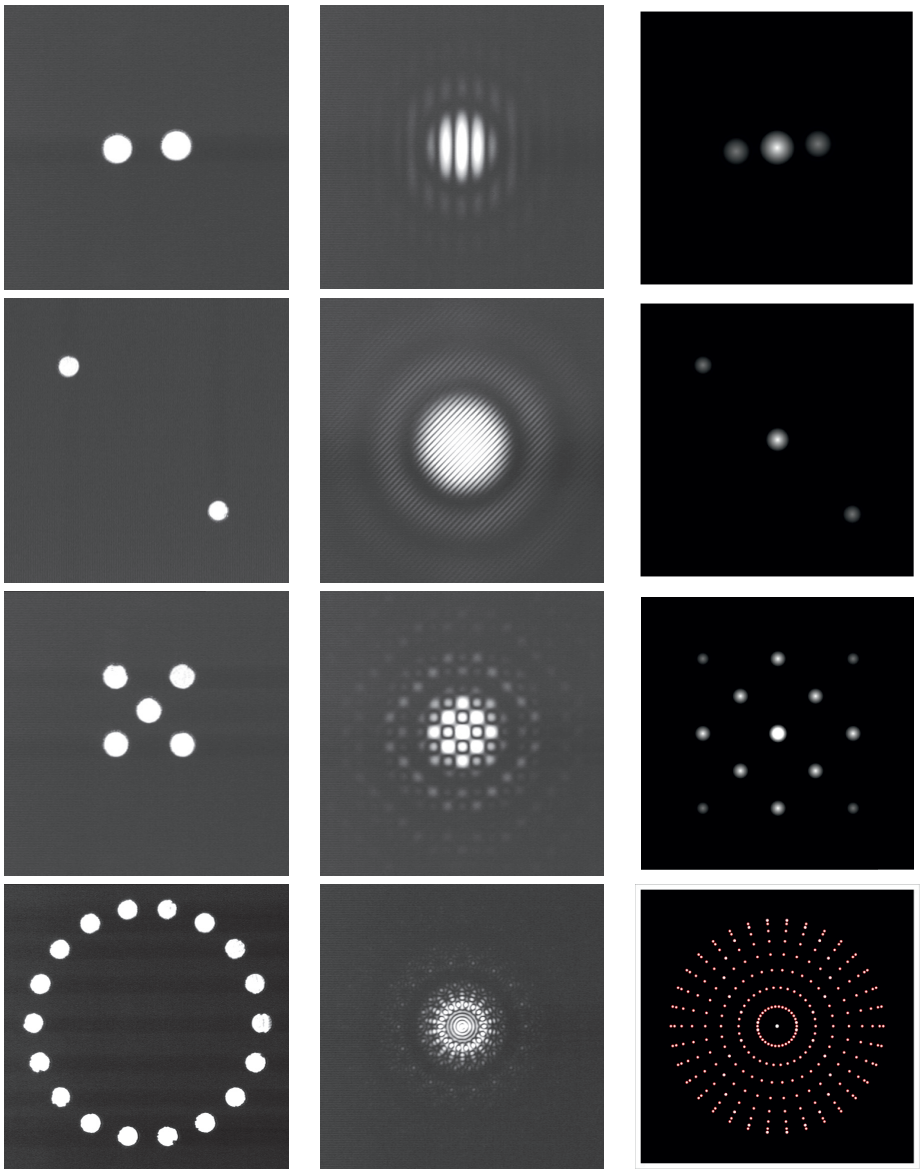


Fig. 14. From *left* to *right*: aperture, PSF and MTF. For the sake of clarity, the MTF corresponding to the 18-aperture interferometer is drawn for smaller elementary apertures than those of the left figure.

we obtain:

$$I(\alpha, \beta) = O(\alpha, \beta) * R(-\alpha, -\beta) \tag{6.3}$$

where $O(\alpha, \beta)$ is the irradiance of the astronomical object. The Fourier transform of $R(-\alpha, -\beta)$ gives the optical transfer function (OTF) $T(u, v)$:

$$T(u,v) = \mathcal{F}[R(-\alpha,-\beta)] = \frac{1}{S}\int\int P(x,y)P^*(x-\lambda u, y-\lambda v)dxdy \qquad (6.4)$$

where u and v are the angular spatial frequencies.

For a perfect circular aperture of diameter D operated at the wavelength λ, the PSF becomes the following radial function of γ:

$$R(\alpha,\beta) = \mathrm{R}(\gamma) = \left(2\frac{J_1(\pi\gamma D/\lambda)}{\pi\gamma D/\lambda}\right)^2 \frac{S^2}{\lambda^2} \qquad (6.5)$$

where $\gamma = \sqrt{\alpha^2+\beta^2}$, and the OTF is the radial function of $w = \sqrt{u^2+v^2}$:

$$T(u,v) = \mathrm{T}(w) = \frac{2}{\pi}\left(\arccos\left(\frac{\lambda w}{D}\right) - \frac{\lambda w}{D}\sqrt{1-\left(\frac{\lambda w}{D}\right)^2}\right). \qquad (6.6)$$

This expression is obtained computing the surface common to two shifted discs. The OTF looks like a Chinese-hat, with a high frequency cutoff $w_c = D/\lambda$.

Examples of PSFs for various apertures presented during the CNRS school are given in Figure 14. The corresponding MTFs shown in the same figure are computed numerically.

7 Conclusion

This presentation aimed at introducing the formalism for Fresnel's diffraction theory, widely used in optics and astronomy.

Besides analytical derivation of basic relationships involving instrumental parameters, visual illustrations using laboratory demonstrations are given, as was presented during the CNRS school. Most of these are basic in the field of image formation and are frequently met in astronomy. A few of them concerning the shadows produced by the screens are seldomly addressed in the astronomical literature up to now, though they presently are emerging topics. Demonstrations are made using laboratory material for students in Physics: a laser and a beam expander, various transmitting or opaque screens and a detector.

The paper begins with a historical background leading to the current context. Then analytical derivations, based on the Huyghens-Fresnel principle, using wavefronts and complex amplitudes are presented, providing expressions for the free space propagation of light. Plenty use is made of convolution relationships and filtering aspects.

Fresnel's diffraction is illustrated through some situations, such as the propagation after a screen with sinusoidal transmission function, or such as shadowing produced by occulters set on the pointing direction of a telescope for coronagraphy. Here are met such effects as the so-called Poisson-Arago spot, and diffraction

by sharp edges (rectangular or circular screens). The use of focusing screens have been considered as well. Along that way, expressions of diffracted amplitudes are given for various shapes of apertures. Then, the Fourier transform properties of lenses and binary screens (made of transparent and opaque zones, *i.e.* transmission function being 0 or 1 accordingly) are presented.

The paper ends with a section describing incoherent imaging in astronomy and dealing with PSFs (intensity response of the instrument to a point-like source) and MTFs (a link with linear filtering). Images of PSFs obtained with the demonstration set-up, are presented for various shapes and configurations of collecting apertures: from single disk to diluted apertures (several sub-pupils) as used in aperture synthesis with several telescopes. Besides, illustrations for associated MTFs are obtained by computation.

The paper could hopefully be used either as a reminder or as an introduction to the basics of the image formation process in the context of diffraction theory.

The authors wish to thank Dr S. Robbe-Dubois for her critical reading of the manuscript.

References

Aime, C., 2005, A&A, 434, 785

Arenberg, J.W., Lo, A.S., Glassman, T.M., & Cash, W., 2007, C.R. Physique, 8, 438

Born, M., & Wolf, E., 2006, Principles of Optics, 7th Ed. (Cambridge University Press), 484

Cash, W., 2011, ApJ, 738, 76

Koechlin, L., Serre, D., & Deba, P., 2009, Ap&SS, 320, 225

Françon, M., 1979, Optical image formation and processing (New York: Academic Press)

Goodman, J.W., 1985, Statistical Optics (New York, NY: John Wiley and Sons)

Goodman, J.W., 2005, Introduction to Fourier Optics (Roberts and Company Publishers)

Koutchmy, S., 1988, Space Sci. Rev., 47, 95

Labeyrie, A., 1970, A&A, 6, 85

Lamy, P., Damé, L., Vivès, S., & Zhukov, A., 2010, SPIE, 7731, 18

Nazarathy, M., & Shamir, J., 1980, J. Opt. Soc. Am., 70, 150

de Senarmont, M., Verdet, E., & Fresnel, L., 1866, Oeuvres complètes d'Augustin Fresnel (Paris, Imprimerie Impériale)

Wolfram Mathematica, 2012, Wolfram Research, Inc., Champaign, IL

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 59–76*

ASTRONOMICAL IMAGING... ATMOSPHERIC TURBULENCE? ADAPTIVE OPTICS!

M. Carbillet[1]

Abstract. This course/paper deals with adaptive optics in the framework of astronomical imaging. It does not pretend to be an exhaustive course of astronomical adaptive optics. It is rather intended to give an introductory overview of it, from my very partial point-of-view.

1 Preamble: Images & turbulence

The image formed at the focus of ground-based telescopes is perturbed mainly by the last 10–20 km traveled by the light from the observed astronomical object, when propagating through the turbulent atmosphere. One has for the resulting image, and at the same time: scintillation, agitation, and spreading.

Scintillation is due to fluctuations of the global intensity of the image, this is the easily observed twinkling of stars. Agitation is the global variation of the photocenter of the formed image, which is due to tip and tilt of the incoming wavefront. Finally, spreading is due to the loss of spatial coherence of the incoming wavefront.

1.1 Object-image relationship

The object-image relationship which links the illumination $I(\boldsymbol{\alpha})$, in the focal plane of the telescope, where α is a bidimensional angular vector describing the line of sight, to the luminance $O(\boldsymbol{\alpha})$ of the object in the sky is a convolution implying the point-spread function (PSF) $S(\boldsymbol{\alpha})$ of the ensemble telescope$\oplus$atmosphere:

$$I(\boldsymbol{\alpha}) = O(\boldsymbol{\alpha}) * S(\boldsymbol{\alpha}). \tag{1.1}$$

This relationship is valid notably at the condition that the system is invariant by translation, *i.e.* everything happens within the isoplanatic domain...

[1] UMR 7293 Lagrange, UNS/CNRS/OCA, Bât. Fizeau, Parc Valrose, 06100 Nice, France; e-mail: marcel.carbillet@unice.fr

DOI: 10.1051/eas/1359004

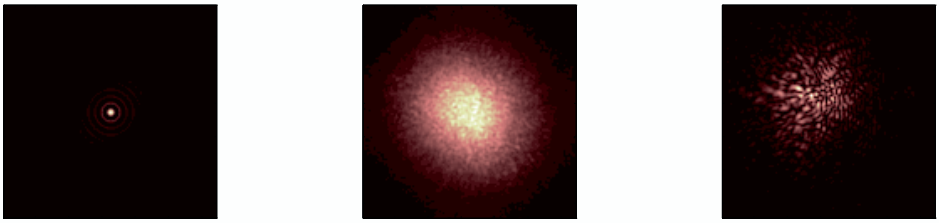


Fig. 1. Example of atmospherically-perturbed PSFs observed at the focus of a large ground-based telescope. From *left* to *right*: "ideal" Airy disc, long-exposure actually-observed PSF, and short-exposure actually-observed PSF. (From Carbillet 1996.)

Figure 1 shows an example of atmospherically-perturbed PSFs that can be observed at the focus of a ground-based telescope. The difference between the expected "ideal" Airy disc, the long-exposure actually-observed PSF (*i.e.* the image of an unresolved object), and the short-exposure actually-observed PSF (a speckle image) is dramatic. The Airy disc is showing a core of full-width at half-maximum (FWHM) λ/D, where λ is the observing wavelength and D the telescope diameter. The long-exposure actually-observed PSF is showing a core of FWHM λ/r_0, where r_0 is the typical size of the spatial coherence cells at the entrance of the telescope pupil (also called Fried parameter and detailed latter on – see next subsection). And the short-exposure actually-observed PSF is showing a speckle pattern which is changing very rapidly due to the time behavior of the turbulence.

1.2 Some basic numbers

Some basic numbers concerning the physical parameters driving the spatial and temporal behaviors of the atmospheric turbulence have to be remembered, in particular with respect to the observing wavelength λ.

Concerning spatial coherence, the basic factor over which everything is then built is the well-known Fried parameter r_0. This fundamental parameter directly gives the resulting angular resolution at the focal plane of the telescope: λ/r_0, quantity which is clearly independent of the telescope diameter D (as far as D is greater than r_0). In addition, r_0 being weakly dependent on the observing wavelength λ (in fact r_0 is proportional to $\lambda^{6/5}$), this angular resolution (*i.e.* the FWHM of the resulting PSF) is roughly independent of λ too. Writing down numbers, a typical r_0 of 10 cm in the visible (at 500 nm) would correspond to 60 cm in the K band (2.2 μm) and both would roughly correspond to a FWHM of the PSF of $\sim$1 arcsec.

Concerning temporal coherence, the basic physical limitation comes this time from atmospheric turbulence layers velocity v, leading to an evolution time $\tau_0 \simeq r_0/v$. As it can be seen, τ_0 is independent of D but strongly dependent on λ. Typically $\tau_0 \simeq 3$ ms at a wavelength of 500 nm and 18 ms at 2.2 μm.

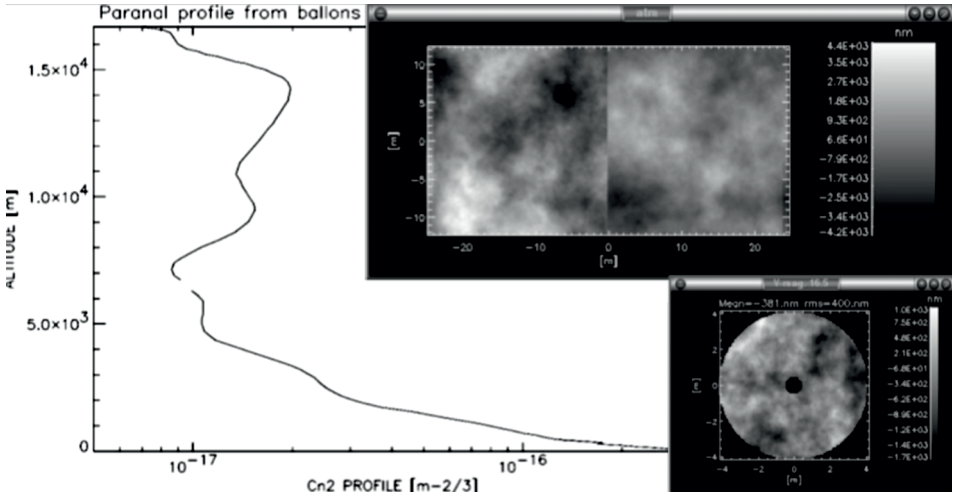


Fig. 2. Atmospheric turbulence. *Left*: typical atmospherical turbulent profile from a Mount Paranal site testing campaign (Sarrazin 1996). *Top right*: two 24 m × 24 m modeled *Kolmogorov/von Kármán* turbulent layers, in terms of optical path difference. *Bottom right*: resulting wavefront propagated through these turbulent layers and to an 8-m telescope pupil. (From Carbillet 2006.)

Figure 2 shows a typical atmospherical turbulence profile, where layers are clearly identifiable, together with the representation of two modeled *Kolmogorov/von Kármán* turbulent layers and the resulting wavefront propagated through these turbulent layers and to the telescope pupil.

1.3 Some basic equations

The wavefront (measured in meters) is, by definition, proportional to the phase $\Phi(\boldsymbol{r})$ (measured in radians) by a factor $\frac{\lambda}{2\pi}$. And $\Phi(\boldsymbol{r})$ is itself linked to the wave $\Psi(\boldsymbol{r})$, which traveled through the turbulent atmosphere, by the relation:

$$\Psi(\boldsymbol{r}) = A(\boldsymbol{r}) \exp \imath\Phi(\boldsymbol{r}), \tag{1.2}$$

where A is the amplitude of the wave and $\boldsymbol{r}$ the bidimensional coordinate. Moreover, the phase $\Phi(\boldsymbol{r})$ can be decomposed on a polynomial basis, like for example the Zernike one, such as:

$$\Phi(\boldsymbol{r}) = \sum_i a_i Z_i(\boldsymbol{r}), \tag{1.3}$$

where $Z_i(\boldsymbol{r})$ represents the i-th Zernike polynomial and a_i its related coefficient.

In addition to this general definition of wavefront and phase, one has to consider at least the principal equations which are ruling the atmospheric turbulence. The

first one concerns the Fried parameter r_0, which is defined as (Roddier 1981):

$$r_0 = 0.185\ \lambda^{\frac{6}{5}} \cos\gamma^{\frac{3}{5}} \left[\int_0^{\infty} C_n^2(z)dz\right]^{-\frac{3}{5}}, \tag{1.4}$$

where γ is the zenith angle and $C_n^2(z)$ is the structure constant of the fluctuations of the air refraction index n, which characterizes the optical energy of turbulence in function of the altitude z.

A number of typical parameters characterizing the resulting speckle pattern can be then deduced from it, such as the typical coherence time τ, defined as (Roddier 1981):

$$\tau_0 = 0.36\ \frac{r_0}{v}, \tag{1.5}$$

or alternatively (Aime *et al.* 1986):

$$\tau_0 = 0.47\ \frac{r_0}{v}, \tag{1.6}$$

where v is the mean velocity of the turbulent layers forming the turbulent atmosphere (weighted by the turbulence profile $C_n^2(z)$); but also the resulting "seeing":

$$\epsilon = 0.98\ \frac{\lambda}{r_0}; \tag{1.7}$$

and the typical isoplanatic patch:

$$\theta_0 = 0.36\ \frac{r_0}{h}, \tag{1.8}$$

where h is the mean height of the turbulent layers (weighted as well by the turbulence profile $C_n^2(z)$).

Finally, the wavefront perturbed by the turbulent atmosphere has a power spectral density which is classically modeled by (within the *Kolmogorov/von Kármán* model):

$$\Phi_\phi(\boldsymbol{\nu}) = 0.0228\ r_0^{-\frac{5}{3}} \left(\nu^2 + \frac{1}{\mathcal{L}_0^2}\right)^{-\frac{11}{6}}, \tag{1.9}$$

where ν is the spatial frequency and $\mathcal{L}_0$ is the outer scale of turbulence (with a typical median value of 20-30 m for mid-latitude sites).

1.4 The craftiness of speckle imaging and Lucky Imaging

Before that the use of adaptive optics (AO) became a common thing for astronomy (since the first very convincing results of the mid-90's of last century), speckle imaging techniques were used in order to obtain high-angular resolution (HAR) images on large ground-based telescopes in the visible and near-infrared domains. A number of results were obtained, using first the pioneering visibility technique proposed by Labeyrie (Labeyrie 1970) and various others in the following – from

the somehow raw shift-and-add technique (Worden *et al.* 1976) to more refined ones offered by bispectral imaging (Weigelt 1977), probability imaging (Aime 1987; Carbillet *et al.* 1998), cross-correlation (Aristidi *et al.* 1997) and others. The main idea under these techniques is that atmospheric perturbations can be frozen if the time exposure is less than τ_0, and then some statistical invariant can be computed on a series of such short-exposure images in order to retrieve informations about the observed object.

Note that a selection of images can be done in order to select the best ones from a series of observations. Such observations were usually made of some thousands of images of a few milliseconds exposure, as many of the object that of an unresolved reference star, in order to obtain an estimate of the quantity which is computed also for the object images – *e.g.* spectrum, bispectrum, high-order probability density function, etc.. This idea is also basically the one under the Lucky Imaging (LI) technique (Baldwin *et al.* 2001) wich is commonly used since the advent of almost-readout-noise-free Electron-Multiplying CCD (EMCCD) detectors, and that is considered also for post-AO images (Mackay *et al.* 2012) for short (visible) wavelengths (were the AO correction is, at least for now, very partial).

2 Adaptive optics

The main problem of the previously described techniques is that the exposure time is limited, especially when considering classical CCD readout-noise-limited detectors, limiting hence sensitivity, signal-to-noise ratio, limiting magnitudes, and the like. Unlike AO, which in principle permits long exposure images (or spectra, or any other kind of data).

2.1 Some basic numbers

AO being designed to compensate atmospheric turbulence, the numbers evoked before (in terms of r_0 and τ_0) are directly the first bricks of any AO instrument study. The typical size d of each correcting element of a deformable mirror (DM) aimed to compensate the turbulence effects on the propagated wavefront, it hence follows that $d \simeq r_0$. As a consequence the total number of correcting elements becomes roughly $(D/r_0)^2$ which, with $D \simeq 10$ cm, is translated into approximately 7500 elements for a correction in the visible band, and 200 elements in K band. The same typical numbers are valid not only for correction (*i.e.* for the DM) but also, indeed, for what concerns the sensing of the incoming wavefront, through a given device (a wavefront sensor – WFS).

Temporal aspects are also very critical, since one would need to sample atmospheric turbulence at, let me say, a tenth of τ_0. This leads to typical temporal frequencies for the whole AO system of 1 kHz at 500 nm and 200 Hz at 2.2 μm.

Figure 3 schematizes the operation of a typical AO system: a perturbed wavefront enters the telescope, is reflected on a DM, sent to a beamsplitter dividing the light dedicated to the scientific device (a CCD, a spectrometer, whatever) and a WFS from which the collected information (*e.g.* spot centroids for computing

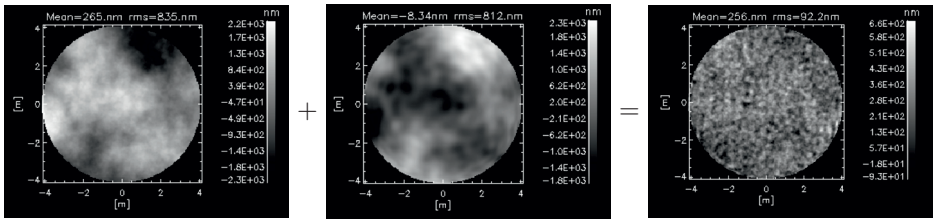


Fig. 3. AO system concept. *Left*: the incoming wavefront. *Center*: the DM shape, resulting from the commands sent from the wavefront reconstructor after analysis of the information collected from the WFS. *Right*: the resulting wavefront, after reflection of the input wavefront on the DM. (From Carbillet 2006.)

local wavefront slopes) is sent to a wavefront reconstructor that will elaborate DM commands from.

2.2 The great variety of AO "concepts" and their observational reasons

We easily understand from the previous section that observing at HAR, at least with a monolithic telescope and not considering speckle techniques, needs an AO system, and that this AO system has to be dimensioned in function of the observing wavelength considered.

This is true but makes abstraction of a number of problems. The first of all is the number of photons necessary for wavefront analysis, or sensing. The analysis temporal frequency being necessarily very high (between 200 Hz and 1 kHz in the illustration numbers given before), very bright stars are mandatory, dramatically reducing the portion of sky available for astrophysical observations (sky coverage).

One goal would then be to overcome this limitation in some way and have a 100% sky coverage. This is the goal of laser-guide-star (LGS) AO systems, which aim is to provide a sufficiently bright star in any direction of the sky, the closer possible to the observed object (Labeyrie & Foy 1985). These artificial AO guide stars, are usually formed either from backscattering of the atmospheric sodium layer (situated at an altitude of 90–100 km) or from Rayleigh scattering of the lower atmosphere (up to $\simeq$40 km). This technique rises a number of additional problems, from the huge necessary power of the employed laser itself, to effects linked to the fact that the star is formed at a finite distance from the telescope, that it is clearly extended, and that tip-tilt is hardly sensed (the same tip-tilt being encountered in the upwards and subsequent downwards propagation).

A second problem which had been darkened till here is the problem of anisoplanatism, and hence very limited field of correction in which observing astrophysical objects around a suitable AO guide star. In order to limit this error and permit to decently observe faint objects a solution is to take advantage from a given number of possible surrounding guide stars nearby the interesting astrophysical object, hence considering multiple-reference AO systems. Such systems can be declined into at least three categories: multi-conjugate AO (MCAO) systems, ground-layer AO (GLAO) systems, and multiple-objects AO (MOAO) systems.

MCAO systems aim at (partially) correct various layers of the turbulent atmosphere via DMs conjugated at different altitudes. At the opposite, GLAO is mono-conjugate and, in a simpler manner, aims at giving images (or spectra) corrected from the ground-layer turbulence only, since a great part of turbulence usually takes place within this layer. More peculiar, MOAO aims at correcting small fields in some directions of interest, within a much larger sensed field, through dedicated mirrors (one per direction of interest) and a global multiple-reference wavefront sensing.

Finally, the need to observe at very high-contrast levels in addition to HAR leads to the so-called "eXtreme" AO (XAO) systems, in which the basic concept is identical to a standard AO system, but each single component is pushed to its ultimate capacities and the whole system needs to break a number of conceptual and technological barriers.

2.2.1 Importance of the observational goal

A given class of astrophysical objects has its own observational priorities, such as the need to be directly detected and possibly spectrally characterized even at very low spectral resolution in the case of exoplanets, or for faint galaxies to obtain its precise morphology. As a consequence, this leads to consider the corresponding dominant AO errors (anisoplanatism in the faint-galaxies case, everything but anisoplanatism in the exoplanets case), and hence implies to consider *ad hoc* AO system concepts... for the present two examples: clearly MCAO, GLAO, or MOAO, possibly LGS-based, for the faint galaxies, and XAO for the exoplanets.

2.3 The post-adaptive-optics error budget

The post-AO error budget, in terms of variance integrated over the whole wavefront, is easily modeled by the following equation:

$$\sigma^2_{\mathrm{post-AO}} = \sigma^2_{\mathrm{atmosphere}} + \sigma^2_{\mathrm{AO\ system}} + \sigma^2_{\mathrm{others}}, \tag{2.1}$$

where three basic quantities are present: the atmospheric error not considered by the AO system ($\sigma^2_{\mathrm{atmosphere}}$), the residual error from the AO system itself ($\sigma^2_{\mathrm{AO\ system}}$), and finally other types of error neither due to the atmosphere nor to the AO system ($\sigma^2_{\mathrm{others}}$). Let me now have a detailed look into these three error terms.

2.3.1 Errors not due to the (limited) adaptive optics correction

Independently from the AO system considered, a number of errors, from both the physics of the (turbulent) atmosphere and the telescope/instrument are present. For what concerns the instrumental part the remaining error can be detailed as follows:

$$\sigma^2_{\mathrm{others}} = \sigma^2_{\mathrm{calibration}} + \sigma^2_{\mathrm{aberrations}} + \dots \tag{2.2}$$

where we see that this error is mainly coming from aberrations within the light path not seen by the AO system ($\sigma^2_{\text{aberrations}}$), but also from possible calibration errors ($\sigma^2_{\text{calibration}}$).

For what concerns the atmospheric effects, a number of them are not corrected at all by a standard AO system, as it is the case simply for scintillation, diffraction effects, chromatic effects, and indeed anisoplanatism, leading to:

$$\sigma^2_{\text{atmosphere}} = \sigma^2_{\text{scintillation}} + \sigma^2_{\text{diffraction}} + \sigma^2_{\text{chromatism}} + \sigma^2_{\text{anisoplanatism}}. \tag{2.3}$$

Note that, in another hand, anisoplanatism is the main enemy when looking for wide-filed images, or simply faint objects far from a bright guide star.

2.3.2 Errors due to the (limited) adaptive optics correction

Within the AO-system error budget, a number of error sources can be identified, leading to the following formulation:

$$\begin{aligned} \sigma^2_{\text{AO system}} &= \sigma^2_{\text{fitting}} + \sigma^2_{\text{aliasing}} + \sigma^2_{\text{measure}} + \sigma^2_{\text{temporal}} \\ &+ \sigma^2_{\text{LGS}} + \sigma^2_{\text{MCAO}}. \end{aligned} \tag{2.4}$$

We will not detail here the last two terms which are strictly relevant to a LGS-based AO system (σ^2_{LGS}) and an MCAO system (σ^2_{MCAO}), respectively. Other specific errors can be defined if a specific AO system is considered.

Fitting Error. The first term of Equation (2.4) concerns the correction itself: $\sigma^2_{\text{fitting}}$. It translates the fact that a limited range of spatial frequencies, and hence atmospheric turbulence modes, can be physically corrected by the mirror, and then the possible mirror modes. The reason is obvious and is simply linked to the total number of actuators building up the considered mirror. This error is consequently expressed in function of the ratio between the inter-actuators mean distance d_{DM} and the Fried parameter r_0 (in the imaging band considered):

$$\sigma^2_{\text{fitting}} \propto \left(\frac{d_{\text{DM}}}{r_0}\right)^{\frac{5}{3}}, \tag{2.5}$$

the exact coefficient of proportionality depending on the mirror construction itself and its ability to mimic atmospheric deformations.

Note that it is worthwhile to look at this error not only in terms of global average over the DM, but also in terms of spatial distribution, especially for segmented DMs. Figure 4 shows an example of the computed fitting error for the adaptive mirror M4 studied for the European Extremely Large Telescope (EELT), for median turbulence conditions (*i.e.*, roughly speaking, the median value of r_0) for the EELT site.

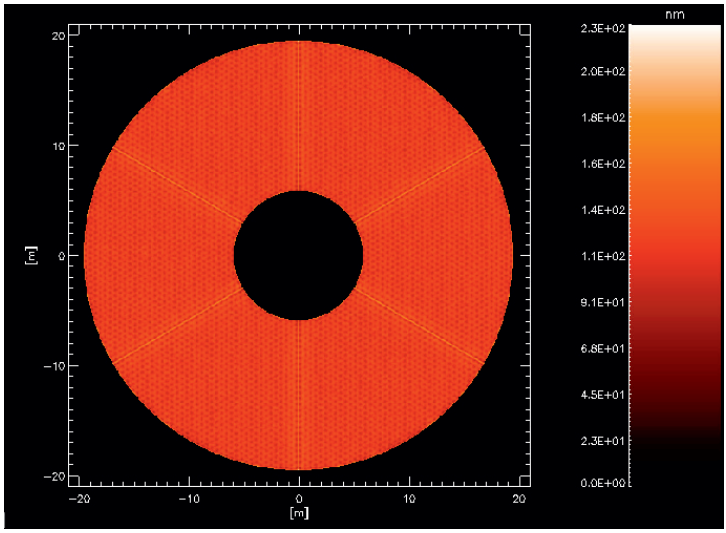


Fig. 4. Mean map of the residual rms wavefront, from which the fitting error can be deduced, for the adaptive mirror M4 of the EELT. (From Carbillet *et al.* 2012.)

Aliasing Error. The second error term of Equation (2.4) regards aliasing and is due to the WFS. Like the DM is limited by its finite number of actuators, the WFS is limited by its finite number of wavefront analysis elements (either the number of lenslets in the case of a Shack-Hartmann Sensor (SHS) or the number of pixels analyzing each of the 4 pupil images in the Pyramid Sensor (PS) case). Hence a problem of aliasing clearly appears because of the unseen spatial frequencies. Supposing that the physical size of the analysis elements of the WFS is $d_{\rm WFS}$, one has here also:

$$\sigma^2_{\rm aliasing} \propto \left(\frac{d_{\rm WFS}}{r_0}\right)^{\frac{5}{3}} . \tag{2.6}$$

Let me note that very often $d_{\rm WFS} \simeq d_{\rm DM}$, but also that the geometry can still be completely different (*e.g.* circular for the DM and square for the WFS).

Measurement Error. The third term of Equation (2.4) is also related to the WFS, and more precisely to the measurement itself done by the WFS. This is a classical problem of light detection by a CCD device, where $\sigma^2_{\rm measure}$ can be written:

$$\sigma^2_{\rm measure} = \sigma^2_{\rm photonization} + \sigma^2_{\rm read-out} + \sigma^2_{\rm dark-current} + \ldots \tag{2.7}$$

where the classical $\sigma^2_{\rm photonization}$ error is clearly inversely proportional to the number of photons available $N_{\rm photons}$:

$$\sigma^2_{\rm photonization} \propto \left(\frac{1}{N_{\rm photons}}\right) , \tag{2.8}$$

and where the read-out noise (RON) error expresses in function of N_{photons} and the associated variance σ_e^2 in terms of electrons/frame/second as:

$$\sigma^2_{\text{read}-\text{out}} \propto \left(\frac{\sigma_e^2}{N^2_{\text{photons}}} \right). \tag{2.9}$$

Note that other minor errors such as the dark-current one can be considered too, and that almost-RON-free detectors such as EMCCDs (Electron-Multiplying CCDs) present in counterpart an "exotic" noise characterized by a Gamma distribution (instead of a Poisson distribution for the photon noise or a Gaussian one for the RON, see Carbillet & Riccardi 2010).

Temporal Error. Last term evoked in Equation (2.4) is the one related to the global AO system temporal error, due to the simple fact that between the instant in which a given wavefront reflects on the DM and the instant in which it can be corrected by it (after measuring by the WFS, computing of the commands by the reconstructor and application of those commands by the DM), some milliseconds are usually gone. This error is indeed dependent on the turbulence coherence time τ_0 and the total AO system "integration $\oplus$ delay" time Δt_{AO}, and can be modeled as:

$$\sigma^2_{\text{temporal}} \propto \left(\frac{\Delta t_{\text{AO}}}{\tau_0} \right)^{\frac{5}{3}}. \tag{2.10}$$

Balancing the Errors. It is clear from this list of errors that the main error sources for which a technological effort has to be done are, at least:

① $\sigma^2_{\text{fitting}}$ when designing the DM,

② $\sigma^2_{\text{aliasing}}$ and $\sigma^2_{\text{measure}}$ when choosing which WFS with which specific options has to be realized,

③ and $\sigma^2_{\text{temporal}}$ when designing the whole AO loop.

Moreover, the critical physical parameters to be optimized are clearly:

① the inter-actuator distance (smaller and smaller),

② the number of analysis elements (higher and higher),

③ the number of photons reaching the WFS (higher and higher),

④ the global measurement variance (smaller and smaller),

⑤ and the global "integration $\oplus$ delay" time (smaller and smaller),

where it is also straightforward that a number of trade-offs will have to be found.

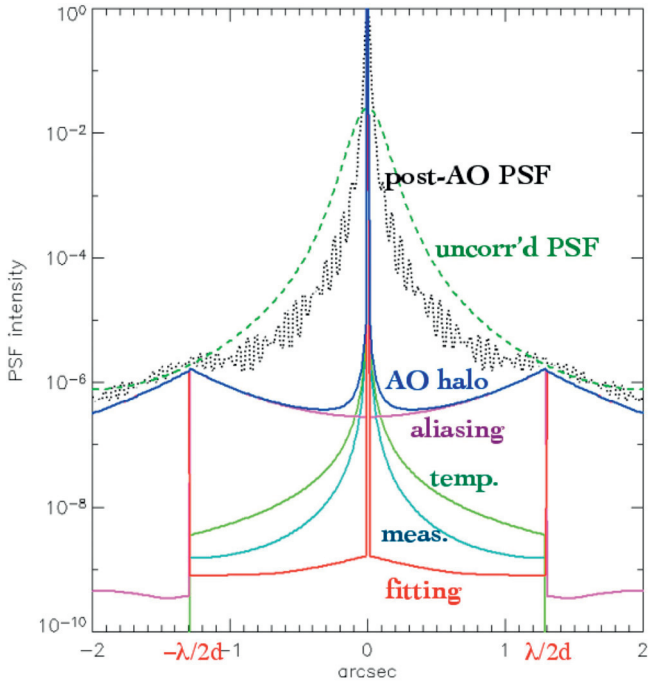


Fig. 5. Spatial/angular distribution of the different post-AO errors in the PSF halo. (From Lardière *et al.* 2005.)

2.4 The (resulting) point-spread function morphology

When separating the effects due to the different error sources from an AO system like I have done in the beginning of the present section, and more precisely looking at the spatial/angular distribution of these errors within the focal plane, *i.e.* within the PSF itself, we obtain what is represented in Figure 5.

The main interesting effect to observe from Figure 5 is the morphology of the fitting error and the aliasing error (here $d_{\rm WFS} = d_{\rm DM} = d$), especially around $\lambda/2d$, which gives this halo ring after which the Airy rings are not visible anymore and hence the benefit from AO correction is no more present. Like aliasing, the measurement error and the temporal error ($\sigma^2_{\rm measure}$ and $\sigma^2_{\rm temporal}$ respectively) also participate for what concerns the angular resolution and to the distribution of error definitely inside the "cleared" λ/d zone.

2.5 Quality of correction?...

The basic quantity permitting to characterize the AO-correction quality is indeed the Strehl ratio (Strehl 1902) (SR), which is defined as:

$$S = \frac{I_{\rm post-AO}[0,0]}{I_{\rm perfect\ case}[0,0]}, \tag{2.11}$$

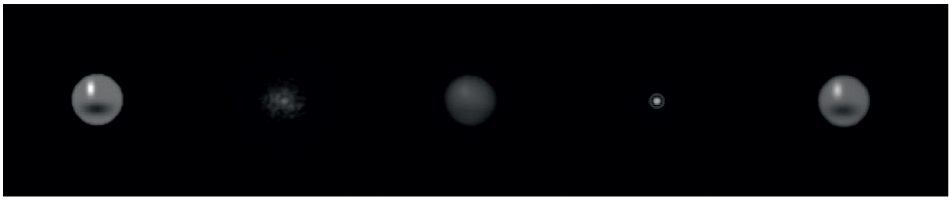


Fig. 6. From *left* to *right*: object, PSF with a SR of 0.07, resulting image, PSF with a SR of 0.93, resulting image.

where $I_{\text{perfect case}}[0,0]$ is the intensity of the ideal PSF in its central point ($[0,0]$) and $I_{\text{post-AO}}[0,0]$ corresponds to the same value but for the post-AO PSF.

Figure 6 shows the different effect that two different levels of attained SR have on the resulting HAR images: while the object is clearly recognizable in its various spatial details with a SR of 0.93, it is almost unrecognizable with a SR of only 0.07.

Nevertheless this could be far from being enough when a detailed study of the observational capabilities of a given instrument, with respect to a given observational goal, is necessary. In many cases alternative more descriptive quantities have to be used, such as:

① the attained FWHM of the PSF (when angular resolution is of main concern),

② the encircled energy for many spectrometric considerations,

③ or for example the post-AO post-coronagraphic PSF wings level for very high-contrast questions.

Indeed all these alternative quantities are linked to the Strehl ratio obtained by a given system in a given observational situation, but not in an obvious linear manner.

More refined criteria for qualifying the AO correction can also be considered, especially when adapted to a given astrophysical goal. For example estimation of the attainable signal-to-noise ratio when dealing with detection problems (exoplanets, faint objects, etc.), or even the capability to obtain well-reconstructed images. In the latter one has typically to consider the whole imaging process: telescope $\oplus$ AO system $\oplus$ instrument $\oplus$ data processing. Figure 7 details such an approach, where the capability for the whole imaging chain (starting here from the Large Binocular Telescope (LBT) in interferometric imaging mode) to obtain astrophysical informations on a given object is estimated through the precision obtained when reconstructing the magnitude difference between the components of the inner close binary star in one hand, and through the fidelity in retrieving the morphology of a very weak circumbinary ring in the other hand.

2.6 The hard side

An introduction on the concept and basic behavior of both the SHS and the PS can be found, *e.g.*, in Campbell & Greenaway (2006). I will focus here on the current duel that is featuring these sensors in particular in the framework of XAO.

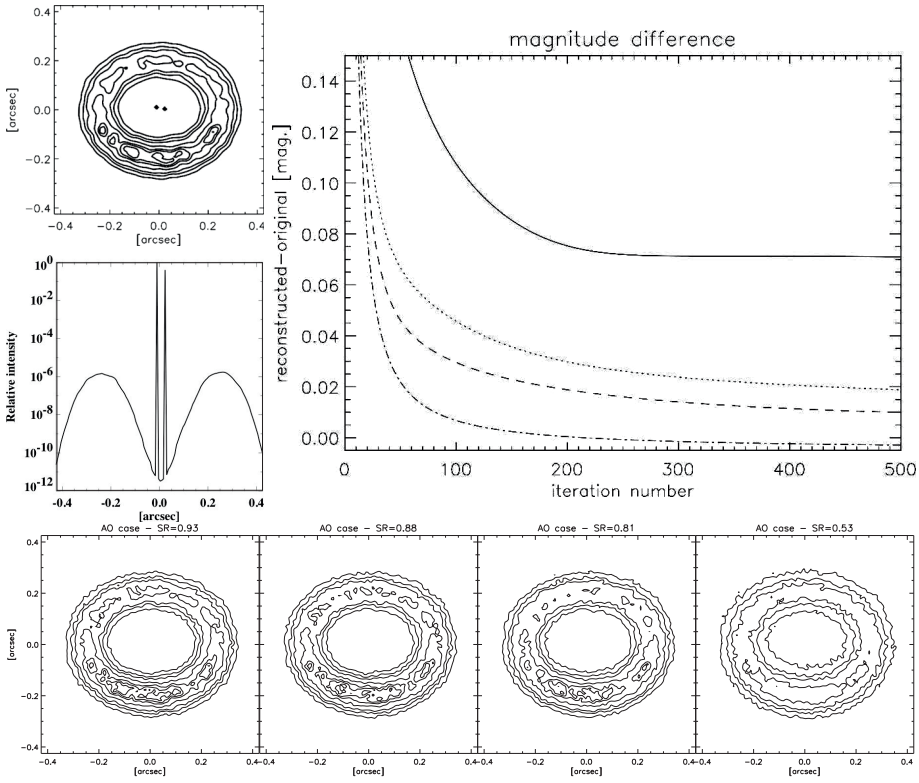


Fig. 7. Quality of post-AO K-band reconstructed images of a given object: a close double star surrounded by a circumbinary ring (more precisely a GG-Tau-like object). *Top left*: observed object (contour plot and cut along the inner binary star axis). *Top right*: quality of the reconstruction in terms of the reconstruction error on the magnitude difference between the components of the inner close binary system, for different Strehl ratios obtained. *Bottom*: quality of the reconstruction for the morphology of the circumbinary ring, for the same different Strehl ratios as before. (Adapted from Carbillet *et al.* 2002.)

Back in 1999 Ragazzoni & Farinato (1999) shown, thanks to an analytic reasoning, that the PS should permit a gain of 2 magnitudes (in terms of limiting magnitude) with respect to its main competitor, the SHS. The analytic reasoning was based on the expression expected for $\sigma^2_{\rm measure}$ for each Zernike component of the perturbed phase, expression derived from the result obtained previously by Rigaut & Gendron (1992) for the SHS.

This was then confirmed by Esposito & Riccardi (2001) by means of numerical simulations modeling AO correction (assuming weak phase perturbations), but in an open-loop regime and without any atmospherical residuals out of the modes corrected by the considered AO system.

Complete end-to-end simulations were then presented by Carbillet *et al.* (2003), considering this time the whole post-AO error $\sigma^2_{\rm post-OA}$, and hence in particular

$\sigma^2_{\text{measure}}$ in one hand and $\sigma^2_{\text{aliasing}}$ in the other hand. A gain was found in all the cases considered: in the photon-starving regime (where $\sigma^2_{\text{measure}}$ dominates) as, at the opposite, in the high-correction regime, where $\sigma^2_{\text{aliasing}}$ dominates.

The wind changed direction when Poyneer & Machintosh (2004) proposed to diminish $\sigma^2_{\text{aliasing}}$ for the SHS by introducing a spatial filtering of each single spot behind the lenslet array. Moreover, Nicolle *et al.* (2004) proposed to diminish also $\sigma^2_{\text{measure}}$ by optimizing the signal calculations made after the SHS.

Finally Vérinaud *et al.* (2005) gave the definitive (but still theoretical) answer: while the PS better performs around the center of the diffraction pattern (*i.e.* around the core itself of the PSF), the (spatially filtered) SHS gives better results towards the edges of the previously evoked "cleared" λ/d zone.

Since then, the PS has performed outstanding and unprecedented results on sky with FLAO, the first-light AO system of the LBT (see Esposito *et al.* 2010 & Riccardi *et al.* 2010). The instrument SPHERE is expected to give similar results aboard the Very Large Telescope (VLT) from the SHS side... but it is still to be proven, at least for now, on sky.

2.7 Deformable mirrors

Different deformable mirrors technologies are being considered for the various AO systems existing or being developed world-wide: piezo-stacked mirrors, piezoelectric mirrors, MOEMS (Micro-Opto-Electro-Mechanical Systems – also called optical MEMS), adaptive secondary mirrors. They are all characterized at the very end by a few basic and fundamental parameters:

① the coefficient before the d_{DM}/r_0 term in Equation (2.5),

② the inter-actuator distance d_{DM} itself,

③ the mirror stroke,

④ the response time necessary for a command to be executed by the mirror.

Concerning the first point listed before, it is completely linked to the morphology of the mirror itself when an actuator is pushed up, as clearly shown in Figure 8, where two different simple mirror technologies are shown to give two different mirror surface shapes and hence two different fitting error coefficients.

At this point a straightforward question has to be raised: how many actuators for a given achievable Strehl ratio? By only considering Equation (2.5) again, actual numbers to be given are (see Brusa *et al.* 1999):

⋆ $S_{\text{fit}} \simeq 0.75 \Rightarrow d \simeq r_0(\lambda)$ gives $N \simeq 350$,

⋆ $S_{\text{fit}} \simeq 0.92 \Rightarrow d \simeq 0.5\ r_0(\lambda)$ gives $N \simeq 1450$,

considering band J and an 8-m class telescope.

The geometry (spatial distribution of the actuators) is another important point, the one that will determine the influence functions of the mirror, and hence its modes, the one that will be applied when a given command will be deduced by the wavefront reconstructor after each WFS measure.

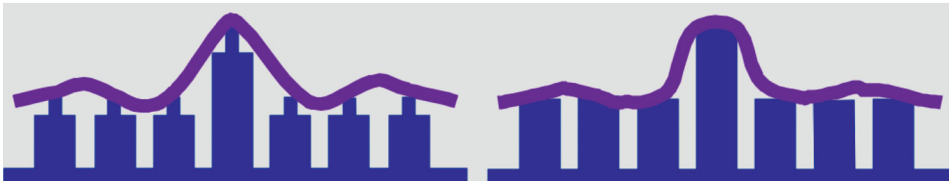


Fig. 8. Deformable mirror fitting illustration. (From Riccardi 2003.)

An interesting case to be further discussed is the adaptive secondary mirror one, for which a number of optical surfaces are eliminated, including the necessity of an additional tip-tilt mirror for usual DMs, leading to a considerable gain in number of photons available for WFS measures, and hence boosting the final performance of an AO system which uses such a type of DM (like it is used for FLAO/LBT, and will be used for the built-in M4 adaptive mirror of the EELT).

2.8 Wavefront reconstruction and command control

In order to have the DM applying the correct commands that will compensate the turbulent wavefront coming from the entrance pupil of the telescope, a wavefront reconstructor has to deduce the command needed to compensate the measured wavefront deformations (slope x- and y-measurements from the WFS), and moreover: a command control has to be considered.

A very basic standard command makes use of a reconstruction process coupled with an integrator control law. The value of the gain of this integrator has to be optimized for a given AO system and a given guide star magnitude (and hence a given number of photons available per temporal unit), together with a number of AO system central parameters such as the WFS integration time and the number of DM modes to be corrected. A step forward consist in optimizing this gain mode by mode, as a function of the signal-to-noise ratio on each mode.

A (more refined) *Kalman* filter approach is also usually considered in order to command the system in an optimal way both for the reconstruction process and the control. The reader is invited to consult the course of J.-P. Folcher within these proceedings (Folcher 2013) for a detailed dissertation about this crucial subject.

3 Going further

3.1 Various improvements are possible

A number of improvements are definitely possible, as long as any term of $\sigma^2_{\mathrm{post-AO}}$ can be diminished in some way. At least the three following possible improvements are currently investigated and seriously considered for AO systems:

① reduction of $\sigma^2_{\mathrm{measure}}$ first: by using EMCCDs for the WFS. These devices have the capability to mimic a very low read-out noise. The counter part of it is nevertheless the addition of an "exotic" dark-current component.

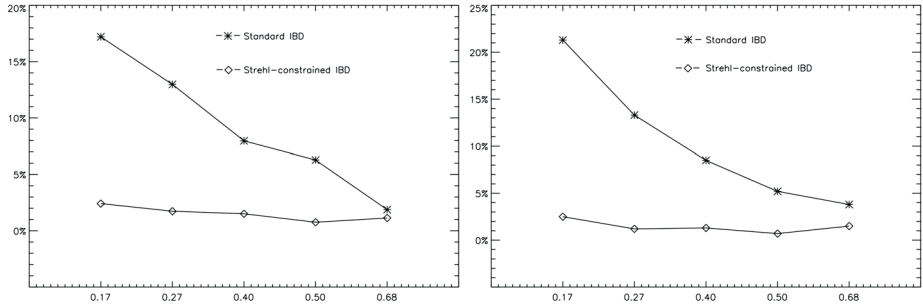


Fig. 9. Final error on the reconstruction of the PSF (*left*) and on the object (*right*), when using an IBD) algorithm. Both plots are made as a function of the SR of the image data and comparing the simple IBD (rhombuses) to the Strehl-constrained IBD (asterisks). A gain of up to a factor ∼10 is achieved for the poorer SR. (From Desiderà & Carbillet 2009.)

② reduction of $\sigma^2_{\rm measure}$ second: by adding a dedicated tip-tilt sensor in addition to the global WFS. Then one has to find an optimal value for the splitting of light between the tip-tilt sensor and the higher-order WFS, and that this splitting of light is still of any advantage with respect to using a single WFS. The answer is not unique but depend on the precise AO system used – and in particular the WFS used (see, *e.g.*, Carbillet *et al.* 2005): PS, SHS, filtered SHS, etc.

③ reduction of $\sigma^2_{\rm measure}$ last: an idea proposed by Le Roux *et al.* (2005) also consider to mask the WFS – toward a coronagraphic WFS?

④ reduction of $\sigma^2_{\rm atmosphere}$: this last error could be the most simple (from the conceptual point-of-view) but the most complicate (from the practical point-of-view) to diminish, since it could imply to consider to install the AO-equipped telescope on a tower in the middle of Antarctica, since it can be seen as the best site on earth when eliminating the very thin turbulence surface layer (see, *e.g.*, Lardière *et al.* 2005; Aristidi *et al.* 2009; Carbillet *et al.* 2010; Giordano *et al.* 2012).

3.2 Post-adaptive-optics object reconstruction

3.2.1 Knowledge of the Quality of Correction ⇒ Even Better Object Reconstruction

Figure 9 shows the advantage of using a constraint on the Strehl ratio when reconstructing the PSF (in the cases where it is unknown or badly known), and hence the object from the obtained image, in the case of an iterative blind deconvolution (IBD) of post-AO data.

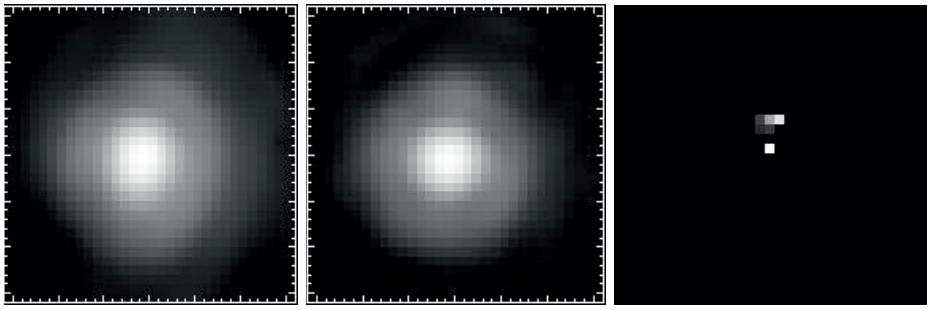


Fig. 10. From *left* to *right*: the observed image, the observed PSF, both pre-processed, and the result of a Lucy-Richardson-based super-resolution algorithm on the image.

3.2.2 Going further in angular resolution

Figure 10 shows a very preliminary (and unpublished) result of the application of a Lucy-Ricardson-based super-resolution algorithm (first evoked in Correia *et al.* 2002 and then described in details in Anconelli *et al.* 2005) on NACO/VLT data of the unresolved binary star HD 87643. It clearly unveils a very close binary star (separation smaller than half an element of resolution λ/D), with some possible matter around the component above, confirming previous AMBER/VLTI observations (Millour *et al.* 2009).

Thanks are due to Armando Riccardi and Olivier Lardière for having kindly provided both of them one of the illustrations presented for this course/paper. Thanks are also due to the organizers of the summer school for which this course/paper was prepared: Céline Theys-Ferrari, David Mary, and Claude Aime.

References

Aime, C., Borgnino, J., Martin, F., *et al.*, 1986, J. Opt. Soc. Am. A, 3, 1001

Aime, C., 1987, J. Optics (Paris), 10, 318

Anconelli, B., Bertero, M., Boccacci, P., & Carbillet, M., 2005, A&A, 431, 747

Aristidi, É., Carbillet, M., Lyon, J.-F., & Aime, C., 1997, A&AS, 125, 139

Aristidi, É, Fossat, É., Agabi, A., *et al.*, 2009, A&A, 499, 955

Baldwin, J.E., Tubbs, R.N., Mackay, C.D., *et al.*, 2001, A&A, 368, L1

Brusa, G., Riccardi, A., Accardo, M., *et al.*, 1999, in Proceedings of the Backaskog workshop on Extremely Large Telescopes, ed. T. Andersen, A. Ardeberg & R. Gilmozzi, ESO Proc., 57, 18

Campbell, H.I. & Greenaway, A.H., 2006, EAS Publication Series, 22, 165

Carbillet, M., 1996, Ph.D. Thesis (Université de Nice–Sophia Antipolis (UNS))

Carbillet, M., Aime, C., Aristidi, É., & Ricort, G., 1998, A&AS, 127, 569

Carbillet, M., Correia, S., Boccacci, P., & Bertero, M., 2002, A&A, 387, 743

Carbillet, M., Vérinaud, C., Esposito, S., *et al.*, 2003, SPIE Proc., 4839, 131

Carbillet, M., Vérinaud, C., Femenía Castellá, *et al.*, 2005, MNRAS, 356, 1263
Carbillet, M., 2006, EAS Publication Series, 22, 121
Carbillet, M., Maire, A.-L., Le Roux, B., *et al.*, 2010, EAS Publications Series, 40, 157
Carbillet, M., & Riccardi, A., 2010, Appl. Opt., 49, JOSA A/App. Opt. Joint Feature Issue on Adaptive Optics, G167–G173
Carbillet, M., Riccardi, A., & Xompero, M., 2012, SPIE Proc., 8447, 844762
Correia, S., Carbillet, M., Boccacci, P., Bertero, M., & Fini, L., 2002, A&A, 387, 733
Desiderà, G., & Carbillet, M., 2009, A&A, 507, 1759
Esposito, S., & Riccardi, A., 2001, A&A, 369, L9
Esposito, S., Riccardi, A., Fini, L., *et al.*, 2010, SPIE Proc., 7736, 773609
Folcher, J.-P., 2013, EAS Publication Series, 59, 93
Giordano, C., Vernin, J., Chadid, M., *et al.*, 2012, PASP, 124, 494
Labeyrie, A., 1970, A&A, 6, 85
Labeyrie, A., & Foy, R., 1995, A&A, 152, L29
Lardière, O., Carbillet, M., Riccardi, A., *et al.*, 2005, EAS Publication Series, 14, 291
Le Roux, B., Coyne, J., & Ragazzoni, R., 2005, Appl. Opt., 44, 171
Mackay, C.D., Rebolo-López, R., Femenía Castellá, B., *et al.*, 2012, SPIE Proc., 8446, 844672
Millour, F., Chesneau, O., Borges Fernandez, M., *et al.*, 2009, A&A, 507, 317
Nicolle, M., Fusco, T., Rousset, G., & Michau, V., 2004, Opt. Lett., 29, 2743
Poyneer, R., & Machintosh, J., 2004, J. Opt. Soc. Am. A, 350, L23
Ragazzoni, R., & Farinato, J., 1999, A&A, 350, L23
Riccardi, A., 2003, in Adaptive Optics Mini-School (Garching-bei-Muenchen, Germany, 19–21 Feb. 2003)
Riccardi, A., Xompero, M., Briguglio, R., *et al.*, 2010, SPIE Proc., 7736, 77362C
Rigaut, F., & Gendron, É., 1992, A&A, 261, 677
Roddier, F., 1981, Progress in Optics XIX, 281, ed. E. Wolf (North-Holland, Pub. Co.)
Sarrazin, M., 1996, ESO Proc. on Topical Meeting on Adaptive Optics
Strehl, K., 1902, Zeit. Instrumenkde, 22, 213
Vérinaud, C., Le Louarn, M., Korkiakoski, V., & Carbillet, M., 2005, MNRAS, 357, L26
Weigelt, G., 1977, Opt. Commun., 21, 55
Worden, S.P., Lynds, C.R., & Harvey, J.W., 1976, J. Opt. Soc. Am., 66, 1243

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 77–92*

INTRODUCTION TO WAVEFRONT CODING FOR INCOHERENT IMAGING

M. Roche[1]

Abstract. We propose in this paper an introduction to the wavefront coding technique for incoherent imaging. Wavefront coding introduces image processing in the conception of an imaging system. It consists in introducing controlled aberrations in the optics able to reduce, after processing, some defaults of the optical system such as defocus, chromaticity. We present the basis of wavefront coding and illustrate them on two images with different characteristics: a spoke pattern and a galaxy image.

1 Introduction

In traditional imaging systems, the design of the optics and the processing of the recorded images are two separate steps. High aperture instruments allow one to obtain images with high resolution, with high signal to noise ratio due to the large amount of light collected and high depth of field. However these instruments are more subject to aberrations like defocus as instrument with smaller aperture size.

In hybrid imaging systems, optics and processing are considered jointly and designed together. These last imaging systems allow one to use optics of lower quality and thus with reduced cost, the quality of the images warrantied not by the quality of the optics but by the processing step. A good example of the interest to associate the image processing to the optics could be the Hubble Space Telescope (HST). It was launched in early 1990, at that time a spherical aberration has been detected, leading to a blurring of the images. The first simple and effective way to solve this degradation was to introduce image processing. Latter in 1993, this default has been corrected by introducing the COSTAR corrective optics in a Shuttle mission.

Wavefront coding was introduced by Dowski & Cathey (1995) for incoherent imaging. They propose to introduce a phase mask in the imaging system, designed

[1] Centrale Marseille, CNRS, Aix-Marseille Université, Fresnel, UMR 7249, 13013 Marseille, France

DOI: 10.1051/eas/1359005

to make the Point Spread Function (PSF) of the instrument insensitive to some aberrations such as chromaticity, or spherical aberrations. It is also used to enhance the depth of field of an instrument by making the imaging system insensitive to defocus aberrations.

Wavefront coding is now used in many domains such as security for iris recognition (Narayanswamy *et al.* 2004) where it is useful for capturing an iris without active user cooperation, thermal imagery (Muyo *et al.* 2004) for controlling thermally induced defocus aberrations in infrared imaging system, fluorescence microscopy (Arnison *et al.* 2007) to increase the depth of field. In astronomy, wavefront coding has been not yet used but it has been shown (Kubala *et al.* 2004) that as telescope performances are limited by aberrations, misalignment, temperature related defaults, this technique will provide improvement of the quality of the images.

This article is an introduction to wavefront coding technique. It does not contain new results on it but presents the different steps that lead, from a degraded image, to an image of higher quality after a post-processing. For the sake of clarity and conciseness, the paper only discuss about defocus default and on the use of a cubic phase mask.

It is organized as follows. In Section 2, the principle of image formation in a classical imaging system is presented. In Section 3, a default of defocus in the optics is introduced and modelized. In Section 4, the wavefront coding is detailed and results are presented on two different images using a cubic phase mask. Section 5 presents the optimization of the parameter of the cubic phase mask. Finally Section 6 discusses on the robustness of wavefront coding in presence of defocus.

2 Image formation in coherent and incoherent imaging

The wavefront coding technique is developped for incoherent illumination. In the following, we first discuss about the coherent illumination of an object, which is needed to explain the incoherent case.

2.1 Image of a point like object

Let us assume that we observe with an optical instrument a point like source, that can be modelized by a Dirac distribution $\delta(x, y)$. When the imaging system is only limited by diffraction, the image amplitude of this point-like source is given by the Fraunhofer diffraction pattern of the exit pupil of the imaging system[2]

$$H(x, y) = \frac{A}{\lambda L} \widehat{P}\left(\frac{x}{\lambda L}, \frac{y}{\lambda L}\right) \tag{2.1}$$

where A is a constant amplitude traducing the attenuation of the amplitude by the imaging system, λ is the wavelength of the light emitted by a point like object,

[2]For detailed calculus see for example Goodman (2005), chapter 6.

L is the distance between the exit pupil and the image plane, $\widehat{P}$ is the Fourier transform of the pupil aperture, x and y the coordinates in the plane of the exit pupil.

In the case of a circular pupil, $H(x, y)$ corresponds to the Airy function of the instrument. For notational convenience, we will assume in the following that $\frac{A}{\lambda L}$ and λL equal unity.

2.2 Impulse function for the observation of an entire object

2.2.1 Coherent illumination

When coherent illumination is considered, all the point of the object emit field whose phasor amplitude vary in unison. Thus the image of the object is obtained by summing all the contributions of the complex amplitude coming from all the point of the object. A coherent imaging system is thus linear in complex amplitude. Assuming that one point of the object is modelized by a Dirac distribution δ_i, it can be shown[3] that the received amplitude from the object is given by:

$$A(x, y) = \sum_i (\delta_i \otimes H)(x, y) = ((\sum_i \delta_i) \otimes H)(x, y) \tag{2.2}$$

where $\otimes$ represents the convolution product, $H(x, y)$ is called the amplitude PSF of the instrument. This can be rewritten:

$$A(x, y) = (O \otimes H)(x, y) \tag{2.3}$$

with $O = \sum_i \delta_i$ the amplitude of the object. The Fourier transform of the amplitude PSF is called the Amplitude Coherent Transfert Function (ACTF). In the case of a symetric pupil (almost the cases encountered), it is easy to show that:

$$ACTF(\mu, \nu) = P(\mu, \nu) \tag{2.4}$$

with P the pupil of the instrument, μ, ν the coordinates in the frequency plane.

For a circular aperture of diameter d in coherent illumination, the instrument behaves as a low-pass filter with cutting frequency $\frac{d}{2}$.

In general, the optic instruments measure intensity that means

$$i(x, y) = |A(x, y)|^2 = |O \otimes H|^2(x, y). \tag{2.5}$$

2.2.2 Incoherent illumination

In the case of incoherent illumination, the phasor amplitudes are totally uncorrelated, the complex amplitude can no more be added. In this case, it can be shown[4] that an incoherent imaging system is linear in intensity:

$$i(x, y) = (o \otimes G)(x, y) \tag{2.6}$$

[3]For detailed calculus see for example Goodman (2005), chapter 6.
[4]See for example Goodman (2005) for detailed calculus, chapter 6.

where $G(x,y) = |H(x,y)|^2$, $i(x,y)$ the intensity of the image observed and $o(x,y)$ the intensity of the object.

$|H(x,y)|^2$ represents the intensity point spread function of the instrument, it will be denoted PSF in the following. The Fourier transform of this PSF is called the Optical Transfert Function (OTF) and its modulus the Modulated Transfert Function (MTF).

3 Analysis of the influence of defocus

Let us assume an incoherent imaging system with a circular pupil $P(\mu,\nu)$ of diameter d, that presents a focus default that is spatially constant over the pupil. This defocus can be modelized in the pupil plane by introducing a supplementary phase:

$$e^{i\Psi_\lambda(\nu^2+\mu^2)} \tag{3.1}$$

where e^a represents the exponential function of a, Ψ_λ corresponds to the defocalisation parameter and $i = \sqrt{-1}$. Ψ_λ depends on the diameter of the pupil d, on the distance between the object and the primary plane of the lens d_0, on the distance between the secondary plane of the lens and the CCD camera d_{ccd} and on the focal distance of the lens $f(\lambda)$ (Dowski *et al.* 1995)

$$\Psi_\lambda = \frac{\pi d^2}{4\lambda}\left(\frac{1}{f(\lambda)} - \frac{1}{d_0} - \frac{1}{d_{ccd}}\right) = \frac{2\pi}{\lambda}W_{20} \tag{3.2}$$

with W_{20} is the traditional defocus aberration constant. The pupil of the imaging system in presence of defocus is thus:

$$P'(\nu,\mu) = P(\nu,\mu)e^{i\Psi_\lambda(\nu^2+\mu^2)}. \tag{3.3}$$

From Equation (2.1), the PSF of this imaging system in the case of incoherent imaging is given by:

$$|H(x,y)|^2 = \left|\widehat{P'}(x,y)\right|^2. \tag{3.4}$$

Figure 1 shows respectively the PSF and the MTF for an instrument of circular aperture when no defocus default is presents a) and c), and when a defocus of parameter $\Psi_{\lambda_3} = 50$ is introduced b) d). The defocus induces a PSF extended with respect to the ideal one (a) and consequently a MTF with an important reduction of the high frequencies. The defocus will introduce a blurring effect in the imaged object.

Figure 2 represents a central cut of the MTF in the case of incoherent imaging system with a circular pupil in presence of different defocus. Three different defocus are considered with $\Psi_{\lambda_1} < \Psi_{\lambda_2} < \Psi_{\lambda_3}$. The circular pupil behaves as a low-pass filter, an increase of the parameter of defocus implies a reduction of the cut-off frequency and introduces oscillations in the MTF with apparition of zeros.

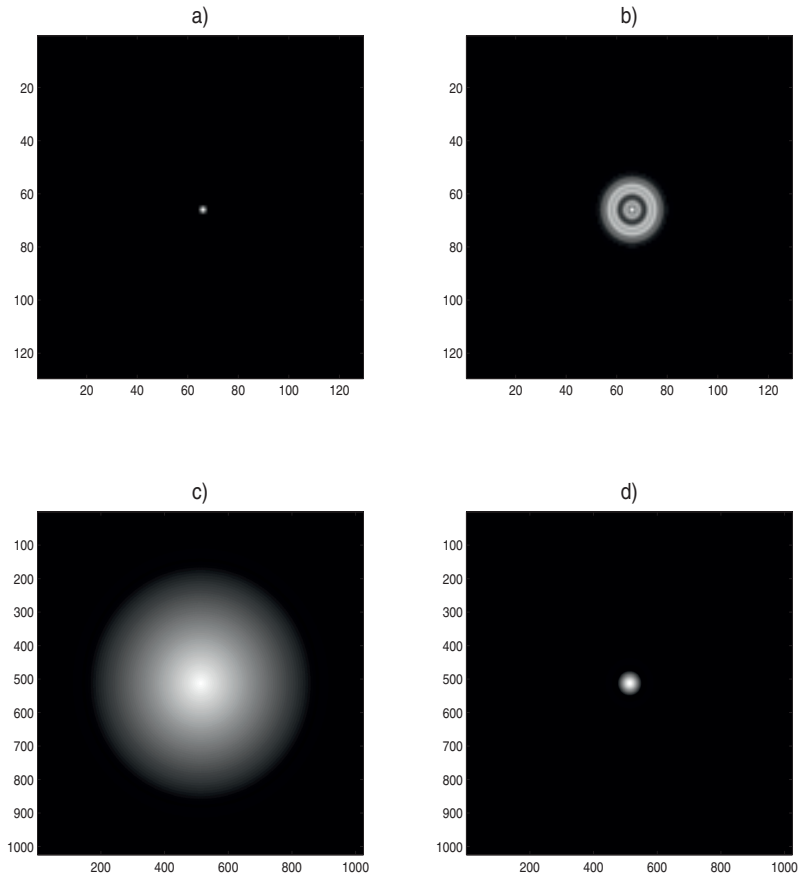


Fig. 1. Representation of the PSF (first line) and MTF (second line) of an instrument with circular aperture when in focus a), c) and when a defocus of parameter $\Psi_{\lambda_3} = 50$ is introduced b), d). In order to improve the visualization, figures a) and b) correspond to a central part of size 128 $\times$128 of the entire PSF (of size 1024 $\times$ 1024).

The effect of the MTF on two different imaged object is obtained from 2.6. The two object considered are respectively, a spoke pattern with high spatial frequencies, and the galaxie UGC 1810 taken by the Hubble Space Telescope[5] containing mostly low spatial frequencies. The results are presented in Figures 3 and 4 which shows the blurring effect appearing in the observed image when the imaging system presents a defocus default of parameter $\Psi_{\lambda_3} = 50$. This blurring effect appears essentially on the edge and on the center of the spoke pattern (Fig. 3c) whereas it is visible in the entire image of the galaxy (Fig. 4c) leading to the disapearrance of the stars (point like object).

[5] `http://hubblesite.org/gallery/album/galaxy/hires/true/`

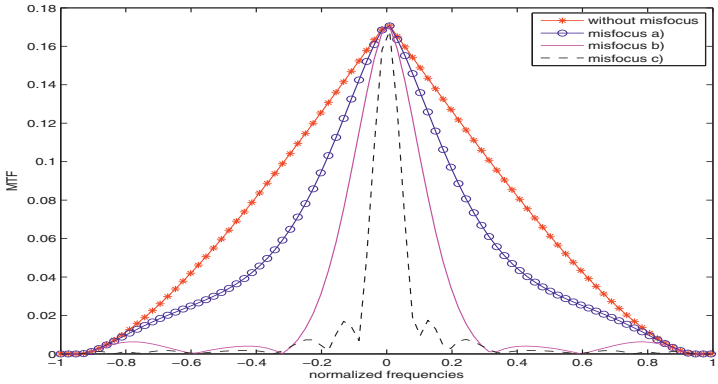


Fig. 2. Effect of different defocus (with a) $\Psi_{\lambda_1} = 10, b)$ $\Psi_{\lambda_2} = 20, c)$ $\Psi_{\lambda_3} = 50$) on the Modulated Transfert Function (MTF) of an incoherent imaging system with a circular aperture as a function of normalized frequencies (the maximum frequency equals one).

4 Correction of the defocus

The observed image (Figs. 3c and 4c) can be processed to reduce the effect of the PSF on the observation. If the PSF is known, classical deconvolution techniques can be implemented allowing to reconstruct an object closed to the true one (Figs. 3a and 4a) (when no noise is present[6]) to obtain the reconstructed image of Figures 3d and 4d. When a defocus is introduced (Figs. 3c and 4c), the image is blurred. The blurring effect can be supressed if it is known (Figs. 3e and 4e). However in most cases this default is not known leading to deconvolved image of Figures 3f and 4f. In these cases, it is evident that the blurring effect has been neither suppressed nor reduced with respect to Figures 3c and 4c.

4.1 Introduction of wavefront coding

The wavefront coding was introduced in Dowski & Cathey (1995). It consists in introducing a phase mask in the pupil. This phase mask is introduced in order to correct the defaults of the imaging system: sphericity, chromatic aberrations (Wach *et al.* 1998), defocus... Moreover, this mask is constructed to avoid the presence of zeros in the corresponding PSF allowing first to preserve frequencies, and to avoid calculus errors in the deconvolution process.

In the pupil plane, the mask can be modelized by:

$$M(\nu, \mu) = e^{i\Phi(\nu,\mu)} \tag{4.1}$$

where $\Phi(\nu, \mu)$ characterizes the shape of the mask, $|\nu| < 1$ and $|\mu| < 1$ are the normalized frequency coordinates.

[6]Of course this hypothesis is not realist but allows to simplify the problem and to present basis on image formation.

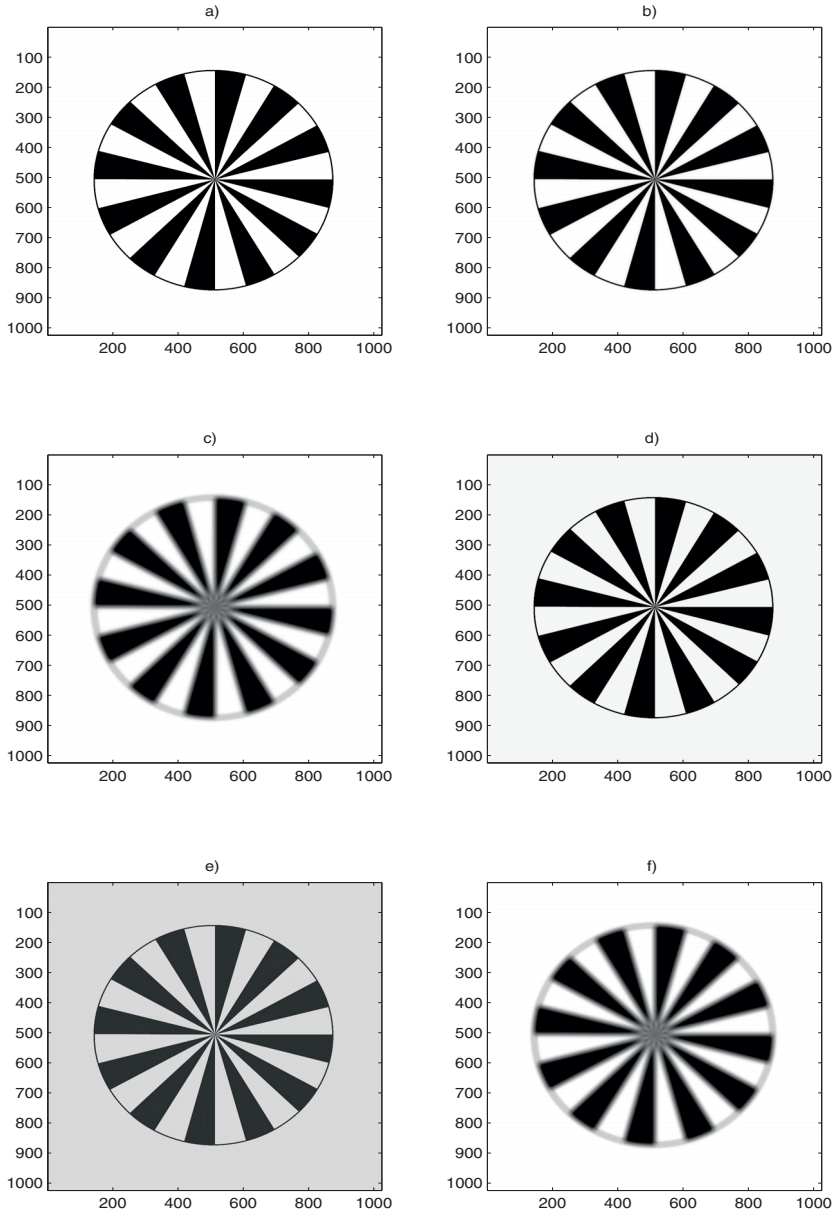


Fig. 3. Effect of different degradation on the initial object a): b) with an ideal circular pupil with a diameter of 480 pixels, c) with a defocus of $\Psi_{\lambda_3} = 50$ on the observation. Deconvolution of the observed image: d) deconvolution of b), e) deconvolution of c) with the defocus known, f) deconvolution of c) with the defocus unknown.

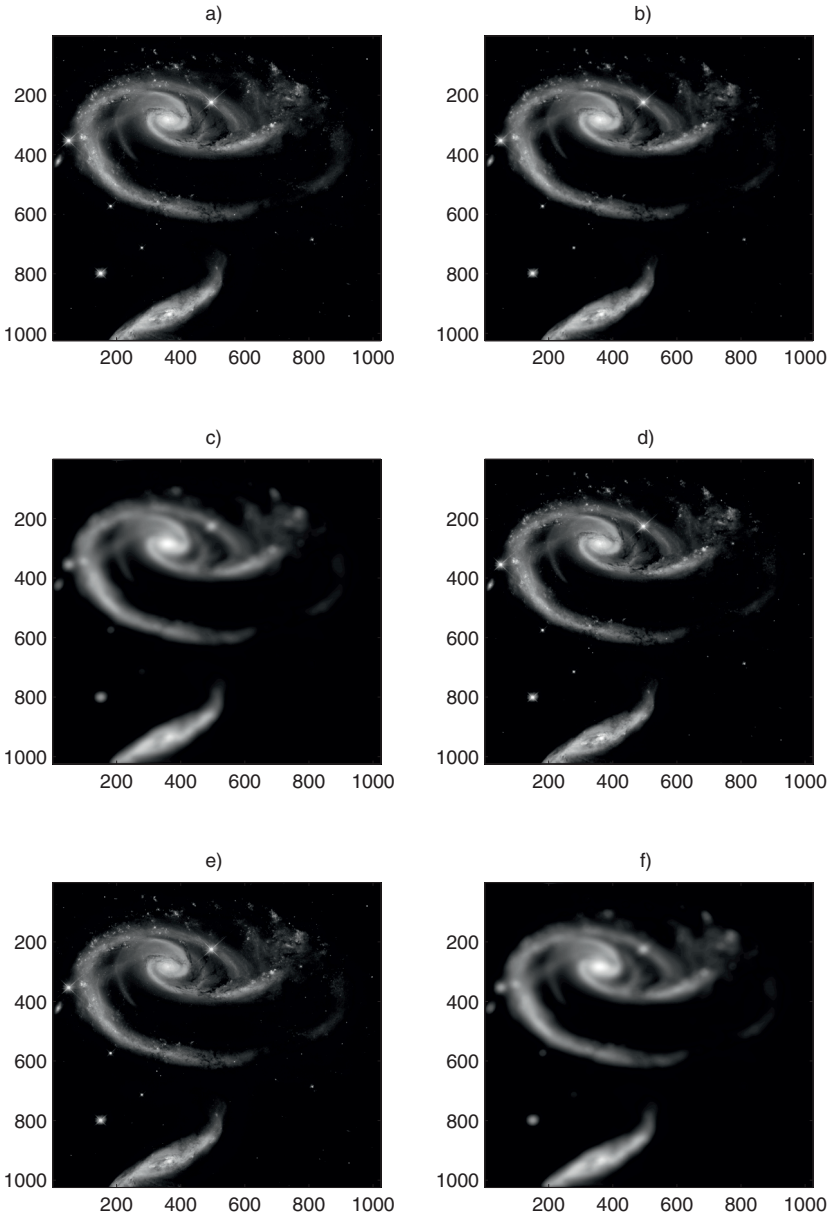


Fig. 4. Effect of different degradation on the initial object a): b) with an ideal circular pupil with a diameter of 480 pixels, c) with a defocus of $\Psi_{\lambda_3} = 50$ on the observation. Deconvolution of the observed image: d) deconvolution of b), e) deconvolution of c) with the defocus known, f) deconvolution of c) with the defocus unknown.

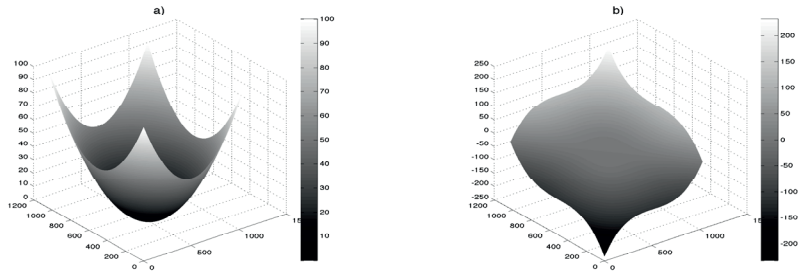


Fig. 5. Representation of the phases of a) the defocus default of parameter $\Psi_{\lambda_3} = 50$ and b) the phase introduced with a cubic phase mask of parameter $\alpha = 116$.

Wavefront coding leads in presence of defocus to a new pupil

$$P_c(\nu, \mu) = P'(\nu, \mu).M(\nu, \mu) = P(\nu, \mu)e^{i\Psi_\lambda(\nu^2+\mu^2)}e^{i\Phi(\nu,\mu)}. \tag{4.2}$$

We focus in this paper on wavefront coding used to increase the depth of field, leading to optical system insensitive to defocus. Different phase mask have been be proposed in the litterature to increase the depth of field: cubic phase mask (Dowski *et al.* 1995), logarithmic (Sherif *et al.* 2004; Zhao *et al.* 2008), fractionnal-power (Sauceda *et al.* 2004), exponentiel (Yang *et al.* 2007), polynomial (Caron *et al.* 2008), asymetric phase mask (Castro *et al.* 2004) and have been compared (Diaz *et al.* 2010; Sherif *et al.* 2004; Yang *et al.* 2007) with respect to different criterion depending on the the aimed application.

These different phase masks are obtained by consideration of different criterion. In Neil *et al.* (2000), an optimization is done to obtain a particular form for the final PSF in the case of confocal microscope. In S. Prasad *et al.* (2004), the authors use the Fisher information and the Strehl ratio to find the mask that reduces the sensitivity of the phase to misfocus.

The cubic phase mask, we consider in the following, proposed by Dowski *et al.* (1995), was obtained by using the ambiguity function and the stationary-phase method.

Let us consider a cubic phase mask of the form

$$\Phi(\nu, \mu) = \alpha(\nu^3 + \mu^3). \tag{4.3}$$

Figure 5 illustrates the phases of a defocus default of parameter $\Psi_{\lambda_3} = 50$ and the phase introduced with a cubic phase mask of parameter $\alpha = 116$.

This mask was constructed to minimize the variation of the OTF with defocus. It presents only one parameter to optimize (α) with respect to the application, leading to a simple mask. Other masks introducing more parameters lead to better results in general, but increase the complexity of the mask.

Figure 6 represents respectively the PSF and the MTF of an instrument with circular aperture with a cubic phase mask of parameter $\alpha = 116$ when no defocus

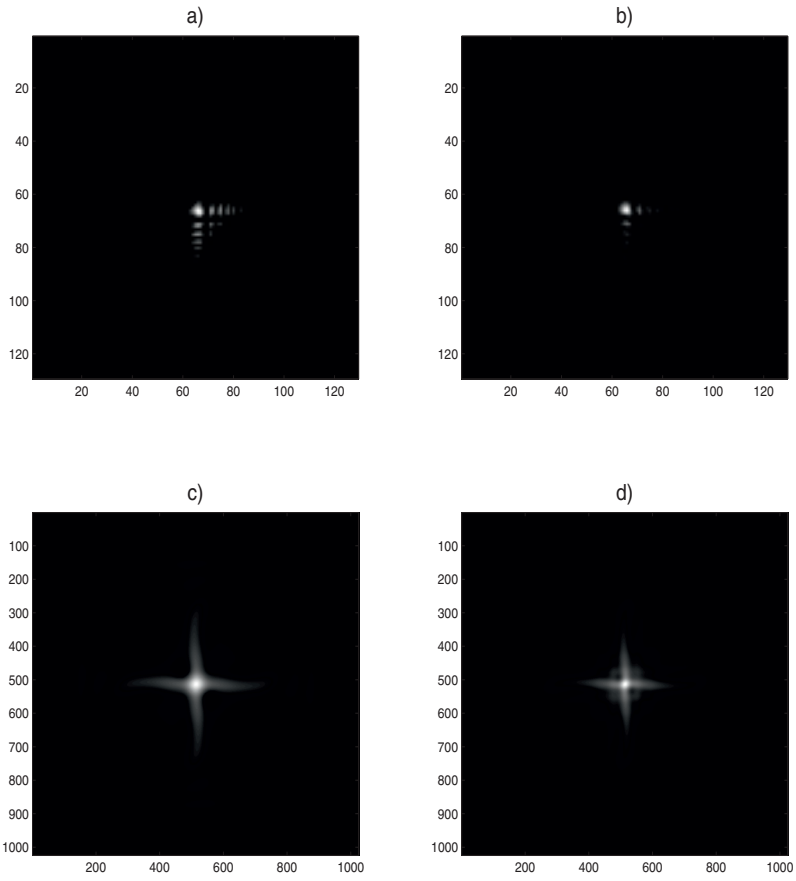


Fig. 6. Representation of the PSF (first line) and MTF (second line) of an instrument with circular aperture with a cubic phase mask of parameter $\alpha = 116$ when no defocus default is present a) c) and when a defocus of parameter $\Psi_{\lambda_3} = 50$ is introduced b) d). In order to improve the visualization, Figures a) and b) correspond to a central part of size 128×128 of the entire PSF (of size 1024×1024).

default is present a), c) and when a defocus of parameter $\Psi_{\lambda_3} = 50$ is introduced b), d). The wavefront coding leads to small noticeable changes in the PSF and in the MTF with defocus.

Figure 7 shows a central cut of the MTFs of Figures 1c, d and 6c, d, representative of different configurations: imaging system with no default, imaging system with defocus default of parameter $\Psi_{\lambda_3} = 50$, imaging system with wavefront coding when no defocus aberration exists and in presence of defocus. A cubic phase mask is used with parameter $\alpha = 116$. The use of wavefront coding allows to increase the cut-off frequency and to reduce the number of zeros. Moreover, the amplitude of the MTF is increased.

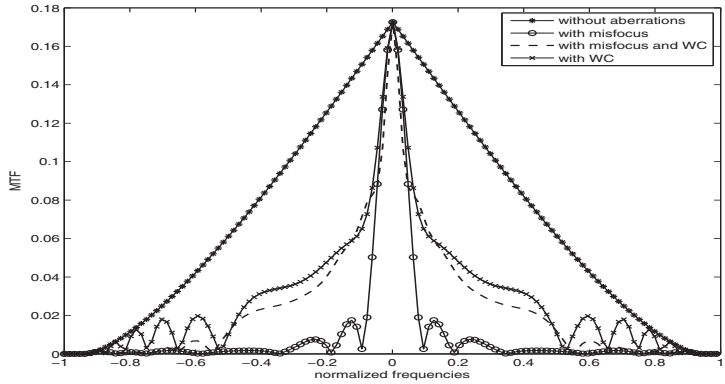


Fig. 7. Effect of different configurations of the imaging system on the MTF as a function of normalized frequencies (the maximum frequency equals one).

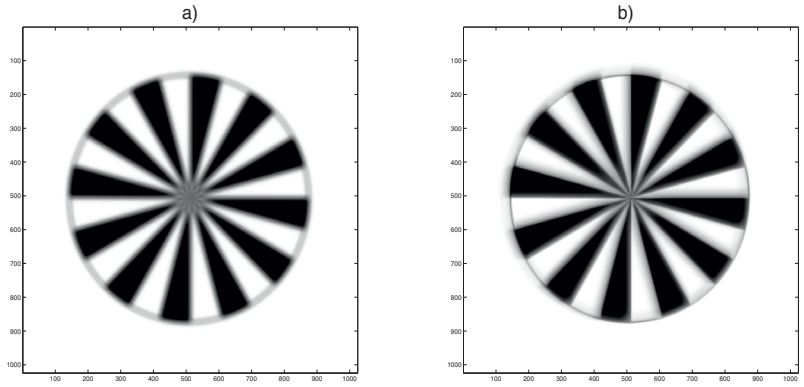


Fig. 8. Image observed in presence of a defocus with $\Psi_{\lambda_3} = 50$ without a) and with wavefront coding b). A cubic phase mask of the form 4.3 is considered with $\alpha = 116$.

4.2 Deconvolution of the images

The influence of the wavefront coding on the observed image is represented on Figures 8b and 9b. It clearly appears that the only introduction of a phase mask allows to reduce the blurring in the observation. The image obtained is then processed in order to still reduce the blurring effect. It is important to notice that the defocus default is not yet known. The deconvolution is thus done considering two configurations of the imaging system: a classical one, and another that introduces the wavefront coding. The results of the deconvolution are presented in Figures 10 and 11. The visual quality is still better when wavefront coding is used and is improved in comparison to the image of Figures 8b and 9b. In particular for the spoke pattern, the region in the center of the image is sharper and for the galaxy, the stars are closed to pointwise object.

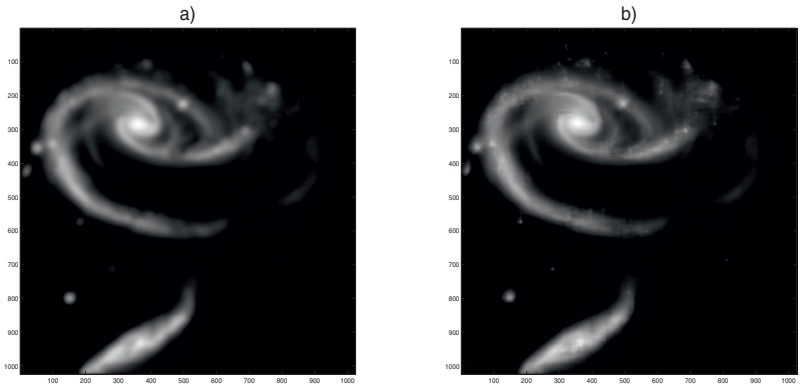


Fig. 9. Image observed in presence of a defocus with $\Psi_{\lambda_3} = 50$ without a) and with wavefront coding b). A cubic phase mask of the form 4.3 is considered with $\alpha = 119$.

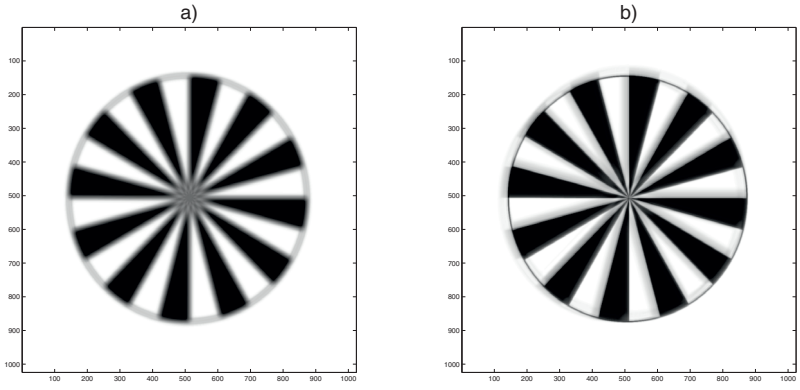


Fig. 10. Deconvolution of the image observed in presence of a defocus with $\Psi_{\lambda_3} = 50$ without a) and with wavefront coding b). A cubic phase mask of the form 4.3 is considered with $\alpha = 116$.

5 Optimization of the parameter of the cubic phase mask

The parameter α of the cubic phase mask must be optimized to obtain a efficient wavefront coding that corrects, after a processing step, the defocus default.

In the results presented in Figures 8b and 10b, the parameter α is chosen equal to 116. This parameter was obtained by considering a quality criterion on the reconstructed image. In our simulation, the chosen criterion is the Mean Square Error (MSE) between the true image of Figure 3a and the reconstructed image when wavefront coding is considered. Other choices of quality criterion can be done, for example the MSE can be averaged on several MSE (Diaz *et al.* 2010) obtained from different values of the defocus, leading to a phase mask robust to

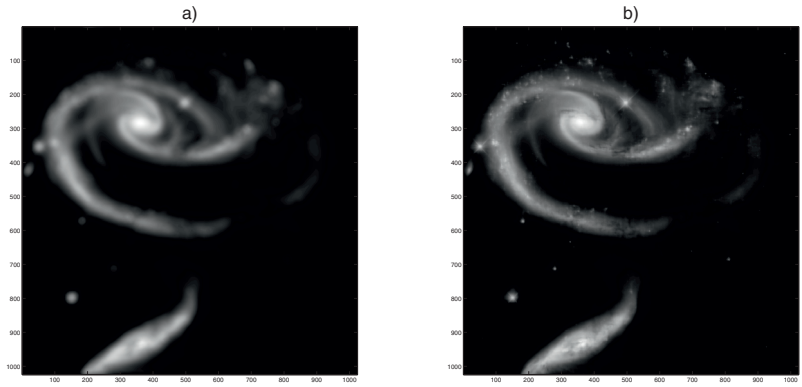


Fig. 11. Deconvolution of the image observed in presence of a defocus with $\Psi_{\lambda_3} = 50$ without a) and with wavefront coding b). A cubic phase mask of the form 4.3 is considered with $\alpha = 119$.

the defocus parameter. These criterions based on the calculus of the MSE can be considered only in simulations when the true object is known.

The MSE is represented in Figure 12 (curve c)) considering a parameter α varying between 1 and 200. The choice $\alpha = 116$ leads to the minimization of the mean square error when a defocus parameter of $\Psi_{\lambda_3} = 50$ is considered.

The curves of MSE obtained for different parameter of defocus (a) $\Psi_{\lambda_1} = 10$, b) $\Psi_{\lambda_2} = 20$) are also represented. It is clear that the value of the parameter α of the cubic mask depends on the defocus parameter, however the choice of α is not so sensitive to the defocus parameter. Indeed, a range of parameter α leads to similar values of the MSE. For example, for the defocus $\Psi_{\lambda_3} = 50$, α can be chosen in the range $[90, 170]$ without leading to significative degradation of the reconstruction.

For the image of the galaxy, the results presented in Figures 9b and 11b are obtained with a mask parameter $\alpha = 119$. The criterion used to optimize this value is still the MSE but computed over a small zone (100×100 pixels) of the image of Figure 4a. The use of the entire image of the galaxy gives a bad criterion of quality as the image is complex. The curves for the MSE in presence of different values of defocus are similar to the one of Figure 12 and are not represented.

6 Robustness of the cubic phase mask with respect to defocus

Once the parameter of the cubic phase mask is optimized, it is interesting to study its robustness with respect to defocus. The curve in Figure 13 represents the MSE between the reconstructed image and the true one when a cubic phase mask of parameter $\alpha = 116$ is chosen, and when the defocus parameter Ψ_λ varies from 0 (no defocus) to 150 (important defocus). The image considered is the spoke pattern represented in Figure 3a.

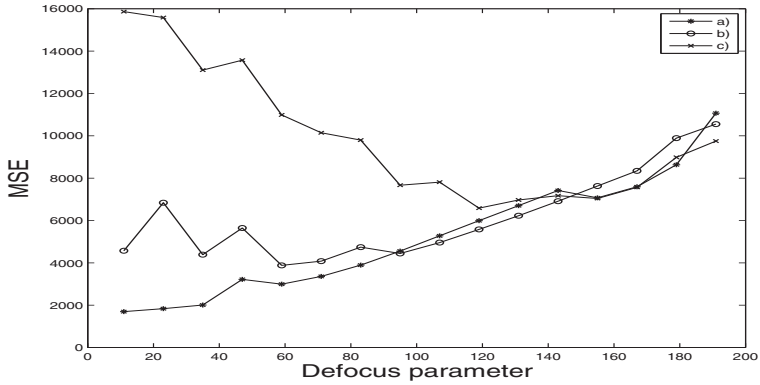


Fig. 12. Influence of the choice of the parameter α in the cubic phase mask for a fixed defocus a)$\Psi_{\lambda_1} = 10, b)\Psi_{\lambda_2} = 20, c)\Psi_{\lambda_3} = 50$.

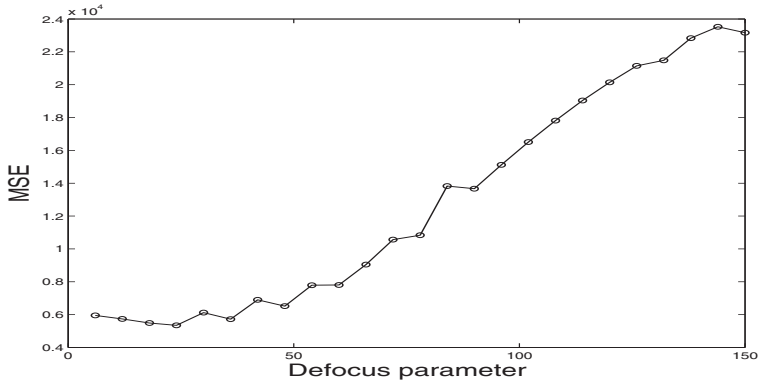


Fig. 13. Robustness of the optimization of the cubic phase mask towards defocus. The parameter α is taken equal to 116 which is the optimal value for a defocus parameter of $\Psi_{\lambda_3} = 50$ in the case of the spoke pattern.

It is clear that the parameter α depends on the defocus factor, however once the parameter α is fixed, similar results are obtained in term of MSE for a defocus parameter in $[0, 50]$. This result is illustrated on Figure 14 where the reconstructions b), d), f) are obtained with the same parameter ($\alpha = 116$) but considering respectively $\Psi_{\lambda_2} = 20$ (first line), $\Psi_{\lambda_3} = 50$ (second line), $\Psi_{\lambda_4} = 100$ (third line).

Once the phase mask is chosen, the optimization is a key point to obtain good quality results. However, the parameter α can takes its value within a range allowing the imaging system to give good results when different values of defocus are introduced. It could be interesting for example when the defocus parameter is not constant over the whole pupil.

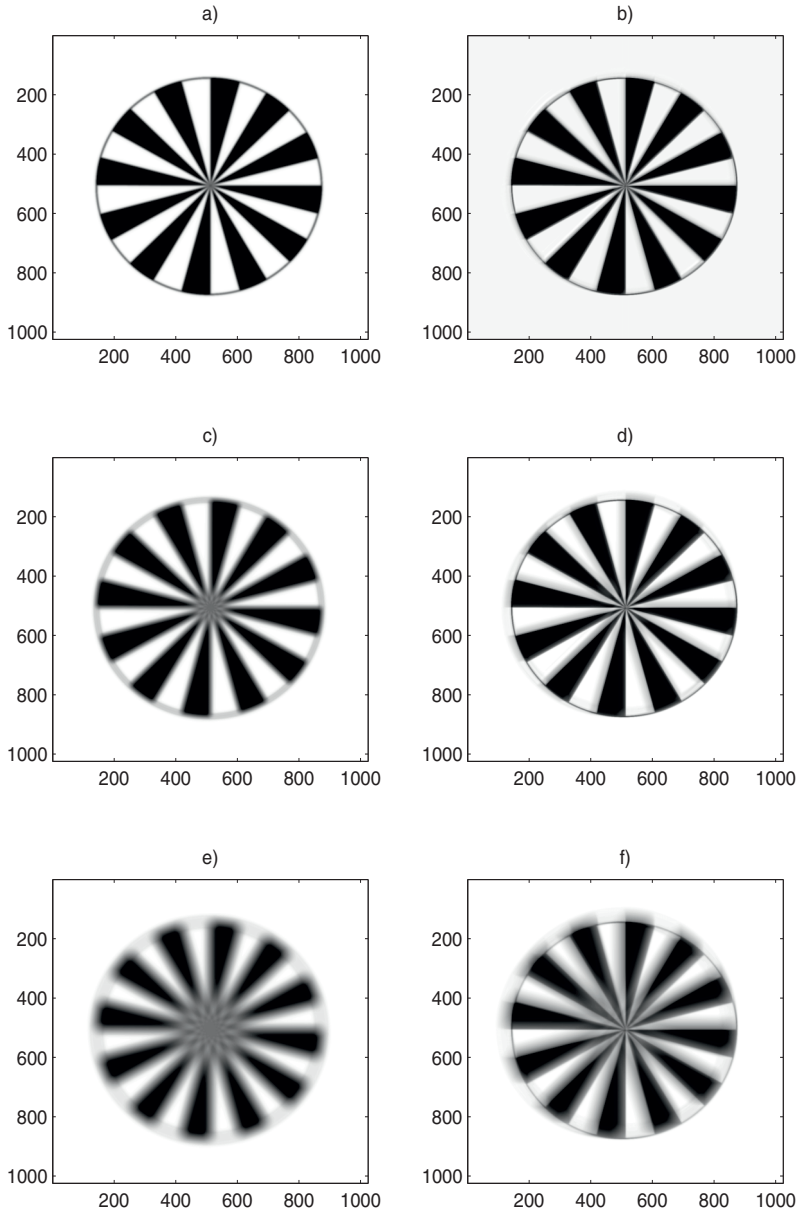


Fig. 14. Robustness of the optimization of the cubic phase mask towards defocus. Figures a), c), e) represent the degraded images with three different parameters of defocus $\Psi_{\lambda_2} = 20$, $\Psi_{\lambda_3} = 50$, $\Psi_{\lambda_4} = 100$. Figures b), d), f) represent the deconvolved images when a cubic phase mask of parameter $\alpha = 116$ is considered.

7 Conclusion

The wavefront coding is a technique that allows one, by introducing a pupil mask, to make insensitive the imaging system to some classical aberrations like defocus leading to increase the depth of field.

The use of wavefront coding, or other techniques that introduce the processing of the images jointly with the optics for the design of the imaging system, is going to increase in the following year. The reduction of the cost of the imaging system associated with the simplification of the conception, leading to high quality images after processing, make the hybrid imaging system of great interest.

The introduction of joint conception of optics and processing will introduce challenging tasks in next years to imagine or associate methods, to define new criterion to qualify the objective to reach, by adapting them to the target application.

References

Arnison, M.R., Cogswell, C.J., Sheppard, C.J.R., & P., Török, 2007, Optical Imaging Micros. Opt. Sci., 87, 169

Caron, N., & Sheng, Y., 2008, Appl. Opt., 47, E39

Castro, A., & Ojeda-Castaneda, J., 2004, Appl. Opt., 43, 3474

Diaz, F., Goudail, F., Loiseaux, B., & Huignard, J.-P., 2009, Opt. Lett., 34, 2970

Diaz, F., Goudail, F., Loiseaux, B., & Huignard, J.-P., 2010, J. Opt. Soc. Am. A, 2123

Dowski, E.R., Jr., & Cathey, A.T., 1995, Appl. Opt., 34, 1859

Cathey, W.T., & Dowski, E.R., 2002, Appl. Opt., 41, 6080

Goodman, J.W., 2005, Fourier Optics, Chapter 6 (Roberts and Company Publishers), 129

Kubala, K.S., Dowski, E., Kobus, J., & Brown, R., 2004, Proc. SPIE, 5524, Novel Optical Systems Design and Optimization VII, 54

Muyo, G., & Harvey, A.R., 2004, Proc. SPIE, 5612, 227

Narayanswamy, R., Johnson, G.E., Silveira, P.E.X., *et al.*, 2004, Appl. Opt., 44, 701

Neil, M.A.A., Kuskaitis, R., Wilson, T., & Laczik, Z.J., 2000, Opt. Lett., 25, 245

Prasad, S., Torgersen, T.C., Pauca, V.P., Plemmons, R.J., & van der Gracht, J., 2004, International Journal of Imaging Systems and Technology Special Issue: High-Resolution Image Reconstruction, 14, 6774

Sauceda, A., & Ojeda-Castaeda, J., 2004, Opt. Lett., 29, 560

Sherif, S., Cathey, T., & Dowski, E., 2004, Appl. Opt., 43, 2709

Wach, H.B., Dowski, E.R., & Cathey, W.T., 1998, Appl. Opt., 37, 5359

Yang, Q., Liu, L., & Sun, J., 2007, Opt. Comm., 272, 56

Zhao, H., Li, Q., & Feng, H., 2008, Opt. Lett., 33, 1171

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 93–130*

ADAPTIVE OPTICS FEEDBACK CONTROL

J.-P. Folcher[1], M. Carbillet[1], A. Ferrari[1] and A. Abelli[1]

Abstract. This paper concentrates on the control aspects of Adaptive Optics (AO) systems and includes a prior exposure to linear control systems from the "classical" point of view. The AO control problem is presented and the well-established optimized modal gain integral control approach is discussed. The design of a controller from a modern control point of view is addressed by means of a linear quadratic Gaussian control methodology. The proposed approach emphasizes the ability of the adaptive optics loop to reject the atmospheric aberration. We derive a diagonal state space system which clearly separates the dynamics of the plant (deformable mirror & wavefront sensor) from the disturbance dynamics (atmospheric model). This representation facilitates the numerical resolution of the problem. A frequency analysis is carried out to check performance and robustness specifications of the multiple-input multiple-output feedback system. The effectiveness of the approach is demonstrated through numerical experiments.

1 Introduction

Among its applications, adaptive optics systems can be used to reduce the effects of atmospheric turbulence on images taken from ground-based telescopes. A Deformable Mirror (DM) is used to spatially compensate the incoming (atmospheric) wavefront as close as possible to a theoretical plane wavefront. The shape of the DM is adjusted in real time using the measurements of a Wavefront Sensor (WFS) which provides the local slopes of the residual wavefront. The AO system imaging performance depends mainly on the WFS and DM characteristics and on the control algorithm efficiency. For an overview of AO, the reader may consult the book of Roddier (1999) and the companion chapter of Carbillet in this book. This paper concentrates on the control aspects of AO systems. Our intended audience includes researchers and research students in astrophysics and in signal processing

[1] UMR 7293, Lagrange Université de Nice Sophia-Antipolis/CNRS/Observatoire de la Côte d'Azur, Parc Valrose, 06108 Nice Cedex 2, France

DOI: 10.1051/eas/1359006

who are not familiar with control engineering. In this context the reader will benefit from a prior exposure to linear control systems from the "classical" point of view. This is the goal of the Section 2 which is an introduction of a lot of fundamental topics in control engineering such as feedback, Laplace transform, transfer function, Bode and Nyquist plots (Franklin *et al.* 1991; Dorf & Bishop 1998) which are illustrated with case studies. In this section we also expose some elements for digital controlled systems such as sampled-data systems, z-transform and discrete time transfer function (Franklin *et al.* 1990; Astrom & Wittenmark 2011) and we present basic case studies. Some paragraphs of the tutorial are selected passages or strongly inspired from the cited books. For instance the automobile cruise control example is presented in the book of Franklin *et al.* (1991). Our goal is not to teach the reader how to design linear controllers (several existing books do a good job for that) but rather to give a comprehensive understanding of feedback systems.

The third section is dedicated to the exposure of the Adaptive Optics control problem. The AO system is modeled as a *multiple-input multiple-output* (MIMO) feedback system using the "classical" control framework. A first category of control strategies: the *optimized modal gain integral control* (OMGI) proposed by Gendron & Léna (1994) and its improvements is discussed, see (Dessenne *et al.* 1998). A static decoupling matrix is inserted in the feedback loop in order to divide the MIMO control problem in a series of *single-input single-output* (SISO) control problem. The design parameters are chosen to ensure stability and a trade-off between disturbance rejection and measurement noise amplification. The main quality of the optimized modal gain integral control, which is the current adaptive optics control system is to express some of the controller's signals in the modal base which facilitates the physical interpretation. Furthermore it is intrinsically a frequency approach: the analysis of the AO feedback system's performance is straightforward. The method can be used when the knowledge of the disturbance temporal dynamics is weak.

The last section contains the design of a controller from a modern control point of view (Kulcsár *et al.* 2006; Looze 2006). This approach was introduced for the first time by Paschall *et al.* (1991), which explicitly tries to minimize the mean-square residual wavefront error (and consequently maximize the imaging performance index as the Strehl ratio). This problem can be formulated as a linear quadratic Gaussian (LGQ) control problem, and the solution consists in the optimal state-feedback control of the DM and the optimal estimation of the atmospheric wavefront. The proposed approach emphasizes the ability of the LQG controller loop to reject the atmospheric aberration. We propose a generic second order autoregressive model to capture the main features of the aberrated wavefront. We derive a diagonal state space system which clearly separates the dynamics of the plant (DM & WFS) from the disturbance dynamics (atmospheric model). Thus, we explicitly consider a disturbance rejection control problem, see (Bitmead *et al.* 1990), which facilitates the numerical resolution of the estimation problem: the order of the estimation discrete time algebraic Riccati equation is reduced. This point is important from a practical point of view for the new generation of

AO systems exhibiting a large number of modes where control laws have to be designed in accordance with real time constraints. Numerical experiments using the Software Package CAOS have been conducted to demonstrate the effectiveness of the proposed approach.

2 Classic feedback control: A tutorial

2.1 Definitions & feedback framework

2.1.1 What is automatic control?

Control is a general concept which refers to a specific interaction between two (or more) devices. Driving an automobile is a typical example: the driver has to control the vehicle to reach a given destination. In such a case, the car is manually controlled. At the opposite, *automatic control* only involves devices: this is the case of automobile cruise control. The rate flow of the fuel/air mixture is adjusted in real time depending on a speedometer measure to obtain a given speed.

2.1.2 What is feedback?

The main idea in control is *feedback control* where the variable/signal being controlled (speed, temperature...) is measured by a sensor and fed back to the process in order to influence the controlled signal. This feedback idea can be illustrated for the automobile cruise control and is described by a component *block diagram* in Figure 1. Main devices of the system are represented by blocks and arrows show interaction from one device to another.

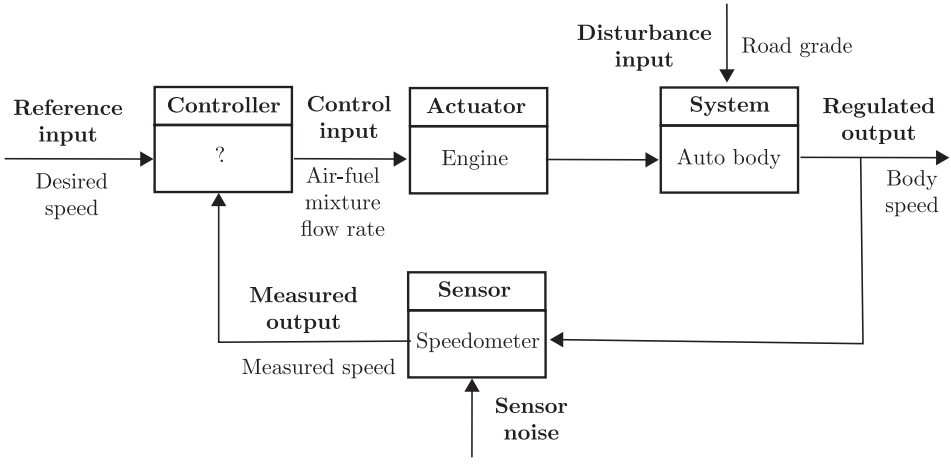


Fig. 1. Component block diagram of automobile cruise control.

Qualitatively, the temporal behavior of this controlled system can be analyzed. Suppose that when an air-fuel mixture is injected in the engine, the actual

measured speed is below the desired speed. Then, the cruise controller will increase the air-fuel mixture flow rate causing an increase of the engine speed and consequently the body vehicle speed. If the actual speed is higher than the desired speed then the cruise controller will decrease the air-fuel mixture flow rate in order to reduce the body vehicle speed. For this example, the generic components of a classic feedback loop are shown in Figure 1. The main component is called the system (or plant or process) where one variable/signal is to be controlled or regulated. In our example the plant is the automobile body and the controlled/regulated output is the vehicle speed. The disturbance input is the road grade which acts on the system. The actuator is the component that influences the regulated variable: here the actuator is the engine. To obtain a feedback, we need to deliver to the controller a measured output which is provided by the sensor. In this case, the sensor is the speedometer. The role of the controller is to generate, using the reference input and the measured output, the control input. Feedback control properties can be demonstrated using quantitative analysis of a simplified model of the automobile cruise control. We will neglect the dynamic response of the car by considering only the steady state case. We will assume that for the range of speed of the vehicle, the approximated relations are linear. For the automobile speed, we measure speed on a level road at 55 kilometers per hour (km/h) and find that a unit change in our control (injection pump input) causes a 10 km/h change in speed. When the grade changes by 1%, we measure a speed change of 5 km/h. The accuracy of the speedometer is sufficient and can be considered exact. These relations permit to obtain the bloc diagram shown in Figure 2.

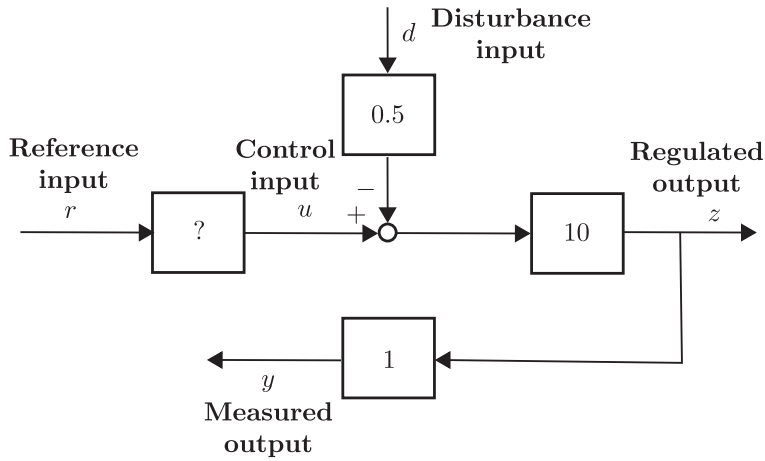


Fig. 2. Block diagram of automobile cruise feedforward control.

Here lines represent signals as regulated output z, control input u, disturbance input d, measured output y and reference input r. Squared/round blocks represent respectively multiplication and summation. In Figure 2, the *feedforward controller* does not use the body speed. A possible control policy consists in inverting the

plant characteristic and the controller sets $u = r/10$. In this case we obtain the regulated output speed

$$\begin{aligned} z &= 10(u - 0.5d) \\ &= 10([r/10] - 0.5d) \\ &= r - 5d. \end{aligned}$$

If $d = 0$ (a level road) and $r = 55$ then the vehicle speed will be $z = 55$ and there will be no error. However if $d = 1$ (a 1% grade) then the speed will be $r = 50$ and we have a 5 km/h error in speed.

In contrast to feedforward control, a *feedback controller* uses the measure of the controlled output (called the feedback signal) as in Figure 3 where the control input is $u = r - y = r - [0.9z]$.

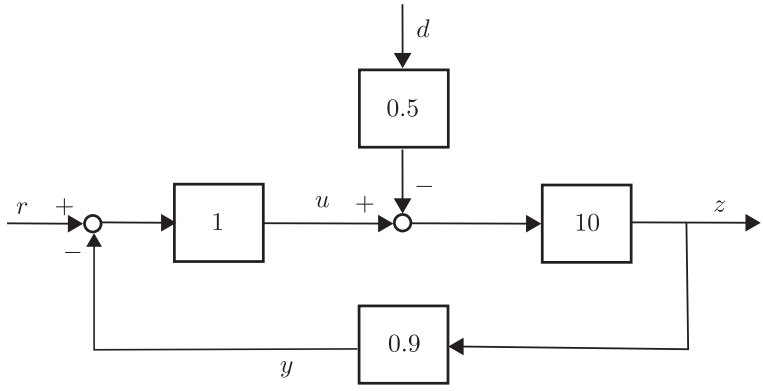


Fig. 3. Component block diagram of automobile cruise feedback control.

The topology of this block diagram include includes a loop: this is a *closed loop control system*. At the opposite the configuration shown in Figure 2 is called *open loop control system*. The equations of the closed loop control system are

$$\begin{aligned} z &= 10(u - 0.5d) \\ &= 10([r - 0.9z] - 0.5d) \\ &= 10r - 9z - 5w \end{aligned}$$

and finally

$$z = r - d/2.$$

In this case, if the reference speed is still $r = 55$ and the grade $d = 1$ then the vehicle speed will be $y = 54.5$ and the error is 0.5 km/h. The effect of feedback is to reduce the speed error by a factor 10! If we include a gain factor for the controller greater than 1 the error will still decrease. But there is a limit for the gain value

due to the power of the engine but more importantly because when the dynamics are introduced, feedback may induce poor temporal response (stability problems). As Stephen P. Boyd contends in (Boyd 1993), "*a bad feedback controller can yield performance much worse than an open loop controller*".

2.1.3 Quantitative analysis in the time domain: A tentative

In order to analyze a feedback controlled system we need to obtain a quantitative mathematical model of the plant. In this paper we assume that the process under study can be considered as *linear* over a reasonably large range of the signals and *time invariant*. That is, a mathematical model is frequently a set of ordinary differential equations and a specific solution can be found using a computer program. The output s of a general time invariant linear system, in the time domain, is given by the *convolution integral*

$$s(t) = (h * e)(t) = \int_0^t h(\tau)e(t-\tau)d\tau, \tag{2.1}$$

where $e(t)$ is the input signal and where $h(t)$ is the *impulse response*. We can use the bloc diagram notation given in Figure 4.

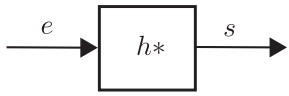


Fig. 4. Block diagram notation of the convolution operation.

This generic block diagram may describe every component of a feedback system as the controller, the actuator and the sensor. We note respectively k, g_1 and g_2 their impulse response. We study now a feedback system shown in Figure 5. The block diagram resembles an automobile cruise block diagram depicted in Figure 3. We require that the regulated output z becomes zero: this is a *disturbance rejection control problem*. Thus the reference signal r is zero and is not represented in the block diagram. We consider a more realistic model of the sensor: an additive sensor noise n is taken into account. We will see later in the paper that this block diagram is a simplified model of an AO control loop.

The equation of the feedback system is

$$z = d - \overbrace{\Big\{ g_1 * \big[\underbrace{k * (\overbrace{g_2 * z + n}^{y})}_{u} \big] \Big\}}^{c}, \tag{2.2}$$

which can be rewritten as

$$z = d - (g_1 * k) * n - (g_1 * k * g_2) * z. \tag{2.3}$$

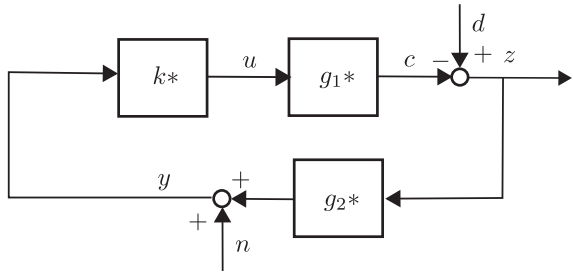


Fig. 5. Convolution based block diagram of the feedback system.

We have a complicated convolution Equation (2.3): the regulated output z is the sum of the disturbance signal d, of the signal $(g_1 * k) * n$ and of the signal $(g_1 * k * g_2) * z$. This last signal is the response of the cascaded system with impulse response $g_1 * k * g_2$ where the input is the regulated input z. The regulated output z depends on itself: this is a feature of the feedback systems. In the time domain we have a complex convolution Equation (2.3) which is not easy to understand or to solve. We will see that in the frequency domain the computation and the interpretation of the transformed equation is straightforward.

2.2 Feedback systems: A frequency approach

2.2.1 Laplace transform & transfer functions

The *Laplace transform* is well suited to find the solution of Equation (2.2) and to give interesting information (settling time, overshoot, final value) of feedback systems. The Laplace transform of a signal $f(t)$ is defined as

$$\mathcal{L}\{f\}(s) = \int_0^\infty f(t)e^{-st}dt. \tag{2.4}$$

A straightforward consequence of convolution integral (2.1) is

$$\mathcal{L}\{s\}(s) = H(s)\mathcal{L}\{e\}(s), \tag{2.5}$$

where $H(s) = \mathcal{L}\{h\}(s)$ is called the *transfer function*. Thus the Laplace transform of the output $\mathcal{L}\{s\}$ is the product of the transfer function H and of the Laplace transform $\mathcal{L}\{e\}$. In the frequency domain Equation (2.5) is the counterpart of convolution integral (2.1) in the time domain. We can use the bloc diagram notation given in Figure 4.

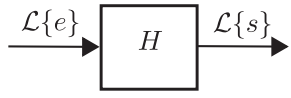


Fig. 6. Block diagram notation of the transfer function.

2.2.2 Feedback system's transfer functions

The feedback system shown in Figure 5 can be "translated" in the frequency domain. We call $G_1(s) = \mathcal{L}\{g_1\}(s)$, $G_2(s) = \mathcal{L}\{g_2\}(s)$ and $K(s) = \mathcal{L}\{k\}(s)$ respectively the actuator transfer function, the sensor transfer function and the controller transfer function. The block diagram is drawn again: the controller's block is moved at the bottom and the sensor's block is displaced at the top.

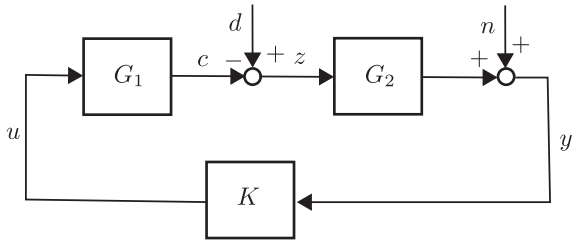


Fig. 7. Transfer function based block diagram of the feedback system.

In the frequency domain blocks $G_1(s)$, $G_2(s)$ and $K(s)$ are simple scaling systems. From block diagram in Figure 7 we obtain

$$\mathcal{L}\{z\} = \mathcal{L}\{d\} - \Big\{ G_1\big[K\big(\overbrace{G_2\mathcal{L}\{z\} + n}^{\mathcal{L}\{y\}}\big)\big]_{\mathcal{L}\{u\}} \Big\}^{\mathcal{L}\{c\}},$$

which can be solved as

$$\mathcal{L}\{z\} = \underbrace{\frac{1}{1 + G_1KG_2}}_{S}\mathcal{L}\{d\} - \underbrace{\frac{G_1K}{1 + G_1KG_2}}_{T}\mathcal{L}\{n\}. \tag{2.6}$$

To understand how controllers ensure relevant properties for the feedback system, the Equation (2.6) is central. We call

$$L = G_1KG_2 \tag{2.7}$$

the *loop transfer function*,

$$S = \frac{1}{1 + L} \tag{2.8}$$

the *sensitivity transfer function*, and

$$T = G_1KS \tag{2.9}$$

the *noise sensitivity transfer function*. For "ideal control" we want $z = 0$ and consequently

$$\mathcal{L}\{z\} \approx 0\mathcal{L}\{d\} + 0\mathcal{L}\{n\}. \tag{2.10}$$

Disturbance rejection is achieved when $S \approx 0$ and noise rejection is ensured when $T \approx 0$. In practice these two transfer function S and T cannot be small at the same values of s and a trade off should be achieved during the design of the controller transfer function K.

2.3 Standard examples

In this section we present two case studies to illustrate the concepts introduced in the preceding paragraph. We will also study the properties ensured both in the frequency domain and in the time domain for classical controllers (proportional and integral).

2.3.1 Case study 1

We suppose that the actuator and the sensor have instantaneous responses:

$$G_1(p) = \alpha \ , \ \ G_2(p) = \beta, \tag{2.11}$$

where α and β are fixed positive scalar. We use a *proportional controller* which produces the control input

$$u(t) = k_P y(t), \tag{2.12}$$

where the scalar k_P is the *proportional gain*. We also consider an *integral controller* which imposes the control input

$$u(t) = k_I \int_0^t y(\tau) d\tau, \tag{2.13}$$

where the parameter k_I is the *integrator gain*. Time domain Equations (2.12) and (2.13) can be cast under the convolution integral form $k * y$ with impulse response $k(t) = k_P \delta(t)$ and $k(t) = k_I$. Hence, controller transfer function K can be calculated. For numerical purpose, we set the actuator's gain $\alpha = 10$ and the sensor's gain $\beta = 1$. We consider a proportional controller with the gain $k_P = 0.2$, an integral controller with the gain $k_I = 0.4$ and another integral controller with the gain $k_I = 1$. These controllers are

$$K^{(a)}(s) = 0.2, \quad K^{(b)}(s) = \frac{0.4}{s}, \quad K^{(c)}(s) = \frac{1}{s}\cdot \tag{2.14}$$

The corresponding sensitivity transfer function, which we denote $S^{(a)}(s)$, $S^{(b)}(s)$, and $S^{(c)}(s)$ respectively, can be computed from (2.8). The closed-loop systems that result from using the controllers $K^{(a)}$, $K^{(b)}$, and $K^{(c)}$ can be compared by examining the sensitivity transfer function $S^{(a)}$, $S^{(b)}$, and $S^{(c)}$. The magnitudes $\left|S^{(a)}(j\omega)\right|$, $\left|S^{(b)}(j\omega)\right|$, and $\left|S^{(c)}(j\omega)\right|$ are plot in Figure 8a. From this figure we can conclude that a low frequency disturbance input will have the least effect in the feedback system with controller $K^{(c)}$ *i.e.* the best disturbance rejection performance. The real disturbance input is usually unknown. A reasonable approach is

to choose a standard test input signal as a step $d(t) = 1$ shown in Figure 9a. This step response checks the ability of the system to perform under normal operating conditions using generic test input signals as a step $d(t) = 1$ shown in Figure 9a. The step responses of the sensitivity transfer function are shown in Figure 8b. From this figure it can be seen that the controller $K^{(c)}$ ensures the faster decay of the transient response.

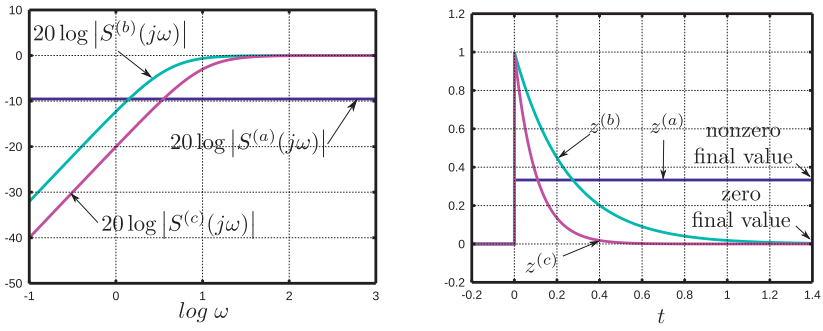


Fig. 8. (a) Magnitude of the sensitivity transfer functions $S^{(a)}$, $S^{(b)}$, and $S^{(c)}$. (b) The step responses from disturbance input d to regulated output z for the sensitivity transfer functions $S^{(a)}$, $S^{(b)}$, and $S^{(c)}$.

The step responses from the disturbance input d to the control input z for the three feedback systems are shown in Figure 9b. For integral controllers $K^{(b)}$ and $K^{(c)}$, final value of their output (control input) is zero when final value of the regulated output z is zero. This is an important feature of integral controllers which ensures zero steady-state error for the actuator/plant/sensor configuration given in (2.11).

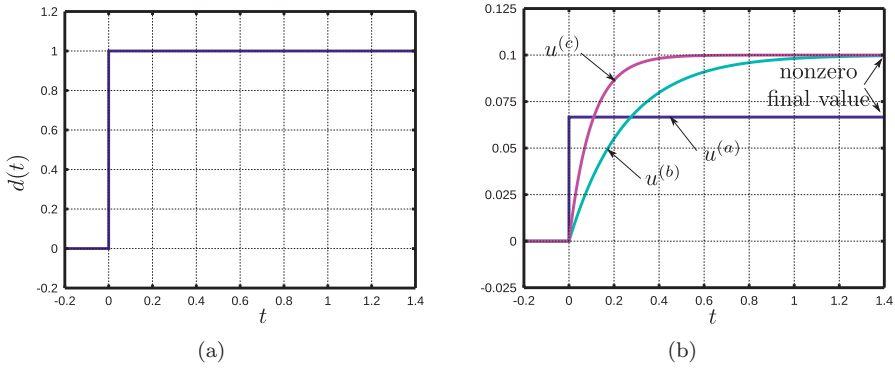


Fig. 9. (a) A step signal d. (b) The step responses from disturbance input d to control input z for the transfer functions $K^{(a)}S^{(a)}$, $K^{(b)}S^{(b)}$, and $K^{(c)}S^{(c)}$.

2.3.2 Case study 2

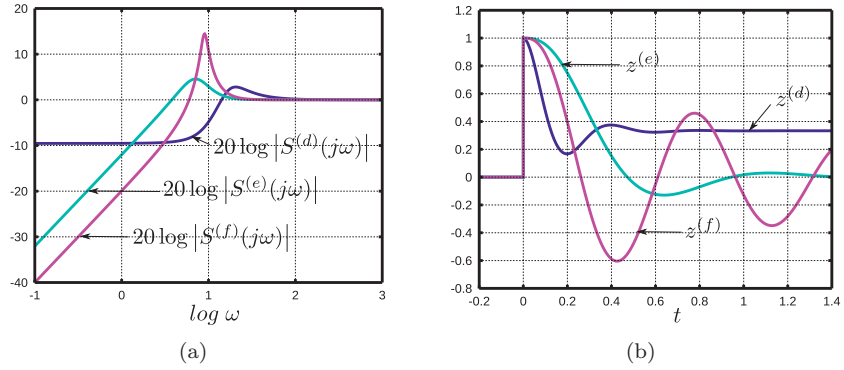
In this section we still consider the standard closed loop system shown in Figure 7. The sensor always has an instantaneous response but here the actuator is a second order dynamical system

$$G_1(p) = \alpha \frac{\omega_n^2}{p^2 + 2\zeta\omega_n p + \omega_n^2}, \quad G_2(p) = \beta. \tag{2.15}$$

For a numerical purpose, we conserve the actuator's gain $\alpha = 10$ and the sensor's gain $\beta = 1$ of Section 2.3.1. We set the damping factor $\zeta = 0.7$ and the natural frequency $w_n = 10$. The controller transfer functions are given in (2.14). The corresponding loop transfer function, which we denote $L^{(d)}(s)$, $L^{(e)}(s)$, and $L^{(f)}(s)$ respectively, can be computed from (2.7). The same notation holds for

- the sensitivity transfer function $S^{(d)}(s)$, $S^{(e)}(s)$, $S^{(f)}(s)$ calculated from (2.8);
- the noise sensitivity transfer function $T^{(d)}(s)$, $T^{(e)}(s)$, $T^{(f)}(s)$ computed from (2.9).

The magnitudes $\left|S^{(d)}(j\omega)\right|$, $\left|S^{(e)}(j\omega)\right|$, and $\left|S^{(f)}(j\omega)\right|$ are plot in Figure 10a. These plots should be compared to the plots depicted in Figure 8a. From this figure we can conclude that a low frequency disturbance input will have the least effect on the feedback system with controller $K^{(c)}$ *i.e.* the best disturbance rejection performance. In the low frequencies domain the remarks in Section 2.3.1 should be similar but there is a large peak of the magnitude $\left|S^{(f)}(j\omega)\right|$. We can conclude that the feedback system with controller $K^{(c)}$ is not stable enough.



(a) (b)

Fig. 10. (a) Magnitude of the sensitivity transfer functions $S^{(a)}$, $S^{(b)}$, and $S^{(c)}$. (b) The step responses from disturbance input d to regulated output z for the sensitivity transfer functions $S^{(a)}$, $S^{(b)}$, and $S^{(c)}$.

The Nyquist plots of the loop transfer function $L^{(d)}$, $L^{(e)}$ and $L^{(f)}$ are shown in Figure 11. The Nyquist plot of $L^{(f)}(j\omega)$ is too close to the -1 point, see

(Franklin *et al.* 1991). We can corroborate that the stability margins are small for the feedback system with the controller $K^{(c)}$. The step responses of the sensitivity transfer function are shown in Figure 10b. From this figure it can be seen that the controller $K^{(c)}$ has a poor transient response. z plot exhibits oscillatory behavior: the damping ratio of the feedback system is weak. Thus controller $K^{(b)}$ is selected to be the operating controller.

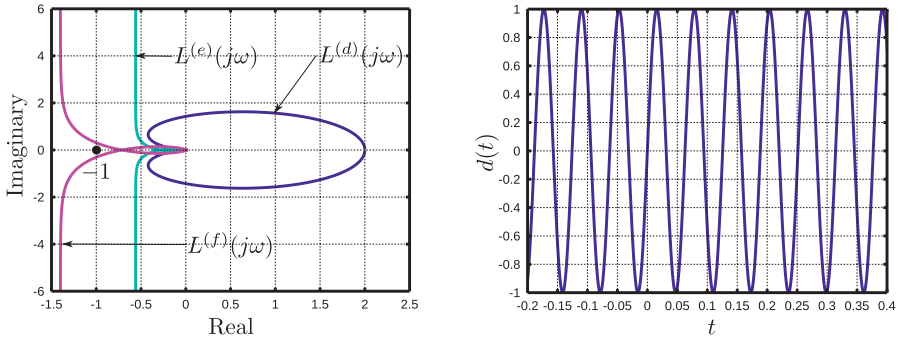


Fig. 11. (a) Nyquist plot for the loop transfer functions $L^{(d)}(s)$, $L^{(e)}(s)$ and $L^{(f)}(s)$. (b) A sinusoidal signal d.

To assess the noise rejection performance we plot the magnitude of the noise sensitivity transfer function $T^{(d)}$, $T^{(e)}$ and $T^{(f)}$. Figure 12 shows $\left|T^{(d)}(j\omega)\right|$, $\left|T^{(e)}(j\omega)\right|$, and $\left|T^{(f)}(j\omega)\right|$, *i.e.*, the magnitudes of the feedback system transfer functions from measurement noise n to regulated output z. From this figure, we can conclude that a high frequency sensor noise will have the greatest effect on z with the controller $K^{(a)}(s)$ and the least effect with the controller $K^{(b)}(s)$. For a given controller, for instance $K^{(b)}(s)$, remark that the magnitude $\left|S^{(e)}(j\omega)\right|$ and $\left|T^{(e)}(j\omega)\right|$ cannot be small in the same frequency domain.

The response of the noise sensitivity transfer function from a sinusoidal disturbance input d plotted in Figure 11b are shown in Figure 10b. From this figure it can be seen that the sinusoidal steady-state response of the feedback system with controller $K^{(b)}(s)$ is the smallest. Controller $K^{(b)}(s)$ ensures the best noise rejection performance.

2.4 Digital controlled systems

2.4.1 Sampled-data feedback system

In practice all control systems that are implemented today are based on a digital computer. A computer controlled system is sketched schematically in Figure 13. This block diagram is very similar to block diagram depicted in Figure 7, except for a digital device which generates the control action. The *analog-to-digital (A/D)*

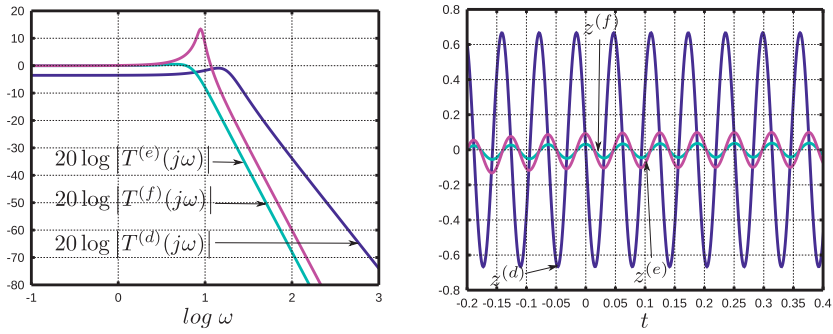


Fig. 12. (a) Magnitude of the noise sensitivity transfer functions $T^{(a)}$, $T^{(b)}$, and $T^{(c)}$. (b) The responses from sinusoidal disturbance input d to regulated output z for the sensitivity transfer functions $T^{(a)}$, $T^{(b)}$, and $T^{(c)}$.

converter shown in Figure 13 is a device that converts the sensor output $y(t)$ to digital numbers read by the computer. We assume that all the numbers arrive with the same fixed period T and we neglect the quantization operation thus

$$y(k) = y(t)|_{t=kT}\,. \tag{2.16}$$

The computer interprets the converted signal, $y(k)$ as a sequence of numbers, processes the measurements using an algorithm, and provides a new sequence of numbers $u(k)$. The *digital-to-analog (D/A) converter* converts the sequence of number $u(k)$ to the physical control signal $u(t)$. In many case the signal $u(t)$ is kept constant between the successive sampling instants

$$u(t) = u(k) \quad kT \le t < (k+1)T. \tag{2.17}$$

We call variables $y(k)$ and $u(k)$ *discrete time signals* to distinguish them from *continuous time signals* $y(t)$ and $u(t)$ which change continuously in time. The computer-controlled system contains both continuous-time signals and discrete-time signals and is called a *sampled-data system.*

For a numerical purpose, we assume that the actuator and the sensor are fading memory systems with transfer function

$$G_1(p) = \frac{\alpha}{0.1s+1}, \quad G_2(p) = \frac{\beta}{0.1s+1}, \tag{2.18}$$

and we retain the actuator's gain $\alpha = 10$ and the sensor's gain $\beta = 1$ of Section 2.3.1. The sampling period is $T = 0.2$ and the disturbance input is a step $d(t) = 1$ shown in Figure 9a. The control sequence $u(k)$ is obtained from the measurement sequence $y(k)$ using the control algorithm

$$u(k) = u(k-1) + k_I T y(k), \tag{2.19}$$

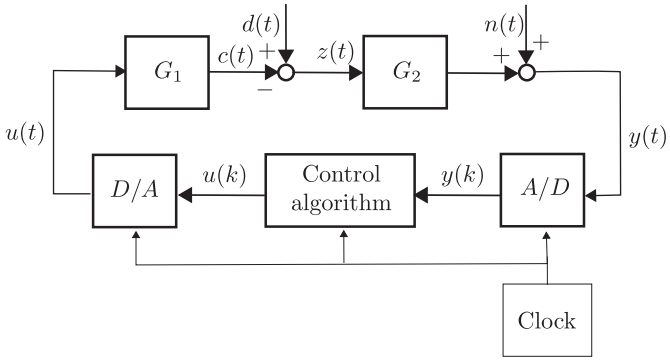


Fig. 13. Sampled-data feedback system.

where $k_I = 0.5$. The behavior of the A/D converter is illustrated in Figure 14. Figure 15a is a plot of the sequence of numbers $u(k)$ obtained from the sequence of numbers $y(k)$ plotted in Figure 14a. Note that the D/A converter keeps the signal $u(t)$ constant between the successive sampling instant kT, see the Figure 15b.

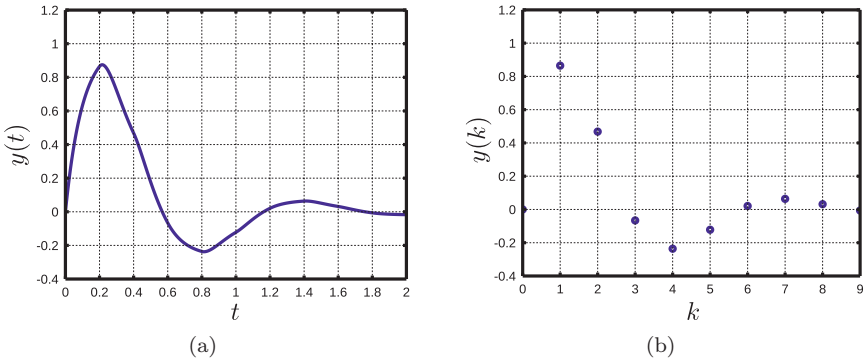


Fig. 14. Analog-to-digital (A/D) converter operation: (a) measured output $y(t)$, (b) control algorithm input $y(k)$.

For the sake of brevity we do not discuss sampling and reconstruction of continuous-time signals. For a comprehensive exposure, the interested reader may consult the book of Astrom & Wittenmark (2011). Remark that to avoid aliasing effect, it is necessary to filter the analog signal $y(t)$ before the A/D converter so that the signal obtained do not have frequencies above the Nyquist frequency. Note that the output of the D/A are rectangular pulses which causes multiple harmonics above the Nyquist frequency. This may cause difficulties for systems with weakly damped oscillatory modes. If needed, the multiple harmonics could be removed with a low pass filter acting as a reconstruction filter. The overall

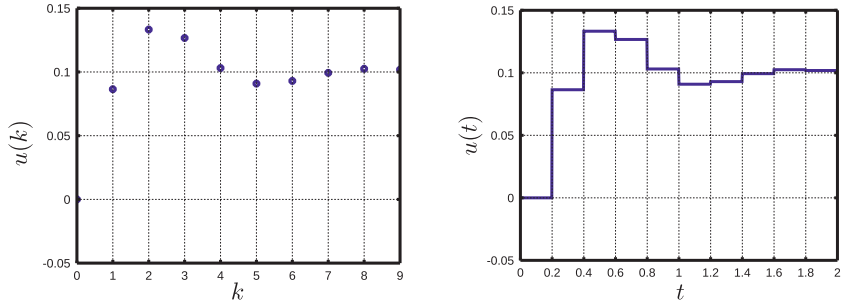


Fig. 15. Digital-to-analog (D/A) converter operation: (a) control algorithm output $u(k)$, (b) control input $u(t)$.

behavior of this hybrid feedback system which incorporates both continuous time signals and discrete time signals can be studied by two different approaches.

1. The first approach, called the *emulation design method*, see (Franklin *et al.* 1991), deals with continuous time transfer function. In this case the digital computer behavior shown in Figure 16 is approximated by an equivalent continuous time system described by transfer function $K(s)$, see Figure 17. The overall feedback system is assumed to be continuous and the continous time framework presented in Sections 2.3.1 and 2.3.2 can be used considering the feedback loop depicted in Figure 7. This approach is discussed in Section 2.4.2.

2. For the latter approach the sampled-data feedback system is transformed into a discrete time feedback system. For this purpose the continuous part of the system is sampled as seen from the digital computer's point of view. The resulting feedback system is characterized by a discrete time transfer function using the z-transform. In this case discrete time controller design methods may be used. An analysis of the feedback discrete time system is performed in Section 2.4.3.

2.4.2 Emulation design method

The output of an integral controller (2.13) at time $t = kT$ is

$$\begin{aligned} u(kT) &= k_I \int_0^{kT} y(\tau)d\tau \\ &= k_I \int_0^{kT-T} y(\tau)d\tau + k_I \int_{kT-T}^{kT} y(\tau)d\tau \\ &= u(kT-T) + k_I \underbrace{\int_{kT-T}^{kT} y(\tau)d\tau}_{I} . \end{aligned}$$

Several approximations of the incremental term I can be chosen as for instance the backward rectangular rule $I \approx Ty(kT)$. Hence we obtain

$$\underbrace{u(kT)}_{u(k)} = \underbrace{u(kT-T)}_{u(k-1)} + k_I T \underbrace{y(kT)}_{y(k)},$$

which is equivalent to Equation (2.19). Thus the digital computer with algorithm defined by Equation (2.19) is a discrete time equivalent to the continuous time controller $K(s) = k_I/s$.

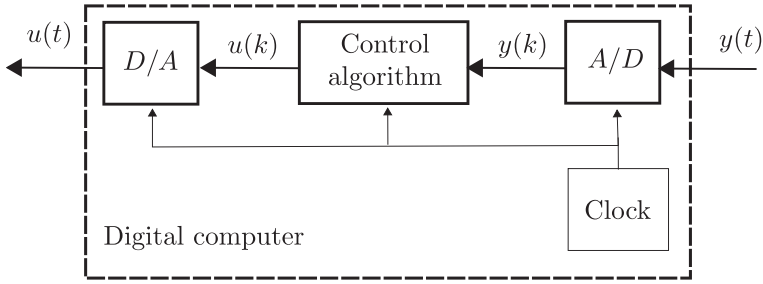


Fig. 16. Association of the A/D converter with the control algorithm and with the D/A converter.

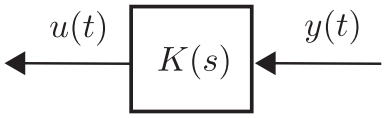
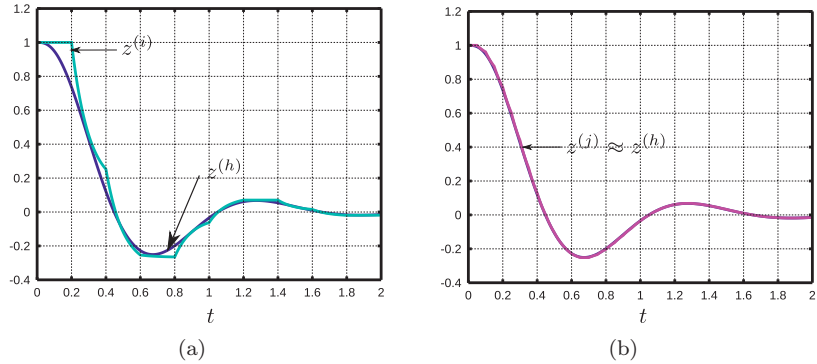


Fig. 17. Equivalent transfer function $K(s)$.

We consider that the feedback system is described by the block diagram shown in Figure 7. For the given transfer functions G_1, G_2 defined by (2.18), standard continuous time design method can be used to obtain the integral controller

$$K^{(h)}(s) = \frac{0.5}{s}.$$

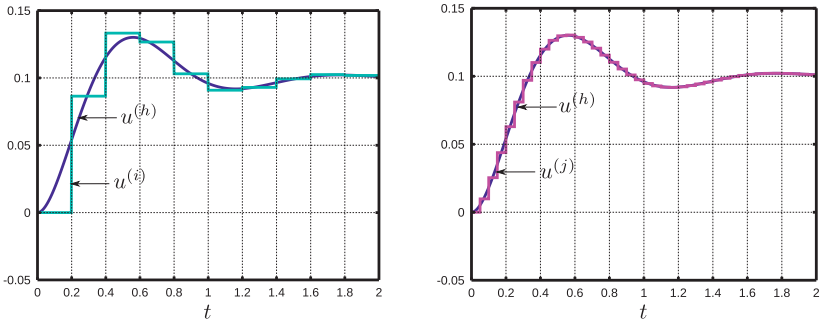
This continuous time controller is approximated with the difference Equation (2.19) and we call $K^{(i)}$ and $K^{(j)}$ the discrete time controller with the sampling period $T = 0.2$ and $T = 0.05$. We assume that the disturbance input d is a step. We called $z^{(h)}$ the "ideal" regulated output response of the continuous time feedback system, $z^{(i)}$ the regulated output response of the sampled-data feedback system when the discrete time controller is $K^{(i)}$, and $z^{(j)}$ the regulated output response of the sampled-data feedback system when the discrete time controller is $K^{(j)}$. These signals are plotted in Figure 18a an Figure 19a. From these figures we can conclude that $z^{(j)}$ is the best approximation of the "ideal" regulated output response $z^{(h)}$.



(a) (b)

Fig. 18. "Ideal" regulated output response $z^{(h)}$ of the continuous time feedback system, regulated output response $z^{(i)}$ of the sampled-data feedback system when $T = 0.2$, and regulated output response $z^{(j)}$ of the sampled-data feedback system when $T = 0.05$.

The "ideal" input response $u^{(h)}$ of the continuous time feedback system, the regulated output response $u^{(i)}$ of the sampled-data feedback system when the discrete time controller is $K^{(i)}$, and the regulated output response $u^{(j)}$ of the sampled-data feedback system when the discrete time controller is $K^{(j)}$ are shown in Figure 18b and in Figure 19b. It can be seen that the response $u^{(j)}$ matches the "ideal" response $u^{(h)}$. We can conclude that clearly the sampling period $T = 0.2$ is too rough and that the sampling period $T = 0.05$ ensures a satisfactory performance. As mentioned by Franklin *et al.* (1991), "*sampling at a rate that is over 20 times faster than the bandwidth is a good, safe rule of thumb*".



(a) (b)

Fig. 19. "Ideal" regulated output response $u^{(h)}$ of the continuous time feedback system, regulated output response $u^{(i)}$ of the sampled-data feedback system when $T = 0.2$, and regulated output response $u^{(j)}$ of the sampled-data feedback system when $T = 0.05$.

2.4.3 Discrete time controller design

In Figure 13 the job of the digital computer is to take the sampled value $y(k)$ and to compute the values $u(k)$ to be sent to the D/A converter. The treatment of the data inside the computer can be expressed as a *linear difference equation* as for example the Equation (2.19), which describes a *discrete time invariant linear system*. In Section 2.2 the Laplace transform plays an important role and permits to introduce the transfer function and frequency interpretation of the closed loop system shown in Figure 7. The discrete-time analog of the Laplace transform is the z-transform which is a convenient tool to study general discrete linear systems. The z-transform of a signal $y(k)$ is defined as

$$\mathcal{Z}\{y\}(z) = \sum_{k=0}^{\infty} y(k) z^{-k}, \tag{2.20}$$

where z is a complex variable. If we multiply (2.19) by z^{-k} and sum over k we obtain

$$\underbrace{\sum_{k=0}^{\infty} u(k) z^{-k}}_{\mathcal{Z}\{u\}(z)} = \sum_{k=0}^{\infty} u(k-1) z^{-k} + k_I T \Big(\underbrace{\sum_{k=0}^{\infty} y(k) z^{-k}}_{\mathcal{Z}\{y\}(z)} \Big). \tag{2.21}$$

In the first term on the right hand side, we let $k-1 = j$ to get $\sum_{k=0}^{\infty} u(k-1) z^{-k} = \sum_{j=1}^{\infty} u(j) z^{-(j+1)} = z^{-1} \mathcal{Z}\{u\}$. Equation (2.20) can be can be rewritten as

$$\mathcal{Z}\{u\}(z) = z^{-1} \mathcal{Z}\{u\}(z) + k_I T \mathcal{Z}\{y\}(z) \tag{2.22}$$

which is simply an algebraic equation in z. The solution is

$$\mathcal{Z}\{u\}(z) = \underbrace{k_I T \frac{z}{z-1}}_{K(z)} \mathcal{Z}\{y\}(z). \tag{2.23}$$

We have obtained

$$\mathcal{Z}\{u\}(z) = K(z) \mathcal{Z}\{y\}(z) \tag{2.24}$$

where $K(z) = k_I T \frac{z}{z-1}$ is called the discrete time *transfer function*. Thus the z-transform of the output $\mathcal{Z}\{u\}$ is the product of the transfer function K and the z-transform $\mathcal{Z}\{y\}$. We can use the bloc diagram notation given in Figure 20.

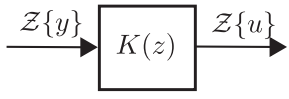


Fig. 20. Block diagram notation of the transfer function.

All the framework presented in Section 2.3 for analyzing continuous time systems can be extended to discrete time systems. We consider a discrete time system

with the associated block diagram shown in Figure 21 where $K(z)$ is the controller transfer function, $G_1(z)$ is the actuator transfer function, and $G_2(z)$ is the sensor transfer function. This block diagram is similar to the block diagram depicted in Figure 7.

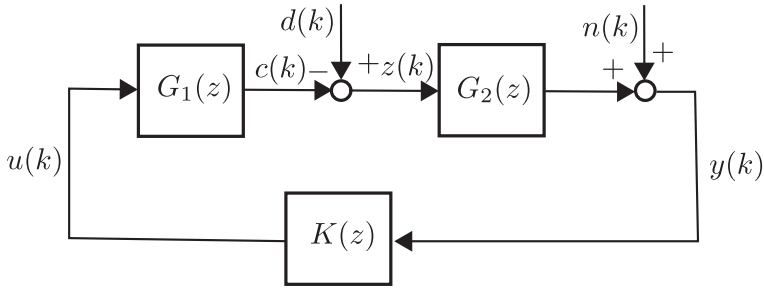


Fig. 21. Discrete time feedback system.

Hence the regulated output response is

$$\mathcal{Z}\{z\} = \underbrace{\frac{1}{1+G_1(z)K(z)G_2(z)}}_{S(z)}\mathcal{Z}\{d\} - \underbrace{\frac{G_1(z)K(z)}{1+G_1(z)K(z)G_2(z)}}_{T(z)}\mathcal{Z}\{n\}\,. \qquad (2.25)$$

We still use the following terminology: (i) $L(z) = G_1(z)K(z)G_2(z)$ is the loop transfer function; (ii) $S(z) = \frac{1}{1+L(z)}$ is the sensitivity transfer function; (iii) $T(z) = G_1(z)K(z)S(z)$ is the noise sensitivity transfer function. All results presented in Section 2.3 for continuous time feedback systems are relevant for discrete time feedback systems.

The main difficulty concerns the correspondence between this block diagram shown in Figure 21 and the block diagram of the "real" sampled data feedback system depicted in Figure 13. This block diagram is redrawn in Figure 22 to make the comparison easier.

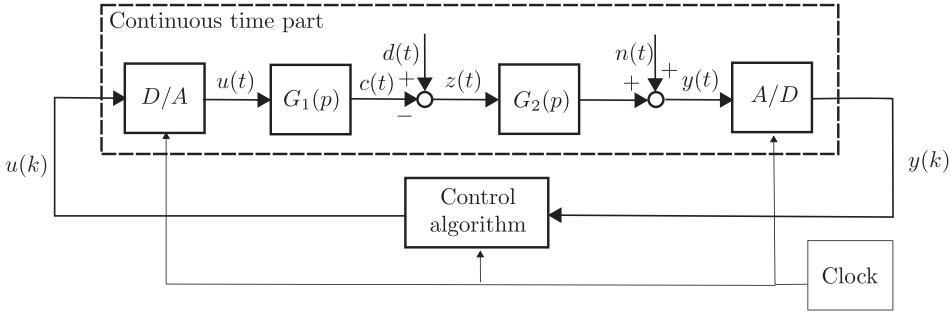


Fig. 22. Sampled-data feedback system.

It is obvious to note that the control algorithm is represented by the transfer function $K(z)$. For a "perfect" connection: (i) the discrete time transfer function $G_1(z)$ should be viewed as the composition of the D/A converter system and the actuator transfer function $G_1(p)$; (ii) the discrete time transfer function $G_2(z)$ should describe the actuator transfer function $G_2(p)$ and the A/D converter. In general this connection is not possible and a deeper analysis should be performed using the pulse transfer function formalism, see (Franklin *et al.* 1990; Astrom & Wittenmark 2011) which is beyond the scope of this tutorial. However in the absence of continuous time disturbance d, the discretization of the continuous time part of sampled data feedback system is a standard result, see Franklin *et al.* (1990) and allows to obtain the aggregated/global transfer function $G_1(z)G_2(z)$. But this global transfer function cannot be split in order to obtain transfer function $G_1(z)$ and transfer function $G_2(z)$. Yet for some special case of sensor transfer function $G_2(p)$ as CCD-based sensor, see (Looze 2005), the connection of the sampled data feedback system's block diagram shown in Figure 22 and the discrete time feedback system's block diagram shown in Figure 21 is faithful.

3 Adaptive optics feedback control

3.1 Problem statement and wavefront spatial discretization

Among its applications, AO systems can be used to reduce the effects of atmospheric turbulence on images taken from ground-based telescopes. The principle of a classical AO system is depicted in Figure 23. The atmospheric wavefront on the telescope aperture, defined at instant t as the two dimensional function $\psi_a(x,t)$, is the input of the feedback system. The deformable mirror introduces a correction denoted by $\psi_m(x,t)$ which is subtracted from the incoming/atmospheric wavefront to obtain the outcoming/residual wavefront

$$\psi_r(x,t) = \psi_a(x,t) - \psi_m(x,t). \tag{3.1}$$

The shape of the DM is adjusted in real time using the measurements y of a wavefront sensor which provides the local slopes of the residual wavefront, see Figure 24.

There exists different type of deformable mirrors and we choose to study the case of the most common one. For additional details on basic principles of adaptive optics, the reader can consult (Roddier 1999). We assume that the frequency bandwidth of the DM is higher than the bandwidth of the A0 loop. Moreover the DM's deformation is sufficiently small to consider a linear response. n_u actuators are used and we denote $a_i(t)$ the stroke of the ith actuator. Thus the DM's shape is modeled as follows

$$\psi_m(x,t) = \sum_{i=1}^{n_u} a_i(t) f_i(x), \tag{3.2}$$

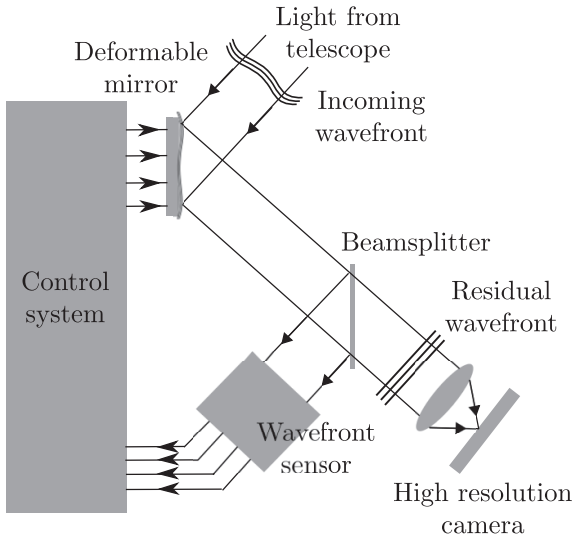


Fig. 23. Adaptive optics system.

where $f_i(x)$ is called the influence function of the ith actuator. We suppose that the DM's actuators and the associated power amplifiers have sufficient fast dynamics such that we assume that

$$a_i(t) = u_i(t). \tag{3.3}$$

We denote $u_i(t)$ the control input which is the power amplifier input of the ith actuator.

Different types of sensors (curvature sensor, pyramid wavefront sensor) may be used to estimate the distortions affecting the outcoming wavefront but the most frequently encountered in existing applications is the Shack-Hartmann (SH) wavefront sensor. The principle of a SH wavefront sensor is shown in Figure 24. The outcoming wavefront is imaged using a lenslet array of size n_w. Each lens takes a small part of the aperture, called sub-pupil, and forms an image of the source recorded by the detector, typically a CCD. If no wavefront aberrations are present, the image pattern is a grid of spots with constant intervals. As soon as the wavefront is distorted, the images are displaced from their nominal positions. Displacements of image centroids in two orthogonal directions u, v are proportional to the average wavefront slopes in u, v over the subapertures. The shift is computed using classic methods (center of gravity algorithms, ...). Thus, a SH sensor measures the wavefront average slopes $(\alpha_{u,i}, \alpha_{v,i})$ for each subaperture i.

A usual representation of wavefront is made through the orthogonal basis, typically Karhunen-Loève functions or Zernike polynomials as defined in (Noll 1976). An infinite number of functions is required to characterize the wavefront, but a truncated basis $\{F_i(x)\}$ of dimension n_b, that we called the *modal basis* is used for

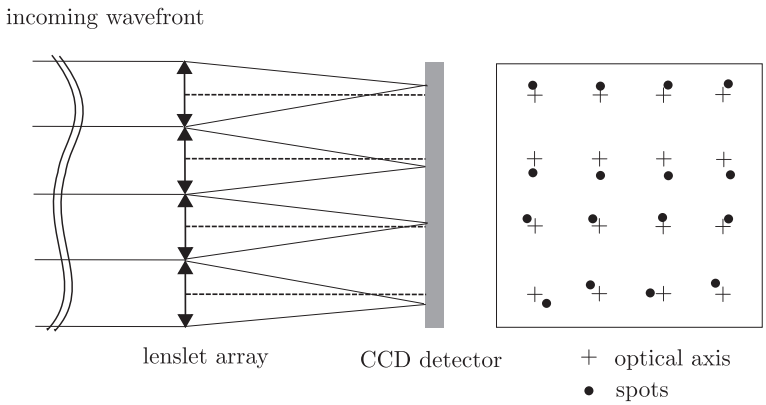


Fig. 24. Shack-Hartmann wavefront sensor principle.

implementation purpose. Thus the atmospheric wavefront ψ_a can be decomposed on the modal basis as follows:

$$\psi_a(x,t) \approx \sum_{i=1}^{n_b} w_{a,i}(t) F_i(x), \tag{3.4}$$

where we denote $w_{a,i}$ the *modal coordinates* which are the coefficients of this decomposition. We collect the scalar coefficient signals $w_{a,1}, ..., w_{a,n_b}$ to form the vector

$$w_a(t) = \begin{bmatrix} w_{a,1}(t) \\ \vdots \\ w_{a,n_b}(t) \end{bmatrix}.$$

The same representation (3.4) is used for the mirror correction ψ_m, and the residual wavefront ψ_r; similarly the coefficient signals are collected to form vector signals w_m and w_r. Control inputs $u_1, ..., u_{n_u}$ and average WFS slopes $\alpha_{u,1}, \alpha_{v,1}, ..., \alpha_{u,n_w}, \alpha_{v,n_w}$ are collected to form the control input vector u and the slope vector s. That is,

$$u(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_{n_u}(t) \end{bmatrix}, \quad s(t) = \begin{bmatrix} \alpha_{u,1} \\ \alpha_{v,1} \\ \vdots \\ \alpha_{u,n_w} \\ \alpha_{v,n_w} \end{bmatrix}.$$

Equations (3.1), (3.2) are translated into modal coordinates using vector notation as

$$w_r(t) = w_a(t) - w_m(t), \tag{3.5}$$

and

$$w_m(t) = M_m u(t), \tag{3.6}$$

where M_m is called the mirror influence matrix. The slope signal s is expressed as

$$s(t) = M_w w_r(t), \tag{3.7}$$

where we denote M_w the WFS matrix. As mentioned by Looze (2005), the output of the CCD detector, intrinsically a discrete-time signal, integrates over the sampling period T the delayed slope

$$\tilde{s}(t) = s(t - \tau). \tag{3.8}$$

We call τ the continuous time measurement delay which is the sum of the CCD's readout time and of the slopes' computation time. Thus, the output of the CCD based sensor is

$$y(t) = \frac{1}{T} \int_{t-T}^{t} \tilde{s}(\sigma) d\sigma + n(t), \tag{3.9}$$

where $n(t)$ is an additive noise caused by the photon fluctuations and by the detector's readout noise.

3.2 Disturbance rejection MIMO feedback loop

If we refer to the feedback block diagram depicted in Figure 13, Equations (3.5), (3.6), (3.7), (3.8), and (3.9) define the continuous time part of the sampled-data feedback system shown in Figure 13. The regulated output is $z(t) = w_r(t)$, the disturbance input is $d(t) = w_a(t)$, and the actuator output is $v(t) = w_m(t)$. The actuator/DM transfer function is simply

$$G_1(p) = M_m.$$

In Figure 13, the sensor/WFS is described by the transfer function

$$G_2(p) = \left(e^{-\tau p} I\right) \left(\frac{1 - e^{-Tp}}{Tp} I\right) M_w.$$

As proposed in the paper (Demerle *et al.* 1994), a first approach, the emulation design method presented in Section 2.4.2, approximates the AO feedback system with the continuous time feedback system shown in Figure 7. In Section 2.4.2 we have considered a *single-input single-output* (SISO) feedback system but here the feedback loop signals may have large dimensions: this is the *multiple-input and multiple-output* (MIMO) *feedback system* depicted in Figure 25.

The Equation (2.6) established for a single-input single-output (SISO) system becomes

$$\begin{aligned} \mathcal{L}\{w_r\}(p) &= \overbrace{(I + G_1(p)K(p)G_2(p))^{-1}}^{S(p)} \mathcal{L}\{w_a\}(p) \\ &\quad - \underbrace{(I + G_1(p)K(p)G_2(p))^{-1} G_1(p)K(p)}_{T(p)} \mathcal{L}\{n\}(p). \end{aligned} \tag{3.10}$$

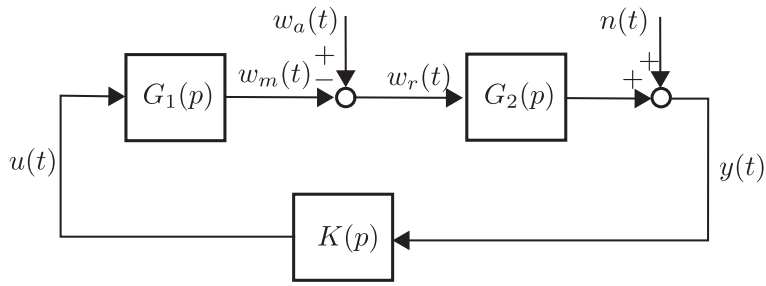


Fig. 25. An approximation of an AO MIMO feedback system.

where the following terminology remains:

(i) $L(p) = G_1(p)K(p)G_2(p)$ is the loop transfer function;

(ii) $S(p) = (I + L(p))^{-1}$ is the sensitivity transfer function;

(iii) $T(p) = S(p)G_1(p)K(p)$ is the noise sensitivity transfer function.

The disturbance rejection performance is entirely determined by transfer functions S and T. At this step no assumption is made for the type of controller (optimized modal controller, linear quadratic Gaussian control, ...) for the set of the perturbation inputs w_a and n. The performance criterion, the "size" of the residual wavefront w_r is not defined either. A possible approach sketched in Section 2.2, involves the frequency response analysis generalized for MIMO systems which provides some crucial information about the system performances (stability, disturbance rejection, command input peak value), see for instance the book 2007. Another way is to evaluate the "size" of the residual wavefront w_r in terms of the variance (mean-square error) $\mathbf{E}\left[w_r(k)^T w_r(k)\right]$ when stochastic signals w_a, n_w are considered zero mean, stationary and independent. The Maréchal approximation (Born & Wolf 1999) can be invoked to show that bounding the mean-square error of the residual wavefront ensures satisfactory imaging performance of AO systems. Thus, in the frequency domain, the variance can be written as

$$\begin{aligned}\mathbf{E}\left[\|w_r(t)\|^2\right] &= \tfrac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{Tr}\left(S(j\omega)\Phi_{w_a}(j\omega T)S(-j\omega)^T\right)d\omega \\ &+\tfrac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{Tr}\left(T(j\omega)\Phi_n(j\omega)T(-j\omega)^T\right)d\omega,\end{aligned} \tag{3.11}$$

where Φ_{w_a} and Φ_n are the power spectral densities of the input signals w_a and n_w. The first term of the right hand side of Equation (3.11) represents the contribution of the atmospheric wavefront and the latter the contribution of the WFS measurement noise. The *optimized modal gain integral control* (OMGI) proposed by Gendron & Léna (1994) and its improvements, see (Dessenne *et al.* 1998), is a well-established method to tackle this control problem.

3.3 Optimized modal gain integral control

The key idea of the approach is to reconstruct the wavefront using the WFS measurement y and to consider the linear relation (3.7). The WFS matrix M_w can be expressed into its singular value decomposition, see (Laub 2004)

$$M_w = U\Sigma V^T \tag{3.12}$$

where U, V are orthogonal matrices. We assume that $\mathbf{rank}\,(M_w) = n_b$ and

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix},\ \Sigma = \begin{bmatrix} S \\ 0 \end{bmatrix},\ \text{with } S = diag\,(\sigma_i)\,,$$

where terms σ_i are positive singular values of the matrix M_w. We define $M_w^\dagger = VS^{-1}U_1^T$ as the Moore-Penrose pseudoinverse of the matrix M_w, see (Laub 2004). We also denote $\mathbf{rank}\,(M_m) = n_b$ and we call $M_m^\dagger$ the Moore-Penrose pseudoinverse of the matrix M_m. An integral (modal) controller can be defined as

$$K(p) = M_m^\dagger \left(\frac{1}{p} K_I\right) M_w^\dagger, \tag{3.13}$$

where K_I is the matrix integrator gain to design. We consider a new atmospheric wavefront signal $\tilde{w}_a$, and a new sensor noise signal $\tilde{n}$, such that

$$w_a = V\tilde{w}_a, \quad n = U_1 S\tilde{n},$$

and a new residual wavefront signal

$$\tilde{w}_r = V^T w_r.$$

The block diagram of the feedback system is depicted in Figure 26.

Despite the complexity of the block diagram, a change of signals allows us to obtain a straightforward expression of the residual wavefront

$$\begin{aligned} \mathcal{L}\{\tilde{w}_r\} = & \underbrace{\left(I + \frac{1}{p}\tilde{K}_I e^{-\tau p}\frac{1-e^{-Tp}}{Tp}\right)^{-1}}_{\tilde{S}} \mathcal{L}\{\tilde{w}_a\} \\ & - \underbrace{\left(I + \frac{1}{p}\tilde{K}_I e^{-\tau p}\frac{1-e^{-Tp}}{Tp}\right)^{-1} \frac{1}{p}\tilde{K}_I}_{\tilde{T}} \mathcal{L}\{\tilde{n}\} \end{aligned} \tag{3.14}$$

where the matrix gain is $\tilde{K}_I = V^T K_I V$. If we fix the matrix gain such that $\tilde{K}_I = diag\left(\tilde{k}_i\right)$, then the MIMO transfer functions $\tilde{S}$ and $\tilde{T}$ are diagonal: the MIMO control problem reduces to n_b independent SISO control problems. We call $\tilde{S}_i$ ($\tilde{T}_i$) the ith diagonal entry of the sensitivity transfer function S (the noise

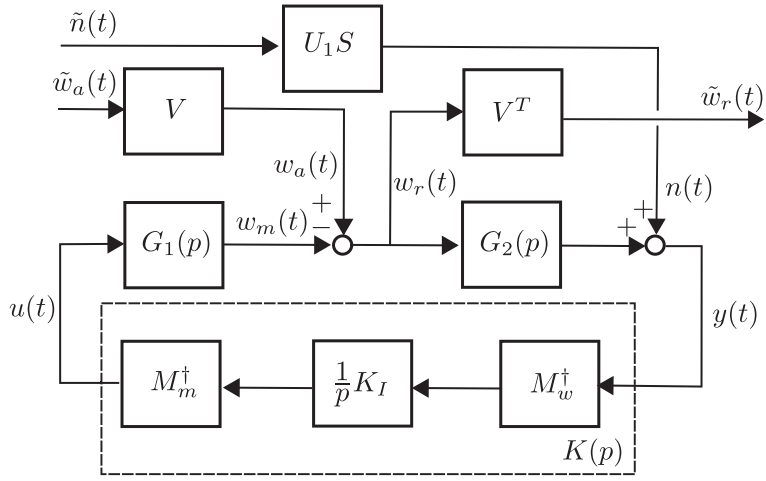


Fig. 26. Modal feedback system.

sensitivity transfer function T). Thus, the variance of each component can be written as

$$\mathbf{E}\left[\|\tilde{w}_{r,i}(t)\|^2\right] = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left|\tilde{S}_i(j\omega)\right|^2 \Phi_{\tilde{w}_{a,i}}(j\omega)d\omega + \frac{1}{2\pi}\int_{-\infty}^{\infty}\left|\tilde{T}_i(j\omega)\right|^2 \Phi_{\tilde{n}_i}(j\omega)d\omega, \tag{3.15}$$

where $\Phi_{\tilde{w}_{a,i}}(j\omega)$ and $\Phi_{\tilde{n}_i}(j\omega)$ are the power spectral densities of the ith component of vector signals $\tilde{w}_a$ and $\tilde{n}$. The integral gain $\tilde{k}_i$ is tuned using the loop shaping approach sketched in Section 2.3.2 to minimize the variance of the ith component which induces the minimization of the variance $\mathbf{E}\left[w_r(t)^T w_r(t)\right]$. Then, the controller matrix gain is computed as

$$K_I = V\tilde{K}_I V^T. \tag{3.16}$$

The main advantage of the optimized modal gain integral control, which explains its success in practice, is to express some of the controller's signals in the modal base which facilitates the physical interpretation. Furthermore it is intrinsically a frequency approach: the analysis of the AO feedback system's performance is straightforward. The well established OMGI control offers interesting abilities. Constant additive disturbances as actuator offset are intrinsically rejected. The real time computational cost is reasonable and induces limited delay. The method can be used when the knowledge of the disturbance temporal dynamics is weak. Some shortcomings have been mentioned in the literature. The integral controller can be transformed into an observer based controller structure, see (Kulcsár *et al.* 2006). The observer is not stable and the control u may blow up. On a simplified SCAO configuration some authors Conan *et al.* (2011) indicated that more advanced control approaches such as linear quadratic Gaussian control ensure better performances.

4 Modern feedback control: LQG method for adaptive optics

4.1 Towards linear quadratic Gaussian control

4.1.1 Adaptive optics feedback loop

The WFS Equations (3.7), (3.8), and (3.9) provide a linear relationship between the temporal average of the residual wavefront over the sampling period T and the discrete time measurement (2.16) corrupted by a measurement noise. Thus, we can write the discrete time residual wavefront $w_r(k)$ as the average of the continuous time residual wavefront $w_r(t)$

$$w_r(k) = \frac{1}{T}\int_{(k-1)T}^{kT} w_r(t)dt. \tag{4.1}$$

The same temporal discretization (4.1) is done for the mirror wavefront $w_m(k)$ and the atmospheric wavefront $w_a(k)$. The WFS Equations (3.7), (3.8), (3.9), and (2.16) are transformed into difference equation. We obtain in the frequency domain

$$\mathcal{Z}\{y\} = \underbrace{z^{-k_y}M_w}_{G_1}\mathcal{Z}\{w_r\} + \mathcal{Z}\{n\}, \tag{4.2}$$

where n is an additive measurement noise and where k_y is the measurement delay such that $\tau = k_y T$. Equations (3.5) and (3.6) become

$$\mathcal{Z}\{w_r\} = \mathcal{Z}\{w_a\} - \underbrace{M_w z^{-k_u}}_{G_2}\mathcal{Z}\{u\}, \tag{4.3}$$

where $k_u \geq 1$ represents the control input delay. We call $G_1(z)$ the DM transfer function and $G_2(z)$ the WFS transfer function. The block diagram of the discrete time AO feedback system is shown in Figure 21. Here the AO loop is a MIMO feedback system. The regulated output response (2.25) established for a SISO system becomes

$$\begin{aligned}\mathcal{Z}\{w_r\}(z) &= \overbrace{(I + G_1(z)K(z)G_2(z))^{-1}}^{S(z)}\mathcal{Z}\{w_a\}(z) \\ &\quad - \underbrace{(I + G_1(z)K(z)G_2(z))^{-1}G_1(z)K(z)}_{T(z)}\mathcal{Z}\{n\}(z).\end{aligned} \tag{4.4}$$

where the following terminology remains:

(i) $L(z) = G_1(z)K(z)G_2(z)$ is the loop transfer function;

(ii) $S(z) = (I + L(z))^{-1}$ is the sensitivity transfer function;

(iii) $T(z) = S(z)G_1(z)K(z)$ is the noise sensitivity transfer function.

Up to now the framework is identical to the approach presented for the continuous time feedback loop in the frequency domain. However we have to keep in mind that here we adopt the point of view of the digital computer and that the regulated output $w_r(k)$ is the temporal average of the "real" regulated output $w_r(t)$. This approach is relevant when the choice of the sampling period T is not critical in regards with the dynamics of the atmospheric wavefront. We assume that signals w_a, n_w are zero mean, stationary and independent stochastic signals. Thus, in the frequency domain, the variance $\mathbf{E}\left[\|w_r(k)\|^2\right]$ can be written as

$$\begin{aligned}\mathbf{E}\left[\|w_r(k)\|^2\right] &= \frac{T}{2\pi}\int_0^{\frac{2\pi}{T}} \mathbf{Tr}\left(S(e^{j\omega T})\Phi_{w_a}(\omega)S(e^{-j\omega T})^T\right)d\omega \\ &\quad + \frac{T}{2\pi}\int_0^{\frac{2\pi}{T}} \mathbf{Tr}\left(T(e^{j\omega T})\Phi_n(\omega)T(e^{-j\omega T})^T\right)d\omega\,,\end{aligned} \tag{4.5}$$

where Φ_{w_a} and Φ_n are the power spectral densities of the input signals w_a and n. The first term of the right hand side of Equation (4.5) represents the contribution of the atmospheric wavefront and the latter the contribution of the WFS measurement noise. Equation (4.5) indicates the frequency range where the frequency responses $S(e^{j\omega T})$ and $T(e^{j\omega T})$ have to be small. Power spectral densities Φ_{w_a} and Φ_n can be seen as weighting functions for performance objective (4.5). The control problem can be formulated as finding the control law that minimizes the variance $\mathbf{E}\left[\|w_r(k)\|^2\right]$ To take into account more accurately the information of the atmospheric wavefront we have to build a model of the temporal evolution of $w_a(k)$.

4.1.2 Identified atmospheric wavefront model

The power spectral densities Φ_{w_a} may be factored as

$$\Phi_{w_a}(w) = G_a(e^{j\omega T})G_a(e^{-j\omega T})^T,$$

and the atmospheric wavefront w_a is assumed to be the output of a causal and stable diagonal transfer function matrice G_a driven by a white noise n_a having a unitary covariance matrix. To take into account the oscillating behavior of $w_a(k)$ a second order diagonal AR model is considered

$$A_0 w_a(k) + A_1 w_a(k-1) + A_2 w_a(k-2) = n_a(k), \tag{4.6}$$

where diagonal matrices (A_0, A_1, A_2) are the AR parameters. The computation of the parameters is carried out with the Burg algorithm, see (Burg 1975), which minimizes the sum of the squares of the forward and backward prediction errors. In the frequency domain we obtain

$$\mathcal{Z}\{w_a\} = \underbrace{\left(A_0 z^2 + A_1 z + A_2\right)^{-1} z^2}_{G_a}\,\mathcal{Z}\{n_a\}\,.$$

which defines the atmospheric wavefront filter G_a. The AO block diagram is depicted in Figure 27 where the different loop signals are mentioned.

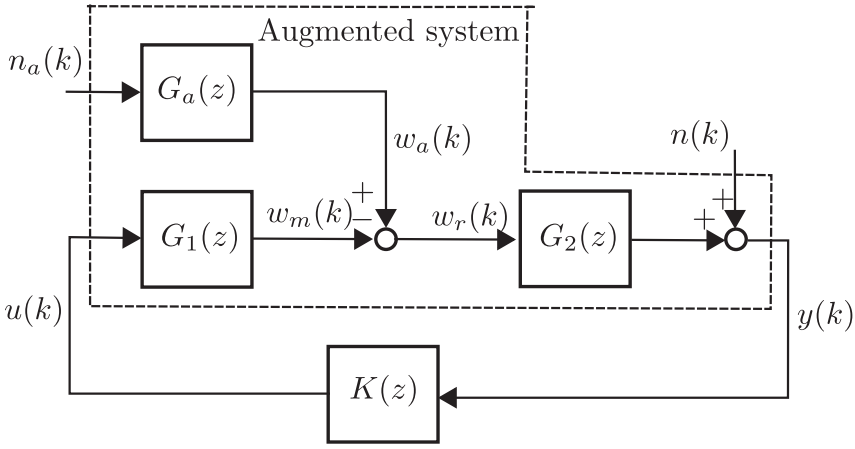


Fig. 27. AO discrete-time system block-diagram including the atmospheric model.

4.1.3 Performance objective in the time domain

In the time domain, the AO control problem can be formulated as finding the control law that minimizes the empirical variance of the residual wavefront, averaged over a large exposure time T_e

$$\mathbf{E}\left[\|w_r(t)\|^2\right] = \lim_{T_e\to\infty} \frac{1}{T_e}\int_0^{T_e} \|w_r(t)\|^2\, dt, \tag{4.7}$$

which is the time domain counterpart of (3.11) for a stationary ergodic process and the "true" imaging performance index. Several authors Kulcsár *et al.* (2006), Looze (2007) demonstrated that the minimization of the residual wavefront variance $\mathbf{E}\left[\|w_r(t)\|^2\right]$ can be performed using the discrete-time model of the hybrid AO system without loss of optimality. Therefore, the performance objective to minimize, in the discrete-time domain is translated as

$$\mathbf{E}\left[\|w_r(k)\|^2\right] = \lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} \|w_r(k)\|^2, \tag{4.8}$$

which is the time domain counterpart of (4.5). This last control objective can be minimized using LQG design approach using a state-space description of the augmented plant (DM, WFS, atmospheric wavefront model) as discussed in Section 4.3.

4.2 LQG control framework

4.2.1 State space equation

The state space method is based on the description of system equation in terms of n first-order difference equations, which may be combined into a first-order

vector-matrix difference equation. The state space equation of a discrete time system can be written

$$\begin{aligned} x(k+1) &= Ax(k) + Be(k) \\ s(k) &= Cx(k) + Du(k). \end{aligned} \tag{4.9}$$

Here $x \in \mathbf{R}^n$ is the *state* of the system, $e \in \mathbf{R}^m$ is the *input*, and $s \in \mathbf{R}^r$ is the *output*. For example consider the AR difference Equation (4.6) when the signals w_a and n_a are scalars

$$a_0 w_a(k+1) + a_1 w_a(k) + a_2 w_a(k-1) = n_a(k),$$

where real scalars a_0, a_1, a_2 are given. To convert this equation into the state space Equation (4.9), we define $x_1(k) = w_a(k)$, $x_2(k) = w_a(k-1)$, $e(k) = n_a(k)$, and $s(k) = w_a(k)$. The first-order difference equations are then

$$\begin{aligned} x_1(k+1) &= w_a(k+1) = -\frac{a_1}{a_0} w_a(k) - \frac{a_2}{a_0} w_a(k-1) + \frac{1}{a_0} n_a(k) \\ &= -\frac{a_1}{a_0} x_1(k) - \frac{a_2}{a_0} x_2(k) + \frac{1}{a_0} e(k) \\ x_2(k+1) &= w_a(k) = x_1(k) \\ s(k) &= w_a(k) = x_1(k). \end{aligned}$$

We can write this in matrix/vector form as

$$\begin{aligned} \begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} &= \begin{bmatrix} -\frac{a_1}{a_0} & -\frac{a_2}{a_0} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \begin{bmatrix} \frac{1}{a_0} \\ 0 \end{bmatrix} e(k) \\ s(k) &= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + 0\, e(k). \end{aligned}$$

If we pose

$$A = \begin{bmatrix} -\frac{a_1}{a_0} & -\frac{a_2}{a_0} \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} \frac{1}{a_0} \\ 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad D = 0,$$

we obtain the state space Equation (4.9). In the general case, the atmospheric wavefront model G_a can be written in state space form as

$$\begin{aligned} x_a(k+1) &= A_a x_a(k) + B_a n_a(k), \\ y_a(k) &= C_a x_a(k), \end{aligned} \tag{4.10}$$

where the state $x_a \in \mathbf{R}^{2n_b}$ is $x_a(k) = \begin{bmatrix} w_a(k)^T & w_a(k-1)^T \end{bmatrix}^T$ and where state space matrices are

$$A_a = \begin{bmatrix} -A_0^{-1} A_1 & -A_0^{-1} A_2 \\ I & 0 \end{bmatrix}, \quad B_a = \begin{bmatrix} A_0^{-1} \\ 0 \end{bmatrix}, \quad C_a = \begin{bmatrix} I & 0 \end{bmatrix}. \tag{4.11}$$

4.2.2 Linear quadratic Gaussian control

The discrete-time LQG control theory considers that the system is linear and that the disturbance (plant noise) and the measurement noise inputs are stochastic. Thus, the system is described by the state-space representation

$$\begin{array}{rcl} x(k+1) & = & Ax(k) + Bu(k) + w(k) \\ y(k) & = & Cx(k) + v(k), \end{array} \tag{4.12}$$

where $x \in \mathbf{R}^n$ is the state vector, $u \in \mathbf{R}^{n_u}$ the command input, $y \in \mathbf{R}^{n_y}$ the measured output, and where $w \in \mathbf{R}^n$ represents the disturbance input and $v \in \mathbf{R}^{n_y}$ is the measurement noise input. We assume that Gaussian noise processes $w(k)$ and $v(k)$ are mutually independent, zero mean white noises with covariance $\mathbf{E}\left[w(k)w^T(l)\right] = W\delta(k-l)$ and $\mathbf{E}\left[v(k)v^T(l)\right] = V\delta(k-l)$, respectively. It is supposed that the pair (A, B), $(A, W^{1/2})$ are stabilizable and the pair (A, C) is detectable.

The LQG control problem is to find the optimal control $u(k)$ for system (4.12) that minimizes the infinite horizon quadratic cost criterion

$$J = \lim_{N\to\infty} \frac{1}{N}\, \mathbf{E}\left[\sum_{k=0}^{N-1} x(k)^T Q x(k) + u(k)^T R u(k)\right], \tag{4.13}$$

with given weighting matrices $Q = Q^T \geq 0$, $R = R^T > 0$ and the pair $(A, Q^{1/2})$ detectable.

The solution of the LQG control problem is then provided by the interconnection of a linear quadratic regulator and a state estimator. This result is known in linear optimal control theory as the *Separation Principle*, see (Kwakernaak & Sivan 1972; Anderson & Moore 1990). The optimal control sequence minimizing the cost function (4.13) is given by the *state-feedback* control law

$$u(k) = -K\hat{x}(k), \tag{4.14}$$

where $\hat{x}$ is the optimal estimate of the state x. The *state-feedback gain* K is a constant matrix

$$K = \left(R + B^T P B\right)^{-1} B^T P A, \tag{4.15}$$

where the matrix $P = P^T$ is the unique positive-semidefinite solution of the control discrete-time algebraic Riccati equation (DARE)

$$P = A^T P A - A^T P B \left(B^T P B + R\right)^{-1} B^T P A + Q. \tag{4.16}$$

Note that the conditions $R > 0$, (A, B) detectable and $(A, Q^{1/2})$ detectable can be relaxed, see (Bitmead & Gevers 1991; Dorato & Levis 1971). The optimal state estimation which minimizes the variance of the estimation error $\mathbf{E}\left[\left|\hat{x}(k) - x(k)\right\|^2\right]$, is performed through a standard Kalman predictor filter with

$$\hat{x}(k+1) = A\hat{x}(k) + Bu(k) + L(y(k) - C\hat{x}(k)), \tag{4.17}$$

where L is the *observer gain*

$$L = AXC^T(CXC^T + B)^{-1}. \tag{4.18}$$

where the matrix $X = X^T$ is the unique positive-semidefinite solution of the estimation DARE

$$X = AXA^T - AXC^T(CXC^T + V)^{-1}CXA^T + W. \tag{4.19}$$

4.3 Application of LQG control to the adaptive optics system

4.3.1 AO state space system

In the sequel we consider a unitary input delay $k_u = 1$ and a unitary output delay $k_y = 1$. The "augmented system", depicted in Figure 27, is described by the state space Equation (4.9) where the signals are defined as follows.

1. The state vector x is split in two parts $x = \begin{bmatrix} x_m^T & x_a^T \end{bmatrix}^T$. The state $x_m(k) = \begin{bmatrix} w_m(k)^T & w_m(k-1)^T \end{bmatrix}^T$ represents the plant dynamics (DM & WFS) and state $x_a(k)$ corresponds to the perturbation dynamics (4.10).
2. The state noise is $w = \begin{bmatrix} 0 \\ B_a \end{bmatrix} n_a$ and the measurement noise is $v = n$.

The state space matrices of the augmented system (DM, WFS, ATM) are defined as

$$A = \begin{bmatrix} A_m & 0 \\ 0 & A_a \end{bmatrix}, \quad B = \begin{bmatrix} B_m \\ 0 \end{bmatrix}, \quad C = M_w \begin{bmatrix} C_m & C_a \end{bmatrix}. \tag{4.20}$$

The state space matrices of the plant are

$$A_m = \begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix}, \quad B_m = \begin{bmatrix} M_m \\ 0 \end{bmatrix}, \quad C_m = \begin{bmatrix} 0 & -I \end{bmatrix}, \tag{4.21}$$

and state-space matrices (A_a, B_a, C_a) are given in (4.11).

The special form of state space matrices (4.20) can be exploited to simplify the resolution of the Riccati equations, see (Bitmead *et al.* 1990). For the presentation of the following results, matrices P, X and Q are partitioned conformally with the matrix A, that is

$$P = \begin{bmatrix} P_m & P_0 \\ P_0^T & P_a \end{bmatrix}, \quad X = \begin{bmatrix} X_m & X_0 \\ X_0^T & X_a \end{bmatrix}, \quad Q = \begin{bmatrix} Q_m & Q_0 \\ Q_0^T & Q_a \end{bmatrix}.$$

4.3.2 Solving the control DARE

The control DARE (4.16) can be simplified to obtain solutions for the individual blocks of P. We have to find the matrix $P_m = P_m^T$ the unique positive-semidefinite solution of the reduced order DARE

$$P_m = A_m^T P_m A_m - A_m^T P_m B_m \left(B_m^T P_m B_m + R\right)^{-1} B_m^T P_m A_m + Q_m. \tag{4.22}$$

The state-feedback gain (4.15) becomes $K = \begin{bmatrix} K_m & K_a \end{bmatrix}$ with

$$K_m = (B_m^T P_m B_m + R)^{-1} B_m^T P_m A_m. \tag{4.23}$$

We search matrix P_0 which is a solution of the following discrete-time Sylvester equation

$$P_0 = (A_m - B_m K_m)^T P_0 A_a + Q_0 \,. \tag{4.24}$$

We obtain

$$K_a = (B_m^T P_m B_m + R)^{-1} B_m^T P_0 A_a. \tag{4.25}$$

The special form of state space matrices (4.21) imply that $K_m = 0$ and that

$$K_a = -\left(R + M_m^T M_m\right)^{-1} M_m^T C_a A_a^2 \,. \tag{4.26}$$

4.3.3 Solving the estimation DARE

The estimation error can be written as $\tilde{x}^T = \begin{bmatrix} \tilde{x}_m^T & \tilde{x}_a^T \end{bmatrix} = \begin{bmatrix} \hat{x}_m^T - x_m^T & \hat{x}_a^T - x_a^T \end{bmatrix}$. The state x_m is a deterministic signal and thus $\tilde{x}_m = 0$ which simplifies the blocks $X_m = 0$, $X_0 = 0$. The estimation DARE (4.19) can be simplified to obtain solutions for the individual blocks of X. Thus the matrix $X_a = X_a^T$ is the unique positive-semidefinite solution of the reduced order DARE

$$X_a = A_a X_a A_a^T - A_a X_a C_a^T \left(C_a X_a C_a^T + V\right)^{-1} C_a X_a A_a^T + B_a B_a^T. \tag{4.27}$$

The observer gain (4.18) becomes $L = \begin{bmatrix} 0 \\ L_a \end{bmatrix}$ with

$$L_a = A_a X_a C_a^T \left(C_a X_a C_a^T + V\right)^{-1}. \tag{4.28}$$

4.3.4 LQG controller

The strictly proper, linear time invariant controller, is described by the state-space equation

$$\begin{aligned} \hat{x}(k+1) &= \hat{A}\hat{x}(k) + \hat{B}y(k) \\ u(k) &= \hat{C}\hat{x}(k) \end{aligned} \tag{4.29}$$

where the matrices $\hat{A}$, $\hat{B}$, $\hat{C}$ are

$$\hat{A} = \begin{bmatrix} A_m & -B_m K_a \\ -L_a M_w C_m & A_a - L_a M_w C_a \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} 0 \\ L_a \end{bmatrix}, \quad \hat{C} = -\begin{bmatrix} 0 & K_a \end{bmatrix}.$$

Note that the LQG controller is equivalently described by the discrete time transfer function

$$K(z) = \hat{C}\left(zI - \hat{A}\right)^{-1} \hat{B},$$

which is a convenient form (i) to analyze the AO feedback system depicted in Figure 27; (ii) to interpret the AO performance index (4.5).

4.4 LQG controller design

We consider an 8-m telescope without obstruction and the 512×512-pixels wavefronts projected over 44 Zernike ($n_b = 44$). The physical modeling has been performed by means of the `Software Package CAOS` (Carbillet *et al.* 2005), developed within the `CAOS` problem-solving environment (PSE), see (Carbillet *et al.* 2010). The computation of the LQG state-space matrices (4.20) is carried out using `Matlab software` and the `Control system toolbox` and involves the following steps.

Step 1: AO discrete-time state-space computation. DM controls perfectly low spatial frequencies with $n_u = 44$ actuators and consequently $M_m = I_{n_b}$. The WFS device is a 8×8 ($\Rightarrow n_y = 52$) subaperture Shack-Hartmann WFS (8×8 $0.2''$ px/subap., $\lambda = 700\,\text{nm} \pm 150\,\text{nm}$, $\Delta t = T = 1\,\text{ms}$.). The WFS influence matrix M_w is determined from the WFS calibration simulation.

Using `Software Package CAOS` $500 \times 1\,\text{ms}$ wavefronts propagated through an evolving 2-layers turbulent atmosphere ($r_0 = 10$ cm at $\lambda = 500$ nm, $\mathcal{L}_0 = 25$ m, wind velocities = 8 & 12 m/s) are obtained. After the projection on the Zernike base, the signal w_a is modeled as the output of an AR system using the approach presented in Section 4.1.2. The computation of the parameters is carried out with the Burg algorithm, see (Burg 1975), using the `Signal Processing Toolbox` of `Matab` and permits to obtain state space matrices (4.11). Then, the computation of the LQG state space matrices is obvious using Equation (4.20).

Step 2: Additive noise covariance estimation. Covariance matrix V for LQG design is a tuning parameter which dictates the performance of the AO control loop. We use the empirical covariance matrix obtained from a photon noise calibration from our CAOS simulations. Note that this needs anyway to be refined for future developments.

Step 3: controller design. To minimize the performance objective $\mathbf{E}\left[\|w_r(k)\|^2\right]$ given in the discrete-time domain (4.8) we consider the LQG performance index J defined by (4.13) with the weigthing parameter $R = 0$ (cheap control case). We have designed two kinds of optimal controller. LQG1 has been designed with the noise covariance matrix V equal to zero, while for LQG2 we use the empirical covariance matrix built in step 2.

4.5 Discussion

4.5.1 A posteriori frequency analysis

In the Figure 28–30 show the singular values of $S(e^{j\omega T})$ in the left part, and the singular values of $T(e^{j\omega T})$ in the right part. The maximum singular values are plotted in plain line, while the minimum singular values are plotted in dashed

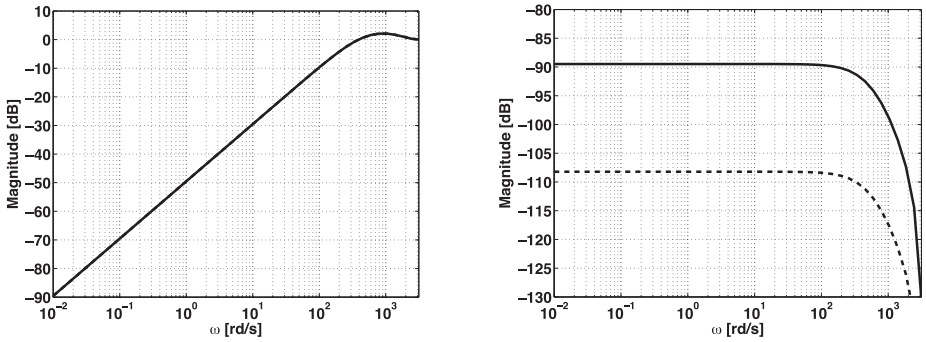


Fig. 28. Plot of singular values of $S(e^{j\omega T})$ in the left part, and the singular values of $T(e^{j\omega T})$ in the right part for integrator case (with a gain of 0.3).

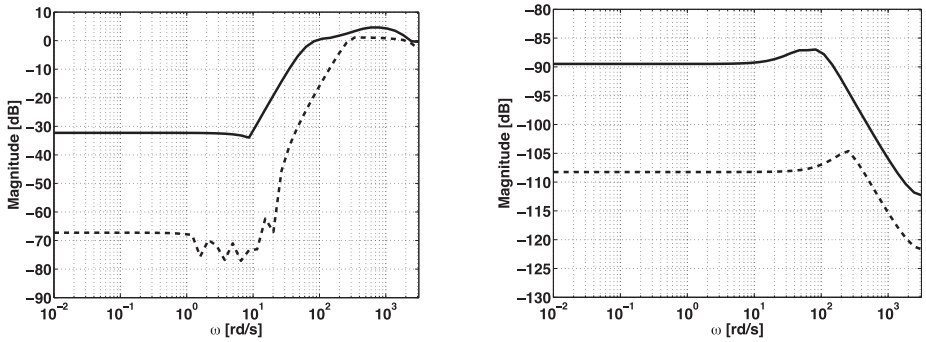


Fig. 29. Plot of singular values of $S(e^{j\omega T})$ in the left part, and the singular values of $T(e^{j\omega T})$ in the right part for LQG1 controller case.

line. The integrator case (with a gain of 0.3) is plotted in Figure 28, the LQG1 controller case in Figure 29, and the LQG2 controller case in Figure 30.

Note that the sensitivity transfer function S for the LQG1 controller case shows that the LQG1 controller ensures a better rejection of the atmospheric wavefront than the LQG2 controller. If we check the frequency response of the noise rejection transfer function T, LQG1 design is more sensitive to noise than LQG2 design. The integrator case exhibits the worst frequency performance. These indications have to be confirmed by using `CAOS` end to end simulation.

4.5.2 Performance comparison

The time simulation has been performed by means of the `Software Package CAOS`. An *ad hoc* module, `SSC`, which stands for "Space-State Control", has been

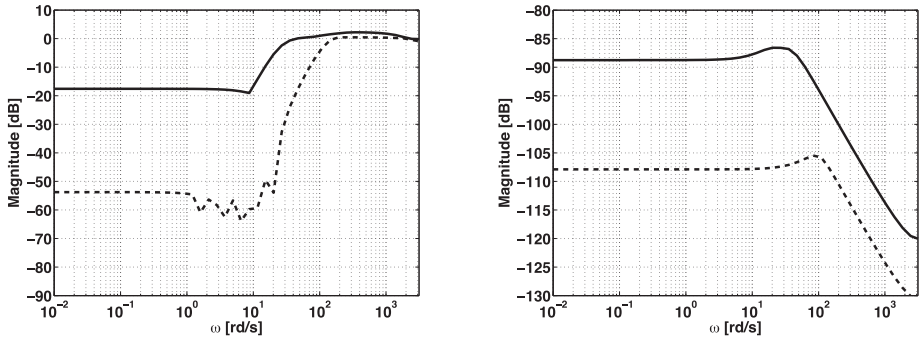


Fig. 30. Plot of singular values of $S(e^{j\omega T})$ in the left part, and the singular values of $T(e^{j\omega T})$ in the right part for LQG2 controller case.

developed especially for this study, also with the goal of making it publicly available with a future upgrade of the `Software Package CAOS`. Figure 31 shows the numerical modeling designed within the `CAOS` PSE.

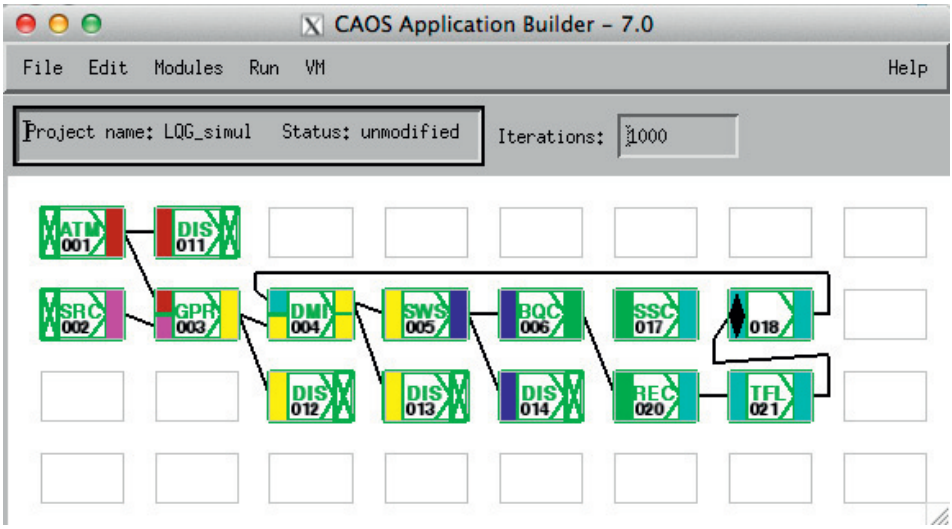


Fig. 31. `CAOS` numerical modeling of the AO system.

Figure 32 represents an example of running simulation. *Left*: the atmospherically-perturbed input wavefront. *Middle*: the corresponding Shack-Hartmann spots. *Right*: the resulting corrected wavefront.

For different operating conditions (star magnitude) we have obtained the following results sum up in Table 1. In bright conditions LGG and integral controllers

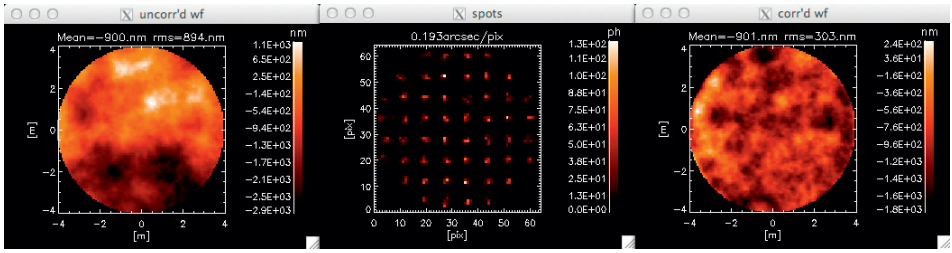


Fig. 32. CAOS running simulation.

Table 1. Obtained residual wavefront rms.

	Photons/subaperture/T	Integrator	LQG1	LQG2
no noise	∞	∼268 nm	**∼267 nm**	∼271 nm
mag 12	∼320	∼269 nm	**∼268 nm**	∼271 nm
mag 14	∼51	∼272 nm	**∼271 nm**	∼273 nm
mag 16	∼8.0	∼296 nm	∼297 nm	**∼284 nm**
mag 17	∼3.2	∼350 nm	∼356 nm	**∼313 nm**
mag 18	∼1.3	∼471 nm	∼475 nm	**∼438 nm**

are equivalent until magnitude 14. In faint conditions (magnitude 16 to magnitude 18) the LQG2 controller induces better performance than the integral controller.

The authors are greatly indebted to the referee Céline Theys, for her helpful and constructive comments and Anthony Schutz for the computer assistance. The first author would like to thank Calypso Barnes for her valuable contribution to improve the quality of the english text.

References

Anderson, B., & Moore, J.B., 1990, Optimal Control: Linear Quadratic Methods (Prentice-Hall)

Astrom, K.A., & Wittenmark, B., 2011, Computer-Controlled Systems: Theory and Design (Dover Publications)

Bitmead, R.R., Gevers, M., & Wertz, V., 1990, Adaptive Optimal Control: the Thinking Man's GPC (Prentice Hall Englewood Cliffs, NJ)

Bitmead, R.R., & Gevers, M., 1991, Riccati Difference and Differential Equations: Convergence, monotonicity and stability, In The Riccati equation, ed. S. Bittanti, A.J. Laub & J.C. Willems (Springer Verlag)

Boyd, S., 1993, Lecture Notes for E105, Introduction to Automatic Control (Stanford University)

Burg, J.P., 1975, Ph.D. Thesis, Maximum Entropy Spectral Analysis (Stanford University)

Born, M., & Wolf, E., 1999, Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of light (Cambridge University Press)

Carbillet, M., Vérinaud, C., Femenía, B., Riccardi, A., & Fini, L., 2005, MNRAS, 356, 1263

Carbillet, M., Desiderà, G., Augier, A., *et al.*, 2010, Proc. SPIE, 7736, 773644

Conan, J.M., Raynaud, H.F., Kulcsár, C., Meimon, S., & Sivo, G., 2011, Are Integral Controllers Adapted to the New Era of ELT Adaptive Optics? In AO4ELT2, Victoria, Canada, September

Dorato, P., & Levis, A., 1971, Optimal Linear Regulators: the Discrete-time Case, IEEE Transactions on Automatic Control, 613

Dorf, R.C., & Bishop, R.H., 1998, Modern Control Systems, Eight edition (Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA)

Demerle, M., Madec, P.Y., & Rousset, G., 1994, Servo-Loop Analysis for Adaptive Optics, In NATO Meeting, Cargèse, France, June 29-July 9, 1993, ONERA, TP, Vol. 423, 73

Dessenne, C., Madec, P.Y., & Rousset, G., 1998, Appl. Opt., 37, 4623

Franklin, G.F., Powell, J.D., & Emami-Naeni, A., 1991, Feedback Control of Dynamic Systems, Second edition (Addison-Wesley)

Franklin, G.F., Powell, J.D., & Workman, M.L., 1990, Digital Control of Dynamic Systems, Second edition (Addison Wesley)

Gendron, E., & Léna, P., 1994, A&A (ISSN 0004-6361), 291

Kulcsár, C., Raynaud, H.F., Petit, C., Conan, J.M., & de Lesegno, P.V., 2006, Appl. Opt, 39, 2525

Kwakernaak, H., & Sivan, R., 1972, Linear Optimal Control Systems (John Wiley & Sons)

Laub, A.J., 2004, Matrix Analysis for Scientists and Engineers, Society for Industrial Mathematics

Looze, D.P., 2005, Realization of Systems with CCD-based Measurements, Automatica, 41

Looze, D.P., 2006, J. Opt. Soc. Am., 23, 603

Looze, D.P., 2007, J. Opt. Soc. Am., 9, 2850

Noll, R.J., 1976, J. Opt. Soc. Am., 66, 207

Paschall, R.N., & Anderson, D.J., 1993, Linear Quadratic Gaussian Control of a Deformable Mirror Adaptive Optics System with Time-delayed Measurements, Appl. Opt., 32

Paschall, R.N., Von Bokern, M.A., & Welsh, B.M., 1991, Design of a Linear Quadratic Gaussian Controller for an Adaptive Optics System, In Proceedings of the 30th IEEE Conference on Decision and Control, 1761

Roddier, F., 1999, Adaptive Optics in Astronomy (Cambridge University Press)

Skogestad, S., & Postlethwaite, I., 2007, Multivariable Feedback Control: Analysis and Design

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 131–140*

SCIROCCO+: SIMULATION CODE OF INTERFEROMETRIC-OBSERVATIONS FOR ROTATORS AND CIRCUMSTELLAR OBJECTS INCLUDING NON-RADIAL PULSATIONS

M. Hadjara[1,2], F. Vakili[2], A. Domiciano de Souza[2], F. Millour[2], R. Petrov[2], S. Jankov[3] and P. Bendjoya[2]

Abstract. The VLTI (Very Large Telescope Interferometer) makes available milli-arcsecond-scale observations in the infrared. It offers new possibilities for constraining stellar structures such as polar jets, equatorial disks and rotationally-flattened photospheres of Be stars. Such constraints allows us to better estimate the stellar fundamental parameters and refine the mechanisms such as mass loss, pulsation and magnetism that govern the variability and evolution of these stars.

In this paper we present a chromatic semi-analytical model of fast rotators, which allows us to study the dynamics and the interaction between the photosphere and the wind of fast rotating stars of O, B, A and F spectral types. Our simple analytical model addresses the oblateness, inclination and position angle of the rotation axis of the star. It produces iso-velocity maps and intensity maps. It includes line profiles, limb-darkening and the von Zeipel effect and the non-radial pulsations.

SCIROCCO+: **S**imulation **C**ode of **I**nterferometric-observations for **RO**tators and **C**ir**C**umstellar **O**bjects including Non-Radial Pulsations, includes all the parameters cited above in order to be fast, powerful and light simulation tool in high angular resolution of rotating objects.

[1] Centre de Recherche en Astronomie, Astrophysique et Géophysique (CRAAG), Route de l'Observatoire, BP. 63, Bouzareah, 16340 Alger, Algérie; e-mail: m.hadjara@craag.dz

[2] Laboratoire J.-L. Lagrange, UMR 7293, Observatoire de la Côte d'Azur (OCA), Université de Nice-Sophia Antipolis (UNS), Centre National de la Recherche Scientifique (CNRS), Campus Valrose, 06108 Nice Cedex 2, France; e-mail: Massinissa.Hadjara@oca.eu

[3] Astronomical Observatory of Belgrade, Volgina 7, PO Box 74, 11060 Belgrade, Serbia

DOI: 10.1051/eas/1359007

1 Introduction

The Be stars with low metallicity are supposed to produce little or no magnetic field. This absence of magnetic field leads to a high spin-up during the contraction then formation of these stars (Martayan *et al.* 2006). This rotation rate can attain more than 80% of the critical, or breakup, velocity $v_c = \sqrt{(GM/R_c)}$ (with R_c the equatorial radius at this velocity) in some cases. These fast-rotating stars are called "fast rotators" and exhibit a number of peculiar characteristics (Domiciano de Souza *et al.* 2003), among which geometrical flattening, coupled with gravitational darkening von Zeipel (1924), making the poles hotter than the equator.

The models from Collins & Sonneborn (1977) indicate a two-components spectral energy distribution (SED) for these stars, with an infrared excess due to gravity darkening. Hence, it is not easy to place these stars in one single spectral classification, as the observed SED depends on its rotational velocity and inclination angle (Maeder & Peytremann 1972).

Furthermore, rapid rotation induce an additional change in the apparent spectral type and class of the star (Collins & Harrington 1966). Indeed, the full widths at half-maximum (FWHM) of UV lines are generally narrower (up 0.2 km/s) than those of the visible lines due again to gravitational darkening (Hutchings *et al.* 1979), since the spectral lines, depending on the temperature and gravity, are not formed uniformly on the star. This has an impact on the estimate of the inclination angle (Hutchings & Stoeckley 1977) and, hence, the estimation of the spectral type of the star. A classification based on the spectral ratio between the widths of these lines would be distorted by this effect (Collins 1974).

In addition, mechanisms such as meridional circulation and/or turbulence may affect the internal structure of the star and its evolution (Meynet 2009). Thus, fast rotators have always been considered as a physics laboratory to study stellar interiors, stellar evolution and primordial stars.

Moreover, Non-Radial Pulsations (NRP) can be a crucial explanation of transient mass ejections in Be stars. The classical observational techniques, as photometry and spectroscopy, suffer from the observational selection of NRP modes that is generally impossible to distinguish from physical selection. For example, the observational selection is different for pair and impair modes (integration of symmetric and asymmetric brightness or velocity distributions) and could explain the fact that only pair (or impair) modes are observed in some Be stars, a phenomenon that can also be due to a physical effect as argued by, for example, Jankov *et al.* (2000). The mechanisms governing the time variations of the mass ejection of Be stars remain largely debated. One possible explanation is the transient combination of several modes of non-radial pulsation (NRP) (Rivinius *et al.* 1999). It depends on the excited modes, which in turn critically depend on the fundamental stellar parameters (Levenhagen *et al.* 2003). The stellar diameter, flattening, rotation velocity, differential rotation and gravity govern the dominant excitation mechanisms. Limb and gravitational darkening have a strong impact on the interpretation of time evolution of spectrophotometric data. However,

differential interferometry yields differential phase information on non resolved objects which allows to measure the diameter, flattening, rotation velocity and differential rotation, and allows much better identification of NRP modes that the spectroscopy and/or photometry alone (*e.g.* Jankov *et al.* 2005). Physical selection mechanisms would select equatorial modes if a high latitudinal differential rotation has a destabilizing effect (Stockley & Buscombe 1987), but the detection of such modes is also favored by observation biases depending from the observation angle. Even if Be stars can be only marginally resolved with the largest VLTI baselines, and structures in the disk are completely unresolved, differential interferometry can extract the displacement with wavelength (λ) of the photocenter $\epsilon(\lambda)$ of an unresolved source from the small variations of the interferometric differential phase through a spectral line (Petrov 1988 and Vakili & Percheron 1991). Recently, we have used this to measure the diameter and the rotation velocity of Achernar (Domiciano de Souza *et al.* 2012). Jankov *et al.* (2001) treated explicitly the case of non-radial stellar pulsations. The photocenter shift delivers the first order moment of the spatial brightness distribution and some stellar regions are reinforced. Consequently, the modes that are observationally canceled in flux spectrum should appear in the spectrally resolved photocenter shift. The full reconstruction of the NRP modes requires a Fourier temporal analysis of the photocenter displacement $\epsilon(\lambda, t)$, in a generalization of the Fourier Doppler Imaging based on the spectrum $S(\lambda, t)$.

In this context, long baseline interferometry using spectral resolution in different bands from the visible to the IR, offers new oppotunities to observe the details of such stars with enough spatial resolution (*e.g.* van Belle 2012) to go beyond the limitation of classical techniques such as spectroscopy, photometry and polarimetry. We hereby describe a numerical model that includes a subset of the different mechanisms explained above: namely fast rotation and stellar pulsation that shape the emergent flux as a function of different parameters such as rotation rate (therefore flatenning), inclination angle to the line of sight, iso-velocity maps across the spectral line among others. The intensity of the differential phase signal critically depends on the characteristics of the observed spectral lines. In HR-K we have access to Brγ, which will be strongly polluted by circumstellar emission and to He lines, which are often strongly affected by atmospheric lines and will also be affected by circumstellar emission, but in a way different than for Br.

Our work including the effect of NRP can be innovative, especially with technological advances in interferometry. Observations campaigns on AMBER / VLTI had been requested and obtained for 2013 by our team, in the hope to validate our numerical model.

2 SCIROCCO+

2.1 Theoretical discription of the model

SCIROCCO+: stands for Simulation Code of Interferometric-observations for rotators and CirCumstellar Objects including non-radial pulsations. It is written

in `Matlab` and makes use of the following semi-analytical approach, adopting the frame depicted in Figure 1 (shown in cartesian reference): a pixellized intensity map is computed independently from a velocity map, and both are combined into a spectrally-resolved intensity image-cube, which can be input in a later step into an interferometric simulation code. This model was inspired by an anterior virsion; Simulation Code of Interferometric-observations for rotators and CirCumstellar Objects (SIROCCO) which does not include the non-radial pulsations effects (Hadjara *et al.* 2012).

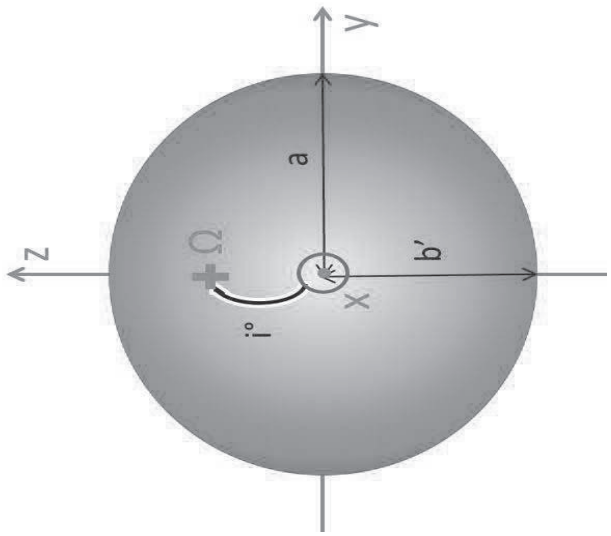


Fig. 1. Adopted reference system for a rotating star (flattened star with major axis a and minor axis b, here the apparent minor axis is $b' = ab/(a + (b - a)\cos i)$; assuming ellipsoid revolution principal/equations). The cross indicates the point where the rotation axis crosses the stellar surface. This rotation axis forms an angle i with the observer's direction (x axis) and its projection onto the sky is parallel to the z axis.

2.1.1 Intensity map

First, an intensity map of the star's photosphere is computed. We can use for example a simple limb-darkened model from (Hestroffer 1997), expressed in the geographical (co-latitude, longitude) coordinates (θ, ϕ):

$$I_c(\theta, \phi) = I_0(1 - \epsilon_\lambda(1 - \mu(\theta, \phi))) \tag{2.1}$$

where I_0 represents the flux at the center of the star, ϵ_λ is the limb darkening parameter, and $\mu(\theta, \phi)$ is the cosine of the angle between the normal to the surface at the point considered and the observer direction (Domiciano de Souza *et al.* 2004). The contour of the star is delimited by an ellipse with the minor axis in the

direction of the rotation axis of the star. The minor-to-major axis depends on the rotation rate following the prescriptions of inclination angle i (see Fig. 1). I_0 can serve as a weighting of the continuum flux as a function of wavelength (λ) using for example a Planck's law:

$$I_0(\lambda, T_{\rm eff}) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda\sigma T_{\rm eff}}} - 1} \tag{2.2}$$

h being Planck's constant, c the speed of light, σ the Boltzmann constant, and $T_{\rm eff}$ the effective temperature of the star. I_0 can also be used to input the von Zeipel's effect into our model, by considering a co-latitude-dependent temperature in the below-mentioned local gravity field equation:

$$I_0(\theta) \propto F(\theta) = \sigma T_{\rm eff}^4(\theta) \tag{2.3}$$

with $T_{\rm eff}(\theta) \propto g^{0.25}(\theta)$, g being the local gravity field, also called the modulus of local effective surface gravity $g = |\nabla\Psi(\theta)|$, with $\Psi(\theta)$ is the stellar equipotential surfaces (Domiciano de Souza *et al.* 2002). An example of intensity map combining rotational flattening and gravity darkening is shown in Figure 2 (top).

2.1.2 Velocity map

SCIRROCO+ produces a velocity map where we consider rotation and non-radial pulsations:

$$V_{proj}(\theta, \phi) = V_{\rm rot}(\theta, \phi) + V_{\rm nrp}(\theta, \phi). \tag{2.4}$$

In this equation we represent the global velocity map combining rotational flattening and non-radial pulsationsis shown in Figure 2 (bottom).

Where non-radial pulsations velocity has been introduced:

$$V_{\rm nrp}(\theta, \phi) = v_{puls} * Y_{lm} = v_{puls} * \sqrt{\frac{(2l+1)}{4\pi} \frac{l-|m|!}{l+|m|!}} P_{l,|m|} \cos(\theta) e^{im\phi}. \tag{2.5}$$

With, v_{puls}: the average velocity pulsation, l: the mode order, m: the mode azimutal order, and $P_{l,|m|}$: the Legendre function.

And rotational velocity be written:

$$V_{\rm rot}(\theta, \phi) = V_{\rm eq} \cos(\phi)(1 - \alpha \sin^2(\theta)) \sin(i). \tag{2.6}$$

Where $V_{\rm eq}$ represent the equatorial rotation velocity, and the parameter α allows us to include a parametric differential rotation law (Domiciano de Souza *et al.* 2004).

An example of the non-radial pulsations velocity map is shown in Figure 3 (top) & another example of pure rotational velocity map is shown in Figure 3 (bottom).

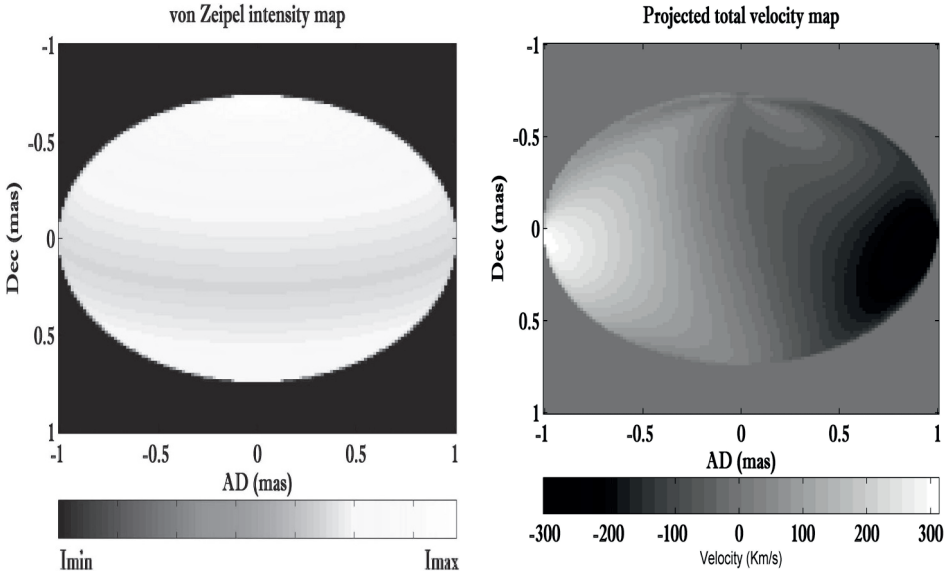


Fig. 2. *Left*: simulated η Centauri intensity map at continuum. The intensity at the poles is greater than at the equator. Here the velocity is upper than 80% of the critical velocity of the star. *Right*: global iso-velocity map (rotation+pulsation) of the same model (inclination 57°). Here the simulated rotation is differential (the velocity of rotation at the equator is 60% larger than at the poles).

2.1.3 Spectrally-resolved image cube

The last step of the modelization process is to compute λ-dependent maps. For that, we need to model the natural line-profile of the considered line: we can assume *e.g.* Gaussian, Lorentzian, or Voigt profile, at the central wavelength λ_0:

$$\begin{cases} H_{\text{Gauss}}(\lambda) = 1 - H_0 \left[-\pi H_0^2 \frac{(\lambda-\lambda_0)^2}{W^2} \right] \\ H_{\text{Lorentz}}(\lambda) = 1 - \left[\frac{H_0}{1+(\frac{\lambda-\lambda_0}{W/2})^2} \right] \\ H_{\text{Voigt}}(\lambda) = (H_{Gauss} * H_{Lorentz})(\lambda). \end{cases} \tag{2.7}$$

Where H_0 and W are the central depths and the equivalent width, respectively. The symbol $*$ represents the convolution product.

The last step calculates the intensity maps of the star as a function of wavelength. For that, we project via the Doppler effect the global velocity map (V_{proj}, Eq. (2.4)) to the intensity map (I_c, Eq. (2.1)), given the line profile (H, Eq. (2.7)) and the work wavelength λ:

$$I(\lambda, \theta, \phi) = H\left(\lambda + \lambda_0 \frac{V_{\text{proj}}(\theta, \phi)}{c}\right) I_c(\theta, \phi). \tag{2.8}$$

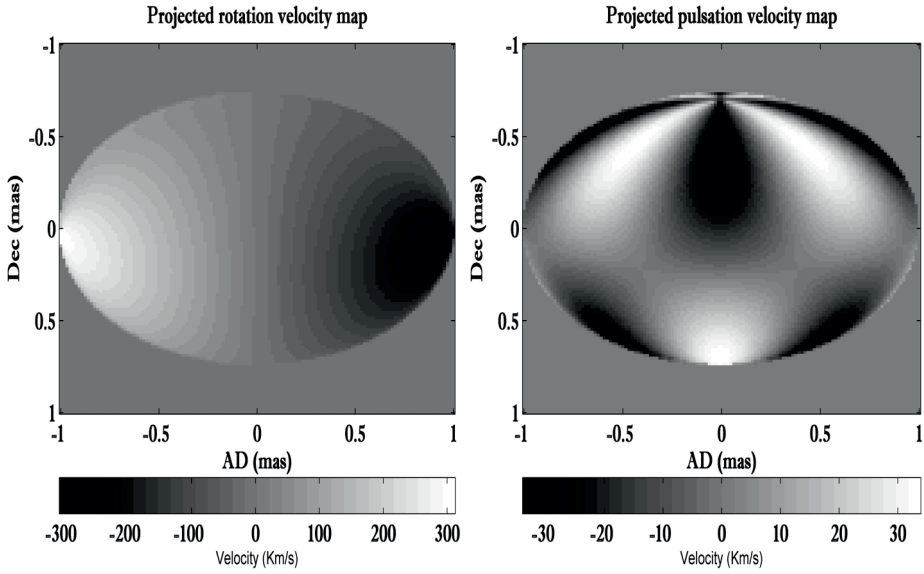


Fig. 3. *Left*: iso-velocity of pure rotation map (inclination 70°, 0° orientation), the direction of rotation thereby is from the left to the right (from the brightness to darkness). *Right*: iso-velocity of pure pulsation map, we note that $m = 4$, the mode azimuthal order (number of nodes lines that pass through the centers of vibration) and that $l = 5$, the mode order (total number of nodes lines). For the both, the color code adopted here, is brightness for the positive velocities and darkness for the negative ones.

We get one intensity map per wavelength of interest around the central wavelength λ_0 of the line (see Fig. 4, left). Once all intensity maps are computed, we synthesize the interferometric observables by Fourier-Transforming each map (see Fig. 4, right). This provides us spectra, visibility amplitudes, phases, and closure phases.

By comparing the observed interferometric measurements to the synthesized quantities, we can access to the parameters of the fast rotating star such as: effective temperature as a function of co-latitude, rotational rate, inclination, angular radius and flattening and, if possible the differential rotation.

2.2 Interferometric observations simulations - Application to η Cen

Assuming the following characteristics (Table 1):

Table 1. η Cen chosen parameters.

Star	*Eta Centauri*	v_{puls} **(km/s)**	*34*	$\mathbf{T}_{eq}$ (K)	*16000*
Spectral type	*Be*	**Orientation** (°)	*0*	$\mathbf{R}_{pole}$ **(R⊙)**	*3-4*
Velocity v **(km/s)**	*340*	**Gravity darkening** β	*0.25*	$\mathbf{R}_{eq}$ **(R⊙)**	*5-6*
Inclination i (°)	*70*	$\mathbf{T}_{pole}$ (K)	*21000*	**Oblateness**	*0.34*

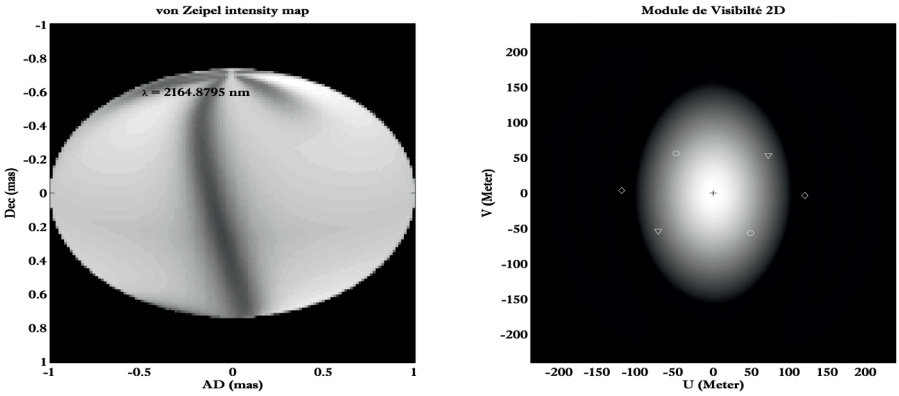


Fig. 4. *Left*: monochromatic intensity map for a given wavelength. *Right*: map of corresponding 2D module visibility, which is represented on the three bases with interferometric which will make the observation (1st base small circles, 2nd small triangles and 3rd small diamonds).

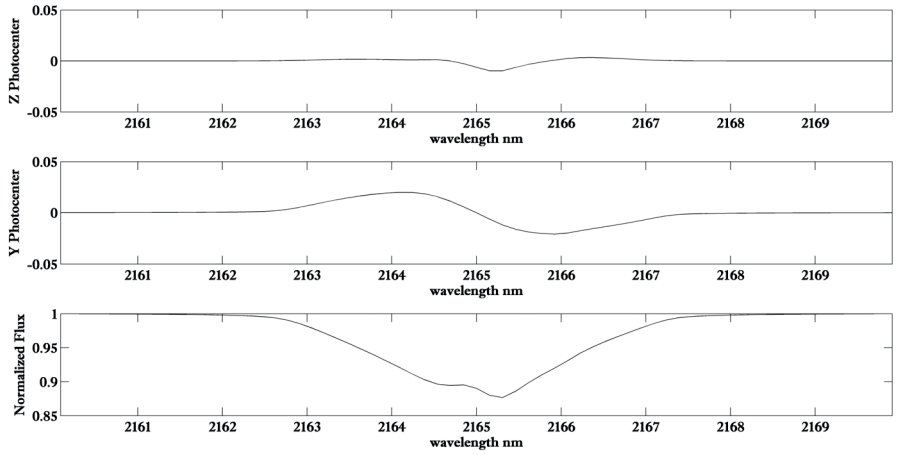


Fig. 5. *Top*: photo-center (or centroid: the first order term of the phase by Mac Lauren development Jankov *et al.* 2001) along the Z axis (see reference adopted in Fig. 1). We note well, here, the influence of the pulsation effect in addition to the inclination effect. *Middle*: photo-center by Y (note that the photo-centers are in radian). *Bottom*: normalized spectrum, we see well that our starting line (with a depth of 0.6, and a $FWHM = 10\Delta\lambda$) has expanded and its depth was decreased (precisely because of the rotation), it is impacted too by the pulsation (the double hump at the bottom of the spectrum).

In addition, we introduce to our model a differential rotation coefficient ($\alpha = 0.6$) and a Voigt intrinsic line profile with a depth of 0.6, and a $FWHM = 10\Delta\lambda$).

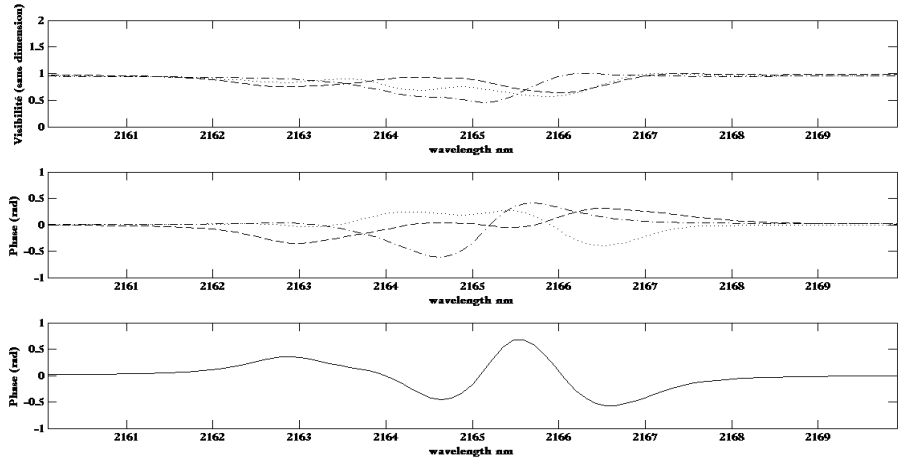


Fig. 6. *Top*: module visibility observed by the interferometric three bases (dot line for the cercles base, dash-dot line for the triangles base and dash line for the diamonds base). *Middle*: the phases observed in the three interferometric bases (dot line for the cercles base, dash-dot line for the triangles base and dash line for the diamonds base), we note that the dash-dot line & the dot line are in opposite phase when their corresponding bases (cercles base & triangles base) are perpendicular. We note although the phase observed by the base close to perpendicular to the axis of rotation of the star (dash-dot line) is the one that has the highest amplitude & inversely that which is close to along the axis of rotation (dot line) is low, without forgetting that the dash line is the lowest that the corresponding interferometric base is outside of the first visibility lobe (diamonds base). *Bottom*: the closure phase; which is classically defined, in AMBER, as: $\Psi = \Phi_{12} + \Phi_{23} - \Phi_{13}$.

We choose to simulate interferometric observations with the AMBER/VLTI instrument on the 3 following interferometric baselines : $K0-G1$ (74.63 m, $-139°$), $G1-A0$ (90.12 m, $-53.6°$), $A0-K0$ (120.6 m, $-91.7°$), around the Brackettγ line (2.165 μm). Without forgetting the pulsation parameters of the star: $m = 4$; m is the azimuthal order of the mode (number of nodes lines that corss the vibration poles) & $l = 5$; l is the mode order (the total number of nodes lines). The star pulsing with a velocity $v_{puls} = 34$ km/h.

The intensity map at continuum & global iso-velocity map are shown in Figure 2, the pure iso-velocity rotation map & pure iso-velocity pulsation map in Figure 3, the monochromatic intensity map at a given wavelength & corresponding 2D visibility amplitude map in Figure 4, the photo-centers & spectrum in Figure 5, and finally visibilities, phases & closure phase in Figure 6.

3 Conclusions & discussions

We presented here a semi-analytical model of fast-rotators including non-radial pulsations whose aim is to interpret interferometric datasets. We are able to produce interferometric observables using a set of physical parameters like the

rotation law, gravity darkening effect, etc., while keeping the computing time reasonable (one set of visibility curves can be computed in 17 s).

Note that for this simulation, we have obtain the same shape of spectrum (in Fig. 5) as observing by Levenhagen *et al.* (2003), with spectroscopy in $H\alpha$, which is encouraging.

The next step is to develop a "model-fitting" approach to compare real datasets with this model.

References

Chelli, A., & Petrov, R.G., 1995, A&A, 109, 401

Collins, G.W., 1974, ApJ, 191, 157

Collins, G.W., & Harrington, J.P., 1966, ApJ, 146, 152

Collins, G.W., & Sonneborn, G.H., 1977, ApJ, 34, 41

Domiciano de Souza, A., Hadjara, M., Vakili, F., *et al.*, 2012, A&A, 545, 130

Domiciano de Souza, A., Zorec, J., Jankov, S., Vakili, F., & Abe, L., 2004, A&A, 418, 781

Domiciano de Souza, A., Kervella, P., Jankov, S., *et al.*, 2003, A&A, 407, L47

Domiciano de Souza, A., Vakili, F., Jankov, S., Janot-Pacheco, E., & Abe, L., 2002, A&A, 393, 345

Hadjara, M., Vakili, F., Domiciano de Souza, A., Millour, F. & Bendjoya, P., 2012, SCIROCCO stands for Simulation Code of Interferometric-observations for rotators and CirCumstellar Objects, SF2A 2012, 533

Hestroffer, D., 1997, A&A, 327, 199

Hutchings, J.B., & Stoeckley, T.R., 1977, PASP, 89, 19

Hutchings, J.B., Nemec, J.M., & Cassidy, J., 1979, PASP, 91, 313

Jankov, S., Petrov, R., Vakili, F., Robbe-Dubois, S., & Domiciano, A., 2005, PASRB, 5, 83

Jankov, S., Vakili, F., Domiciano de Souza, A., & Janot-Pacheco, E., 2001, A&A, 377, 721

Jankov, S., Janot-Pacheco, E., & Leister, N.V., 2000, ApJ, 540, 535

Levenhagen, R.S., Leister, N.V., Zorec, J., *et al.*, 2003, A&A, 400, 599

Maeder, A., & Peytremann, E., 1972, A&A, 21, 279

Martayan, C., Frémat, Y., Hubert, A.-M., *et al.*, 2006, A&A, 452, 273

Meynet, G., 2009, Lect. Notes Phys., 765, 139

Petrov, R.G., 1988, Diffraction-Limited Imaging with Very Large Telescopes, ed. D.M. Alloin & J.M. Mariotti (Kluwer), 249

Rivinius, Th., Štefl, S., & Baade, D., 1999, MNRAS, 227, 801

Stockley, T.R., & Buscombe, W., 1987, MNRAS, 227, 801

Vakili, F., & Percheron I., 1991, Rapid Variability of OB-Stars: Nature and Diagnosis Value, ed. D. Baade, 15-17 Oct. (ESO, GArching, Germany), 77

van Belle, G.T., 2012, A&ARv, 20, 51

von Zeipel, H., 1924, MNRAS, 84, 665

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 141–154*

HIGH ANGULAR RESOLUTION AND YOUNG STELLAR OBJECTS: IMAGING THE SURROUNDINGS OF MWC 158 BY OPTICAL INTERFEROMETRY

J. Kluska[1], F. Malbet[1], J.-P. Berger[2], M. Benisty[1], B. Lazareff[1], J.-B. Le Bouquin[1] and C. Pinte[1]

Abstract. In the course of our VLTI young stellar object PIONIER imaging program, we have identified a strong visibility chromatic dependency that appeared in certain sources. This effect, rising value of visibilities with decreasing wavelengths over one base, is also present in previous published and archival AMBER data. For Herbig AeBe stars, the H band is generally located at the transition between the star and the disk predominance in flux for Herbig AeBe stars. We believe that this phenomenon is responsible for the visibility rise effect. We present a method to correct the visibilities from this effect in order to allow "gray" image reconstruction software, like `Mira`, to be used. In parallel we probe the interest of carrying an image reconstruction in each spectral channel and then combine them to obtain the final broadband one. As an illustration we apply these imaging methods to MWC158, a (possibly Herbig) B[e] star intensively observed with PIONIER. Finally, we compare our result with a parametric model fitted onto the data.

1 Introduction

The processes that lead to the formation of exoplanets are important to understand. Stars form after a collapse of a giant cloud of dust and gas. After a million year, a protoplanetary disk is forming around the star, believed to be the birthplace of planets.

[1] Institut de Planétologie et d'Astrophysique de Grenoble, UMR 5274, BP. 53, 38041 Grenoble Cedex 9, France

[2] ESO, Santiago Office, Alonso de Cordova 3107, Vitacura, Casilla 19001, Santiago de Chile, Chile

DOI: 10.1051/eas/1359008

A young star is surrounded by an active environment with which it interacts. Accretion disks (Monnier & Millan-Gabet 2002), inner gaseous disks (Benisty *et al.* 2010; Eisner *et al.* 2009; Tannirkulam *et al.* 2008), infalling envelop renmands, winds (Cabrit *et al.* 2010; Dougados *et al.* 2005; Malbet *et al.* 2007; Tatulli *et al.* 2007) and jets (Cabrit 2003; Dougados *et al.* 2004) are the main components of such environments. There are several types of young stellar objects. The complexity of physical phenomena at play requires direct observation at the astronomical unit (A.U.) scale. Optical interferometry is able to bring such informations, because it can observe both in the near infrared, where the hot dust and hot gas nearby the star are emitting, and resolve the first A.U., which correspond to milliarcsecond scale at the distance of star formation regions.

Interferometry consists in combining the light of 2 or more telescopes in order to measure the complex degree of coherence. For that purpose, the interferometer measures interference fringes. The amplitude of the fringes yields the norm, and its position the phase of a complex quantity called visibility $V(u, v)$. Thanks to the van Cittert-Zernicke theorem we know that the Fourier transform of the visibilities in the Fourier plan (u, v) gives us the intensity distribution $I(x, y)$ of the source. Unfortunately, in the near infrared (NIR) the atmosphere blurs the phases of the visibilities. The hint then, is to measure a quantity that is the sum of the phases over a baseline triangle. In that case, the atmospheric influence vanishes and we obtain only an astrophysical quantity called the closure phase. So, in practice, there are two interferometric measurements: the squared amplitude of the visibilities V^2 and the closure phases.

We noticed in several datasets that the visibility is higher at short wavelengths. If B is a baseline length projected on the sky plan and λ the wavelength, we can plot the squared visibilities V^2 in function of B/λ (which is the spatial frequency). For a monochromatic object, we expect the points to follow a general trend in the visibility curve, since B/λ represents only the spatial frequency. But, this is not the case. Indeed, we can see (Fig. 1) that the rising curve of visibilities per base is not fitting the general trend of the data for different baselines. First, it was seen in AMBER (Petrov *et al.* 2007) data, but it was considered as an instrumental defect. Now, the same effect has been observed with PIONIER (Le Bouquin *et al.* 2011). We try to explain this effect astrophysically, claiming that the image of the object is varying through the different spectral channels inside the same spectral band, and we propose three techniques in order to take it into account and to be able to reconstruct images.

These methods will be applied to an astrophysical object. They are useful to analyze MWC158 (also known as HD 50138). This star is a Be star known to have the B[e] phenomenon and presents a strong variability (Andrillat & Houziaux 1991; Borges Fernandes *et al.* 2009; Hutsemekers 1985; Pogodin 1997), which complexify the evolutionary stage identification of the source. Its distance is poorly constrained ($d = 500$ pc $\pm$ 150 pc, van Leeuwen 2007).

In Section 2 we will describe the chromatic effects in the visibilities and the Section 3 will show the different methods to deal with them. Finally we will apply them to the astrophysical case of MWC 158 in the Section 4.

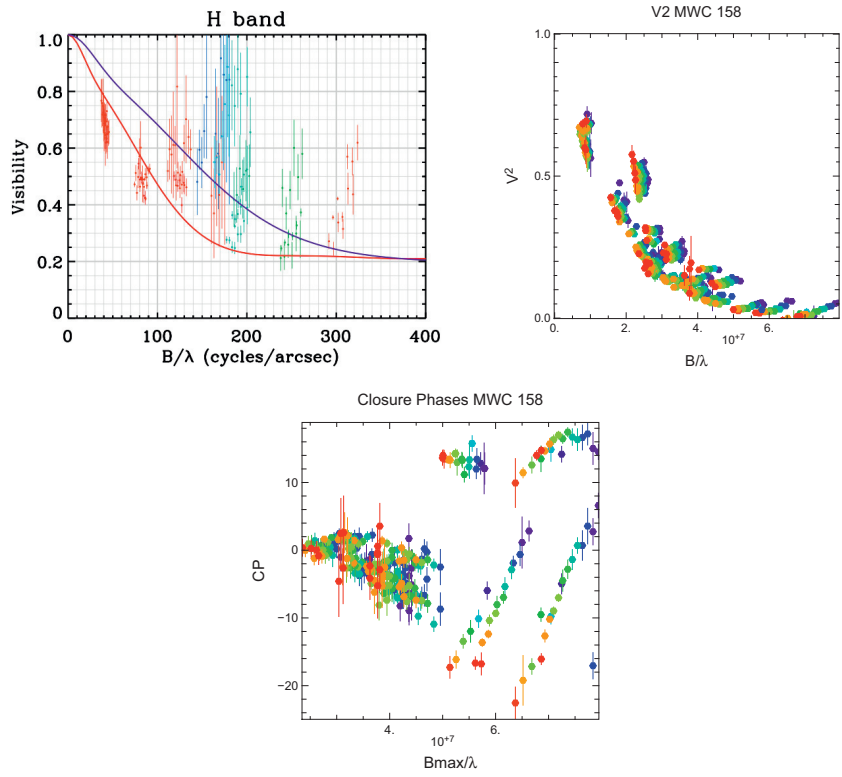


Fig. 1. Data on MWC 158. *Left*: AMBER (Petrov *et al.* 2007) data from Borges Fernandes *et al.* (2011). *Center*: PIONIER (Le Bouquin *et al.* 2011) squared visiblities. *Right*: PIONIER closure phases. For PIONIER data the gray scale varies in function of the wavelength. We can see on the visibility graphs than a "coma" trend appear for visibilities belonging to the same baseline.

2 Chromatism

Since interferometric instruments with spectral dispersion exist, we need to take into account the flux variations with the wavelength in order to correctly analyze the data and have access to the spectral super synthesis. In the case of Young Stellar Objects (YSOs), we noticed that the visibilities have a strong spectral dependence such as the geometrical shape of the object could not explain it.

For Herbig AeBe star, the chromatic effect explained in the Section 1 exists typically for the Near Infrared interferometry. In the following, we explore the possibility that this effect is caused by a different spectral index between the central star and its surrounding media.

In order to confirm that, we made a simple model with a central star and its dusty disk.

2.1 The star

In our model, the star is considered to be unresolved. This hypothesis is justified for the young objects we are looking at. To simplify our model, we assume $V_{star} = 1$.

For the star we have 3 parameters: the radius (R_*), the distance (d) and the temperature (T_*). If we asumme a Herbig AeBe star with a temperature of 12 000 K radiating as a black body, we know that in NIR we will look on the Rayleigh-Jeans regime of a black body (see Fig. 2). That means that the spectral curve is proportional to a power law: $F_\lambda^{star} \propto \lambda^{-4}$ (F being the luminous flux).

2.2 The disk

The disk model is simple: it is a geometrically thin optically thick passive disk. Its temperature is a function of the radius:

$$T(r) = T(r_0)\left(\frac{r}{r_0}\right)^{-q} \tag{2.1}$$

with:

$$q = \frac{3}{4} \tag{2.2}$$

see Adams *et al.* (1988); Lynden-Bell & Pringle (1974).

The disk will be sampled on several rings, each ring having its own temperature as a function of its distance to the star. We will use a lot of rings (more than 100) to model the disk. We can then approximate that each ring has the fourier transform of an infinitesimal width ring. However, for the flux, we will take each ring width into account. The other geometrical parameters are the inclination (inc), the inner and outer rims radii (R_{in}, R_{out}), and the temperature of the inner rim (T_{in}). The flux of each ring will be a black body at the temperature of the disk. The ring visibility is defined as follows (Berger 2003):

$$V_{ring} = J_0\left(2\pi r\frac{B}{\lambda}\right). \tag{2.3}$$

Then to obtain the Fourier transform of the whole disk we have to add the flux of each ring and sum every contributions:

$$V_{disk} = \sum_i^{n_{ring}} J_0(2\pi r_i \frac{B}{\lambda}) B_\lambda(T_i) S_i. \tag{2.4}$$

S_i being the surface of the i-th ring ($S_i = 2\pi r_i w_i$; w_i being the width of the i-th ring and r_i its radius).

In the results shown in Figure 2 the chromatic effect which tends to look like the data shown in Figure 1. The visibilities have the particularity to rise for every base even though the general trend of the visibilities is to decrease. It is similar

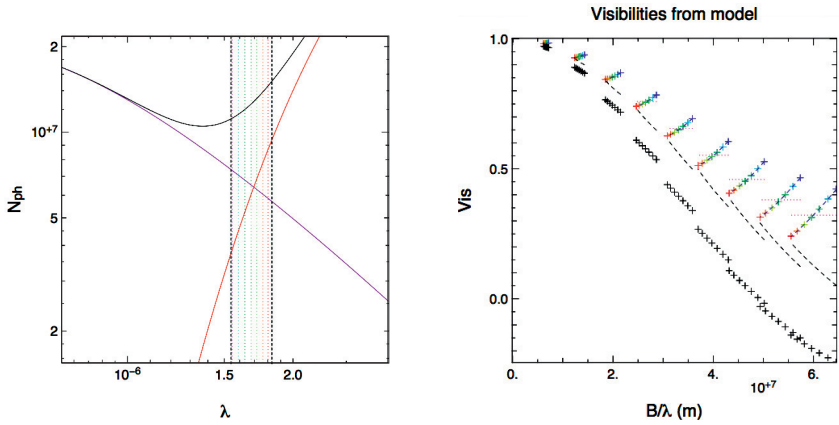


Fig. 2. On the *left*, we show the location of the PIONIER spectral channels on the Spectral Energy Distribution (SED) of the model. The component on the left is the stellar photosphere approximated by a black body and the component on the right is the environment. The black line is the sum of the two components. We can see that they are located at the crossing between the stellar and the dust fluxes. On the *right*, we can see that the chromatic phenomenon is reproduced in the visibilities. For any base the visibility is rising with the spatial frequency (B/λ). On black we have the visibilities of the environment only.

to that we observe in real data. We can reproduce this effect by the chromatism that exists between the star and its environment. We can then suggest that the chromatic effect is not instrumental but astrophysical.

The effect is dominated by the flux ratio which is changing through the different spectral channels. If we compute the total correlated flux we have (given that the Fourier transform is linear):

$$V_{\rm tot}(B/\lambda)F_{\rm tot}(\lambda) = F_*(\lambda) + V_{\rm disk}(B/\lambda)F_{\rm disk}(\lambda) \tag{2.5}$$

with:

$$F_{\rm tot}(\lambda) = F_*(\lambda) + F_{\rm disk}(\lambda). \tag{2.6}$$

If we introduce the stellar to total flux ratio f_*, we obtain the mathematical description of the chromatic phenomena:

$$V_{\rm tot}(B/\lambda) = f_*(\lambda) + V_{\rm disk}(B/\lambda)(1 - f_*(\lambda)) \tag{2.7}$$

with

$$f_*(\lambda) = \frac{F_*(\lambda)}{F_{\rm tot}(\lambda)}. \tag{2.8}$$

In the next section we will discuss the different methods to overcome the chromatic effect.

3 Methods

Our goal is to be able to analyze chromatic data. We developed three complementary methods to do that: gray image reconstructions, data modification and parametric fit. The first two methods are based on image reconstruction and the last one is model fitting. We are mostly interested in the disk around the star and we are looking for informations on the resolved geometry and the strength of the chromatic effects.

3.1 Image reconstruction per spectral channel

Once we are aware of the chromatic effect, one can make image reconstructions selecting only one wavelength per reconstruction (see Fig. 3). In that case the gray image reconstruction is justified. The technique is to have one image per wavelength and to stack all the images in order to have the final broadband one.

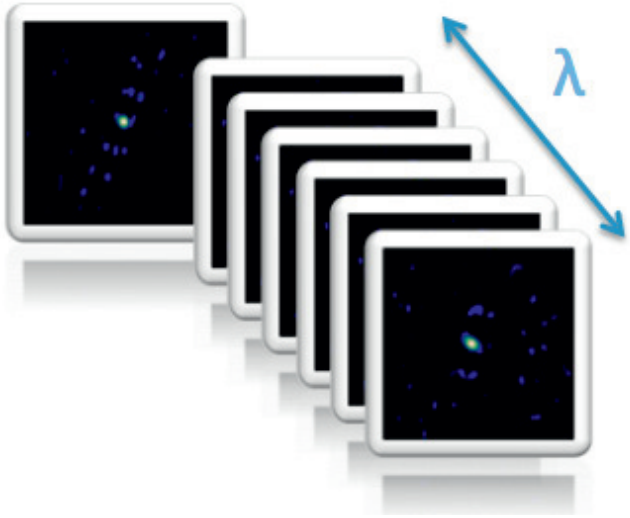


Fig. 3. An image reconstruction is made for each spectral channel of the instrument. Then all the images are stacked together in order to obtain the final image.

The presence of various components of different spectral indexes prevents from using a gray emission approximation in the image reconstruction process. As a consequence, since we need to work on a per-spectral-channel basis the (u, v) coverage quality is severely affected.

3.2 Modification of the data

We want to have access to the disk visibilities. From the Equation (2.7), if we know the SED and then the stellar flux ratio $f_*(\lambda)$ and its variation through the

wavelengths, we can compute the disk visibilities as:

$$V_{\rm disk}(B/\lambda) = \frac{V_{\rm tot}(B/\lambda) - f_*(\lambda)}{1 - f_*(\lambda)}. \tag{3.1}$$

We can apply the modification described in Equation (3.1) to one of the interferometric measurements which is the power spectrum ($VV^* = |V|^2$). In summary, our measurements are $|V_{\rm data}|^2$ and we want to recover $|V_{disk}|^2$. Using the Equation (3.1), we have:

$$V^2_{\rm disk}(B/\lambda) = \left(\frac{\sqrt{|V_{\rm data}(B/\lambda)|^2} - f_*(\lambda)}{1 - f_*(\lambda)} \right)^2. \tag{3.2}$$

One of the problem is the value that we take for $\sqrt{|V_{\rm data}(B/\lambda)|^2}$; we must choose between the positive (phase $\phi = 0$) and the negative one ($\phi = \pi$). But it could be solved analyzing more precisely the data and other interferometric observables like the phase of the bispectrum (also called the closure phase). The goal is to find where the visibilities are reaching the first "zero", where there is a jump of π in the phase (and the in the closure phase). Moreover the chromatic parameter f_* must be estimated from other observations (*e.g.* photometry).

It is not possible to retrieve the bispectrum phase of the dust from the data because we are loosing the phase of each pair of telescopes. The equations lead to a solution where we need the phase (Ragland *et al.* 2012).

3.3 Parametric model

In this section we have attempted to model the object. The model is geometrical and includes the chromatic effect as described in the Sections 1 and 2. Our model is composed of multiple components and was developed when chromatic data was fitted.

3.3.1 Geometric part of the fit

The first component of the model is an unresolved star (a dirac in the image space) which can be shifted compared to the image photo center (that will produce a rise of closure phases). The second component is a Gaussian ring. This shape is close to the shape of a puffed-up inner disk rim model. (Isella & Natta 2005). In the Fourier space the ring is defined as in Equation (2.3) but using $\sqrt{u^2 + v^2}$ for the spatial frequencies (B/λ) and their orientations that we want to solve.

We take into account the Position Angle (*P.A.*), which is defined from the North to the East, and the inclination (*inc*). One of the parameters of this shape is the ring radius r. But this will define a ring with a infinitely small width. In order to have a Gaussian width we have to convolve the ring formulae by a Gaussian, in other words, to multiply the visibility of the ring by the visibility of the Gaussian function with the correspondent width w. Once we have the Gaussian ring, we will

add some azimuthal modulations of the ring intensity to be closer to the physics of an inner rim. The modulations are functions in cosinus and sinus of the azimuthal angle (α) of the ring which starts at its major axis. We have included two sorts of modulation: one on 2 π (c_1, s_1) and the second on π (c_2, s_2). They are described on Figure 4 We add a Gaussian width to the ring, with r_{gauss} being the Half Width at Half Maximum (HWHM) of the Gaussian function.

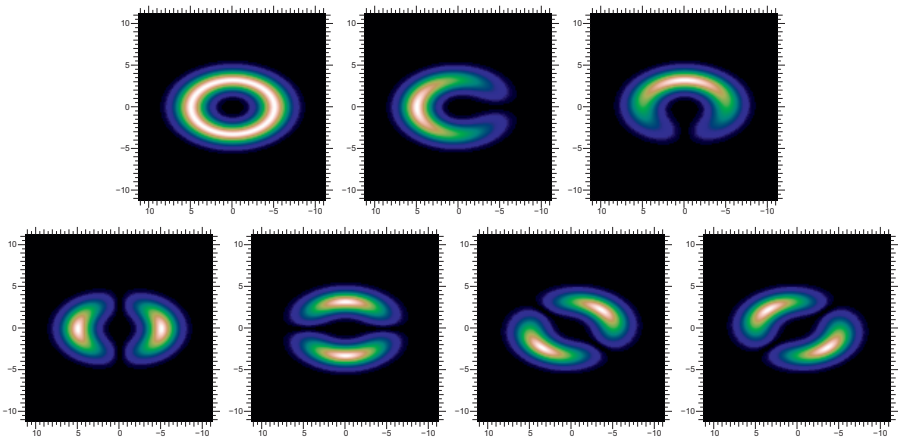


Fig. 4. Description of the azimuthal modulations. The first figure is the original gaussian ring slightly inclined. The second figure represent a π-modulation with $c1 = 1$ (modulation through the long axis) and the third with $s1 = 1$ (modulation through the short axis). The second line represent the 2π-modulation with respectively $c2 = 1$, $c2 = -1$, $s2 = 1$, $s2 = -1$. The model can combine the modulations and express them with slighter intensity (below 1).

We have a model with three components: the star, the Gaussian ring and a second Gaussian ring or a Gaussian function.

The total visibilities are depending on all these components weighted by their flux.

3.3.2 Modeling the chromatism

To obtain the model visibilities we use the linearity property of the Fourier transform.

$$F_{\text{tot}} V_{\text{tot}} = F_* V_* + F_1 V_1 + F_2 V_2. \tag{3.3}$$

The fluxes are the ones recieved by the interferometric instrument. So we have:

$$F = \int_{\lambda - \frac{\Delta\lambda}{2}}^{\lambda + \frac{\Delta\lambda}{2}} F_\lambda d\lambda = \int_{\nu - \frac{\Delta\nu}{2}}^{\nu + \frac{\Delta\nu}{2}} F_\nu d\nu \tag{3.4}$$

with $\Delta\lambda$ and $\Delta\nu$ being the spectral width of one spectral channel in wavelength (λ) and frequency (ν).

We will use the approximation that the channel spectral width is constant and that the flux is constant in one spectral band. The flux is then equal to the value of F_λ at the central wavelength of a spectral channel. From the Equation (3.3), we see that we can determine a flux ratio at one wavelength and to deduce the ratios on the other wavelength by the laws that we assume for each component. PIONIER is operating in the NIR in the H band. At this wavelength, we can assume that Herbig stars are in their Rayleigh-Jeans regime. That means that their flux F_λ is proportional to the wavelength at the power of -4. The laws for the environment are more difficult to find. We can fit a power-law in wavelength or to a black body variation if we are resolving a thermal emitting region. Since the dust temperature is supposed to be below 2000 K (Dullemond & Monnier 2010), we can assume that it is in its Wien regime. Then if we assume black body regimes we obtain (from Eq. (3.3)):

$$V_{\rm tot}(B,\lambda) = f_*^0\big(\frac{\lambda}{\lambda_0}\big)^{-4} + f_1^0\frac{BB(\lambda,T_1)}{BB(\lambda_0,T_1)}V_1(BB,\lambda) + f_2^0\frac{BB(\lambda,T_2)}{BB(\lambda_0,T_2)}V_2(BB,\lambda) \tag{3.5}$$

with f^0 the flux ratios at λ_0, T the temperature of a component and B the baseline and

$$BB(\lambda,T) = \frac{2hc\lambda^{-5}}{\exp\frac{hc}{k_B\lambda T}-1} \tag{3.6}$$

is the black body function, with h the Planck constant, c the light speed, and k_B the Boltzmann constant.

The variations of the flux ratios through the observational band will build the chromatic effect that we want to take into account in our fit.

Once we get all our tools to investigate data with chromatic effect, let us apply them on an astrophysical case: MWC 158.

4 The case of MWC 158

The interest on this object came with the data we get with PIONIER (Le Bouquin *et al.* 2011) a 4 telescopes interferometric, visitor instrument operating at the VLTI and which observes in the H band.

4.1 Image reconstructions

We were interested into this data (see Fig. 1) because it shows clearly signs of chromatism. As the u,v-plan is sufficiently covered we can reconstruct images. We use the `Mira` algorithm (Thiébaut 2008), but as many image reconstruction algorithms it does not take into account the chromatism. Since it extrapolates and interpolates the Fourier space, the chromatism makes him "guessing" badly and many artifacts appear. We then use the monochromatic reconstructions per spectral channel. It means that we select every spectral channel one by one and

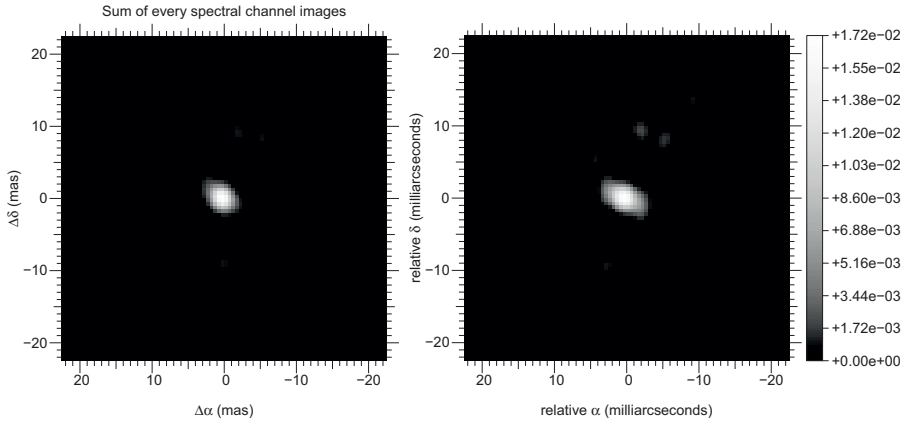


Fig. 5. *Left*: the stack of image reconstructions per spectral channels. *Right*: image reconstruction after modifying the squared visibilities.

make one reconstruction by channel. We also use the visibility correction method. These method (described Sect. 3.2) is not modifying the closure phases (they should be stronger). The results are showed on the Figure 5.

We can see that there is a second resolved component. We can also see the orientation of the smallest extended component. Both of the reconstruction methods shows similar patterns. That brings us to the idea to fit two extended components.

4.2 Parametric fit

The fit bring us an idea on the geometry and the light emission from the source but with a strong *a priori* which is the model we want to fit. That is why we took the geometries suggested by the image reconstructions. We can see a central extended part which is composed of the star and its environment which seems to have a P.A. and an inclination.

We have begun the fit with one extended component which is a Gaussian function or a Gaussian ring. Both of the fits gave us the more or less the same inclinations and P.A. which are consistent with the image reconstructions. But the data was not entirely fitted: the short baselines indicates that there is a more extended component as showed by image reconstruction. We then add another component to our fit. In order to fit the strong closure phase signal we add azimuthal modulation to the ring. It appears not to be sufficient, and the best fit was to shift the central star. It is the only solution to fit the closure phases.

In the end, and adding the different parameters, we ended with 15 parameters and a χ^2 of 3.5. In the current state of the data processing and interpretation, we believe that the best fit is presented Figure 6. The parameters are in the Table 1. We can see that the best fit is done with two Gaussian rings.

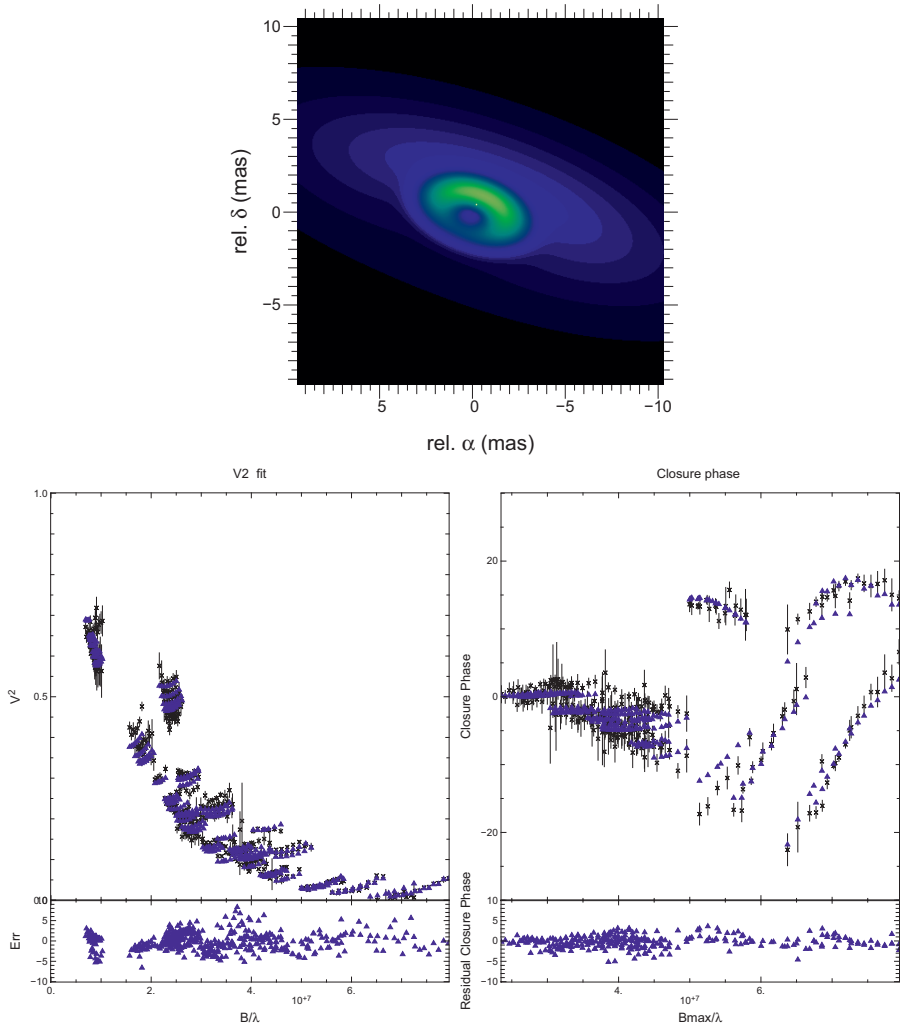


Fig. 6. The best fit results are presented. *Left*: the image corresponding to the best fit. *Center*: the fit on the V2. Left the fit on the Closure Phases. The triangles are the fit result, the crosses are the data points.

The geometrical fit suggests a star, with a relatively close Gaussian ring (radius of 1.5 mas) with a lot of flux (≈60%). We interpret that as the resolution of the inner rim of the dusty disk. Its azimuthal modulation is strong in the semi minor axis direction which leads us to deduce that it is due to the inclination. Moreover, the star is shifted towards the most brilliant part of the inner rim. It indicates that the inner rim has a non-negligible height. The outer ring suggests the continuation of the disk, or a part of the disk which is not self shadowed, or a halo. The constraints are poor so we can not conclude on its origins.

Table 1. The parametric fit results. rr_i is the i-th ring radius, x_* and y_* are the star shift with respect to the rings. The other acronyms are described in the Section 3.3.

χ^2	3.53							
Star			Ring 1			Ring 2		
Parameter	Value	Error	Param.	Value	Error	Param.	Value	Error
f_*^0	18.0%	±0.7%	f_1^0	58.7%	±1.1%	f_2^0	23.3%	±1.8%
			T_1	1482 K	±79 K	T_2	1326 K	±48 K
x_*	−0.22 mas	±0.01 mas	rr_1	1.76 mas	±0.02 mas	rr_2	3.90 mas	±0.02 mas
y_*	0.40 mas	±0.01 mas	w_1	1.94 mas	±0.13 mas	w_2	13.44 mas	±0.32 mas
			P.A.	71.67	±0.64	*P.A.*	71.67	±0.64
			inc_1	52.6	±1.3	inc_2	67.7	±1.4
			c_1	0.126	±0.013	c_1	0.126	±0.013
			s_1	−0.593	±0.033	s_1	−0.593	±0.033

Table 2. The previous interferometry results on MWC 158. Some results were complete on instruments watching at longer wavelengths (10.7 or 2.2 μm). The *P.A.* are consistent and the inclinations i also. The references are: 6: Borges Fernandes *et al.* (2011) and 20: Monnier *et al.* (2009).

FWHM	FWHM2	i	*P.A.*	χ^2	λ_0	Ref.
66 ± 4		45	63 ± 6	1	10.7 μm	20
64.7 ± 0.6		70.1 ± 0.7	59.1 ± 1.7	5.1	10.7 μm	6
35.2 ± 1.5	131.4 ± 11.2	56.7 ± 0.4	65.9 ± 2.0	1.9	10.7 μm	6
4.4 ± 0.5		54 ± 8	66 ± 9	40.8	2.2 μm	6
3.0 ± 0.4	$\geq$14.0	54 ± 8	77 ± 2	13.3	2.2 μm	6

The results are shown in Table 2. They are closed to the images get by reconstruction. The results are also consistent with that found with previous observations (Borges Fernandes *et al.* 2011; Monnier *et al.* 2009). The authors found similar *P.A.* with close values of the inclination of the most luminous extended object.

The fit of the chromatism, indicates us a black body temperature of the inner rim of $\approx$1500 K (see Table 1). This is approximately the dust sublimation temperature found in the litterature (Dullemond & Monnier 2010; Duschl *et al.* 1996).

5 Conclusions

The chromatic effect due to the flux predominances of two objects of different sizes is well understood and can be used in order to find astrophysical information of the object. In the case of Herbig AeBe stars we are able to have an approximation of the temperature of the components. If the chromatic information is given, we can perform gray disk image reconstructions. They are contributing to the astrophysical analysis of the object because they show the P.A. and the inclinations of the disk. Moreover, in the case of MWC 158 it brought us the idea of the second extended component, even if it is poorly constrain (we can have the information on the flux ratio). By the fit we were able to find a value for the inner rim radius and its temperature and to compare what we found with the data from photometry. We bring the first confirmation of the dust sublimation temperature at the inner rim. The information taken from the NIR interferometry and the chromatic effect argue in favor of a young nature of MWC 158.

The main challenge is to be able to make chromatic Young Stellar Objects image reconstructions keeping the super spectral synthesis and without information on the total flux variation. It seems to be degenerated and we need information from photometry. One of the thing which is in process of testing, is the adaptation of the `Mira` algorithm to the case of young stellar object. The "gray" `mira` free parameters are the image pixels intensities. If we define the image as the

image of the dust at λ_0, the start can be represented by a dirac at the center of the image. Hence, we can put the stellar flux and a stellar relative spectral power law as additional parameters to the fit. Since the regularization will tend to smooth the Fourier plan, the algorithm will favor the added parameters to fit the fixture. Indeed, the parameters will not be constrained by the regularization and the gradient will be stronger on it than on the pixels values.

For these objects, there is a need to include parametric models in the image reconstruction algorithms and to develop a global chromatic image reconstruction algorithm.

This work is supported by the French ANR POLCA project (Processing of pOLychromatic interferometriC data for Astrophysics, ANR-10-BLAN-0511).

References

Adams, F.C., Lada, C.J., & Shu, F.H., 1988, ApJ, 326, 865

Andrillat, Y., & Houziaux, L., 1991, IAU Circ., 5164, 3

Benisty, M., Natta, A., Isella, A., *et al.*, 2010, A&A, 511, A74

Berger, J.-P., 2003, EAS Publications Series, 6, 23

Borges Fernandes, M., Kraus, M., Chesneau, O., *et al.*, 2009, A&A, 508, 309

Borges Fernandes, M., Meilland, A., Bendjoya, P., *et al.*, 2011, A&A, 528, A20

Cabrit, S., 2003, Ap&SS, 287, 259

Cabrit, S., Ferreira, J., Dougados, C., & Garcia, P., 2010, Highlights of Astronomy, 15, 261

Dougados, C., Bouvier, J., Ferreira, J., & Cabrit, S., 2005, IAU Symp., 226, 491

Dougados, C., Cabrit, S., Ferreira, J., *et al.*, 2004, Ap&SS, 293, 45

Dullemond, C.P., & Monnier, J.D., 2010, ARA&A, 48, 205

Duschl, W.J., Gail, H.-P., & Tscharnuter, W.M., 1996, A&A, 312, 624

Eisner, J.A., Graham, J.R., Akeson, R.L., & Najita, J., 2009, ApJ, 692, 309

Hutsemekers, D., 1985, A&AS, 60, 373

Isella, A., & Natta, A., 2005, A&A, 438, 899

Le Bouquin, J.-B., Berger, J.-P., Lazareff, B., *et al.*, 2011, A&A, 535, A67

Lynden-Bell, D., & Pringle, J.E., 1974, MNRAS, 168, 603

Malbet, F., Benisty, M., de Wit, W.-J., *et al.*, 2007, A&A, 464, 43

Monnier, J.D., & Millan-Gabet, R., 2002, ApJ, 579, 694

Monnier, J.D., Tuthill, P.G., Ireland, M., *et al.*, 2009, ApJ, 700, 491

Petrov, R.G., Malbet, F., Weigelt, G., *et al.*, 2007, A&A, 464, 1

Pogodin, M.A., 1997, A&A, 317, 185

Ragland, S., Ohnaka, K., Hillenbrand, L., *et al.*, 2012, ApJ, 746, 126

Tannirkulam, A., Monnier, J.D., Harries, T.J., *et al.*, 2008, ApJ, 689, 513

Tatulli, E., Isella, A., Natta, A., *et al.*, 2007, A&A, 464, 55

Thiébaut, E. 2008, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conf. Ser., Vol. 7013

van Leeuwen, F., 2007, A&A, 474, 653

Physical Models and Data Processing

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 157–187*

PRINCIPLES OF IMAGE RECONSTRUCTION IN INTERFEROMETRY

É. Thiébaut [1]

Abstract. Image reconstruction from interferometric data is an inverse problem. Owing to the sparse spatial frequency coverage of the data and to missing Fourier phase information, one has to take into account not only the data but also prior constraints. Image reconstruction then amounts to minimizing a joint criterion which is the sum of a likelihood term to enforce fidelity to the data and a regularization term to impose the priors. To implement strict constraints such as normalization and non-negativity, the minimization is performed on a feasible set. When the complex visibilities are available, image reconstruction is relatively easy as the joint criterion is convex and finding the solution is similar to a deconvolution problem. In optical interferometry, only the power-spectrum and the bispectrum can be measured and the joint criterion is highly multi-modal. The success of an image reconstruction algorithm then depends on the choice of the priors and on the ability of the optimization strategy to find a good solution among all the local minima.

The best angular resolution of a telescope is given by the diffraction limit λ/D (with D the diameter of the primary mirror and λ the wavelength). For an astronomical interferometer, this limit is λ/B (with B the separation of the telescopes projected in a plane perpendicular to the line of sight). In the optical, the largest telescopes have a diameter $D \approx 10\,\mathrm{m}$; thus, with baselines up to $B \approx 600\,\mathrm{m}$, astronomical interferometers resolve much smaller angular scales, below the milliarcsecond in the H band ($1.65\,\mu\mathrm{m}$). This unrivaled resolution has however a cost: an interferometer measures only a single spatial frequency per baseline, while a monolithic telescope harvests all spatial frequencies (up to its diffraction limits) in a single exposure. The data collected by an interferometer are thus very sparse and image reconstruction is a mandatory tool to build an image in spite of the voids in the spatial frequency coverage.

[1] Centre de Recherche Astronomique de Lyon, Université Claude Bernard Lyon I, École Normale Supérieure de Lyon, France; e-mail: `eric.thiebaut@univ-lyon1.fr`

DOI: 10.1051/eas/1359009

Inverse problem approach is a very powerful tool for extracting meaningful information from available data. In particular, it is the method of choice for image reconstruction from interferometric observables. A power of the inverse approach is to relax the constraint that the model of the observables be invertible and thus let us exploit a realistic model. To benefit from this potential, the data model has to be wisely written knowing the instrument and making relevant approximations. The direct model of the interferometric observable is developed in the first sections of this paper. From the instantaneous output of an interferometer (Sect. 1), time averaging (Sect. 2) yields the expression of the complex visibilities integrated during an exposure. In the most simple case, that is when complex visibilities are directly measurable, image reconstruction amounts to solving a deconvolution problem (Sect. 3). In optical interferometry, atmospheric turbulence introduces unknown random optical path perturbations which prevent to directly measure complex visibilities and imposes to integrate observables such as the powerspectrum and the bispectrum which are insensitive to such perturbations (Sect. 4).

Owing to the sparsity of the interferometric data and to the missing of part of the Fourier phases, prior information must be taken into account to solve the image reconstruction problem in a stable and robust way. Without loss of generality, image reconstruction can be stated as an optimization problem over a feasible set (Sect. 5). The penalty to minimize is the sum of a likelihood term (Sect. 6) which enforces fidelity to the measurements and a regularization term (Sect. 7) which favors the priors. Finally, it remains to design an optimization algorithm to effectively solve the image reconstruction problem (Sect. 8).

1 Instantaneous output of an interferometer

In its simplest form, a stellar interferometer (see Fig. 1) consists in two telescopes (or antennae for an array of radio-telescopes) pointing at the astronomical target and coherently recombined. By varying the optical path delay between the two arms of the interferometer, one observes interference fringes. The contrast of the fringes and their phase are the amplitude and phase of the so-called *complex visibility* which is related to the observed object by:

$$V_{j_1,j_2}(\lambda,t) = g^*_{j_1}(\lambda,t)\, g_{j_2}(\lambda,t)\, \widehat{I}_\lambda(\boldsymbol{b}_{j_1,j_2}(t)/\lambda) \tag{1.1}$$

with j_1 and j_2 the indexes of the interfering telescopes, λ the wavelength, t the time, $g_j(\lambda,t)$ the instantaneous complex amplitude transmission for the jth telescope, $g^*_j(\lambda,t)$ its complex conjugate, $\widehat{I}_\lambda(\boldsymbol{\nu})$ the angular Fourier transform of the specific brightness distribution $I_\lambda(\boldsymbol{\theta})$ of the observed object in angular direction $\boldsymbol{\theta}$, and $\boldsymbol{b}_{j_1,j_2}(t)$ the projected *baseline*:

$$\boldsymbol{b}_{j_1,j_2}(t) = \boldsymbol{r}_{j_2}(t) - \boldsymbol{r}_{j_1}(t)$$

where $\boldsymbol{r}_j(t)$ is the position of the jth telescope projected on a plane perpendicular to the line of sight. The amplitude of the complex transmission $g_j(\lambda,t)$ accounts for

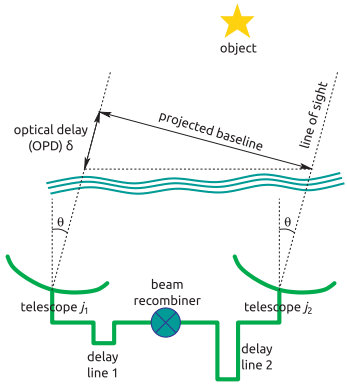


Fig. 1. Interferometer.

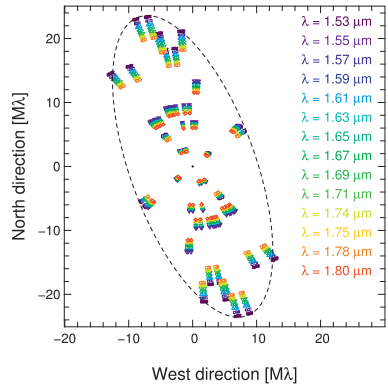


Fig. 2. (u, v) coverage with IOTA 3-telescope interferometer in the H band (from: Lacour *et al.* 2008). The spatial frequencies $\boldsymbol{\nu}$ are given as the projected baselines $\boldsymbol{b}$ in mega-wavelength units (symbol Mλ) that is $10^6 \times \boldsymbol{b}/\lambda = 10^6 \times \boldsymbol{\nu}$.

the efficiency of the transfer of the light from the jth telescope to the recombiner, the phase of $g_j(\lambda, t)$ accounts for the optical delay along this travel.

Equation (1.1) shows that a stellar interferometer samples the Fourier transform of the brightness distribution $\widehat{I}_\lambda(\boldsymbol{\nu})$ at the spatial frequency:

$$\boldsymbol{\nu}_{j_1,j_2}(\lambda, t) = \boldsymbol{b}_{j_1,j_2}(t)/\lambda = \big(\boldsymbol{r}_{j_2}(t) - \boldsymbol{r}_{j_1}(t)\big)/\lambda\,.$$

A single exposure yields one measurement of $\widehat{I}_\lambda(\boldsymbol{\nu})$ per pair of recombined telescopes per spectral channel. For N_{tel} telescopes in a non-redundant configuration, there is a maximum of $N_{\mathsf{tel}}\,(N_{\mathsf{tel}} - 1)/2$ different baselines. Thanks to Earth rotation, the sampling of the spatial frequencies – the so-called (u, v) plane – by a given configuration of telescopes varies with the time, this is called *super-synthesis*. The sampled frequencies also depend on the wavelength: the longer the wavelength the shorter the sampled frequency. Because of the limited number of telescopes for current optical interferometers ($2 \leq N_{\mathsf{tel}} \leq 6$), even by combining all these possible measurements, the sampling of the (u, v) plane remains very sparse and uneven (*cf.* Fig. 2).

2 Averaging during exposures

The previous equations consider the instantaneous and monochromatic case: they are given for continuously varying time t, wavelength λ and projected telescope positions $\boldsymbol{r}_j(t)$. In practice, a finite number of measurements are obtained for given exposure times, spectral channels and telescope combinations. In the sequel, we use the index m to label the available data: for the m-th measurement (possibly complex), the exposure time is denoted t_m, λ_m is the effective wavelength of the

spectral channel and there are up to three interfering telescopes numbered $j_{m,1}$, $j_{m,2}$ and $j_{m,3}$. Of course different measurements, say m and m', may have the same observing times ($t_{m'} = t_m$) or may share the same telescopes and the same spectral channel.

Because of the finite exposure time and spectral bandwidth, the instantaneous and monochromatic complex visibility in Equation (1.1) must be averaged to give the *effective* complex visibility:

$$V_m = \langle V_{j_{m,1},j_{m,2}}(\lambda,t)\rangle_m = \left\langle g^*_{j_{m,1}}(\lambda,t)\, g_{j_{m,2}}(\lambda,t)\, \widehat{I}_\lambda\big(\boldsymbol{b}_{j_{m,1},j_{m,2}}(t)/\lambda\big)\right\rangle_m \tag{2.1}$$

where $\langle\ldots\rangle_m$ denotes averaging (or integrating) during the exposure and inside the spectral channel corresponding to the m-th measurement:

$$\langle f(\lambda,t)\rangle_m \stackrel{\text{def}}{=} \frac{1}{\Delta t_m}\int_{t_m-\Delta t_m/2}^{t_m+\Delta t_m/2} \frac{1}{\Delta\lambda_m}\int s_m(\lambda)\, f(\lambda,t)\,\mathrm{d}\lambda\,\mathrm{d}t \tag{2.2}$$

with Δt_m the duration of the exposure, $s_m(\lambda)$ the transmission of the spectral channel, and $\Delta\lambda_m \stackrel{\text{def}}{=} \int s_m(\lambda)\,\mathrm{d}\lambda$ the effective spectral bandwidth.

To measure interference patterns, the effective bandwidth $\Delta\lambda_m$ must be such that the complex amplitude transmissions are approximately constant in each spectral channel and the exposure duration Δt_m must be short enough to neglect the temporal variation of the baselines. Under these conditions, the double integral which results from combining Equations (2.1) and (2.2) becomes separable:

$$\begin{aligned} V_m &= \frac{1}{\Delta t_m}\int_{t_m-\Delta t_m/2}^{t_m+\Delta t_m/2} \frac{1}{\Delta\lambda_m}\int s_m(\lambda)\, g^*_{j_{m,1}}(\lambda,t)\, g_{j_{m,2}}(\lambda,t) \\ &\qquad\qquad \times \widehat{I}_\lambda\big(\boldsymbol{b}_{j_{m,1},j_{m,2}}(t)/\lambda\big)\,\mathrm{d}\lambda\,\mathrm{d}t \\ &\approx \frac{1}{\Delta t_m}\int_{t_m-\Delta t_m/2}^{t_m+\Delta t_m/2} g^*_{j_{m,1}}(\lambda_m,t)\, g_{j_{m,2}}(\lambda_m,t)\,\mathrm{d}t \\ &\quad\times \frac{1}{\Delta\lambda_m}\int s_m(\lambda)\,\widehat{I}_\lambda\big(\boldsymbol{b}_{j_{m,1},j_{m,2}}(t_m)/\lambda_m\big)\,\mathrm{d}\lambda \\ &= \widehat{h}_m\,\widehat{I}_m(\boldsymbol{\nu}_m) \end{aligned} \tag{2.3}$$

with:

$$\widehat{h}_m \stackrel{\text{def}}{=} \frac{1}{\Delta t_m}\int_{t_m-\Delta t_m/2}^{t_m+\Delta t_m/2} g^*_{j_{m,1}}(\lambda_m,t)\, g_{j_{m,2}}(\lambda_m,t)\,\mathrm{d}t\,, \tag{2.4}$$

$$\widehat{I}_m(\boldsymbol{\nu}) \stackrel{\text{def}}{=} \frac{1}{\Delta\lambda_m}\int s_m(\lambda)\,\widehat{I}_\lambda(\boldsymbol{\nu})\,\mathrm{d}\lambda \tag{2.5}$$

$$\approx \widehat{I}_{\lambda_m}(\boldsymbol{\nu}) \tag{2.6}$$

$$\boldsymbol{\nu}_m \stackrel{\text{def}}{=} \boldsymbol{b}_m/\lambda_m\,, \tag{2.7}$$

$$\boldsymbol{b}_m \stackrel{\text{def}}{=} \boldsymbol{r}_{j_{m,2}}(t_m) - \boldsymbol{r}_{j_{m,1}}(t_m), \tag{2.8}$$

respectively the effective interferometric transfer function, the Fourier transform of the specific brightness distribution integrated in the spectral channel, the spatial frequency and the effective baseline for the m-th observed complex visibility. The approximation in Equation (2.6) applies for spectral bandwidths narrower than the spectral features of the specific brightness distribution. To simplify the notations but without loss of generality, we will assume that this is the case in what follows.

When the complex amplitude transmissions are stable during and exposure, the effective interferometric transfer function can be further simplified:

$$\widehat{h}_m \approx g^*_{j_{m,1}}\, g_{j_{m,2}} \tag{2.9}$$

where:

$$g_{j_{m,i}} \stackrel{\text{def}}{=} \frac{1}{\Delta t_m} \int_{t_m-\Delta t_m/2}^{t_m+\Delta t_m/2} g_{j_{m,i}}(\lambda_m, t)\,\mathrm{d}t \approx g_{j_{m,i}}(\lambda_m, t_m). \tag{2.10}$$

Thus, for monochromatic observations with an interferometer composed of N_{tel} telescopes and under stable observing conditions, the effective transfer function only depends on $N_{\mathsf{tel}} - 1$ complex numbers (one complex amplitude transmission can be chosen arbitrarily) per exposure while there are $N_{\mathsf{tel}}\,(N_{\mathsf{tel}} - 1)/2$ measured complex visibilities. Depending on the number of interfering telescopes, the amount of information needed to estimate the transfer function may be much smaller than the amount of measurements. This open the possibility to perform *self-calibration* (Cornwell & Wilkinson 1981; Schwab 1980).

3 Easy case: image reconstruction $\sim$ deconvolution

Considering only complex visibilities for a given effective wavelength λ, we can combine them to form the distribution:

$$\widehat{d}_\lambda(\boldsymbol{\nu}) \stackrel{\text{def}}{=} \sum_{m\in\mathbb{S}_\lambda} V_m\,\delta(\boldsymbol{\nu} - \boldsymbol{\nu}_m) \tag{3.1}$$

with $\mathbb{S}_\lambda = \{m\colon \lambda_m = \lambda\}$ and $\delta(\cdot)$ the Dirac's distribution. Using the definition of the observed complex visibilities V_m in Equation (2.3) and the approximation in Equation (2.6), $\widehat{d}_\lambda(\boldsymbol{\nu})$ can be expanded as follows:

$$\begin{aligned}
\widehat{d}_\lambda(\boldsymbol{\nu}) &= \sum_{m\in\mathbb{S}_\lambda} \widehat{h}_m\,\widehat{I}_{\lambda_m}(\boldsymbol{\nu}_m)\,\delta(\boldsymbol{\nu} - \boldsymbol{\nu}_m) \\
&= \widehat{I}_\lambda(\boldsymbol{\nu}) \sum_{m\in\mathbb{S}_\lambda} \widehat{h}_m\,\delta(\boldsymbol{\nu} - \boldsymbol{\nu}_m) \\
&= \widehat{I}_\lambda(\boldsymbol{\nu})\,\widehat{h}_\lambda(\boldsymbol{\nu}),
\end{aligned} \tag{3.2}$$

with:

$$\widehat{h}_\lambda(\boldsymbol{\nu}) = \sum_{m\in\mathbb{S}_\lambda} \widehat{h}_m\,\delta(\boldsymbol{\nu} - \boldsymbol{\nu}_m). \tag{3.3}$$

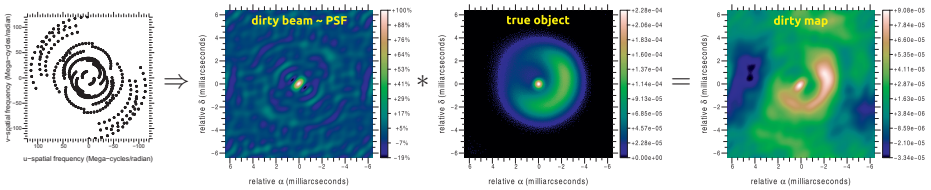


Fig. 3. From *left* to *right*: spatial frequency sampling, dirty beam, object brightness distribution and dirty image.

Taking the inverse Fourier transform of $\widehat{d_\lambda}(\boldsymbol{\nu})$, we obtain a 2D angular distribution called the *dirty image*:

$$\begin{aligned} d_\lambda(\boldsymbol{\theta}) &\stackrel{\text{def}}{=} \iint \widehat{d_\lambda}(\boldsymbol{\nu})\, \mathrm{e}^{+\mathrm{i}\,2\,\pi\,\langle\boldsymbol{\theta},\boldsymbol{\nu}\rangle}\, \mathrm{d}^2\boldsymbol{\nu} \\ &= \iint \widehat{h_\lambda}(\boldsymbol{\nu})\, \widehat{I_\lambda}(\boldsymbol{\nu})\, \mathrm{e}^{+\mathrm{i}\,2\,\pi\,\langle\boldsymbol{\theta},\boldsymbol{\nu}\rangle}\, \mathrm{d}^2\boldsymbol{\nu} \\ &= (h_\lambda * I_\lambda)\,(\boldsymbol{\theta}) \end{aligned} \tag{3.4}$$

where $\langle\boldsymbol{\theta},\boldsymbol{\nu}\rangle$ is the 2D scalar product of $\boldsymbol{\theta}$ by $\boldsymbol{\nu}$ and the symbol $*$ denotes the convolution product of the brightness distribution $I_\lambda(\boldsymbol{\theta})$ by the so-called *dirty beam*:

$$\begin{aligned} h_\lambda(\boldsymbol{\theta}) &\stackrel{\text{def}}{=} \iint \widehat{h_\lambda}(\boldsymbol{\nu})\, \mathrm{e}^{+\mathrm{i}\,2\,\pi\,\langle\boldsymbol{\theta},\boldsymbol{\nu}\rangle}\, \mathrm{d}^2\boldsymbol{\nu} \\ &= \sum_{m\in\mathbb{S}_\lambda} \widehat{h}_m\, \mathrm{e}^{+\mathrm{i}\,2\,\pi\,\langle\boldsymbol{\theta},\boldsymbol{\nu}_m\rangle}. \end{aligned} \tag{3.5}$$

In words, the dirty image $d_\lambda(\boldsymbol{\theta})$, synthesized from the observed complex visibilities, is simply the convolution of the specific brightness distribution $I_\lambda(\boldsymbol{\theta})$ by the dirty beam $h_\lambda(\boldsymbol{\theta})$. Figure 3 shows, for given (u,v)-coverage and observed object, the resulting dirty beam and dirty image. The dirty beam $h_\lambda(\boldsymbol{\theta})$ is the analogous of the point spread function (PSF) in conventional imaging; it is however not a probability density function, in particular, when super-synthesis is exploited, $h_\lambda(\boldsymbol{\theta})$ is not a normalized non-negative distribution (*cf.* the negative lobes of the dirty beam in Fig. 3).

To summarize, when the observables are the complex visibilities V_m and the transfer function $\widehat{h}_m$ properly calibrated, Equation (3.4) shows that image reconstruction amounts to a deconvolution problem (Cornwell 1995). There are however many unmeasured values – the voids in the coverage of the (u,v)-plane – thus the problem is, at least, ill-posed and other constraints than the data are required to warrant the uniqueness and the stability of the solution. The principles of image reconstruction developed in the remaining sections of this paper can be applied to solve this inverse problem.

So far, no considerations have been made regarding the quality of the measurements which may be very variable. In practice, *regridding* techniques (Sramek & Schwab 1989; Thompson & Bracewell 1974) are implemented to synthesize a dirty image with proper weighting of the data according to their confidence levels. By inverse Fourier transforming the expression of $\widehat{d}_\lambda(\boldsymbol{\nu})$ given by Equation (3.1), the dirty image can be directly synthesized from the complex visibilities data V_m:

$$d_\lambda(\boldsymbol{\theta}) = \sum_{m\in\mathbb{S}_\lambda} V_m \, \mathrm{e}^{+\mathrm{i}\,2\,\pi\,\langle\boldsymbol{\theta},\boldsymbol{\nu}_m\rangle}. \tag{3.6}$$

To take into account the variable quality of the measurements, one can use statistical weights and synthesize the dirty image as:

$$d_\lambda(\boldsymbol{\theta}) = \sum_{m\in\mathbb{S}_\lambda} w_m \, V_m \, \mathrm{e}^{+\mathrm{i}\,2\,\pi\,\langle\boldsymbol{\theta},\boldsymbol{\nu}_m\rangle}. \tag{3.7}$$

where the weights w_m are computed according to the variance of the noise. The corresponding dirty beam then writes:

$$h_\lambda(\boldsymbol{\theta}) = \sum_{m\in\mathbb{S}_\lambda} w_m \, \widehat{h}_m \, \mathrm{e}^{+\mathrm{i}\,2\,\pi\,\langle\boldsymbol{\theta},\boldsymbol{\nu}_m\rangle}. \tag{3.8}$$

The somewhat idealized case considered here is relevant for radio-astronomy for which the complex amplitude transmissions $g_j(\lambda,t)$ are stable during an exposure and can be calibrated. We will see next (Sect. 4) that, due to the atmospheric turbulence, these assumptions cannot be made in the optical where the situation is much more involved. In terms of complexity, an intermediate situation arises when the transfer function $\widehat{h}_m$ cannot be calibrated. *Self calibration* methods (Cornwell & Wilkinson 1981) have been developed to cope with this case and consist in jointly recovering the complex amplitude transmissions $g_j(\lambda_m,t_m)$, see Equation (2.10), and the image of the object from uncalibrated complex visibilities. Self calibration is the analogous of *blind deconvolution* in conventional imaging (Campisi & Egiazarian 2007).

4 The effects of turbulence

The *atmospheric turbulence* induces random variations of the refractive index along the path traveled by the light (Roddier 1981). These fluctuations affect the modulus and the phase of the complex transmissions $g_j(\lambda,t)$ during an exposure. For instance, for an instrument like AMBER (Petrov *et al.* 2007), the modulus $|g_j(\lambda,t)|$ fluctuates due to the boiling of the speckle pattern in the focal plane of the telescopes which changes the amount of coherent light injected in the optical fibers which feed the instrument and perform the spatial filtering. The turbulence also induces random delays in the optical path which affect the phase $\phi_j(\lambda,t)$ of $g_j(\lambda,t)$. The variations of the modulus of the complex transmissions can be estimated or calibrated, *e.g.* by the photometric channels of AMBER. But it is much more

difficult to estimate the phase errors. The situation is about to improve with the development of recombinators with phase reference (Delplancke *et al.* 2003) but, for now, there are no reliable means to estimate the phase $\phi_j(\lambda, t)$. This has a profound impact on the kind of measurements provided by an optical interferometer.

Because of the fluctuations of the complex transmissions $g_j(\lambda, t)$ during an exposure, the approximation in Equation (2.9) no longer applies: the effective transfer function $\widehat{h}_m$ is given by Equation (2.4). Then, if the fluctuations of the phase $\phi_j(\lambda, t)$ of $g_j(\lambda, t)$ are too important during the exposure, the integrand in Equation (2.4) becomes randomly distributed around zero and the averaging during the exposure yields:

$$\widehat{h}_m \approx 0. \tag{4.1}$$

This means that the complex visibilities cannot be measured when the unknown random phase fluctuations are too large during an exposure. This is the case at optical wavelengths. Even if the phase fluctuations are not so important, the effective transfer function cannot be described by a small number of complex transmissions. This forbids the use of self-calibration to guess the effective transfer function: in order to directly exploit the mean complex visibilities, $\widehat{h}_m$ must be calibrated simultaneously to the observations. For these reasons, astronomers have to integrate observables which are *insensitive to phase delay errors*.

Using very short exposure durations, typically $\sim$1 ms, compared to the evolution time of the atmospheric effects, the instantaneous complex visibilities can be measured but with unknown phase terms. The interferometric observables are then computed by forming, from simultaneously observed complex visibilities, quantities which are insensitive to the phase of the complex transmissions. These observables are the *powerspectrum*:

$$\begin{aligned} P_m &\stackrel{\text{def}}{=} \langle |V_{j_{m,1},j_{m,2}}(\lambda,t)|^2 \rangle_m \\ &\approx \underbrace{\langle |g_{j_{m,1}}(\lambda,t)|^2 \, |g_{j_{m,2}}(\lambda,t)|^2 \rangle_m}_{>0} \, |\widehat{I}_{\lambda_m}(\boldsymbol{\nu}_m)|^2 \end{aligned} \tag{4.2}$$

and the *bispectrum*:

$$\begin{aligned} B_m &\stackrel{\text{def}}{=} \left\langle V_{j_{m,1},j_{m,2}}(\lambda,t)\, V_{j_{m,2},j_{m,3}}(\lambda,t)\, V_{j_{m,3},j_{m,1}}(\lambda,t) \right\rangle_m \\ &\approx \underbrace{\langle |g_{j_{m,1}}(\lambda,t)|^2 \, |g_{j_{m,2}}(\lambda,t)|^2 \, |g_{j_{m,3}}(\lambda,t)|^2 \rangle_m}_{>0} \, \widehat{I}^{(3)}_{\lambda_m}(\boldsymbol{\nu}_m, \boldsymbol{\nu}'_m) \end{aligned} \tag{4.3}$$

where:

$$\begin{aligned} \boldsymbol{\nu}_m &= (\boldsymbol{r}_{j_{m,2}}(t_m) - \boldsymbol{r}_{j_{m,1}}(t_m))/\lambda_m \,, \\ \boldsymbol{\nu}'_m &= (\boldsymbol{r}_{j_{m,3}}(t_m) - \boldsymbol{r}_{j_{m,2}}(t_m))/\lambda_m \,, \end{aligned}$$

and:

$$\widehat{I}^{(3)}_{\lambda}(\boldsymbol{\nu}, \boldsymbol{\nu}') \stackrel{\text{def}}{=} \widehat{I}_{\lambda}(\boldsymbol{\nu})\, \widehat{I}_{\lambda}(\boldsymbol{\nu}')\, \widehat{I}^{*}_{\lambda}(\boldsymbol{\nu} + \boldsymbol{\nu}') \tag{4.4}$$

is the bispectrum of the brightness distribution of the object. To be able to measure the powerspectrum, given by Equation (4.2), two telescopes ($j_{m,1}$ and $j_{m,2}$) have to be coherently recombined; while, to measure the bispectrum, given by Equation (4.3), three telescopes ($j_{m,1}$, $j_{m,2}$ and $j_{m,3}$) have to be coherently recombined.

Note that, being non-linear quantities, the empirical powerspectrum and bispectrum have bias terms which are not shown here to simplify the equations. Dainty & Greenaway (1979) and Wirnitzer (1985) give the expressions of unbiased estimators for the powerspectrum and for the bispectrum respectively at low light levels (photon counting mode).

5 Inverse problem approach for image reconstruction

Given the interferometric observables, we want to recover an image, that is an approximation of the object specific brightness distribution at a given wavelength. Before going into the details of a method to tackle this problem, we can anticipate a number of issues and make some preliminary remarks. (i) Due to voids in the spatial frequency coverage, we are dealing with very *sparse data* (with typically a few tens of baselines, see Fig. 2). (ii) Avoiding the turbulence effects implies to use *non-linear data* (powerspectrum or bispectrum) which is more difficult to fit than, say, the complex visibilities. (iii) Compared to the $N_{\mathsf{tel}}\,(N_{\mathsf{tel}}-1)/2$ sampled frequencies per exposure, the powerspectrum provides no Fourier phase information while the bispectrum only provides $(N_{\mathsf{tel}}-1)\,(N_{\mathsf{tel}}-2)/2$ *phase closures*, so there are missing phase data (with only 3 telescopes, 2/3rd of the phases are missing). (iv) There may be calibration problems which means that there are additional unknown factors in the data.

For the sake of simplicity, we will consider in the following the case of monochromatic image reconstruction (at a given wavelength λ) and assume that we are working with debiased and calibrated data. That is, all the effective transfer functions are assumed to be equal to unity and the main problem is to deal with the sparsity of the data, the missing Fourier phase information and the non-linearity of the estimators. The possible types of measurements that may be available are:

- complex visibilities: $V_m \approx \widehat{I}_\lambda(\boldsymbol{\nu}_m)$;
- powerspectrum data: $P_m \approx \left|\widehat{I}_\lambda(\boldsymbol{\nu}_m)\right|^2$;
- bispectrum data: $B_m \approx \widehat{I}^{(3)}_\lambda(\boldsymbol{\nu}_m, \boldsymbol{\nu}'_m)$;

where the $\approx$ symbol is used because of omitted error terms.

As all measured quantities are related to the Fourier transform of the specific brightness distribution, we first need a model of the complex visibilities. This is the subject of Section 5.1.

On the one hand, due to the noise, exactly fitting the data is pointless and we expect some discrepancy between actual data and their model given the sought image. On the other hand, owing to the amount of missing information (sparse

sampling of the spatial frequencies and, perhaps, only partial Fourier phase information), the data alone cannot uniquely define an image: additional *priors* are required. Image reconstruction is then a compromise between fidelity to the data and to the priors; the different formulations of this inverse problem are introduced in Section 5.2.

We will see that solving the image restoration problem amounts to minimizing the sum of two terms: a likelihood term to enforce data fidelity and a regularization term to promote agreement with the priors. Bayesian inference (Sect. 5.3) can be invoked to formally derive these terms. Practical derivation of the likelihood term is discussed in Section 6. The regularization is developed in Section 7. At least because of the necessary flexibility of the regularization[2], choosing the regularization and its tuning parameters is needed. This is briefly discussed in Section 7.3.

Finally it remains to effectively solve the problem, that is to find the best image parameters which minimize the given penalized likelihood. Numerical optimization is introduced in Section 8.

5.1 Image and complex visibilities models

Because of the noise and of the limited number of measurements, it is hopeless to aim at recovering the specific brightness distribution $I_\lambda(\boldsymbol{\theta})$ of the observed object exactly. Instead, a realistic objective is to seek for a good estimate of an approximation $i(\boldsymbol{\theta})$ of $I_\lambda(\boldsymbol{\theta})$ which depends on a finite number of parameters. To that end, the specific brightness distribution in angular direction $\boldsymbol{\theta}$ can be approximated by:

$$i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum\nolimits_n x_n \, b_n(\boldsymbol{\theta}) \approx I_\lambda(\boldsymbol{\theta}) \tag{5.1}$$

with $\{b_n(\boldsymbol{\theta})\colon \mathbb{R}^2 \mapsto \mathbb{R}\}_{n=1}^N$ a basis of functions and $\boldsymbol{x} \in \mathbb{R}^N$ the *image parameters*. This general parametrization accounts, for instance, for a pixelized image, for a wavelet decomposition, etc.. For image reconstruction, it may be the most convenient to use a shift-invariant basis of functions defined by:

$$b_n(\boldsymbol{\theta}) = b(\boldsymbol{\theta} - \boldsymbol{\theta}_n) \tag{5.2}$$

where $b(\boldsymbol{\theta})\colon \mathbb{R}^2 \mapsto \mathbb{R}$ is a single basis function and $\mathbb{G} = \{\boldsymbol{\theta}_n \in \mathbb{R}^2 \mid n = 1, \ \ldots, N\}$ is a grid of evenly spaced positions. If $b(\boldsymbol{\theta})$ is an interpolation function (Thévenaz *et al.* 2000), then the image parameters sample the brightness distribution:

$$x_n = i(\boldsymbol{\theta}_n) \approx I_\lambda(\boldsymbol{\theta}_n).$$

The advantage of approximating the specific brightness distribution by the linear expansion $i(\boldsymbol{\theta})$ given in Equation (5.1) is that its exact Fourier transform is also linear with respect to the image parameters $\boldsymbol{x}$:

$$\widehat{i}(\boldsymbol{\nu}) = \sum\nolimits_n x_n \, \widehat{b}_n(\boldsymbol{\nu}) \approx \widehat{I}_\lambda(\boldsymbol{\nu}), \tag{5.3}$$

[2]Such flexibility is required because the object of interest is unknown.

where the hat $\widehat{\ }$ denotes the Fourier transformed distribution and $\boldsymbol{\nu}$ is the spatial frequency conjugate of the angular position $\boldsymbol{\theta}$. For any sampled spatial frequency $\boldsymbol{\nu}_m$ the model complex visibility thus writes:

$$y_m \stackrel{\text{def}}{=} \widehat{i}(\boldsymbol{\nu}_m) = \sum\nolimits_n \widehat{b}_n(\boldsymbol{\nu}_m)\, x_n = \sum\nolimits_n H_{m,n}\, x_n \approx \widehat{I}_\lambda(\boldsymbol{\nu}_m)\,,$$

with $H_{m,n} = \widehat{b}_n(\boldsymbol{\nu}_m)$. In matrix notation:

$$\boldsymbol{y} = \mathbf{H} \cdot \boldsymbol{x}, \tag{5.4}$$

where $\boldsymbol{y} \in \mathbb{C}^M$ collects the model complex visibilities at all sampled frequencies and $\mathbf{H} \in \mathbb{C}^{M\times N}$ is a sub-sampled Fourier transform operator. The memory requirement to store the coefficients of the operator $\mathbf{H}$ and the computer time needed to apply $\mathbf{H}$ (or its adjoint) both scale as $O(M \times N)$. Fast approximations of $\mathbf{H}$ based on the FFT can be used (Fessler & Sutton 2003; Potts *et al.* 2001) when $M \times N$ is too large. To use these fast approximations, the image model must be defined with shift-invariant basis functions, see Equation (5.2), on an evenly spaced grid $\mathbb{G}$.

5.2 Inverse problem formulations

As stated before, image reconstruction is a compromise between various constraints resulting from the measurements and from prior knowledge. The first of these constraints is that the image must be *compatible with the available data*. This is asserted by comparing the measurements with their model given the image parameters $\boldsymbol{x}$. To keep the maximum flexibility and since the model of all the measured quantities depend on the model complex visibilities $\boldsymbol{y} = \mathbf{H} \cdot \boldsymbol{x}$, we postulate that, to be compatible with the measurements, the image parameters must satisfy the following criterion:

$$f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x}) \leq \eta \tag{5.5}$$

where $f_{\mathsf{data}}(\boldsymbol{y})$: $\mathbb{C}^M \mapsto \mathbb{R}_+$ is a measure of the distance between the model complex visibilities $\boldsymbol{y} = \mathbf{H} \cdot \boldsymbol{x}$ and the actual data. The threshold η is chosen to set how close to the data should be the model. As $f_{\mathsf{data}}(\boldsymbol{y})$ is a distance, the smaller η the closer the model to the data. However taking $\eta = 0$ would mean that the model should exactly match the data and thus *fit the noise* which is undesirable. So we always want $\eta > 0$, depending on the exact definition of $f_{\mathsf{data}}(\boldsymbol{y})$, the value of the threshold may vary with, *e.g.*, the noise level and the number of measurements.

The level of agreement with the prior knowledge can be expressed in the same spirit by specifying a distance $f_{\mathsf{prior}}(\boldsymbol{x})$ and requiring that this distance be as small as possible providing that the model remains compatible with the data. Formally, this writes:

$$\boxed{\boldsymbol{x}^\star = \underset{\boldsymbol{x}\in\mathbb{X}}{\arg\min}\, f_{\mathsf{prior}}(\boldsymbol{x}) \quad \text{s.t.} \quad f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x}) \leq \eta,} \tag{5.6}$$

where the *feasible set* $\mathbb{X} \subset \mathbb{R}^N$ is introduced to impose strict constraints such as the non-negativity of the image. For instance, using bilinear interpolation for

the approximation in Equation (5.1), the specific brightness distribution $i(\boldsymbol{\theta})$ is non-negative and normalized if and only if the parameters $\boldsymbol{x}$ are non-negative and their sum is equal to $\xi \overset{\text{def}}{=} \iint i(\boldsymbol{\theta})\,\mathrm{d}^2\boldsymbol{\theta}$:

$$i(\boldsymbol{\theta}) \geq 0 \quad \text{and} \quad \iint i(\boldsymbol{\theta})\,\mathrm{d}^2\boldsymbol{\theta} = \xi \quad \Longleftrightarrow \quad \boldsymbol{x} \in \mathbb{X}$$

with:

$$\mathbb{X} = \{\boldsymbol{x} \in \mathbb{R}^N \mid \boldsymbol{x} \geq 0, \mathbf{1}^\top \cdot \boldsymbol{x} = \xi\}, \tag{5.7}$$

where the inequality $\boldsymbol{x} \geq 0$ is taken componentwise and where $\mathbf{1}$ is the vector of $\mathbb{R}^N$ with all components equal to 1:

$$\begin{aligned} \boldsymbol{x} \geq 0 \quad &\Longleftrightarrow \quad \forall n, x_n \geq 0 \\ \mathbf{1} = (1,\ \ldots, 1)^\top \quad &\Longrightarrow \quad \mathbf{1}^\top \cdot \boldsymbol{x} = \sum\nolimits_n x_n. \end{aligned}$$

The constrained problem (5.6) is usually solved via the Lagrangian (Nocedal & Wright 2006):

$$\mathcal{L}(\boldsymbol{x}; \ell) = f_{\mathsf{prior}}(\boldsymbol{x}) + \ell\, f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x})$$

with $\ell \geq 0$ the Lagrange multiplier for the inequality constraint $f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x}) \leq \eta$. Assuming that $\mathcal{L}(\boldsymbol{x}; \ell)$ has a unique reachable minimum on the feasible set, we can define:

$$\boldsymbol{x}^+_{\mathcal{L}}(\ell) \overset{\text{def}}{=} \underset{\boldsymbol{x} \in \mathbb{X}}{\arg\min}\, \mathcal{L}(\boldsymbol{x}; \ell)\,,$$

and seek for the value $\ell^\star \geq 0$ of the multiplier such that the solution $\boldsymbol{x}^\star = \boldsymbol{x}^+_{\mathcal{L}}(\ell^\star)$ complies with the constraints. Obviously, we want $\ell^\star > 0$ otherwise the data play no role in the determination of the solution. Intuitively, having the solution strictly closer to the data than required, *i.e.* $f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x}^\star) < \eta$, yields a worst value of $f_{\mathsf{prior}}(\boldsymbol{x}^\star)$ than having $f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x}^\star) = \eta$. Thus, unless the a priori solution:

$$\boldsymbol{x}_{\mathsf{prior}} \overset{\text{def}}{=} \underset{\boldsymbol{x} \in \mathbb{X}}{\arg\min}\, f_{\mathsf{prior}}(\boldsymbol{x})$$

is such that $f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x}_{\mathsf{prior}}) \leq \eta$, in which case the solution is $(\boldsymbol{x}^\star, \ell^\star) = (\boldsymbol{x}_{\mathsf{prior}}, 0)$, the solution to the problem (5.6) is given by $\boldsymbol{x}^\star = \boldsymbol{x}^+_{\mathcal{L}}(\ell^\star)$ with $\ell^\star > 0$ such that $f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x}^\star) = \eta$.

Since the solution is obtained for a Lagrange multiplier strictly positive, we can take $\mu = 1/\ell$ and alternatively define the solution to be given by minimizing another penalty function:

$$\boldsymbol{x}^+_f(\mu) = \underset{\boldsymbol{x} \in \mathbb{X}}{\arg\min}\, f(\boldsymbol{x}; \mu) \quad \text{with:} \quad f(\boldsymbol{x}; \mu) = f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x}) + \mu\, f_{\mathsf{prior}}(\boldsymbol{x})\,. \tag{5.8}$$

The solution is then $\boldsymbol{x}^\star = \boldsymbol{x}^+_f(\mu^\star)$ where the optimal weight $\mu^\star > 0$ for the priors is such that $f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x}^\star) = \eta$. The two different formulations are equivalent and yield the same solution of the constrained problem (5.6).

We shall now see how to derive the expression of the *distances* $f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x})$ and $f_{\mathsf{prior}}(\boldsymbol{x})$.

5.3 Bayesian inference

The previous considerations may found strong theoretical justification in a Bayesian framework where probabilities represent any available information. For instance, in a *maximum a posteriori* (MAP) approach, the best image parameters $\boldsymbol{x}_{\mathsf{MAP}}$ are the most likely ones given the data $\boldsymbol{z}$:

$$\boldsymbol{x}_{\mathsf{MAP}} = \arg\max_{\boldsymbol{x}} \Pr(\boldsymbol{x}|\boldsymbol{z}),$$

where $\Pr(\boldsymbol{x}|\boldsymbol{z})$ denotes the probability (or the probability density function) of $\boldsymbol{x}$ given $\boldsymbol{z}$. Note that the data $\boldsymbol{z}$ collects all measurements; in our case, $\boldsymbol{z}$ may include complex visibilities, powerspectra and bispectra. Using Bayes theorem[3], discarding terms which do no depend on $\boldsymbol{x}$ and noting that $-\log(p)$ is a strictly decreasing function of p yields:

$$\begin{aligned}\boldsymbol{x}_{\mathsf{MAP}} &= \arg\max_{\boldsymbol{x}} \frac{\Pr(\boldsymbol{z}|\boldsymbol{x})\ \Pr(\boldsymbol{x})}{\Pr(\boldsymbol{z})} \\ &= \arg\max_{\boldsymbol{x}} \Pr(\boldsymbol{z}|\boldsymbol{x})\ \Pr(\boldsymbol{x}) \\ &= \arg\min_{\boldsymbol{x}} -\log(\Pr(\boldsymbol{z}|\boldsymbol{x})) - \log(\Pr(\boldsymbol{x})).\end{aligned}$$

Hence:

$$\boldsymbol{x}_{\mathsf{MAP}} = \arg\min_{\boldsymbol{x}} f_{\boldsymbol{z}|\boldsymbol{x}}(\boldsymbol{x}) + f_{\boldsymbol{x}}(\boldsymbol{x}), \tag{5.9}$$

with:

$$f_{\boldsymbol{z}|\boldsymbol{x}}(\boldsymbol{x}) = -\log(\Pr(\boldsymbol{z}|\boldsymbol{x})) \tag{5.10}$$

$$f_{\boldsymbol{x}}(\boldsymbol{x}) = -\log(\Pr(\boldsymbol{x})). \tag{5.11}$$

In words, the MAP solution $\boldsymbol{x}_{\mathsf{MAP}}$ is a compromise between maximizing the likelihood of the data $\boldsymbol{z}$ given the model parameters $\boldsymbol{x}$ and maximizing the prior probability of the model. Said otherwise, the compromise is between fitting the data, *i.e.* minimize $f_{\boldsymbol{z}|\boldsymbol{x}}(\boldsymbol{x})$, and agreement with prior knowledge, *i.e.* minimize $f_{\boldsymbol{x}}(\boldsymbol{x})$.

Finally, the solution $\boldsymbol{x}_f^+(\mu)$ of the problem (5.8) is also the MAP solution $\boldsymbol{x}_{\mathsf{MAP}}$ if we take:

$$f_{\mathsf{data}}(\mathbf{H}\cdot\boldsymbol{x}) = c_0' + c_1\, f_{\boldsymbol{z}|\boldsymbol{x}}(\boldsymbol{x}) \tag{5.12}$$

$$\mu\, f_{\mathsf{prior}}(\boldsymbol{x}) = c_0'' + c_1\, f_{\boldsymbol{x}}(\boldsymbol{x}) \tag{5.13}$$

with c_0', c_0'' and $c_1 > 0$ any suitable real constants. From this close relationship, we deduce a possible way to define the penalty functions $f_{\mathsf{data}}(\mathbf{H}\cdot\boldsymbol{x})$ and $f_{\mathsf{prior}}(\boldsymbol{x})$. This is the subject of the next two sections.

[3]Bayes theorem states that the joint probability of A and B writes:

$$\Pr(A,B) = \Pr(A)\ \Pr(B|A) = \Pr(B)\ \Pr(A|B).$$

6 Likelihood of the data

Ideally, the likelihood should be strictly based on the noise statistics of the data:

$$f_{\mathsf{data}}(\mathbf{H}\cdot\boldsymbol{x}) \stackrel{\text{def}}{=} c_0' - c_1 \log(\Pr(\boldsymbol{z}|\mathbf{H}\cdot\boldsymbol{x})).$$

If the measurements have Gaussian statistics, then for $c_1 = 2$ and for an appropriate choice of c_0', the likelihood term is a so-called χ^2 given by:

$$f_{\mathsf{data}}(\mathbf{H}\cdot\boldsymbol{x}) = [\boldsymbol{z} - \widetilde{\boldsymbol{z}}(\mathbf{H}\cdot\boldsymbol{x})]^\top \cdot \mathbf{W} \cdot [\boldsymbol{z} - \widetilde{\boldsymbol{z}}(\mathbf{H}\cdot\boldsymbol{x})],$$

where $\widetilde{\boldsymbol{z}}(\mathbf{H}\cdot\boldsymbol{x})$ is the model of the measurements $\boldsymbol{z}$ and $\mathbf{W}$ is a weighting matrix equal to the inverse of the covariance of the measurements: $\mathbf{W} = \mathrm{Cov}\{\boldsymbol{z}\}^{-1}$. Our notation accounts for the fact that the model of the measurements only depends on the model complex visibilities $\mathbf{H}\cdot\boldsymbol{x}$ and assumes that all measurements are real valued (any complex valued data has to be considered as a pair of reals).

A first difficulty is that the statistics of real interferometric measurements is not well known and may not be Gaussian at all. For instance, Figure 4 shows the empirical distribution of bispectrum data. At low signal to noise ratio (SNR), the distribution may be well approximated by a Gaussian distribution; while, at high SNR, the *banana shaped* distribution of the data suggests that the amplitude and phase of the complex bispectrum may be independent variables. Figure 5 shows that this banana shaped distribution can only be grossly approximated by a Gaussian with respect to the real and imaginary parts of the bispectrum data.

A second difficulty is that not all statistical information is provided with the data. Generally, only estimates of the error bar (standard deviation) of each measurement is available. In particular no information is stored about the correlation of the measurements. This is the case of data stored into the OI-FITS format, a data exchange standard for optical interferometry (Pauls *et al.* 2005). Without any measured correlations, one is obliged to assume that measurements are independent variables (for the powerspectrum data) or pairs of variables (for complex data like the complex visibilities and the bispectra). The likelihood term is then a sum of terms, one for each independent subset of data:

$$f_{\mathsf{data}}(\mathbf{H}\cdot\boldsymbol{x}) = \sum_m f_m(\boldsymbol{z}_m - \widetilde{\boldsymbol{z}}_m(\mathbf{H}\cdot\boldsymbol{x}))$$

where each elementary datum $\boldsymbol{z}_m$ is either a real or a pair of reals (amplitude and phase or real and imaginary parts of a complex measurement).

In the most simple case, the data consists in independent calibrated complex visibilities with independent and identically distributed (i.i.d.) real and imaginary parts (the so-called Goodman approximation, Goodman 1985). The likelihood term then writes:

$$f_{\mathsf{data}}(\mathbf{H}\cdot\boldsymbol{x}) = \sum_m w_m \, |\boldsymbol{z}_m - (\mathbf{H}\cdot\boldsymbol{x})_m|^2$$

with $w_m = 1/\,\mathrm{Var}\{\mathrm{Re}\{\boldsymbol{z}_m\}\} = 1/\,\mathrm{Var}\{\mathrm{Im}\{\boldsymbol{z}_m\}\}$ and $|\boldsymbol{z}_m - (\mathbf{H}\cdot\boldsymbol{x})_m|$ the modulus of the complex residuals. In matrix notation and providing the Argand

representation[4] of the complex visibilities is used, the likelihood can be put in the form of a quadratic cost function with respect to the unknowns $\boldsymbol{x}$:

$$f_{\mathsf{data}}(\mathbf{H}\cdot\boldsymbol{x}) = (\boldsymbol{z} - \mathbf{H}\cdot\boldsymbol{x})^\top \cdot \mathbf{W} \cdot (\boldsymbol{z} - \mathbf{H}\cdot\boldsymbol{x})$$

where $\mathbf{W}$ is block diagonal matrix with 2×2 blocks. This is suitable for radio-astronomy data but not for current optical interferometers. See, for instance, Meimon *et al.* (2005a) and Thiébaut (2008) for various approximate expressions of the likelihood term. Note that Goodman approximation would give circular isocontours in Figure 5.

For complex data $\boldsymbol{z}_m = \rho_m \exp(\mathrm{i}\,\varphi_m)$ in polar form with independent modulus and phase, Meimon *et al.* (2005a) suggested to use a quadratic approximation of the likelihood:

$$f_m(\mathbf{H}\cdot\boldsymbol{x}) = \boldsymbol{e}_m(\mathbf{H}\cdot\boldsymbol{x})^\top \cdot \begin{pmatrix} W_m^{\mathsf{rr}} & W_m^{\mathsf{ri}} \\ W_m^{\mathsf{ri}} & W_m^{\mathsf{ii}} \end{pmatrix} \cdot \boldsymbol{e}_m(\mathbf{H}\cdot\boldsymbol{x}), \tag{6.1}$$

with the weights:

$$W_m^{\mathsf{rr}} = \frac{\cos^2\varphi_m}{\mathrm{Var}\{\rho_m\}} + \frac{\sin^2\varphi_m}{\rho_m^2\ \mathrm{Var}\{\varphi_m\}}, \tag{6.2}$$

$$W_m^{\mathsf{ri}} = \left(\frac{1}{\mathrm{Var}\{\rho_m\}} - \frac{1}{\rho_m^2\ \mathrm{Var}\{\varphi_m\}}\right) \cos\varphi_m\ \sin\varphi_m\,, \tag{6.3}$$

$$W_m^{\mathsf{ii}} = \frac{\sin^2\varphi_m}{\mathrm{Var}\{\rho_m\}} + \frac{\cos^2\varphi_m}{\rho_m^2\ \mathrm{Var}\{\varphi_m\}}, \tag{6.4}$$

and the complex residuals:

$$\boldsymbol{e}_m(\mathbf{H}\cdot\boldsymbol{x}) = \begin{pmatrix} \rho_m\ \cos\varphi_m - \widetilde{\rho}_m(\mathbf{H}\cdot\boldsymbol{x})\ \cos\widetilde{\varphi}_m(\mathbf{H}\cdot\boldsymbol{x}) \\ \rho_m\ \sin\varphi_m - \widetilde{\rho}_m(\mathbf{H}\cdot\boldsymbol{x})\ \sin\widetilde{\varphi}_m(\mathbf{H}\cdot\boldsymbol{x}) \end{pmatrix} \tag{6.5}$$

where the tilde indicates the model of a given measurement. The expression of the likelihood in Equation (6.1) can be used for complex visibilities V_m or bispectrum data B_m in polar form as provided by OI-FITS format. However note that this yields a non-quadratic penalty for the bispectrum.

Some algorithms ignore the measured amplitudes of the bispectrum and only consider the bispectrum phase $\beta_m = \arg(B_m)$ to provide Fourier phase information for the image reconstruction, the Fourier amplitude information being provided by the powerspectrum data. In this case, practical expressions of the likelihood with respect to such kind of data must be introduced. In MiRA algorithm (Thiébaut 2008), powerspectrum data are treated as independent Gaussian variables, the likelihood for the measured powerspectrum P_m then writes:

$$f_m(\mathbf{H}\cdot\boldsymbol{x}) = \frac{\left(P_m - \widetilde{P}_m(\mathbf{H}\cdot\boldsymbol{x})\right)^2}{\mathrm{Var}\{P_m\}}. \tag{6.6}$$

[4]Real and imaginary parts.

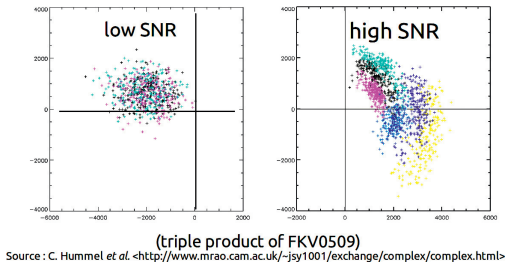


Fig. 4. Empirical distribution of complex bispectrum data at low (left and high (right) signal to noise ratio (SNR). Horizontal axis is the real part, vertical axis is the imaginary part.

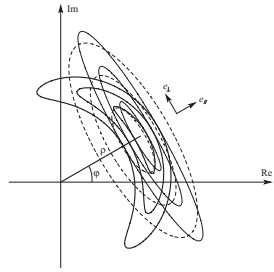


Fig. 5. Convex quadratic approximations of the true distribution of errors for a complex measurement. Thick lines: χ^2 isocontours (at 1, 2 and 3 rms levels) for a complex data with independent amplitude and phase. Dashed lines: isocontours for the global quadratic approximation. Thin lines: isocontours for the local quadratic approximation. (Meimon *et al.* 2005a).

In order to account for phase wrapping and to avoid excessive non-linearity, the term related to the phase closures data is defined by MiRA to be the weighted quadratic distance between the complex phasors rather than between the phases closures:

$$f_m(\mathbf{H}\cdot\boldsymbol{x}) = \frac{1}{\mathrm{Var}\{\beta_m\}}\left|\mathrm{e}^{\mathrm{i}\,\beta_m} - \mathrm{e}^{\mathrm{i}\,\widetilde{\beta}_m(\mathbf{H}\cdot\boldsymbol{x})}\right|^2. \tag{6.7}$$

In the limit of small phase closure errors, the penalty becomes:

$$f_m(\mathbf{H}\cdot\boldsymbol{x}) \approx \frac{\left[\beta_m - \widetilde{\beta}_m(\mathbf{H}\cdot\boldsymbol{x})\right]^2}{\mathrm{Var}\{P_m\}} \tag{6.8}$$

which is readily the χ^2 term that would be obtained for Gaussian phase statistics. This justifies the weighting used in Equation (6.7). Other expressions of the likelihood with respect to phase data have been proposed to cope with the phase wrapping (Haniff 1991; Lannes 2001) but, in practice, they give penalties which slow down or even prevent the convergence of the optimization algorithm.

For optical interferometry which only provides powerspectrum and bispectrum data, the likelihood term $f_{\mathsf{data}}(\mathbf{H}\cdot\boldsymbol{x})$ is highly non-quadratic, *e.g.* see Equations (6.6) and (6.7). This will give rise to optimization issues when fitting the data. Before tackling these issues, let us discuss the second penalty term, that is the regularization.

7 Regularization

In principle, the regularization penalty could be derived from Bayesian considerations (see Sect. 5.3):

$$\mu\, f_{\mathsf{prior}}(\boldsymbol{x}) = c_0'' - c_1\, \log(\Pr(\boldsymbol{x})). \tag{7.1}$$

with c_0'' any real constant and $c_1 > 0$ the same constant as in the previous section. However, introducing a prior probability density function of the parameters which is sufficiently general for all possible observed objects would yield highly uninformative priors which do not really help finding a satisfying image. To be effective, the regularization has to be more restrictive which implies to make more specific assumptions about the object brightness distribution. Besides, even if we knew the object quite exactly, we would like that the prior penalty be at least insensitive to the observing conditions, thus to the position of the object, its orientation and its distance (*i.e.* its angular size and its integrated brightness).

7.1 Simple quadratic regularization

Let us examine the consequences of these elementary considerations. To simplify our reasoning, we consider the pixel-oriented image model:

$$x_n = i(\boldsymbol{\theta}_n) \approx I_\lambda(\boldsymbol{\theta}_n),$$

and assume that the parameters $\boldsymbol{x}$ follow a Gaussian distribution[5], then:

$$\begin{aligned} f_{\boldsymbol{x}}(\boldsymbol{x}) &= -\log(\Pr(\boldsymbol{x})) \\ &= c + (1/2)\,(\boldsymbol{x} - \overline{\boldsymbol{x}})^\top \cdot \mathbf{C}_{\boldsymbol{x}}^{-1} \cdot (\boldsymbol{x} - \overline{\boldsymbol{x}}) \end{aligned} \tag{7.2}$$

where $\overline{\boldsymbol{x}} = \mathrm{E}\{\boldsymbol{x}\}$ is the expected value of $\boldsymbol{x}$, $\mathbf{C}_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x})$ its covariance and

$$c = \frac{1}{2}\, \log\left[\det\left(\frac{\mathbf{C}_{\boldsymbol{x}}}{2\,\pi}\right)\right]$$

is a constant due to the normalization of $\Pr(\boldsymbol{x})$ and which does not depend on $\boldsymbol{x}$.

From the principle that the regularization shall be shift-invariant, the covariance $(\mathbf{C}_{\boldsymbol{x}})_{n,n'}$ between the nth and the n'th pixels must only depend on their relative position $\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n'}$; moreover, since the regularization shall be isotropic, it must only depend on the relative distance $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n'}\|$. This is also true for the inverse of the covariance matrix, thus:

$$\left(\mathbf{C}_{\boldsymbol{x}}^{-1}\right)_{n,n'} = \alpha\, \zeta(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n'}\|/\Omega) \tag{7.3}$$

[5]Which cannot be really true because of the non-negativity and, perhaps, normalization constraints.

where $\alpha > 0$ is a scaling factor, $\zeta\colon \mathbb{R}_+ \mapsto \mathbb{R}$ is a function of the relative angular separation between the pixels and Ω is a typical angular size. From the requirements that the prior shall not depend on the absolute brightness of the object, nor on its angular size, the factor α shall scale as the reciprocal of the square of the object brightness and Ω shall scale as the angular size of the object.

As the regularization shall be shift-invariant, the mean must not depend on the pixel index, hence:

$$\overline{\boldsymbol{x}} = \beta\,\mathbf{1}, \tag{7.4}$$

where β is the mean pixel brightness. Noting that:

$$\begin{aligned}(x_n - x_{n'})^2 &= [(x_n - \beta) - (x_{n'} - \beta)]^2 \\ &= (x_n - \beta)^2 + (x_{n'} - \beta)^2 - 2\,(x_n - \beta)\,(x_{n'} - \beta) \\ \Longrightarrow \quad (x_n - \beta)\,(x_{n'} - \beta) &= (1/2)\,[(x_n - \beta)^2 + (x_{n'} - \beta)^2 - (x_n - x_{n'})^2]\,,\end{aligned}$$

the prior penalty $f_{\boldsymbol{x}}(\boldsymbol{x})$ in Equation (7.2) writes:

$$\begin{aligned}f_{\boldsymbol{x}}(\boldsymbol{x}) &= c + (1/2)\,(\boldsymbol{x} - \beta\,\mathbf{1})^\top \cdot \mathbf{C}_{\boldsymbol{x}}^{-1} \cdot (\boldsymbol{x} - \beta\,\mathbf{1}) \\ &= c + \frac{1}{2}\sum_{n,n'} \left(\mathbf{C}_{\boldsymbol{x}}^{-1}\right)_{n,n'} (x_n - \beta)\,(x_{n'} - \beta) \\ &= c + \frac{\mu_0}{2}\sum_n (x_n - \beta)^2 + \frac{1}{2}\sum_{n<n'} \mu_{n,n'}\,(x_n - x_{n'})^2 \end{aligned} \tag{7.5}$$

with:

$$\mu_0 = \sum_n \left(\mathbf{C}_{\boldsymbol{x}}^{-1}\right)_{n,n'} = \sum_{n'} \left(\mathbf{C}_{\boldsymbol{x}}^{-1}\right)_{n,n'} \tag{7.6}$$

$$\mu_{n,n'} = -\left(\mathbf{C}_{\boldsymbol{x}}^{-1}\right)_{n,n'} \tag{7.7}$$

where the two equivalent expressions for μ_0 come from the fact that the covariance matrix is symmetrical and so is its inverse.

Taking $c_1 = 2$ (as for the likelihood in Sect. 6) and $c_0'' = -c_1\,c$, yields the quadratic regularization term:

$$\mu\, f_{\mathsf{prior}}(\boldsymbol{x}) = \mu_0 \sum_n (x_n - \beta)^2 + \sum_{n<n'} \mu_{n,n'}\,(x_n - x_{n'})^2\,. \tag{7.8}$$

These simple and general considerations lead us to the quadratic regularization in Equation (7.8) which has the required properties (shift-invariance, isotropy, etc.) and which is parametrized by so-called *hyper-parameters*: α, β (both related to the object brightness), Ω (the size of the object) and $\zeta\colon \mathbb{R}_+ \mapsto \mathbb{R}$ the relative weighting function. If we take $\mu = \mu_0 > 0$, $\beta = 0$ and $\mu_{n,n'} = 0$, $\forall (n, n')$, then we obtain the most simple form of Tikhonov's regularization (Tikhonov & Arsenin 1977):

$$f_{\mathsf{prior}}(\boldsymbol{x}) = \sum_n x_n^2 = \|\boldsymbol{x}\|_2^2.$$

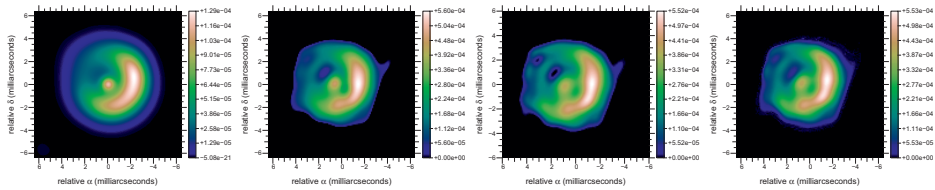


Fig. 6. Image reconstruction with various types of regularization. From *left* to *right*: (a) original object smoothed to the resolution of the interferometer (FWHM ~ 15 mas); (b) reconstruction with a quadratic regularization given by Equation (7.9) and which imposes a compact field of view; (c) reconstruction with edge-preserving regularization as in Equation (7.10); (d) reconstruction with maximum entropy regularization as in Equation (7.12). All reconstructions by algorithm MiRA (Thiébaut 2008) and from the powerspectrum and the phase closures data of the 2004' Imaging Beauty Contest (Lawson *et al.* 2004).

Whereas if we take $\mu_0 = 0$ and $\mu_{n,n'} \geq 0$ a decreasing function of the distance between the nth and the n'th pixels, then we obtain a regularization which favors solutions where nearby pixels have similar values hence the smoothness of the restored image.

7.2 A marketplace for regularization

The Gaussian assumption for the prior distribution of the image parameters yields quadratic regularizations, like the one in Equation (7.8), which are easy to minimize numerically. However such regularizations alone[6] are not very efficient to interpolate missing data when dealing with sparse interferometric data. They are also not the best choice to restore some features of the observed objects, in particular point-like sources or sharp edges. Non-quadratic regularizations have been proposed which may be more suitable for sparse data and images with sharp structures.

The most useful regularizations for image restoration are shift-invariant, (approximately) isotropic and parametrized by a few hyper-parameters. However, in the case of optical interferometry data where the observables (powerspectrum and bispectrum) are insensitive to the position of the object, it may be useful to introduce a shift-variant regularization to fix this degeneracy (see the *compactness* regularization below proposed by le Besnerais *et al.* 2008).

It is impossible to give an exhaustive list of regularizations, but for image restoration, in particular from interferometric data, the following prior penalties have been used with some success:

Quadratic smoothness is imposed by minimizing the differences between close pixels. This is achieved with:

$$f_{\mathsf{prior}}(\boldsymbol{x}) = \|\mathbf{D} \cdot \boldsymbol{x}\|_2^2$$

[6] Without the strict constraints imposed by the feasible set $\mathbb{X}$.

where $\mathbf{D}$ is a finite difference operator. For instance, in 1-D:

$$(\mathbf{D} \cdot \boldsymbol{x})_n = x_{n+1} - x_n$$

and in 2-D:

$$(\mathbf{D} \cdot \boldsymbol{x})_{n_1,n_2} = \begin{pmatrix} x_{n_1+1,n_2} - x_{n_1,n_2} \\ x_{n_1,n_2+1} - x_{n_1,n_2} \end{pmatrix}.$$

This regularization is specific instance of Equation (7.8) with $\mu_0 = 0$ and

$$\mu_{n,n'} = \mu\,[\delta(n_1 + 1 - n_1')\,\delta(n_2 - n_2') + \delta(n_1 - n_1')\,\delta(n_2 + 1 - n_2')].$$

A similar result can be obtained with:

$$f_{\mathsf{prior}}(\boldsymbol{x}) = \|\boldsymbol{x} - \mathbf{S} \cdot \boldsymbol{x}\|_2^2$$

where $\mathbf{S}$ is a smoothing operator.

Compactness can be achieved with

$$f_{\mathsf{prior}}(\boldsymbol{x}) = \sum\nolimits_n w_n^{\mathsf{prior}} x_n^2, \tag{7.9}$$

where $w_n^{\mathsf{prior}} \geq 0$ are given weights. If the weights increase with the distance to a given position (for instance, $w_n^{\mathsf{prior}} \propto \|\boldsymbol{\theta}_n\|^\beta$ with $\beta > 0$), this regularization favors a compact brightness distribution with its flux concentrated around this position. In the Fourier domain, this yields *spectral smoothness* which may be very helpful to interpolate the voids in the (u, v)-coverage.

If the weights are all strictly positive, it can be shown (le Besnerais *et al.* 2008) that the default solution:

$$\boldsymbol{x}^{\mathsf{prior}} \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{x} \in \mathbb{X}} \sum\nolimits_n w_n^{\mathsf{prior}} \boldsymbol{x}_n^2$$

on the feasible set $\mathbb{X}$ given in Equation (5.7) is simply:

$$x_n^{\mathsf{prior}} \propto 1/w_n^{\mathsf{prior}}$$

where the constant of proportionality is such that the normalization constraint is satisfied.

Non-linear smoothness can be imposed with the following general expression:

$$f_{\mathsf{prior}}(\boldsymbol{x}) = \sum\nolimits_n \sqrt{\|\nabla x_n\|^2 + \epsilon^2} \tag{7.10}$$

where $\|\nabla x_n\|^2$ is the squared magnitude of the spatial gradient in the image at nth pixel and $\epsilon \geq 0$. Taking $\epsilon = 0$ yields the so-called *total variation* (TV) regularization which favors flat regions separated by sharp edges (Rudin *et al.* 1992). Otherwise, taking $\epsilon > 0$ yields *edge-preserving smoothness* (Charbonnier *et al.* 1997) which behaves as a quadratic smoothness prior

in region where the spatial gradient of the image is smaller than ϵ in magnitude, while preserving sharp edges elsewhere. The actual expression in Equation (7.10) depends on the approximation of the spatial gradient which is usually implemented via a finite difference operator: $\nabla x_n = \mathbf{D}_n \cdot \boldsymbol{x}$ (Chambolle *et al.* 2011). There are also other possibilities to achieve edge-preserving regularization (see *e.g.*, Charbonnier *et al.* 1997).

Spatial sparsity can be imposed thanks to separable ℓ_p norms:

$$f_{\mathsf{prior}}(\boldsymbol{x}) = \sum\nolimits_n |x_n|^p, \tag{7.11}$$

with $p \geq 0$. If $p < 1$, minimizing the ℓ_p norm favors sparse distribution, while $p = 2$ corresponds to regular *Tikhonov regularization* (Tikhonov & Arsenin 1977) and favors flat distributions. Note that p must be greater or equal 1 to have a convex criterion. Taking the smallest such p, that is $p = 1$, is what is advocated in *compress sensing* (Donoho 2006).

Maximum entropy methods (MEM) have been proposed for radio-astronomy and exploit a separable non-linear regularization with the general form:

$$f_{\mathsf{prior}}(\boldsymbol{x}) = -\sum\nolimits_n h(x_n|\overline{x}_n). \tag{7.12}$$

Here the prior is to assume that the image is drawn toward a prior model $\overline{\boldsymbol{x}}$ according to a non-quadratic potential h, called the *entropy*. Various entropy terms have been proposed in the literature (Narayan & Nityananda 1986):

$$\begin{aligned}
&\text{MEM-sqrt:} && h(x|\overline{x}) = \sqrt{x};\\
&\text{MEM-log:} && h(x|\overline{x}) = \log(x);\\
&\text{MEM-prior:} && h(x|\overline{x}) = x - \overline{x} - x\,\log\left(x/\overline{x}\right).
\end{aligned}$$

Being separable, the expression in Equation (7.12) assumes that the pixel values are uncorrelated. To impose some level of smoothness in the solution, Horne (1985) has proposed a non-separable MEM regularization by defining the prior model $\overline{\boldsymbol{x}}$ as a smoothed version of the model $\boldsymbol{x}$, for instance: $\overline{\boldsymbol{x}} = \mathbf{S} \cdot \boldsymbol{x}$ with $\mathbf{S}$ a smoothing operator.

7.3 Choosing and tuning the regularization

As we have just seen, there are many different possible expressions for the regularization term. Since the exact statistics of the sought object is seldom known, the regularization has to be chosen on the basis of general properties that one expect to see in the sought image. In the case of interferometric imaging, Renard *et al.* (2011) have compared the regularization methods presented in the previous section. As expected they concluded that the best prior depends on the object of interest. However, non-linear smoothness, in Equation (7.10), and compactness combined with non-negativity constraints, in Equation (7.9), are the most

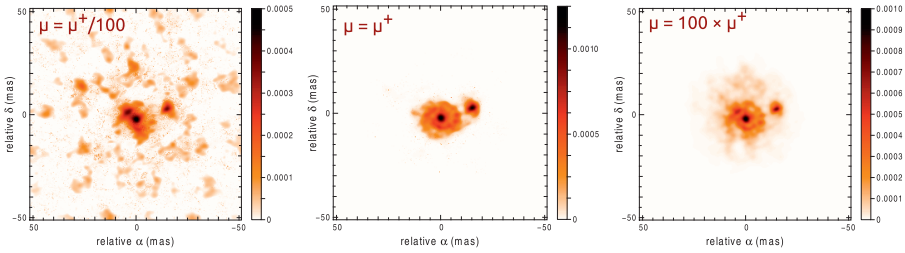


Fig. 7. Image reconstruction with ℓ_2 compactness and for various levels of regularization. The optimal regularization level is μ^+ (source: Renard *et al.* 2011).

successful regularizations in general. Figure 6 shows that images restored with different types of regularization are fairly similar. This is a general observation: providing there are sufficient data and the hyper-parameters are correctly set (see below), the restored image either succeeds to approximate the object or clearly fails (Renard *et al.* 2011). In practice, it is fruitful to exploit the variety of regularization types to determine which one is most adapted to the object of interest. Comparing images obtained under different priors is also useful to disentangle between artifacts and real features. One must however keep in mind that, among other properties, the priors must be able to lift the degeneracies of the inverse problem and to regularize it, that is to warrant a unique and stable solution with respect to small perturbations such as those due to the noise.

In addition to the choice of the form of the regularization itself, there are tuning parameters: the weight μ of the regularization, and perhaps some other hyper-parameters (*e.g.* the relaxation parameter ϵ in the edge-preserving regularization below). Ideally one would like to set these hyper-parameters automatically according to some objective criterion. Although several unsupervised methods have been proposed for setting the hyper-parameters, this is still a vivid research subject and no methods is at the same time robust and easy to apply. When there are few hyper-parameters, visual assessment of the result is often sufficient to correctly set these parameters. For instance, Figure 7 shows the effects of tuning the level of regularization μ. Compared to the optimal setting (central panel in Fig. 7), if the weight of the regularization is too small, many artifacts due to the voids in the (u, v) coverage contaminate the image (left panel in Fig. 7). On the contrary, if the weight of the regularization is too important, the image becomes too flat (right panel in Fig. 7). Although this depends on the particular regularization implemented.

8 Optimization strategy

We have seen that image reconstruction amounts to solving:

$$\min_{\boldsymbol{x}\in\mathbb{X}} \underbrace{\{\mu\, f_{\mathsf{prior}}(\boldsymbol{x}) + f_{\mathsf{data}}(\mathbf{H}\cdot\boldsymbol{x})\}}_{f(\boldsymbol{x})}\cdot \tag{8.1}$$

In the case of optical interferometric data, this constrained optimization problem depends on a very large number of parameters (the image pixels), is highly non-linear[7] and multi-modal (has multiple minima). Solving such a problem requires *global optimization* or a good starting point followed by continuous optimization. It is remarkable that existing image reconstruction algorithms implement not only different priors but also different strategies to search the solution.

CLEAN (Högbom 1974) was initially developed for radio-interferometry (*i.e.* for complex visibility data) and exploits a matching pursuit algorithm to iteratively build the image by modifying a single pixel at every iteration. The *building-blocks* method (Hofmann & Weigelt 1993) is an adaptation of the CLEAN algorithm to deal with bispectrum data. The assumption made by these two methods is that the object of interest mainly consists in point-like sources. Using the regularization given by Equation (7.11) with $p = 1$ (*i.e.* taking the ℓ_1 norm of the pixels as the prior penalty) yields a similar result and produces a spatially sparse solution. Introducing such a continuous regularization, although not smooth, gives the opportunity to use optimization strategies much more efficient than matching pursuit algorithms (Thiébaut *et al.* 2012).

WISARD (Meimon *et al.* 2005b) implements a kind of self-calibration strategy alternating between (i) estimating the missing Fourier phases given the object and the phase closures to complete the data and produce pseudo-complex visibility data, and (ii) image reconstruction given these pseudo-data and the priors.

MACIM (Markov Chain Imager, Ireland *et al.* 2008) generates a stochastic sampling of the posterior probability

$$\Pr(\boldsymbol{x}|\boldsymbol{z}) \propto \Pr(\boldsymbol{z}|\boldsymbol{x})\ \Pr(\boldsymbol{x})$$

by means of a Monte-Carlo Markov Chain (MCMC) algorithm. The image samples can then be used to find the mode of the distribution (which gives the most likely solution) or to compute the posterior mean of the sought image (which gives the image which minimizes the mean quadratic error). For large size problems, MCMC may however take prohibitive computational time to generate good samples of the posterior distribution.

WIPE (Lannes *et al.* 1997), BSMEM (Baron & Young 2008; Buscher 1994) and MiRA (Thiébaut 2008) directly minimize the penalty in Equation (8.1) by means of non-linear conjugate gradient algorithm, sub-space method (Skilling and Bryan 1984) or quasi-Newton methods (Nocedal & Wright 2006). These optimization algorithms can deal with non-linear penalties with very large number of parameters and, possibly, with constraints such as non-negativity. A change of variables can be introduced to implement the normalization constraint (le Besnerais *et al.* 2008). To my knowledge, WIPE can only cope with complex visibility data and has not been adapted to deal with optical interferometry data.

In an attempt to unify direct optimization and self-calibration approaches to solve the image reconstruction problem (8.1), we describe next another

[7] Which means that the joint criterion $f(\boldsymbol{x})$ is *non-quadratic*.

optimization strategy that can be adapted to any type of data and priors. The method follows the Alternating Direction Method of Multipliers (ADMM, Gabay & Mercier 1976) and consists in alternatively minimizing the two terms $f_{\mathsf{prior}}(\boldsymbol{x})$ and $f_{\mathsf{data}}(\boldsymbol{y})$ subject to the constraint $\boldsymbol{y} = \mathbf{H} \cdot \boldsymbol{x}$.

8.1 Augmented Lagrangian

Solving the image reconstruction problem (8.1) by *direct minimization* is exactly the same as solving the *constrained problem*:

$$\min_{\boldsymbol{x} \in \mathbb{X}, \boldsymbol{y}} \{\mu\, f_{\mathsf{prior}}(\boldsymbol{x}) + f_{\mathsf{data}}(\boldsymbol{y})\} \quad \text{s.t.} \quad \mathbf{H} \cdot \boldsymbol{x} = \boldsymbol{y} \tag{8.2}$$

where the *model complex visibilities* $\boldsymbol{y} = \mathbf{H} \cdot \boldsymbol{x}$ have been explicitly introduced as *auxiliary variables*. This will allow us to treat separately the specificity of $f_{\mathsf{prior}}(\boldsymbol{x})$ and $f_{\mathsf{data}}(\boldsymbol{y})$, in particular their non linearity or lack of smoothness.

A standard approach to solve the constrained problem (8.2) is to use the Lagrangian of the problem:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}) = \mu\, f_{\mathsf{prior}}(\boldsymbol{x}) + f_{\mathsf{data}}(\boldsymbol{y}) + \boldsymbol{u}^\top \cdot (\mathbf{H} \cdot \boldsymbol{x} - \boldsymbol{y})\,,$$

with $\boldsymbol{u}$ the Lagrange multipliers associated to the constraints $\mathbf{H} \cdot \boldsymbol{x} = \boldsymbol{y}$. For a solution $\{\boldsymbol{x}^\star, \boldsymbol{y}^\star, \boldsymbol{u}^\star\}$ of the problem, the necessary conditions of optimality, the so-called Karush-Kuhn-Tucker (KKT) conditions, write:

$$\mathbf{H} \cdot \boldsymbol{x}^\star = \boldsymbol{y}^\star \tag{8.3}$$

$$\mathbf{0} \in \partial_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{y}^\star, \boldsymbol{u}^\star) \tag{8.4}$$

$$\mathbf{0} \in \partial_{\boldsymbol{y}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{y}^\star, \boldsymbol{u}^\star) \tag{8.5}$$

where ∂ denotes the subdifferential operator Boyd *et al.* (2010) which only contains the gradient of its argument if it is differentiable. For instance, if the Lagrangian is differentiable with respect to variables $\boldsymbol{x}$, the second KKT condition in Equation (8.4) becomes:

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{y}^\star, \boldsymbol{u}^\star) = \mathbf{0}.$$

Using the Lagrangian involves searching the optimal multipliers $\boldsymbol{u}^\star$ such that minimizing the Lagrangian with respect to the variables $(\boldsymbol{x}, \boldsymbol{y})$ given the multipliers yields a solution matching the constraints. However, finding the optimal multipliers requires to solve a system of M (the number of observed baselines) non-linear equations which is much more involved than finding a single root as required by the constrained problem in Section 5.2.

Unless a closed form solution exists, it is easier to solve the constrained problem (8.2) by using the *augmented Lagrangian* (Hestenes 1969; Powell 1969):

$$\mathcal{L}_{\mathsf{A}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}; \rho) = \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}) + (\rho/2)\, \|\mathbf{H} \cdot \boldsymbol{x} - \boldsymbol{y}\|_2^2 \tag{8.6}$$

with $\rho > 0$ the weight of the augmented penalty to reinforce the constraints. Obviously for any variables matching the constraints, *i.e.* such that $\mathbf{H} \cdot \boldsymbol{x} = \boldsymbol{y}$,

the Lagrangian and the augmented Lagrangian are equal; thus they both yield the same solution. Solving the constrained problem (8.2) via the augmented Lagrangian however has a number of practical advantages compared to using the Lagrangian: (i) it provides an explicit update formula for the multipliers (see Eq. (8.7) in Algorithm 1), (ii) it owns strong convergence properties for ρ large enough even for non-smooth penalties, (iii) it can be exploited to derive a simple yet efficient algorithm based on alternate minimization (see Algorithm 2).

Solving the image reconstruction problem (8.2) via the augmented Lagrangian and simply considering the variables $\boldsymbol{x}$ and $\boldsymbol{y}$ as a single group of variables yields the following algorithm:

Algorithm 1: *Augmented Lagrangian algorithm for solving (8.2).* Choose initial multipliers $\boldsymbol{u}_0$. Then, for $k = 0, 1, \ldots$, repeat the following steps until convergence:

1. Choose augmented penalty parameter $\rho_k > 0$ and improve the variables:

$$\left\{\boldsymbol{x}_{k+1}, \boldsymbol{y}_{k+1}\right\} \approx \underset{\boldsymbol{x}\in\mathbb{X},\boldsymbol{y}}{\arg\min}\, \mathcal{L}_{\mathsf{A}}\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}_k; \rho_k\right).$$

2. Update the multipliers:

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \rho_k \left(\mathbf{H}\cdot\boldsymbol{x}_{k+1} - \boldsymbol{y}_{k+1}\right). \blacksquare \tag{8.7}$$

8.2 Alternating direction method of multipliers

Algorithm 1 involves minimizing the likelihood and the regularization at the same time which has not much practical interest compared to directly minimizing Equation (8.1) with respect to $\boldsymbol{x}$. The minimization becomes easier if one considers the penalties $f_{\mathsf{prior}}(\boldsymbol{x})$ and $f_{\mathsf{data}}(\boldsymbol{y})$ separately. To that end, Step 1 of Algorithm 1 can be implemented thanks to alternating minimization, for instance:

$$\boldsymbol{x}_{k+1} = \underset{\boldsymbol{x}\in\mathbb{X}}{\arg\min}\, \mathcal{L}_{\mathsf{A}}(\boldsymbol{x}, \boldsymbol{y}_k, \boldsymbol{u}_k; \rho_k),$$

followed by

$$\boldsymbol{y}_{k+1} = \underset{\boldsymbol{y}}{\arg\min}\, \mathcal{L}_{\mathsf{A}}(\boldsymbol{x}_{k+1}, \boldsymbol{y}, \boldsymbol{u}_k; \rho_k).$$

This imposes to choose an initial value $\boldsymbol{y}_0$ for the auxiliary variables $\boldsymbol{y}$. If an initial image $\boldsymbol{x}_0$ is available, the order of updating $\boldsymbol{x}$ and $\boldsymbol{y}$ can be exchanged. Alternating minimization yields the following algorithm:

Algorithm 2: *Alternate Direction Method of Multipliers (ADMM) algorithm for solving (8.2).* Choose initial multipliers $\boldsymbol{u}_0$ and initial complex visibilities $\boldsymbol{y}_0$. Then, for $k = 0, 1, \ldots$, repeat the following steps until convergence:

1. **Image Reconstruction Step.** Choose the augmented penalty parameter $\rho_k > 0$ and approximately find the best image given the complex visibilities and the Lagrange multipliers:

$$\boldsymbol{x}_{k+1} \approx \arg\min_{\boldsymbol{x}\in\mathbb{X}} \mathcal{L}_{\mathsf{A}}\left(\boldsymbol{x}, \boldsymbol{y}_k, \boldsymbol{u}_k; \rho_k\right).$$

2. **Self Calibration Step.** Approximately find the best complex visibilities given the image and the Lagrange multipliers:

$$\boldsymbol{y}_{k+1} \approx \arg\min_{\boldsymbol{y}} \mathcal{L}_{\mathsf{A}}\left(\boldsymbol{x}_{k+1}, \boldsymbol{y}, \boldsymbol{u}_k; \rho_k\right).$$

3. **Updating of the Lagrange Multipliers.** Apply the following formula to update the multipliers:

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \rho_k \left(\mathbf{H}\cdot\boldsymbol{x}_{k+1} - \boldsymbol{y}_{k+1}\right). \blacksquare \tag{8.8}$$

Before going into the details of the algorithm, let us remark that by elementary manipulations, the augmented Lagrangian can be rewritten as:

$$\begin{aligned}\mathcal{L}_{\mathsf{A}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{u};\rho) &= \mu\, f_{\mathsf{prior}}(\boldsymbol{x}) + f_{\mathsf{data}}(\boldsymbol{y}) + \boldsymbol{u}^\top\cdot(\mathbf{H}\cdot\boldsymbol{x}-\boldsymbol{y}) + \frac{\rho}{2}\left\|\mathbf{H}\cdot\boldsymbol{x}-\boldsymbol{y}\right\|_2^2 \\ &= \mu\, f_{\mathsf{prior}}(\boldsymbol{x}) + f_{\mathsf{data}}(\boldsymbol{y}) + \frac{\rho}{2}\left\|\mathbf{H}\cdot\boldsymbol{x}-\boldsymbol{y}+\boldsymbol{u}/\rho\right\|_2^2 - \frac{1}{2\,\rho}\left\|\boldsymbol{u}\right\|_2^2. \end{aligned} \tag{8.9}$$

8.2.1 Image reconstruction step

Discarding in Equation (8.9) terms which do not depend on the variables $\boldsymbol{x}$, Step 1 of Algorithm 2 consists in improving $\boldsymbol{x}$ given the other variables and writes:

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \arg\min_{\boldsymbol{x}\in\mathbb{X}} \mathcal{L}_{\mathsf{A}}(\boldsymbol{x},\boldsymbol{y}_k,\boldsymbol{u}_k;\rho_k) \\ &= \arg\min_{\boldsymbol{x}\in\mathbb{X}} \mu\, f_{\mathsf{prior}}(\boldsymbol{x}) + (\rho_k/2)\left\|\mathbf{H}\cdot\boldsymbol{x}-\boldsymbol{y}_k+\boldsymbol{u}_k/\rho_k\right\|_2^2 \\ &= \arg\min_{\boldsymbol{x}\in\mathbb{X}} (\mu/\rho_k)\, f_{\mathsf{prior}}(\boldsymbol{x}) + (1/2)\left\|\mathbf{H}\cdot\boldsymbol{x}-\boldsymbol{v}_k\right\|_2^2 \end{aligned} \tag{8.10}$$

$$\text{with:}\quad \boldsymbol{v}_k = \boldsymbol{y}_k - \boldsymbol{u}_k/\rho_k\,, \tag{8.11}$$

which is the analogous of *image reconstruction* given *pseudo-complex visibilities* $\boldsymbol{v}_k = \boldsymbol{y}_k - \boldsymbol{u}_k/\rho_k$ with i.i.d. Gaussian noise of variance $\propto \mu/\rho_k$. Note that, if the feasible set is just $\mathbb{R}^N$, the right hand side of Equation (8.10) is the value returned by the proximity operator[8] of $(\mu/\rho_k)\, f_{\mathsf{prior}}$ at $\boldsymbol{v}_k$ (Combettes & Pesquet 2011).

[8]The proximity operator of f: $\mathbb{R}^N \mapsto \mathbb{R}$ is defined by:

$$\mathrm{prox}_f(\boldsymbol{v}) = \arg\min_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) + (1/2)\left\|\boldsymbol{x}-\boldsymbol{v}\right\|_2^2 \right\}.$$

Depending on the particular regularization $f_{\mathsf{prior}}(\boldsymbol{x})$, a specific algorithm may be designed to efficiently solve this problem. If the regularization is quadratic, Equation (8.10) is a large scale quadratic problem which can be solved by existing methods like the *gradient projection conjugate gradient* algorithm (GPCG by Moré & Toraldo 1991). Otherwise, for a number of non smooth $f_{\mathsf{prior}}(\boldsymbol{x})$, there exist closed form solutions of Equation (8.10) with $\mathbb{X} = \mathbb{R}^N$ (Combettes & Pesquet 2011) which can be adapted to account for non negativity constraint (Thiébaut *et al.* 2012).

8.2.2 Updating the complex visibilities

Discarding in Equation (8.9) terms which do not depend on the auxiliary variables $\boldsymbol{y}$, Step 2 of Algorithm 2 consists in improving $\boldsymbol{y}$ given the other variables and writes:

$$\begin{aligned}\boldsymbol{y}_{k+1} &= \arg\min_{\boldsymbol{y}} \mathcal{L}_{\mathsf{A}}(\boldsymbol{x}_{k+1}, \boldsymbol{y}, \boldsymbol{u}_k; \rho_k) \\ &= \arg\min_{\boldsymbol{y}} f_{\mathsf{data}}(\boldsymbol{y}) + (\rho_k/2)\,\|\mathbf{H}\cdot\boldsymbol{x}_{k+1} - \boldsymbol{y} + \boldsymbol{u}_k/\rho_k\|_2^2 \\ &= \arg\min_{\boldsymbol{y}} f_{\mathsf{data}}(\boldsymbol{y}) + (\rho_k/2)\,\|\boldsymbol{y} - \boldsymbol{w}_k\|_2^2 \qquad (8.12)\end{aligned}$$

$$\text{with:}\quad \boldsymbol{w}_k = \mathbf{H}\cdot\boldsymbol{x}_{k+1} + \boldsymbol{u}_k/\rho_k \qquad (8.13)$$

which enforces the complex visibilities $\boldsymbol{y}$ to be a compromise between the actual data and the *shifted* model complex visibilities $\boldsymbol{w}_k = \mathbf{H}\cdot\boldsymbol{x}_{k+1} + \boldsymbol{u}_k/\rho_k$. If there are missing data (for instance, incomplete Fourier phases when working with the bispectrum or the phase closures and the powerspectrum), this step is nevertheless a well posed problem thanks to the augmented term $(\rho_k/2)\,\|\boldsymbol{y} - \boldsymbol{w}_k\|_2^2$.

8.3 Conclusions about optimization strategy

Steps 1 and 2 of Algorithm 2 are the analogous of the image reconstruction and self-calibration steps in self-calibration methods (Cornwell & Wilkinson 1981; Meimon *et al.* 2005b; Schwab 1980). However, to really mimic these latter methods, these steps should be carried out in Algorithm 2 with the Lagrange multipliers always equal to zero. Formally, this means that standard self-calibration methods do not consistently solve a well defined optimization problem. This is not the case of the proposed approach where the self-calibration step accounts for the Lagrange multipliers which are associated to the constraints that $\mathbf{H}\cdot\boldsymbol{x} = \boldsymbol{y}$.

Although global optimization is in principle required to solve Equation (8.1), the most successful algorithms proposed for optical interferometry BSMEM (Baron & Young 2008) and MiRA (Thiébaut 2008) use direct optimization. They however implement numerical optimization algorithms designed for smooth penalties[9].

[9] *Smooth* means here twice continuously differentiable.

Thanks to the variable splitting trick, Algorithm 2 handles separately the specificities of $f_{\mathsf{prior}}(\boldsymbol{x})$ and $f_{\mathsf{data}}(\boldsymbol{y})$. As a consequence, it can efficiently cope with non-smooth penalties such as the ones used to impose sparsity. Moreover, the augmented penalty term introduces a simple quadratic term which regularizes the minimization of $f_{\mathsf{prior}}(\boldsymbol{x})$ and that of $f_{\mathsf{data}}(\boldsymbol{y})$. This makes theses sub-optimization problems well posed and may speed up their numerical solving.

9 Summary and perspectives

After describing the type of measurements which can be acquired with an interferometer and the specific issues due to the turbulence. We addressed the inverse problem of synthesizing an image from these data. The inverse approach provided us a useful framework to derive a kind of recipe for image reconstruction. This recipe involves:

1. A **direct model** of the observables $\boldsymbol{z}$ given the image parameters $\boldsymbol{x}$. This model implements an approximation of the brightness distribution $I_\lambda(\boldsymbol{\theta})$ and its Fourier transform $\widehat{I}_\lambda(\boldsymbol{\nu})$ from which is derived the linear relationship $\boldsymbol{y} = \mathbf{H} \cdot \boldsymbol{x}$ between the sampled complex visibilities $y_m = \widehat{I}_\lambda(\boldsymbol{\nu}_m)$ and the image parameters.

2. A **criterion** to be minimized to determine a unique and stable solution. This criterion takes the form $f(\boldsymbol{x}) = f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x}) + \mu\, f_{\mathsf{prior}}(\boldsymbol{x})$ and reflects the compromise between fidelity to the data, *i.e.* minimizing $f_{\mathsf{data}}(\mathbf{H} \cdot \boldsymbol{x})$, and to the priors, *i.e.* minimizing $f_{\mathsf{prior}}(\boldsymbol{x})$. The hyper-parameter $\mu > 0$ is used to tune this trade-off. Eventually, a feasible set $\mathbb{X}$ can be introduced to account for strict constraints such as non negativity or normalization of the solution.

3. An **optimization strategy** to solve the constrained optimization problem.

The same general framework can been used to describe most (if not all) interferometric image reconstruction algorithms (le Besnerais *et al.* 2008; Thiébaut & Giovannelli 2010; Thiébaut 2009) so the issues encountered while cooking the recipe are also general and have their counterparts in all proposed methods.

In this short presentation, we mainly focused on the so-called *analysis approach* to reconstruct a non-parametric model of the brightness distribution. An alternative, the *synthesis approach*, is to describe the image as the combination of a number of elementary atoms (Elad *et al.* 2007). In the synthesis approach, the regularization is achieved by imposing to use the smallest number of atoms to explain the data. As described in our presentation, this sparsity constraint may be introduced via an ℓ_1 norm penalty and the problem solved by specific algorithms to cope with continuous but non-smooth criteria. It is also possible to try to mimic the effects of using an ℓ_0 norm penalty with *greedy algorithms*. The CLEAN algorithm (Högbom 1974) mentioned in Section 8 can be seen as a precursor of the synthesis approach where the atoms have all the same shape (they are point-like sources) which are only allowed to have different brightnesses and positions.

The ADMM strategy implemented by Algorithm 2 was introduced for pedagogical proposes to make a link between constrained optimization and self-calibration methods and to exhibit some of the issues of solving the optimization part of the image restoration problem. We have argued that the proposed strategy is more consistent than existing self-calibration methods and more flexible than using algorithms restricted to smooth penalties. Introducing variables splitting and ADMM strategy was also motivated by the effectiveness of a similar approach for multi-spectral interferometric data. In this case, the reconstruction algorithm was designed to deal with complex visibilities and exploits structured sparsity regularization to favor point-like sources in the image (Thiébaut *et al.* 2012). To deal with current optical interferometry data, it remains to demonstrate whether such an approach has the ability to find a path to a good solution at a lower cost than a stochastic global optimization method like MACIM (Ireland *et al.* 2008).

As mentioned along this presentation, optimization is not the only direction of research to improve interferometric imaging. Perhaps first of all, multi-spectral image reconstruction is now required to fully exploit the spectral resolution of the existing interferometers. Indeed, it has been clearly demonstrated that spatio-spectral regularization drastically improves the quality of the restored images (Soulez *et al.* 2008). Hence existing algorithms must be extensively modified to globally account for multi-variate data and not just reused to perform independent reconstructions at given wavelengths (le Bouquin *et al.* 2009). In spite of its unrivaled angular resolution, stellar interferometry is not as popular as, say adaptive optics, in the astronomical community. This is partially due to the difficulty to interpret the interferometric data. Making state of the art image reconstruction algorithms available to non-specialists may be a good way to promote interferometric observations. To that end, the methods must be not only robust but also relatively easy to use. Developing unsupervised methods to automatically tune the hyper-parameters of image reconstruction algorithms is therefore of particular interest.

References

Baron, F., & Young, J.S., 2008, Image reconstruction at cambridge university. In Society of Photo-Optical Instrumentation Engineers (SPIE) Conf. Ser., Vol. 7013, 70133X, DOI: 10.1117/12.789115

Boyd, S., Parikh, N., Chu, E., Peleato, Bo., & Eckstein, J., 2010, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends in Machine Learning, 3, 1, DOI: 10.1561/2200000016, http://www.stanford.edu/~boyd/papers/pdf/admm_distr_stats.pdf

Buscher, D.F., 1994, Direct maximum-entropy image reconstruction from the bispectrum, ed. J.G. Robertson & W.J. Tango, IAU Symp. 158: Very High Angular Resolution Imaging, p. 91

Campisi, P., & Egiazarian, K., 2007, Blind image deconvolution: theory and applications (CRC Press), ISBN 9780849373671

Chambolle, A., Levine, S.E., & Lucier, B.J., 2011, SIAM J. Imaging Sciences, 4, 277

Charbonnier, P., Blanc-Féraud, L., Aubert, G., & Barlaud, M., 1997, IEEE Trans. Image Process., 6, 298

Combettes, P.L., & Pesquet, J.-C., 2011, Proximal splitting methods in signal processing, chapter Fixed-Point Algorithms for Inverse Problems in Science and Engineering (Springer, New York), 185

Cornwell, T., 1995, Imaging concepts, ed. J.A. Zensus, P.J. Diamond & P.J. Napier, ASP Conf. Ser. 82, 39

Cornwell, T.J., & Wilkinson, P.N., 1981, MNRAS, 196, 1067

Dainty, J.C., & Greenaway, A.H., 1979, J. Opt. Soc. Am., 69, 786

Delplancke, F., Derie, F., Paresce, F., *et al.*, 2003, Ap&SS, 286, 99,

Donoho, D., 2006, Comm. Pure Appl. Math., 59, 907

Elad, M., Milanfar, P., & Rubinstein, R., 2007, Inverse Probl., 23, 947

Fessler, J.A., & Sutton, B.P., 2003, IEEE Trans. Signal Process., 51, 560

Gabay, D., & Mercier, B., 1976, Comput. Math. Applications, 2, 17

Goodman, J.W., 1985, Statistical Optics (John Wiley & Sons), ISBN 0-471-01502-4

Haniff, C., 1991, J. Opt. Soc. Am. A, 8, 134

Hestenes, M.R., 1969, J. Optimiz. Theory Applications, 4, 303

Hofmann, K.-H., & Weigelt, G., 1993, A&A, 278, 328

Horne, K., 1985, MNRAS, 213, 129

Högbom, J.A., 1974, A&AS, 15, 417

Ireland, M.J., Monnier, J., & Thureau, N., 2008, Monte-Carlo imaging for optical interferometry, ed. J.D. Monnier, M. Schöller & W.C. Danchi, Advances in Stellar Interferometry, Vol. 6268, p. 62681T1, SPIE, DOI: 10.1117/12.670940

Lacour, S., Meimon, S., Thiébaut, É., *et al.*, 2008, A&A, 485, 561

Lannes, A., Anterrieu, E., & Maréchal, P., 1997, A&AS, 123, 183

Lannes, A., 2001, J. Opt. Soc. Am. A, 18, 1046

Lawson, P.R., Cotton, W.D., Hummel, C.A., *et al.*, 2004, BAAS, 36, 1605

le Besnerais, G., Lacour, S., Mugnier, L.M., *et al.*, 2008, IEEE J. Selected Topics Signal Process., 2, 767

le Bouquin, J.-B., Lacour, S., Renard, S., *et al.*, 2009, A&A, 496, L1

Meimon, S., Mugnier, L.M., & le Besnerais, G., 2005a, J. Opt. Soc. Am. A, 22, 2348

Meimon, S., Mugnier, L.M., & le Besnerais, G., 2005b, Opt. Lett., 30, 1809

Moré, J., & Toraldo, G., 1991, SIAM J. Optim., 1, 93, http://locus.siam.org/SIOPT/volume-01/art_0801008.html

Narayan, R., & Nityananda, R., 1986, ARA&A, 24, 127

Nocedal, J., & Wright, S.J., 2006, Numerical Optimization, 2nd edition (Springer Verlag), http://www.zla-ryba.cz/NumOpt.pdf

Pauls, T.A., Young, J.S., Cotton, W.D., & Monnier, J.D., 2005, PASP, 117, 1255

Petrov, R.G., Malbet, F., *et al.*, 2007, A&A, 464, 1

Potts, D., Steidl, G., & Tasche, M., 2001, Modern Sampling Theory: Mathematics and Applications, chapter Fast Fourier transforms for nonequispaced data: A tutorial (Birkhauser, Boston), 249

Powell, M.J.D., 1969, Optimization, chapter A method for nonlinear constraints in minimization problems (Academic Press), 283

Renard, S., Thiébaut, É., & Malbet, F., 2011, A&A, 533, A64

Roddier, F., 1981, The effects of atmospheric turbulence in optical astronomy, Vol 19 (North-Holland Publishing Company, Amsterdam), 281

Rudin, L.I., Osher, S., & Fatemi, E., 1992, Physica D, 60, 259

Schwab, F., 1980, Proc. SPIE, 231, 18

Skilling, J., & Bryan, R.K., 1984, MNRAS, 211, 111

Soulez, F., Thiébaut, É., Gressard, A., Dauphin, R., & Bongard, S., 2008, Heterogeneous multidimensional data deblurring, In 16th European Signal Processing Conference (EUSIPCO), Lausanne, Suisse, http://hal-ujm.ccsd.cnrs.fr/ujm-00293660/en/

Sramek, R., & Schwab, F., 1989, Imaging, ed. Richard A. Perley, Frederic R., Schwab & Alan H. Bridle, Synthesis Imaging in Radio Astronomy, Vol. 6, 117

Thiébaut, É., & Giovannelli, J.-F., 2010, IEEE Signal Process. Mag., 27, 97

Thiébaut, É., 2008, MiRA: an effective imaging algorithm for optical interferometry, ed. Françoise Delplancke Markus Schöller, William C. Danchi. Astronomical Telescopes and Instrumentation, Vol. 7013, 70131I–1, SPIE

Thiébaut, É., 2009, New Astron. Rev., 53, 312

Thiébaut, É., Soulez, F., & Denis, L., 2012, accepted for publication in J. Opt. Soc. Am. A, http://arxiv.org/abs/1209.2362

Thompson, A.R., & Bracewell, R.N., 1974, AJ, 79, 11

Thévenaz, P., Blu, T., & Unser, M., 2000, IEEE Trans. Medical Imag., 19, 739, http://bigwww.epfl.ch/publications/thevenaz0002.html

Tikhonov, A.N., & Arsenin, V.Y., 1977, Solution of Ill-posed Problems, Scripta Series in Mathematics (Winston & Sons, Washington), ISBN 0-470-99124-0

Wirnitzer, B., 1985, J. Opt. Soc. Am. A, 2, 14

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 189–201*

IMAGING TECHNIQUES IN MILLIMETRE ASTRONOMY

M. Bremer[1]

Abstract. Compared to optical astronomy, millimetre radio astronomy experiences not only a different and complementary aspect of the universe but also different perturbations and limitations from Earth's atmosphere that are mostly imposed by atmospheric water vapour and its dynamics. After discussing the physics behind the refractive index variations and possible correction schemes, a small introduction into the basics of radio interferometry and image reconstruction with the CLEAN algorithm is given.

1 Introduction

Millimetre Astronomy is a powerful tool to observe the cold, molecular gas in space, ranging from nearby objects in the Solar system over targets in our Galaxy to galaxies so remote that their light has been travelling for more than 90% of the age of the universe before reaching Earth.

Radio interferometry differs in a number of important points from optical interferometry. Real and imaginary parts of the incoming signal are detected and correlated electronically, observations can be performed at daytime and night time with only weather-imposed limitations, and Earth's atmosphere becomes a bright background against which the astronomical source must be detected.

2 Atmosphere

In the optical, astronomical observations are perturbed by seeing: a large number of small ($\sim$10 cm), rapid refractive index fluctuations in the atmosphere leads to multiple splitting of an image into speckles that change with a rate of about 100 Hz (Lohman *et al.* 1983). The influence of speckles diminishes in the near infrared, but so does the atmospheric transmission when moving further to longer wavelengths: Extended absorption line systems of mainly water vapour and carbon dioxide block ground-based astronomical observations. When observations

[1] Institut de Radio Astronomie Millimétrique (IRAM), 300 rue de la Piscine, 38406 Saint-Martin-d'Hères, France

DOI: 10.1051/eas/1359010

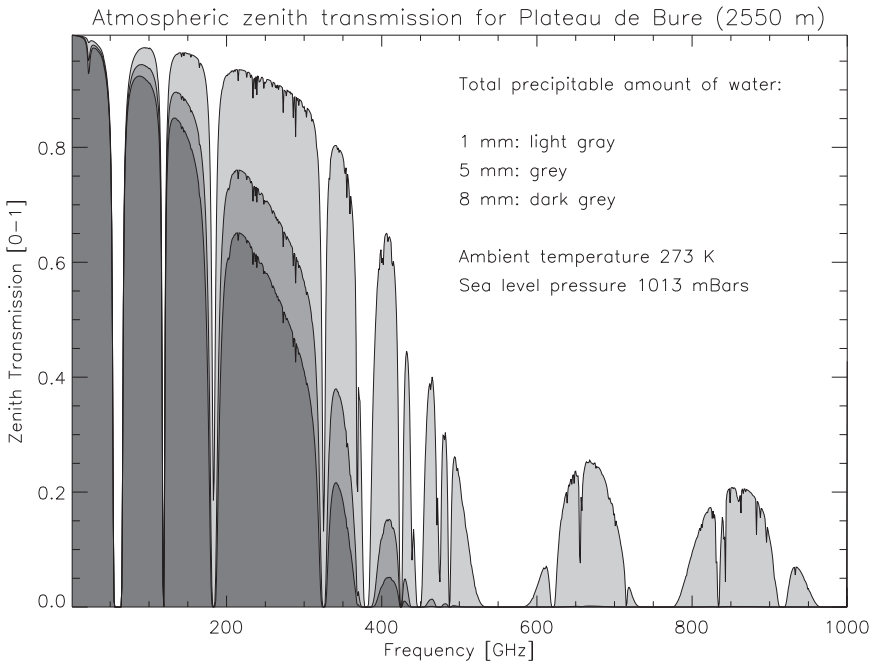


Fig. 1. Atmospheric transmission on Plateau de Bure, France calculated with the ATM model (Pardo *et al.* 2001) for different amounts of total precipitable water. Water vapour has an exponential scale height of only 2 km (dry air has 8.4 km) allowing high altitude sites to extend the edge of the radio window to higher frequencies.

become possible again at the short wavelength edge of the radio window (see Fig. 1), seeing has completely changed its character. It is now slow (0.1 – 0.01 Hz) and moves the whole image around instead of splitting it into speckles. Obviously, radio waves become sensitive to an atmospheric component that does not influence optical wavelengths in the same way. This component is water vapour. Its extended absorption line systems between the optical and radio windows do more than locally absorb emission. For H_2O a total of 114, 241, 164 spectral lines are recorded in the HITRAN V13.0 database (Rothman *et al.* 2009, the number of H_2O transitions is given on the associated web pages); each line introduces a long-range step into the refractive index according to the Kramers-Kronig relation that links the real and imaginary parts of the dielectric constant. When solving the Kramers-Kronig relation for the whole spectrum is impractical, local approximations based on experimental measurements can be employed. Numerical atmospheric radiative transfer codes like ATM (Pardo *et al.* 2001, also available in the ASTRO software within the GILDAS-IRAM package[2]) or AM (Paine 2012) often use fitted pseudo-continua to approximate distant line wings for opacities,

[2] `http://www.iram.fr/IRAMFR/GILDAS`

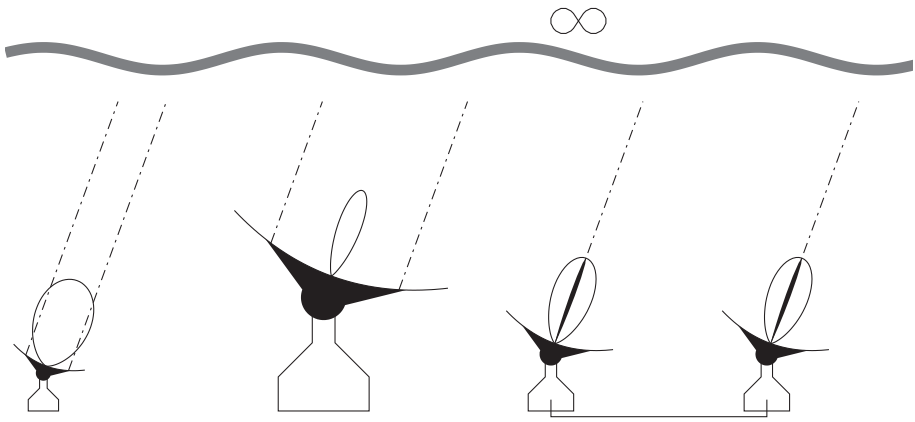


Fig. 2. Ground-based radio observatories observing the (nearly) infinite sky. At a given frequency, small antennas have a wide beam, large antennas a fine beam, and interferometers wide individual beams but narrow synthesised beams.

and may use an empirical step in the refractive index calculation when they reach the edge of their line database (15 THz for AM 7.2). For numerical values for the refractive index in the radio range (without using a model code) see *e.g.* Hill & Cliffort (1981). The dependency of the total refractive index on ambient temperature, dry pressure and partial vapour pressure can be taken into account over the Smith - Weintraub equation (Smith & Weintraub 1953, see also Thompson *et al.* 1986).

Water vapour has several unusual physical properties. Under terrestrial conditions it can change between its gaseous, liquid and solid states, act as an important heat transfer medium, is lighter than dry air but stays in the lower atmosphere due to the negative tropospheric temperature gradient. Other than most atmospheric gasses, water vapour mixes badly with the dry atmosphere and tends to form bubbles of some metres to kilometres in size. Those bubbles have a tendency to merge but are opposed in this by the action of turbulence, which tends to create a weather-dependent power law size distribution of the bubbles and their associated refractive index perturbations that levels off to a constant high level when the outer scale of turbulence is reached. This "phase screen" can move horizontally with the wind, making wind speed an important parameter for phase noise even when the total amount of precipitable water is low. The refractive index variations impact radio observations on two levels (Fig. 2). In single-dish telescopes, it is known as "anomalous refraction" (see Altenhoff *et al.* 1987; Downes & Altenhoff 1990) and typically indicates that the single beam of a telescope seems to move around the requested position. Large radio telescopes have a finer beam resolution than small antennas at the same frequency, and are therefore more sensitive to this phenomenon. In extreme cases this can perturb observations so much that no meaningful data can be obtained: the beam can wander over neighbouring parts of

an extended source, stay an unknown fraction of time off-source for an unresolved target, and can cause severe errors during the pointing and focus calibrations.

For interferometers with their extended baselines, phase variations can be noticed well before the beams of the individual antennas become perturbed (Fig. 3). Also for the synthesised beam, the source seems to wander around slowly (Fig. 4); but other than the single dish anomalous refraction that is sensitive to the total fluctuation of the water vapour column, the phase on a baseline is only sensitive to the difference of the fluctuations along the lines of sight of the connected antennas. This distinction is advantageous for compact interferometer configurations under summer conditions, when the outer scale of turbulence can be in the kilometre range (PdBI baseline lengths in compact 6Dq configuration are $24 - 97$ m).

An important impact of phase noise on the integrated amplitude is expressed in the formula $\|V_{obs}\| = \|V_{real}\| \cdot exp(-\phi^2/2)$ where ϕ is the phase noise in radians, V_{obs} the observed integrated visibility, and V_{real} the visibility in the absence of atmospheric phase noise. As the phase noise scales to good approximation linearly with observing frequency, observing conditions can be prohibitive for band 3 and 4, marginal for band 2, and correct for band 1 (see Fig. 5 for the frequency ranges of the bands). Telescopes that operate in service mode can choose the project that is best adapted to meteorological conditions. But when only fixed telescope time slots are offered to observers, this flexibility does not exist.

Statistically, atmospheric parameters like temperature, wind speed and refractive index variations can be treated as non-stationary random processes. This means that not only individual measurement values but also their averages wander around in time. Classical averages and their variations are in such a situation ill defined. An elegant method to characterise those processes are structure functions, and the view of turbulence as an energy transport mechanism from large to small scales, where the continuously sub-dividing turbulent eddies become finally small enough to dissipate their kinetic energy as heat (Tatarski 1961; Kolmogorov 1941a,b, 1991a,b).

3 Water vapour radiometry

The most reliable way to correct for the phase noise generated by turbulent water vapour is real-time remote monitoring close to the observed line of sight, with a time resolution of the order of telescope diameter divided by typical wind speed. A radiometer is a receiver that is sensitive to the thermal emission of water vapour. In order to distinguish between clouds and gaseous vapour, radiometers are employed that monitor the contrast between the cloud-generated continuum and a water vapour line. The droplets in clouds contribute significantly to the detected emission, but only little to the optical path; their rejection is therefore essential for a meaningful phase correction in the presence of clouds. The increased opacity on a water vapour line causes a stronger coupling to the vertical temperature distribution in the atmosphere, and thus an increase in the power received by the radiometer. Figure 5 shows how precisely the atmospheric emission needs to be measured (in Kelvin) as a function of monitoring frequency to obtain a given

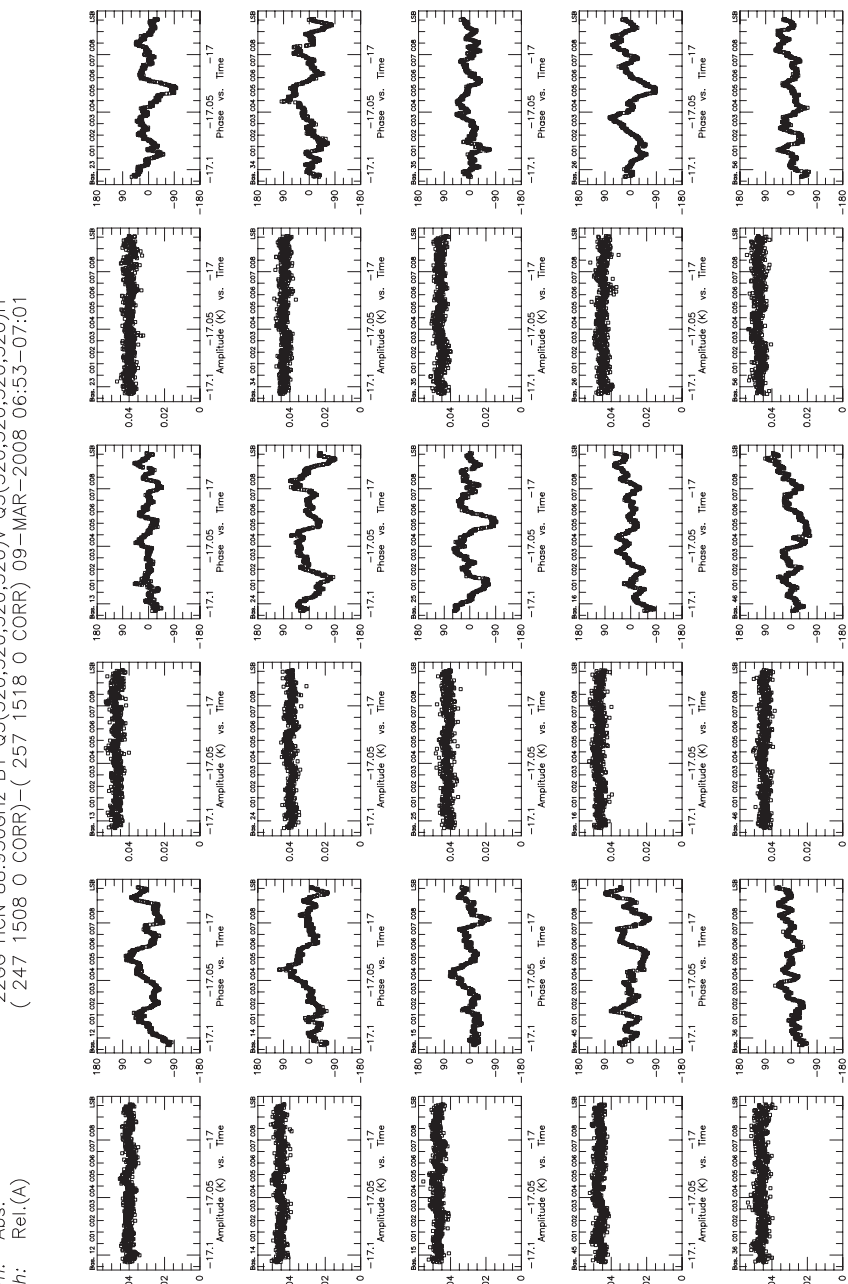


Fig. 3. Amplitudes and phases on a strong point source during 495 s. The stability of the amplitudes shows that individual antenna beams do not wander around on the sky. The phases, however, move significantly.

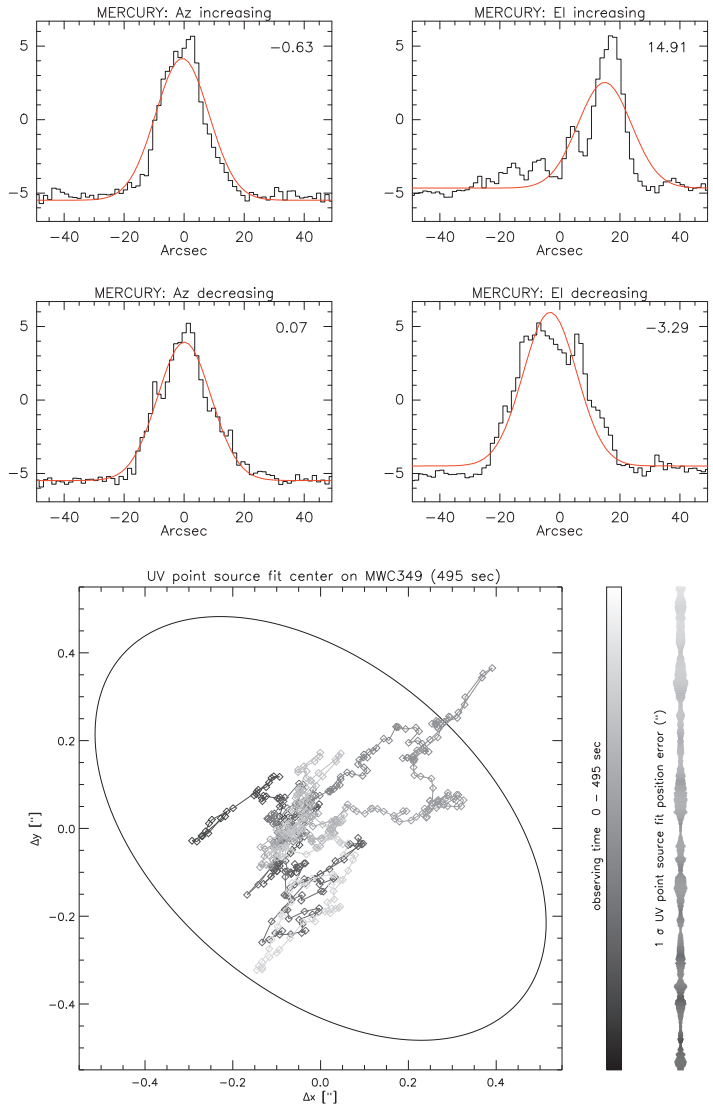


Fig. 4. *Top*: pointing scan at the IRAM 30-m telescope under conditions of anomalous refraction. The scan consists of two drifts in azimuth and elevation over Mercury. *Bottom*: the amplitudes and phases of Figure 3 expressed as the movement of MWC 349 (time sequence coded in grey levels) observed at 88.950 GHz. The 1-sigma error bar (rightmost vertical trace) as a function of time shows that the movement is significant. The ideal clean beam half-power contour is indicated by an ellipse.

path rms. Other criteria must also be taken into account: electronic components of high stability, low price and ambient temperature operation are available for

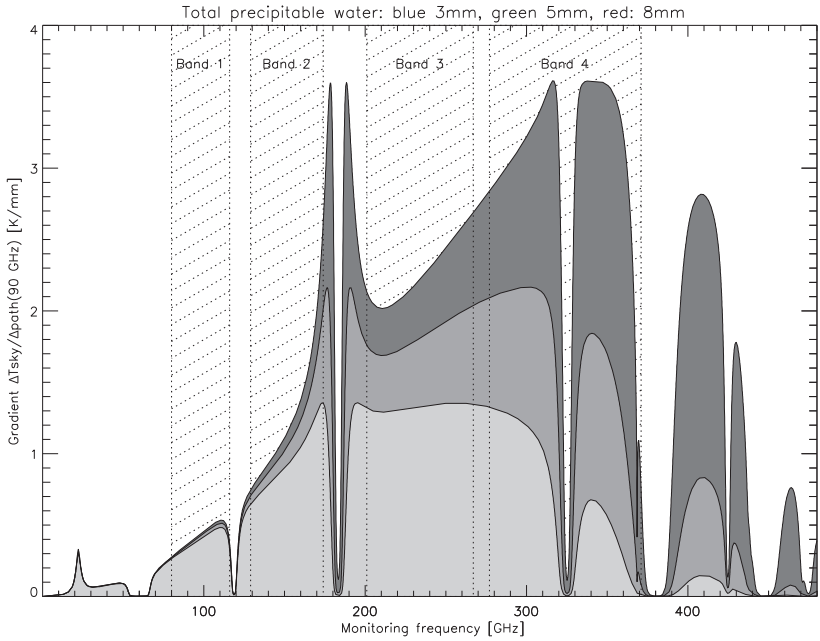


Fig. 5. Gradient $\Delta T_{sky}/\Delta path$ for 3 mm (dark grey) 5 mm (grey) and 8 mm (light grey) total precipitable water, calculated as a function of monitoring frequency for a fixed observing frequency of 90 GHz at Plateau de Bure (2550 m). The observing bands 1–4 of the interferometer are indicated.

lower frequencies while components for high frequencies may require cooling, or need more elaborate calibration techniques to counter drifts in performance. The smaller opacity τ at frequencies lower than the astronomical receiver bands allows to measure the whole tropospheric line of sight under observing conditions, while high monitoring frequencies might reach their $\tau = 1$ layer not far from the antennas and thus stay insensitive to an important fraction of the observing band phase fluctuations. And finally, interference by satellite emitters or telecommunication relays is more common at lower frequencies, with a tendency to rise over the years as main-stream technologies evolve and the requests for increased communication bandwidth become more pressing. Each observatory must therefore carefully choose the type of radiometer that is adapted to its needs. For the Plateau de Bure, the total power signal of the 1 mm astronomical receivers was used between 1995 and mid-2004, after mid-2004 the system was fully switched to dedicated 22 GHz radiometers with cloud correction. These radiometers use three 1 GHz large channels centred on 19.2 GHz, 22.0 GHz and 25.2 GHz and operate at ambient temperature. Figure 6 illustrates the benefit of the radiometric phase correction: The phase noise is not only lower, but also less dependent on baseline length, which helps to conserve the angular resolution of the resulting map. In 2012, the radiometers on Bure were only perturbed by two satellite beacons, but

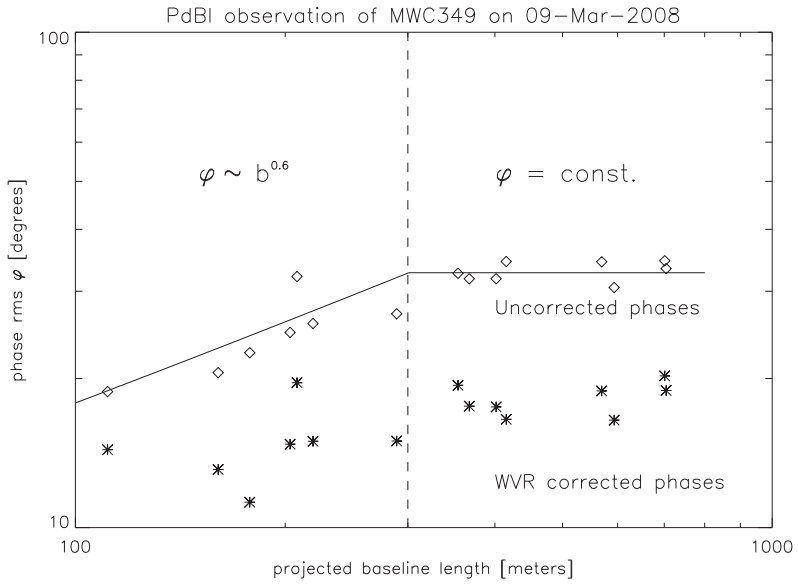


Fig. 6. Baseline-specific phase rms for the observations in Figure 3 *versus* projected baseline length, with and without 22 GHz radiometric phase correction. The vertical dashed line indicates the approximate outer scale size of the turbulent water vapour cells on that day.

it required extended negotiations and the goodwill of several telecommunication operators to keep powerful ground-based signal relays in the 16–24 GHz range away from the neighbourhood of the observatory.

4 Image formation in radio interferometry

Radio interferometry is one of the few domains in science where the Fourier transform of a desired quantity is observed, and needs to be inversely transformed (van Cittert-Zernike theorem, Fraunhofer diffraction). This subject has been treated in depth in many excellent textbooks and conference proceedings (*e.g.* Thompson *et al.* 1986; Perley *et al.* 1986; Taylor *et al.* 1989, 1999). Combining radio telescopes into an interferometer requires a number of basic conditions:

- The antennas must observe the source simultaneously.
- The detected signal is a complex visibility vector that can be expressed as an amplitude and phase, and which is baseline-specific. As a consequence, both antennas must share a common frequency reference against which the phase can be measured. This can be done by either distributing a signal from one master reference to all antennas (connected element interferometer), or to use frequency standards of sufficient stability on each antenna (*e.g.* very long baseline interferometry (VLBI)).

- Individual antennas have their primary beam that corresponds to a single pixel detector, with a diffraction-limited resolution defined by the reflector diameter. However, within this primary beam each combination of two antennas can resolve details with a diffraction-limited resolution of the antenna spacing. As each combination of antennas counts, the number of simultaneous baselines is given by $N_{bas} = N_{ant} \cdot (N_{ant} - 1)/2$, *i.e.* adding one antenna to an array of five increases the number of simultaneous baselines by 50%. As the astronomical source rises, culminates and sets, the apparent orientation and projected length of each baseline changes and thus traces a curved line in the Fourier plane, which is also called UV plane (Earth rotation synthesis).

- Antennas cannot be spaced closer than the average of their diameters, and that means that the baseline "zero" that defines the integral flux cannot be accessed. This is the so-called "zero spacing" problem.

The higher the observing frequency and the larger the observing bandwidth, the more demanding become the technical requirements for the receivers, the signal transport with its delay compensation, and the correlator that needs to process the full spectral bandwidth of all baselines. The WIDEX wide-band correlator currently in service on the Plateau de Bure has a capacity of 914 Tera-operations per second (M. Torres, priv. com.).

5 Gridding and image restoration techniques

During the construction of a local interferometer, some effort has to be invested into the planning of station positions (also called "pads") where a mobile antenna can clamp down, connect to the system and start observing (see *e.g.* Kogan 2000; Boone 2001 & Cohanim *et al.* 2004). The Earth rotation synthesis tracks of each baseline need to fill the UV plane as efficiently as possible for a variety of source declinations, while the total cost for building and cabling the stations must stay within reasonable limits. Over a year an interferometer will undergo a limited number of configuration changes, with closer spaced antennas during turbulent summer conditions and extended spacing during stable and dry winter conditions. An observer can apply to observe his/her source in several configurations to study it in different spatial resolutions, and to combine the result into a single map.

The resulting UV data are not regularly sampled and have no measurements at the centre; also, they are limited in their maximum UV coverage. In order to use fast Fourier Transform (FFT) methods, the observations need to be interpolated to a regular grid. That is typically done by convolving the UV data with a gridding function; the smoothing effect of the convolution is welcome because the individual visibilities are noisy samples of a locally smooth distribution. In order to avoid information loss, the grid spacing needs to be at least Nyquist sampled, and preferentially by a factor of two finer (for a smoother gridded image). The grid extend must be at least two times larger than the useful field of view to avoid folding back part of the valuable image information onto itself.

With the inverse Fourier transform, the image corresponding to the UV data can be obtained. This image is called "dirty image" for a good reason: It shows positive and negative structure, and an integral of zero. Clearly, this data needs further treatment before it can be interpreted scientifically. One possibility is the fitting of UV models to the visibilities; in this case a careful study of the fit residuals is necessary but the resulting fit parameters have well-defined errors. The alternative is image restoration: the image is deconvolved, but then convolved again with an appropriate elliptical Gaussian beam.

In order to perform the deconvolution step, the instrumental point spread function (PSF) needs to be known. In radio interferometry this is the Fourier transform of an ideal point source, seen through the UV tracks of the real observation and thus precisely known. It has the same inconveniences as the observed image (zero integral, positive and negative structure) and is therefore named "dirty beam". Numerous techniques exist to obtain a deconvolved solution.

The most popular algorithm for the deconvolution and restoration of interferometer maps is CLEAN (Högbom 1974). It decomposes the source into a number of point-like components; its steps can be summarised as follows:

1. **Deconvolution:** First, set the residuals map to the dirty map, reset the list of point source components to zero, and choose a loop gain $\gamma = 0.1 \ldots 0.3$ that stays the same during the following steps.
2. Identify the maximum *absolute* value $\|I_{max}\|$in the residual map.
3. Add $\gamma \cdot I_{max}$ and its position to the list of point source components.
4. Subtract $\gamma \cdot \mathrm{PSF}$ at this position from the residual map.
5. Go back to point (2) until the convergence criterion is reached. This can be the maximum number of iteration, a relative noise level expressed in terms of the maximum of the map, or an absolute noise level.
6. **Restoration:** The obtained result needs to be convolved with a *clean beam*, which is defined by fitting an elliptical Gaussian to the central part of the PSF. On scales smaller than the clean beam the deconvolution may create artifacts (no free "super resolution"), and this convolution will remove them.
7. Add the residual map to the clean map. This step is essential to allow noise estimates, and allows to reduce the effect if the cleaning was not deep enough.

It must be clearly stated that the result will be ambiguous and won't have a well-defined error estimate, but it will be closer to the real source structure than the dirty image and therefore a better basis for scientific analysis.

An important feature is to allow negative point source components. They can either appear as intermediate steps in the iteration, or can be a real part of the source structure if there is an absorption against an extended background emission (which may have been suppressed due to the spatial filtering). It is typically a good idea to keep track of the cumulative flux of positive and negative point source

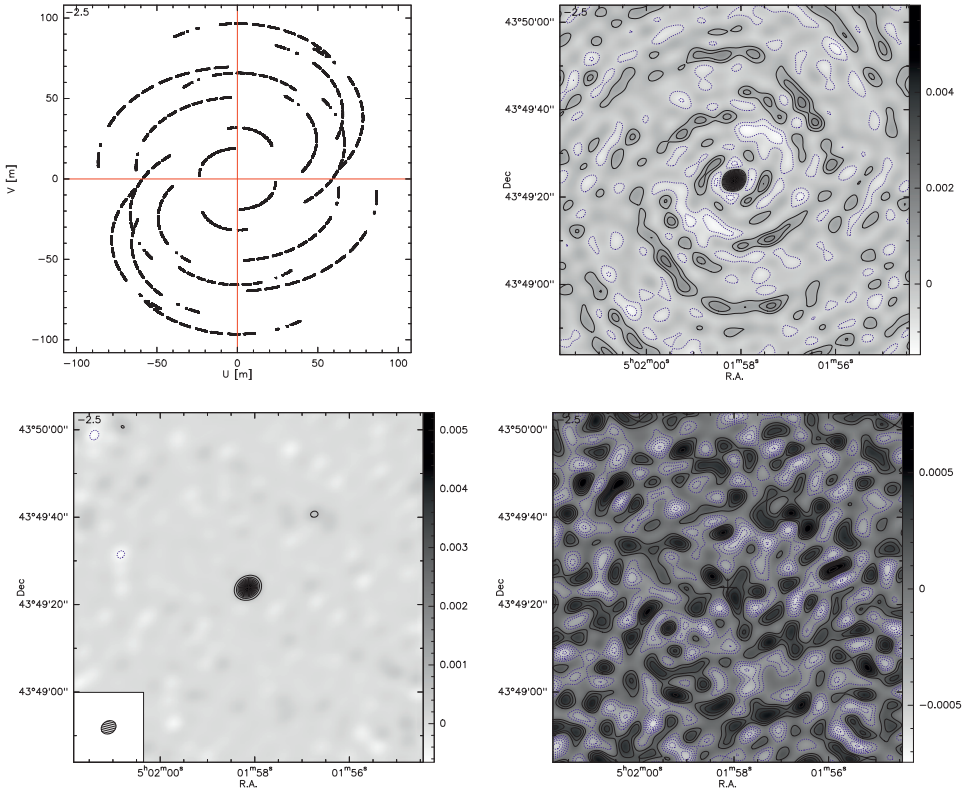


Fig. 7. *Top row*: UV coverage of a 5-antenna configuration, and the resulting dirty image of a point source. *Bottom row*: cleaned image (*left*) and residuals of a point source fit to the phase centre.

components during the iterations: if it reaches a stable level it is a good indicator that the deconvolution has converged. If it starts to oscillate strongly or diverges, the convergence criteria need to be adjusted.

It is possible to observe sources with an interferometer that are more extended than the primary beam of the individual antennas. In this case, the whole interferometer observes repeatedly a grid of pointings called "mosaic" on the sky (again, Nyquist sampling should be used). This technique allows to obtain several fields of view under comparable conditions. In this case it is necessary to obtain information on UV spacings shorter than the interferometer provides. This can be done with single dish observations and (if possible) with observations by an interferometer composed of smaller antennas, otherwise extended source structure will be missed. Already if the source is larger than 1/3 of the primary telescope beam, its interferometer map starts to be affected. Fortunately IRAM operates not only an interferometer but also a large (30-m) single dish telescope that allows to obtain zero spacing observations in an uncomplicated way.

Variants of CLEAN exist that address particular problems, sometimes in the context of overlapping multiple maps (Mosaics) or with an improved restoration of extended structure (see Schwab 1984; Steer *et al.* 1984; Wakker & Schwarz 1988 and Cornwell 2008). Maximum entropy methods (see Richardson 1972; Lucy 1974; Narayan & Nityananda 1986) are often not well adapted to cases where negative values are present in the dirty map. Their use may only be attempted if the dirty beam is nearly Gaussian.

6 Conclusion

Advances in signal transport, receiver and computer technology are important cornerstones of today's knowledge in millimetre astronomy. While the data acquisition and interpretation are less direct than for optical observations and require some experience, modern radio observatories provide an important support (pipeline-reduced data, local contact astronomers) for the scientists who want to use these tools. Over the Internet sites of many observatories (in the case of IRAM, `http://www.iram-institute.org`), astronomers can obtain the latest information on the further evolving instrumental capabilities and download dedicated data processing software. There still remain many secrets to be discovered beyond our planet, sometimes they are just one large step in observing frequency away.

The interferometric imaging, deconvolution and restoration sections of this article discuss material that was presented in much greater depth by J. Pety and F. Gueth in the context of the 7th IRAM Interferometry school (2010) (`http://www.iram-institute.org/EN/content-page-212-7-67-182-212-0.html`).

References

Altenhoff, W.J., Baars, J.W.M., Wink, J.E., & Downes, D., 1987, A&A, 184, 381

Boone, F., 2001, A&A, 377, 368

Cohanim, B.E., Hewitt, J.N., & de Weck, O., 2004, ApJS, 154, 705

Cornwell, T.J., 2008, IEEE J. Selected Topics Signal Proc., 2, 793

Downes, D., & Altenhoff, W.J., 1990, URSI/IAU Symposium on Radio Astronomical Seeing, 31

Hill, R.J., Cliffort, S.F., 1981, Radio Science, 16, 77

Högbom, J.A., 1974, A&AS, 15, 417

Kogan, L., 2000, IEEE Trans. Antennas Propagation, 48, 1075

Kolmogorov, A.N., 1941a, Proc. USSR Acad. Sci., 30, 299 (Russian)

Kolmogorov, A.N., 1941b, Proc. USSR Acad. Sci., 32, 16 (Russian)

Kolmogorov, A.N., 1991a, Proceedings of the Royal Society of London, Series A: Math. Phys. Sci., 434, 9 (translated to English by V. Levin)

Kolmogorov, A.N., 1991b, Proceedings of the Royal Society of London, Series A: Math. Phys. Sci., 434, 15 (translated to English by V. Levin)

Lohmann, A.W., Weigelt, G., & Wirnitzer, B., 1983, Appl. Opt., 22, 4028

Lucy, L.B., 1974, AJ, 79, 745

Narayan, R., & Nityananda, R., 1986, ARA&A, 24, 127

Paine, S., 2012, SMA Technical Memo, No. 152

Pardo, J.R., Cernicharo, J., & Serabyn, E., 2001, IEEE Trans. Antennas Propagation, 49, 1683

Perley, R.A., Schwab, F.R., Bridle, A.H., 1986, "Synthesis Imaging", NRAO Workshop, No. 13

Richardson, W.H., 1972, J. Opt. Soc. Am. (1917-1983), 62, 55

Rothman, L.S., Gordon, I.E., Barbe, A., *et al.*, 2009, J. Quant. Spectrosc. Radiative Transfer, 110, 533 (web pages `http://www.cfa.harvard.edu/hitran/`)

Schwab, F.R., 1984, AJ, 89, 1076

Smith, E.K., & Weintraub, S., 1953, Proc. IRE, 41, 1035

Steer, D.G., Dewdney, P.E., & Ito, M.R., 1984, A&A, 137, 159

Tatarski, V.I., 1961, "Wave propagation in a turbulent medium" (McGraw-Hill New York, Toronto, London)

Taylor, G.B., Carilli C.L., & Perley, R.A., 1989, Astron. Soc. Pacific Conf. Ser., Vol. 6

Taylor, G.B., Carilli C.L., & Perley, R.A., 1999, Astron. Soc. Pacific Conf. Ser., Vol. 180

Thompson, A.R., Moran, J.M., & Swenson, G.W., 1986, "Interferometry and Synthesis in Radio Astronomy"

Wakker, B.P., & Schwarz, U.J., 1988, A&A, 200, 312

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 203–212*

SMOS-NEXT: A NEW CONCEPT FOR SOIL MOISTURE RETRIEVAL FROM PASSIVE INTERFEROMETRIC OBSERVATIONS

Y. Soldo[1,2], F. Cabot[1,2], B. Rougé[2], Y.H. Kerr[1,2], A. Al Bitar[1] and E. Epaillard[1]

Abstract. Present soil moisture and ocean salinity maps retrieved by remote sensing are characterized by a coarse spatial resolution. Hydrological, meteorological and climatological applications would benefit greatly from a better spatial resolution. Owing to the dimensions of the satellite structure and to the degradation of the instrument's radiometric sensitivity, such improvement cannot be achieved with classical interferometry. Then, in order to achieve this goal an original concept for passive interferometric measurements is described. This concept should allow to achieve a much finer spatial resolution, which can be further improved with the application of disaggregation methods. The results will then allow the integration of global soil moisture maps into hydrological models, a better management of water resources at small scales and an improvement in spatial precision for various applications.

1 Introduction

During the last decades the need for a global estimation with high temporal resolution of key environmental variables such as soil moisture and ocean salinity has grown greatly (Robock *et al.* 2000; Dai *et al.* 2004; Roemmich *et al.* 2000).

Satellites represent the best mean for satisfying such need, and several instruments have been launched onboard European and American satellites with the intent of retrieving large-scale soil moisture and ocean salinity maps.

These instruments are based on different principles. They may involve radiometers (Njoku *et al.* 2003), scatterometers (Bartalis *et al.* 2007), interferometric radiometers (Kerr *et al.* 2001), or they may rely on both passive and active elements (LeVine *et al.* 2007; Entekhabi *et al.* 2010).

[1] Centre d'Études Spatiales de la Biosphère (CESBIO), 18 avenue Edouard Belin, bpi 2801, 31401 Toulouse Cedex 9, France; e-mail: yan.soldo@cesbio.cnes.fr

[2] Centre National d'Études Spatiales (CNES), 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France

DOI: 10.1051/eas/1359011

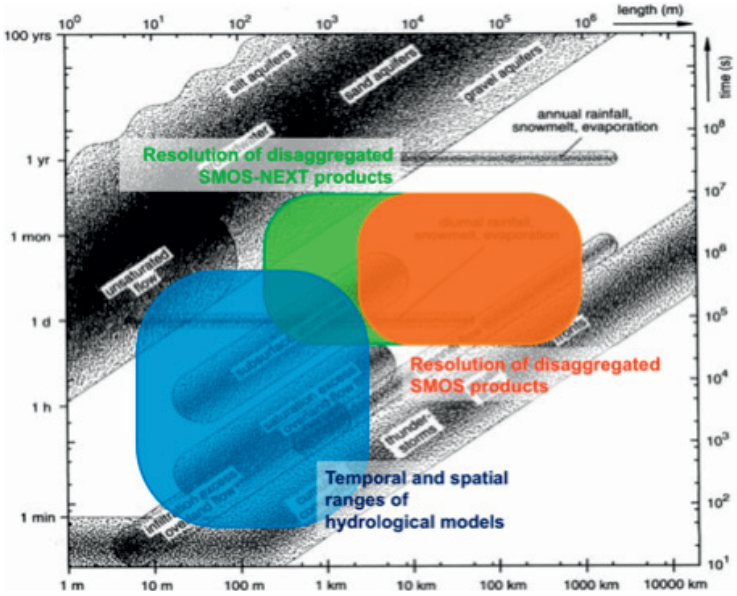


Fig. 1. Characterization of spatial and temporal scales for hydrological models and disaggregated remote sensed data.

Although different technologies were adopted, all these instruments are limited by a spatial resolution of few tens of kilometers.

In order to be able to make use of these data in hydrological models, and for many other applications, like the survey of water resources at the scale of irrigated zones, a better spatial resolution must be achieved, typically it should be improved by an order of magnitude.

To assure continuous monitoring of soil moisture and ocean salinity, while attaining an unprecedented fine spatial resolution, an original concept was proposed in Cabot *et al.* (2012) and Kerr *et al.* (2010), which aims at achieving a spatial resolution of few kilometers while maintaining roughly the same radiometric resolution.

2 The need for high spatial resolution

The management of water resources is already, and will be even more in the future, a critical issue (Alcamo *et al.* 2000; Döll *et al.* 2003). Properly dealing with this issue cannot be done without a deep understanding of the processes involved in the water cycle, which are studied using hydrological models.

Temporal and spatial scales can vary significantly from one model to the other depending on the processes they focus on. However, by improving the spatial resolution, the interaction between remotely sensed global soil moisture maps and hydrological models will be more effective and will result in an improved

environmental knowledge. Figure 1 shows how disaggregation of SMOS-NEXT retrievals will result in a better cooperation with hydrological models.

Combined with weather models, remotely sensed global soil moisture maps can help achieve more accurate forecasting predictions, as well as to assess the risk for fires or floods on specific areas. Storms or heavy precipitations are of course more likely to cause floods over moist soils, and winds over very dry soils will increase the risk for fires.

Over ocean, the salt content will be retrieved. Indeed, ocean salinity's annual and inter-annual variations are crucial for monitoring and understanding of climate and climate changes, as they influence ocean currents and water evaporation from oceanic surfaces. A better resolution will improve the capability to follow in more details how currents vary with time as well as how river plumes interact with these oceanic currents.

Other than the nominal uses, SMOS has proved to be a versatile satellite, as its data has been used also for applications like wind speed estimation inside tornadoes (Grodsky *et al.* 2012) or the monitoring of the extent of sea ice sheets (Kaleschke *et al.* 2012). Naturally all these applications will benefit from a finer spatial resolution.

3 Operating frequency and spatial resolution

The maximum sensitivity to both soil moisture and ocean salinity is close to the protected 1400–1427 MHz band, and atmospheric disturbances are negligible at these frequencies (Wigneron *et al.* 2000), thus for a passive instrument like SMOS-NEXT, this wave band is clearly the best choice in term of operating frequency.

Even though artificial emissions are forbidden in this band to allow passive observations of both Earth and sky (ITU Radio Regulations 1996), after the first SMOS' data retrievals, the presence of contaminating unlawful sources was noticed (Anterrieu & Khazaal 2011), so a strategy has been developed to deal with these radio frequency interferences that should provide a cleaner signal. The detailed description of this strategy is out of the scope of this contribution.

Once the operating frequency has been fixed, there are only two other parameters that define the spatial resolution (R_s)

$$R_s = \frac{H\lambda}{d} \tag{3.1}$$

H, the satellite altitude, and d, the diameter of the equivalent real aperture antenna, that in our case is equal to the maximum baseline (λ is the wavelength of the central operating frequency).

The spatial resolution could be improved by reducing the altitude of the satellite. However, the choice of altitude is also driven by constraints linked to the width of the swath and to the density of the atmosphere which determines the fuel consumption, and hence the weight at launch. All similar missions are orbiting, or are planned to orbit, at altitudes of about 700 km, so it is safe to assume that a similar value will be chosen.

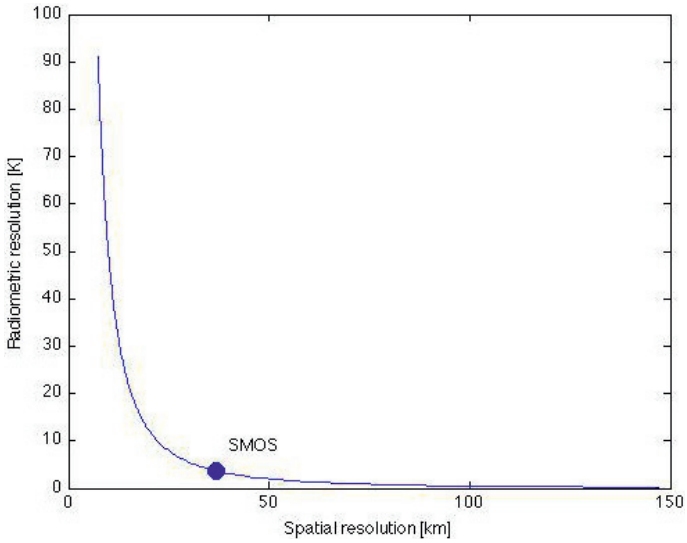


Fig. 2. Interdependence between spatial resolution and radiometric sensitivity; SMOS' maximal baseline is already close to the optimal trade off.

Consequently, the only way to improve the resolution is to increase the length of the baselines.

However, longer baselines means bigger surfaces of the equivalent real aperture antenna, and that is detrimental to the radiometric resolution (ΔT) (Camps *et al.* 1998) according to:

$$\Delta T = A \frac{T_A + T_{rec}}{\sqrt{B t_i}} \sqrt{N_V} \tag{3.2}$$

where A is the pixel area, T_A is the antenna temperature, T_{rec} is the receiver temperature, B is the spectral bandwidth, t_i is the integration time interval and N_V is the number of points sampled by the array in the Fourier domain.

To improve the spatial resolution by an order of magnitude means to have baselines ten times bigger. For SMOS this would lead to three 40 m long arms, which represents obvious feasibility difficulties. Moreover, as the spatial resolution vary linearly with d, the radiometric resolution is proportional to the square of d, through N_V (for a Y-shaped instrument $N_V = 6N_{el}^2 + 6N_{el} + 1$, with N_{el} the number of receivers per arm); hence longer baselines would lead to a loss of radiometric sensitivity, which is unacceptable with respect to the very stringent requirements for oceanic observations (Berger *et al.* 2002).

Because of this relation between spatial resolution and radiometric sensitivity, this classical approach can hardly lead to the improvement of one without the degradation of the other (see Fig. 2).

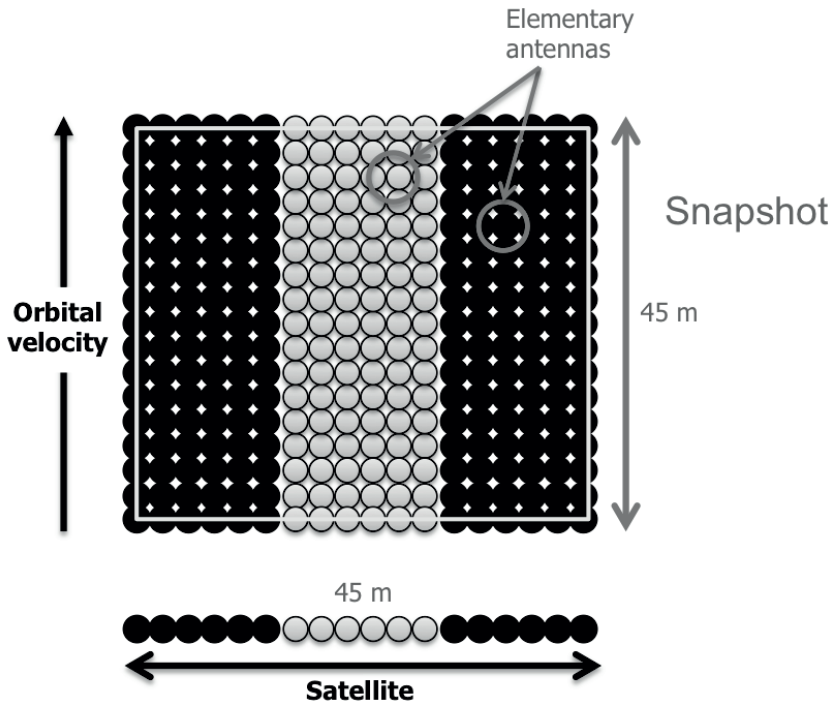


Fig. 3. Principle of spatio-temporal aperture synthesis: the phase differences are due to the difference in space and time between antennas.

4 Spatio-temporal interferometry

In order to improve the spatial resolution while maintaining roughly the same radiometric resolution, an original solution was proposed in Cabot *et al.* (2012) and Kerr *et al.* (2010). It consists in using observations made by a set of antennas at different times. The temporal coherence of a signal with spectral bandwidth B is defined as:

$$\tau = \frac{1}{B}. \tag{4.1}$$

With a fine filtering it is possible to select bandwidths as small as 100 Hz, which results in coherence time of 0,01 s. At the orbital speed, the satellite travels, within the coherence time, a distance of 75 m, *i.e.* more than what it is required for a snapshot.

In this condition it is possible for the satellite to observe at two different times, two signals that are coherent with one another. Based on this principle, that was studied to depth in Braun 2011, SMOS-NEXT would be a 1-D interferometric radiometer, whose second dimension is given by the movement of the satellite on its orbit. In other words, signal acquired by the $i-th$ elementary antenna at the

instant t_0 would be correlated (also) to the acquisitions of the same antenna at later times $t_0 + \Delta t$ (Fig. 3).

In this way the relation between spatial and radiometric resolution is less strict and it is possible to meet the requirements for both.

In fact with the elementary antennas along the arm and the satellite's motion, a set of both real and virtual antennas (*i.e.* real antennas at later times) is created, and while the spatial resolution is assured by the baselines in the two directions (along the satellite's arm and along the movement of the satellite), the radiometric resolution is only function of the real elementary antennas along the arm.

5 Preliminary design

SMOS-NEXT requires long physical baselines. But a satellite with three 40 m long arms disposed in a Y-shape is not a technologically realistic solution for the time being. Nevertheless it is possible to launch and deploy on orbit a single 40–45 m long arm. This is one of the two design options considered today, the other solution consisting in two satellites flying in formation.

Both solutions represent technical difficulties but the second one would imply a spatial resolution that depends on the relative distance between the satellites, which shortens at high latitudes, when the orbital planes cross. So we will consider only the first solution here on.

From the satellite's altitude, we can calculate its mean velocity:

$$V_{sat} = \sqrt{\frac{\mu}{a}} \tag{5.1}$$

where μ is the standard gravitational parameter equal to $398\,600\ \mathrm{km^3\,s^{-2}}$, and a is the semi-major axis of the orbit. For a circular orbit and an altitude of 700 km, V_{sat} is roughly equal to $7500\ \mathrm{ms^{-1}}$.

The arm will be filled with elementary antennas spaced, as for SMOS, by 0,875 wavelengths. If the maximum redundancy configuration is chosen the total number of antennas will be roughly 250.

Sampling frequency (F_s) will be such that the spacing between real and virtual antennas ($\Delta s_{antennas}$) is at least equal to the spacing between real antennas, that is:

$$F_s \geq \frac{V_{sat}}{\Delta s_{antennas}} = \frac{7500\ \mathrm{ms^{-1}}}{0,875\lambda} = \frac{7500\ \mathrm{ms^{-1}}}{0,875 \cdot 0,21\ \mathrm{m}} \approx 40\ \mathrm{kHz}. \tag{5.2}$$

A snapshot is then defined as the sequence of acquisitions that the satellite makes in a time interval corresponding to a displacement of about 45 m. The time interval required for the satellite to cover this distance (0,006 s) must be lower than the coherence time for the chosen bandwidth (0,01 s). Each snapshot will then be composed by 240 or more acquisitions.

6 Cross-correlations between antennas

The expression of an electromagnetic signal for a set of virtual and real antennas can be expressed in function of time and space as in:

$$\frac{S_0(s,\nu)}{\rho(s)} = e^{-2j\pi\nu(t-\frac{\rho(s,t)}{c})} \tag{6.1}$$

where S_0 is the electromagnetic signal, ν is the signal's frequency, $\rho(s,t)$ is the source-antenna distance, t is the time and s indicates the source's position.

Under the assumption that sources are spatially incoherent the cross-correlations between the two electromagnetic fields at the antennas' positions can be written as

$$\langle S_0(s,\nu)S_0^*(s,\nu)\rangle \frac{e^{-2j\pi\nu(t_1-\frac{\rho_1(s,t_1)}{c})}}{\rho_1(s)} \frac{e^{-2j\pi\nu(t_2-\frac{\rho_2(s,t_2)}{c})}}{\rho_2(s)} \tag{6.2}$$

where indexes 1 and 2 indicate the two antennas.

Even though the source-antenna distance varies with time due to Earth's oblateness and orbital ellipticity we can consider it independent from time during a snapshot.

If we consider the case in which 1 and 2 represent the same antenna at different times, then we have:

$$\langle S_0(s,\nu)S_0^*(s,\nu)\rangle \frac{e^{-2j\pi\nu(t_1-\frac{\rho_1(s,t_1)}{c})}}{\rho_1(s)} \frac{e^{-2j\pi\nu(t_1+\Delta t-\frac{\rho_2(s,t_1+\Delta t)}{c})}}{\rho_2(s)} \tag{6.3}$$

$$\langle S_0(s,\nu)S_0^*(s,\nu)\rangle \frac{e^{-2j\pi\nu(t_1-\frac{\rho_1(s,t_1)}{c})}}{\rho_1(s)} \frac{e^{-2j\pi\nu(t_1-\frac{\rho_2(s,t_1+\Delta t)}{c})}}{\rho_2(s)} e^{2j\pi\nu\Delta t}. \tag{6.4}$$

We have then obtained the expression of the van Cittert-Zernike theorem multiplied by an exponential term.

In fact the term $\langle S_0(s,\nu)S_0^*(s,\nu)\rangle$ is simply the intensity of the electromagnetic radiation of the source, noted T_B, and by applying the far field approximation and the quasi monochromatic approximation, the product of the phase terms can be expressed as a function of the direction cosines (ξ,η) as follows:

$$\frac{e^{-2j\pi\nu(t_1-\frac{\rho_1(s,t_1)}{c})}\rho_1(s)e^{-2j\pi\nu(t_1-\frac{\rho_2(s,t_1+\Delta t)}{c})}\rho_2(s)}{\rho_1(s)\rho_2(s)} \simeq e^{-2j\pi\nu\frac{(d_1\xi+d_2\eta)}{c}}. \tag{6.5}$$

Integrating over the observation area we have:

$$V = \iint\limits_{\xi^2+\eta^2<1} \frac{T_B(\xi,\eta)}{\sqrt{1-\xi^2-\eta^2}} e^{-2j\pi\nu\frac{(d_1\xi+d_2\eta)}{c}} e^{2j\pi\nu\Delta t} \mathrm{d}\xi\mathrm{d}\eta = \tag{6.6}$$

$$= e^{2j\pi\nu\Delta t} \iint\limits_{\xi^2+\eta^2<1} \frac{T_B(\xi,\eta)}{\sqrt{1-\xi^2-\eta^2}} e^{-2j\pi\nu\frac{(d_1\xi+d_2\eta)}{c}} \mathrm{d}\xi\mathrm{d}\eta \tag{6.7}$$

Aside from the term $e^{2j\pi\nu\Delta t}$, in this expression the visibility (V), corresponding to the cross-correlation between antennas, is expressed as the two-dimensional Fourier transform of the so-called modified brightness temperature map of the source, which is the brightness temperature divided by the obliquity factor $\sqrt{1-\xi^2-\eta^2}$ (Camps *et al.* 1998).

So far we only considered the central frequency of the signal. The integration of (6.7) on the filter's bandwidth (B) can be written dropping the double integral since it is independent from frequency.

$$\int_{\nu-\mathrm{B}/2}^{\nu+\mathrm{B}/2} e^{2j\pi\nu\Delta t} = e^{2j\pi\nu\Delta t} sinc(B\frac{u\xi+v\eta}{\nu}) \tag{6.8}$$

where

$$sinc(B\frac{u\xi+v\eta}{\nu}) = \tilde{r}(\xi,\eta) \tag{6.9}$$

is called the Fringe Washing Function. By integrating 6.8 and 6.9 in 6.7 we obtain

$$= e^{2j\pi\nu\Delta t} \iint_{\xi^2+\eta^2<1} \frac{T_B}{\sqrt{1-\xi^2-\eta^2}} \tilde{r} e^{-2j\pi\nu\frac{(d_1\xi+d_2\eta)}{c}} \mathrm{d}\xi\mathrm{d}\eta. \tag{6.10}$$

As soon as we consider real antennas, their radiation patterns ($F(\xi,\eta)$) and their corresponding solid angles (Ω) must be taken into account. In the case under study (same antenna, different times) the final expression is then written as follows

$$V = e^{2j\pi\nu\Delta t} \iint_{\xi^2+\eta^2<1} \frac{F(\xi,\eta)^* F(\xi,\eta)}{\Omega(\xi,\eta)} \frac{T_B(\xi,\eta)}{\sqrt{1-\xi^2-\eta^2}} \tilde{r} e^{-2j\pi\nu\frac{(d_1\xi+d_2\eta)}{c}} \mathrm{d}\xi\mathrm{d}\eta \tag{6.11}$$

where $F^*(\xi,\eta)$ represents the complex conjugate of $F(\xi,\eta)$.

7 Detemporalization

Previously we made the choice of selecting a 100 Hz band. Even though possible, this solution represent technical difficulties, and in order to use wider range of the protected band several bandwidths of this amplitude would be needed, thus multiplying the quantity of information to be downlinked to the ground stations.

A different approach is therefore proposed. It consists in using larger bandwidths and applying a temporal shift to the signal received by one of the antennas.

This approach is called *detemporalization*.

This is implemented by multiplying by $e^{-2j\pi\nu\Delta t}$ the phasor describing the electromagnetic field at time t_2 that appear in the precedent expression. Following the same development, we obtain hereafter the expression for the visibility function:

$$V = \frac{1}{\Omega} \iint_{\xi^2+\eta^2<1} F^* F \frac{T_B}{\sqrt{1-\xi^2-\eta^2}} \tilde{r} e^{-2j\pi\nu\frac{(d_1\xi+d_2\eta)}{c}} \mathrm{d}\xi\mathrm{d}\eta. \tag{7.1}$$

That is the fundamental relationship between visibility and brightness temperature used for SMOS, before considering the effects of antenna radiation patterns, and before integration with respect to the frequency.

8 Disaggregation

This information on ground will then provide brightness temperature global maps, and using several observations the soil moisture can be retrieved. The spatial resolution of these maps will be, using the data explicated above:

$$R_s = \frac{H\lambda}{d} < 4 \text{ km}. \tag{8.1}$$

This result represents a significant improvement with respect to the data available for the time being, but still it is not sufficient for integration with the hydrological models. In order to do so, this data needs to be downscaled further. Disaggregation methods allow downscaling of soil moisture microwave measurements, by making use of the knowledge of the evaporative fractions over specific areas, that are retrieved by optical, near-infrared or thermal infrared measurements (Merlin *et al.* 2008).

9 Conclusions

The objective of improving the spatial resolution of soil moisture and ocean salinity maps by an order of magnitude can be achieved with the use of a long baseline spatio-temporal interferometer.

The detemporalization technique was then introduced to ease the technical constraints of such instrument.

The resolution obtained is not yet sufficient for the implementation in hydrological models and in future weather models, in which the spatial resolution will be improved. Then disaggregation methods can then be used to downscale further space borne microwave soil moisture retrievals.

Theoretical studies have been conducted to study the principle of the spatio-temporal aperture synthesis, and experimental campaigns are going to be carried out in the near future.

References

Alcamo, J., Henrichs, T., & Rösch, T., 2000, "World Water in 2025 - Global modeling and scenario analysis for the World Commission on Water for the 21st Century", Kassel World Water Series 2, Center for Environmental Systems Research (University of Kassel, Germany)

Anterrieu, E., & Khazaal, A., 2011, "One Year Of RFI Detection And Quantification With L1a Signals Provided By SMOS Reference Radiometers", Geoscience and Remote Sensing Symposium (IGARSS) 2011 IEEE International, 2245

Bartalis, Z., Wagner, W., Naeimi, V., *et al.*, 2007, Geophys. Res. Lett., 34

Berger, M., Camps, A., Font, J., *et al.*, 2002, "Measuring Ocean Salinity with ESA's SMOS Mission", ESA Bulletin, No. 111, 113

Braun, D., 2011, "Physical analysis of spatial and temporal correlations for SMOS-next"

Cabot, F., Kerr, Y.H., Anterrieu, E., *et al.*, 2012, "SMOS-Next: New perspectives for Soil Moisture and Ocean Salinity from space", Microrad (Frascati, IT, 5–9 March 2012)

Camps, A., Corbella, I., Bara, J., & Torres, F., 1998, IEEE Trans. Geosci. Remote Sensing, 36, 680

Dai, A., Trenberth, K.E., & Qian, T., 2004, J. Hydrometeor., 5, 1117

Döll, P., Kaspar, F., & Lehner, B., 2003, J. Hydrology, 270, 105

Entekhabi, D., Njoku, E.G., O'Neill, P.E., *et al.*, 2010, Proc. IEEE, 98, 704

Grodsky, S.A., Reul, N., Lagerloef, G., *et al.*, 2012, Geophys. Res. Lett., L20603

ITU Radio Regulations, 1996, Frequency Allocation National Bulletin, Official Bull. Span State, Aug. 9

Kaleschke, L., Tian-Kunze, X., Maass, N., Mäkynen, M., & Drusch, M., 2012, Geophys. Res. Lett., L05501

Kerr, Y.H., Waldteufel, P., Wigneron, J.-P., *et al.*, 2001, Geosc. Remote Sensing, IEEE Trans., 39, 1729

Kerr, Y.H., Rougé, B., Cabot, F., *et al.*, 2010, "SMOS NEXT: The new generation", Microrad (Washington DC, USA, 1–4 March 2010)

LeVine, D.M., Lagerloef, G.S.E., Colomb, F.R., Yueh, S.H., & Pellerano, F.A., 2007, Geosc. Remote Sensing, IEEE Trans., 45, 2040

Merlin, O., Chehbouni, A., Walker, J.P., Panciera, R., & Kerr, Y.H., 2008, Geosc. Remote Sensing, IEEE Trans., 46, 3

Njoku, E.G., Jackson, T.J., Chan, T.K., & Nghiem, S.V., 2003, Geosc. Remote Sensing, IEEE Trans., 41, 215

Robock, A., Vinnikov, K.Y., Srinivasan, G., *et al.*, 2000, Bull. Amer. Meteor. Soc., 81, 1281

Roemmich, D., & Owens, W.B., 2000, Oceanography, 13, 45

Wigneron, J.P., Chanzy, A., Waldteufel, P., *et al.*, 2000, "Retrieval capabilities of L-Band 2-D interferometric radiometry over land surfaces (SMOS Mission)" (VSP, The Netherlands)

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 213–262*

FORMATION, SIMULATION AND RESTORATION OF HYPERTELESCOPES IMAGES

D. Mary[1], C. Aime[1] and A. Carlotti[2]

Abstract. This article first provides a historical and detailed introduction to the image formation models for diluted pupils array and their densified versions called *hypertelescopes*. We propose in particular an original derivation showing that densification using a periscopic setting like in Michelson's 20− foot interferometer, or using inverted Galilean telescopes are fully equivalent. After a review based on previous reference studies (Tallon & Tallon-Bosc 1992; Labeyrie 1996; Aime 2008 and Aime *et al.* 2012), the introductory part ends with a tutorial section for simulating optical interferometric images produced by cophased arrays. We illustrate in details how the optical image formation model can be used to simulate hypertelescopes images, including sampling issues and their effects on the observed images.

In a second part of the article, we address the issue of restoring hypertelescope images and present numerical illustrations obtained for classical (constrained Maximum Likelihood) methods. We also provide a detailed survey of more recent deconvolution methods based on sparse representations and of their spread in interferometric image reconstruction.

The last part of the article is dedicated to two original and numerical studies. The first study shows by Monte Carlo simulations that the restoration quality achieved by constrained ML methods applied to photon limited images obtained from a diluted array on a square grid, or from a densified array (without spectral aliasing) on a grid, are essentially equivalent. The second study shows that it is possible to recover in hypertelescopes images quasi point sources that are not only far outside the clean field, but also superimposed on the replicas of other objects. This is true at least for the considered pupil array and in the limit of vanishing noise.

[1] Laboratoire Lagrange, UMR 7293, Université de Nice Sophia-Antipolis, CNRS, Observatoire de la Côte d'Azur, Campus Valrose, 06108 Nice Cedex 02, France

[2] Princeton University, Mechanical & Aerospace Engineering, Olden Street, Princeton, NJ 08544, USA

DOI: 10.1051/eas/1359012

1 Introduction

The History of Science is that of a continuous quest for a better understanding of Nature. In particular, the history of Astronomy reflects the breakthroughs which have lead to our modern conception of the Universe.

As researchers, the participants of the Fréjus School are probably all aware of the uncountable efforts that must be spent on various tasks − huge bibliographical studies, long and difficult theoretical calculations or ultra sensitive, and thus ultra irritating instrumental experiments − in order to obtain a single useful, correct and well understood result. These efforts constitute an invisible but necessary sand, which is paved by the bright success stories published in the official History of Sciences.

Success stories often exert on researchers the ambiguous attraction of perfect things: their essence is so bright, exemplary and rare that it appears often discouraging, not to say ridiculous, to imagine any comparison between such major achievements and research of its own. On the other hand, these stories tell us that research efforts need to sum up substantially before major works really happen to culminate. Besides, past successes constantly diffuse, as a constant background Moon for research groping in the darkness, a faint light of scientific glory that shines down into the deepest and most obscure offices of every research laboratory. We start with two such stories, which are connected to the topics of this article, Hypertelescopes.

One of the fundamental concern of Astronomy, of which we can find traces in Mesopotamian, Egyptian, Greek and Arabic Astronomy, is that of high angular resolution. A higher resolution of celestial objects means that the objects can be better understood because they are better seen. High resolution once allowed Mankind to discover that the celestial objects change and move. But Mankind had to wait a very long time before it was able to prove by observations that planets do not move in perfect circles, and that stars are not fixed on an hypothetic celestial sphere beyond which, as Aristotle once wrote, nothing shall exists, not even space or time. This brings us to our first story.

Probably because his own observations of the stars' positions relatively to each other did not match those of his predecessors, and because he had guessed that stars' positions and magnitudes could be variable[3], the Greek astronomer and mathematician Hipparchus (2$^{\text{nd}}$ Century BC) collected the positions and apparent magnitudes of about a thousand stars. This catalogue, which constitutes one of the most audacious legacy to future research, was transmitted to Arabic and European astronomers via copies of the *Almagest* of Ptolemy.

This major book[4] was written four centuries after Hipparchus' time, in the 2$^{\text{nd}}$ Century AD. After the prestigious, several-century-old Library of Alexandria was

[3] According for instance to Halley (1715).

[4] Ptolemy's geocentric model of the Universe, using spheres and epicycles, was based on previous models from Eudoxus of Cnidus (4$^{\text{th}}$ Century BC), Appolonius of Perga (3$^{\text{rd}}$ Century BC) and Hipparchus. This model constituted the standard model and was continuously made more complex until N. Copernic (1473-1543).

definitively destroyed, the Almagest was copied, studied and enriched in various places and epochs during one millenium, from the new capital Constantinopolis of the Orient Roman Empire, through Arabic observatories as Maragha or the active translation centers of Andalousia.

Quite a few years after Hipparchus' times – in the 1710 s – the British astronomers Halley and Flamsteed could show, by comparing their observations to those of Hipparchus, that the position of Sirius in the sky had moved (by proper motion) of an angle of about one lunar diameter since the Hellenistic period (as well as Arcturus and Aldebaran (Mignard & Martin 1997)). After almost 2000 years, Hipparchus had won his bet: his observations had finally served his successors in proving facts he could only suspect. Patient and repeated high angular resolution observations eventually forced the minds to open on a Universe totally different from what the most brilliant scientists of the Antiquity could imagine.

The history of Hipparchus' catalog is not a success, it is an absolute and multiple triumph. First, contributing to the observational evidence that stars could be moving and variable was bringing strong arguments against the old vivid idea that they were inherently immobile and ever lasting; this discovery had deep, cosmological and philosophical implications about the size and nature of the Universe. Second, the careful attention paid by each link of this long chain to the knowledge of his predecessors (Hipparchus to the Mesopotamian tables, Ptolemy to Hipparchus, Al-Tusi, many other Arabic astronomers, Halley and Flamsteed to Ptolemy) examplifies how new science may succeed in building carefully on the experience of the past. Third, the project realized by Hipparchus with his catalogue comes as a striking remembrance that research thought in the long term (several centuries in this case) is not necessarily wasted research.

The second story is also about breakthroughs in experimental high angular resolution and about progressive accumulation of knowledge. It starts one day of 1868, at the Academy of Sciences in Paris, where H. Fizeau reports about a treatise on the "directions of ether vibrations in polarized light".

At the end of his reading, Fizeau mentions that interference fringes can arise from two interfering apertures only if the source has a very small angular dimension. "Hence", Fizeau pursues, "to mention this briefly, we might hope that by using this principle, and by forming, for instance by means of two large separated slits, interference fringes in the focal plane of instruments aimed at observing stars, it might become possible to obtain some new insights on the angular diameter of these stars"[5] (Fizeau 1868). The name of the author of the treatise reported by Fizeau was covered, and it is not reported in the Comptes-Rendus of the Academy. The treatise was deposed under the title: *Sine experientia nihil sufficienter sciri potest* – without experience, nothing can be known sufficiently.

Five years later, in a letter to Fizeau communicated by the same Academy, É. Stephan (1874) textually recalls Fizeau words quoted above, in which he recognizes "a totally new path that might lead to results otherwise inaccessible to

[5]The english translation is ours.

the methods currently available in Astronomy". He formalizes the principle of the corresponding experiment and reports the first tentative measurements of the diameter of Sirius (undertaken at Marseille). The following year, in 1874, Stephan reports to the same assembly extensive measurements from which he concludes that all observed stars have diameter (much) less than 158 milliarcseconds (158 mas, Stephan 1874). These experiments were achieved by placing two slits on an 80 cm aperture telescope. Fizeau's 1868 comment had lead to the first generation of stellar interferometers.

As early as in the 1880 s, A. Michelson had used an interferometer of Fizeau's type to measure the diameters of four satellites of Jupiter (≈100 mas, Michelson 1891). But according to Michelson later on, the method was not tested on stellar objects for the thirty following years, probably for two reasons. First, the success of such experience was supposed to require ideal seeing conditions. Second, diameters of the order of 10 mas would have required a distance between the apertures of 10 meters or more − a size entirely out of question at that time (Michelson 1920).

In 1919 however, A. Michelson discovers by tests that fringes can be obtained even with bad seeing conditions (Michelson 1920); and J. Anderson manages in 1920 to measure the separation of two components of Capella's system (54 mas) on a 100−inch (2.5 m) reflector (Anderson 1920). Stimulated by these results, Michelson proposes in the same article to use a setting that is similar to that of Fizeau, but where a periscopic mounting is introduced. In this Michelson stellar interferometer, the holes or slits of the Fizeau-Stephan mask are replaced by apertures that can be moved on the same mounting, allowing much larger separations than Fizeau's setting. The apertures' beams are redirected on a smaller telescope, in the focal plane of which the interferences fringes appear.

The experimental discovery that relatively steady fringes could be obtained, along with the successes of the experiments of 1920 led to the building of the famous 20−foot (about 6 m) Michelson stellar interferometer at Mount Wilson. This second generation interferometer allowed to determine the diameter of Betelgeuse as 47 mas, within 10% (Michelson 1921).

About 50 years later, at the Observatory of Nice, a third generation of interferometers appeared when A. Labeyrie obtained for the first time, using two independent telescopes separated by 12 m, optical interference fringes on Vega (Labeyrie 1975). This experimental success ignited the modern developments of optical high angular resolution interferometry, whose most exploited instruments are today the Very Large Telescope Interferometer (VLTI) in Chile and the Center for High Angular Resolution Astronomy (CHARA) array at Mount Wilson. These systems allow to create interference fringes from 4 to 6 independent telescopes, which are cophased pairwise.

The high angular resolution optical systems described in this paper inherit from the characteristics of the stellar interferometers mentioned above, of which they constitute a further evolution. These systems use a possibly very large number of apertures (an array), which are simultaneously cophased. The interference pattern of the whole array is recorded so as to allow *direct* imaging of the objects. They can be used either in Fizeau mode (in which case the array can be seen

as a huge masked aperture) or in Michelson mode (where the relative size of the apertures is changed with respect to their separation by the periscopic setting, a process called *subpupil densification*). Densification can also be obtained by using inverted Galilean telescopes. *Hypertelescopes* (Labeyrie 1996) refer generically to densified interferometers.

Our second story, which started in 1868, is thus not over yet. Actually, it even reaches a critical point because next generation high angular resolution interferometric arrays are currently subject to in in depth comparative studies. These studies will allow to choose which observing technology should be pushed in the next decades.

Three main types of next generation optical interferometric arrays emerge, the apertures' number and configurations of which will tend to be similar to current radioastronomical arrays (see M. Bremer's article on radiointerferometry in these proceedings). First, a direct extension of the VLTI system: few (possibly extremely) large telescopes that remain relatively compactly disseminated (in the 10^2 m range). Second, few relatively large (8 m) telescopes separated by kilometric distances. Third, a large number (in the hundreds) of small telescopes disseminated on kilometric distances.

In the three cases, such optical systems are expected to reduce our uncertainty and maybe to solve questions whose cosmological and philosophical implications are comparable to those evoked at the beginning of this Introduction. The first of those is the existence of life in distant stellar systems, but many other fields can reveal important discoveries, like stellar physics (through spatio-spectral studies of their atmospheres), or Active Galactic Nuclei (see on these issues the articles of A. Labeyrie, D. Mourard, M. Hadjara and J. Kluska in these proceedings).

The paper continues with issues and topics that are also echoed in several other articles of this volume. Section 2 describes more precisely the differences in the optical models of Fizeau and hypertelescopes configurations. We address in this section Michelson's periscopic interferometer. The Appendix derives the densification operated by inverted Galilean telescopes, and uses for that purposes results from the theory of light propagation that are detailed in the article of Aime in this Volume. Section 3 can be seen as a tutorial to numerically simulate these systems. Section 4 turns to methods aimed at improving the images recorded in the focal plane of such instruments, and proposes a survey of restoration methods based on sparse representations. Articles of these proceedings connected to optimization issues arising in image restoration are those of M. Bertero, C. Theys, and É. Thiébaut. A substantial part of A. Bijaoui's article is in addition dedicated to methods based on sparsity. Section 5 proposes a comparison of Fizeau versus hypertelescopes configurations in a specific case. Section 6 presents original simulations aimed at detecting an object that is small, faint, and far from a central object. The last section summarizes and concludes the paper.

2 The *Fizeau* and *hypertelescopes* configurations

In a Fizeau configuration, the ratio between the distance between any subpupil and their diameter is the same for the input and the output pupil (input and

output pupils are homothetic). In hypertelescopes this ratio is allowed to change. This is illustrated in Figure 1. In the hypertelescope configuration using inverted Galilean telescopes, the diameter of the output pupils is magnified relatively to their separations, which remain unchanged. In the "periscopic Michelson" hypertelescope configuration, the diameters are unchanged but the relative separation is smaller. Indeed, all three configurations are unchanged by applying a global arbitrary scaling factor.

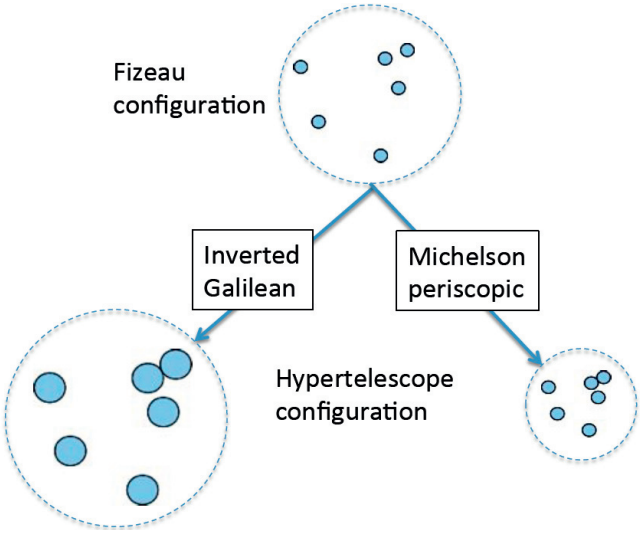


Fig. 1. Illustration of Fizeau (*top*) and hypertelescopes (*bottom*) configurations.

We will consider in this paper that the subpupils are all the same: circular, with diameter D. We will also consider that the atmospheric turbulence is negligible, and that the plane wave is monochromatic with wavelength λ.

We shall now see for both systems how we model the intensity distribution in the focal plane (the Image) that is obtained from a given celestial scene. The perfect (geometric) image of the celestial scene is called the Object. This section considers the continuous model (spatial and frequency variables are continuous). The discrete setting comes into play when images are sampled, and in numerical simulations of optical systems. Discretization will be addressed in Section 3.

The synthesis proposed below is a summary of Tallon & Tallon-Bosc (1992), Aime (2008) and Aime *et al.* (2012), papers to which we refer for a detailed treatment of the periscopic Michelson mode. The hyperterlescope mode using inverted Galilean telescopes is detailed in the Appendix and leads to the same results.

2.1 Fizeau configuration

The image formation mechanism in Fizeau mode is described by the general and standard equations of convolutional optics. The Point Spread Function

(PSF, denoted by R) in its normalized form (*i.e.*, summing to one) can be written as a function of the angular coordinates $\boldsymbol{\beta}(\beta_x, \beta_y)$ on the sky as:

$$R(\boldsymbol{\beta}) = \frac{1}{S\lambda^2} \left| \widehat{P}\left(\frac{\boldsymbol{\beta}}{\lambda}\right) \right|^2, \tag{2.1}$$

where superscript $\widehat{\ }$ denotes the Fourier Transform (FT), $\widehat{P}(\frac{\boldsymbol{\beta}}{\lambda})$ is the scaled FT of the telescope aperture transmission $P(\mathbf{r})$ with $\mathbf{r}(r_x, r_y)$ the vector of position, λ is the light wavelength, S is the total surface of the telescope aperture. As for classical (monolithic) telescope, the cophased optical system acts as a linear (bandpass) filter in the Fourier space. The transfer function T of this filter is the FT of the PSF. By the Wiener-Kintchine theorem, T corresponds to the spatial autocorrelation function of the input diluted pupil. If we denote by $\mathbf{u}(u, v)$ the angular frequency vector, the normalized optical transfer function (OTF) $T(\mathbf{u})$ is defined by

$$T(\mathbf{u}) = \frac{1}{S} \iint P(\mathbf{r}) P^*(\mathbf{r} - \lambda\mathbf{u}) \mathrm{d}\mathbf{r}. \tag{2.2}$$

The Object-Image relation, relating the object O to the image I^F in the Fizeau mode, is a convolution in the direct space

$$\underbrace{I^F(\boldsymbol{\beta}) = O(\boldsymbol{\beta}) \star R(\boldsymbol{\beta})}_{Angular\ Convolution}, \tag{2.3}$$

and a multiplication in the Fourier space

$$\underbrace{\widehat{I^F}(\mathbf{u}) = \widehat{O}(\mathbf{u})\ T(\mathbf{u})\ .}_{Frequency\ Filtering} \tag{2.4}$$

Let us have a close look at the structure of the transfer function in the case of a diluted pupil composed of K subpupils of diameter D. In this case, the normalized OTF of each subpupil $T_0(\mathbf{u})$ (see Eq. (7.8)) corresponds to the autocorrelation function of a disk, which is sometimes called a "chinese hat" function. The support of $T_0(\mathbf{u})$ is a disk of diameter $2D/\lambda$.

Let the centers of the K subpupils be at spatial positions $\mathbf{r}_k$, $k = 1, \ldots, K$. The autocorrelation function of the centers defines a set of central frequencies $\mathbf{u}_{kl} = (\mathbf{r}_k - \mathbf{r}_l)/\lambda$. The optical system composed of the diluted pupil samples frequencies located within a disk of radius $2D/\lambda$ around the central frequencies $\mathbf{u}_{kl}$. The transfer function of the diluted pupil can thus be written as

$$T(\mathbf{u}) = T_0(\mathbf{u}) + \frac{1}{K} \sum_{l=1}^{K} \sum_{k \neq l}^{K} T_0(\mathbf{u} - \mathbf{u}_{kl}), \tag{2.5}$$

where the double sum collects the contributions around the frequencies $\mathbf{u}_{kl}$, $k \neq l$.

The Fizeau image has frequency content

$$\widehat{I^F}(\mathbf{u}) = \widehat{O}(\mathbf{u}) T(\mathbf{u}) = \widehat{O}(\mathbf{u})\ T_0(\mathbf{u}) + \frac{1}{K} \sum_{l=1}^{K} \sum_{k \neq l}^{K} \widehat{O}(\mathbf{u})\ T_0(\mathbf{u} - \mathbf{u}_{kl}). \tag{2.6}$$

The image I^F formed in the focal plane of a Fizeau interferometer is the inverse FT of (2.6). Clearly, a lot of information about O is missing in I^F because a lot of frequencies are zero in $\widehat{I^F}(\mathbf{u})$. While for monolithic telescopes the missing frequencies are the high frequencies (at low frequencies, the transfer function has no "hole" or zero value), for a diluted pupil the Fourier coverage may present voids in any frequency region. In the Fourier space, the transfer function may be seen as an "archipelago of emerged islands" in the middle of a black sea where no measurement is available (see Fig. 2, bottom images).

Here appears an aspect that is crucial for the comprehension of this problem: we see precisely what I^F is missing to be O. This calls for five remarks.

- 1. For a given frequency sampling (*i.e.* for a given pupil array) the problem of recovering O is both an interpolation and an extrapolation problem in the Fourier space. This suggests that the image restoration methods can be designed, or least interpreted, as methods controlling the way the voids of the Fourier space are filled in, while preserving the observed frequencies.
- 2. The restoration quality will be object- and sampling- dependent. To see this, imagine two objects observed via the same pupil array (fixed sampling). Assume the first object has the most energetic part of its frequency content in the support of $T(\mathbf{u})$, while the most energetic part of the second object falls in the "sea" (that is, outside the support of $T(\mathbf{u})$). Clearly, this array is good for the first object but bad for the second, or the first object is good for this sampling but not the second object. This dependance is true in general but fortunately, it cannot be arbitrarily uncontrollable. The reason is that most natural objects have their frequency contents mostly located at low frequencies. Hence, ensuring a fair coverage at low frequencies guarantees that at least some useful information will be sampled for most objects.
- 3. To best sample the object we would like to maximize the Fourier coverage of the pupil array (or, to refer back to our image, to minimize the area of the black sea). Hence, for a fixed number K, we would like to maximize the frequency support of $T(\mathbf{u})$ (ignoring the effects of noise). This leads to configurations that are called *non redundant* (Kopilovich & Sodin 2001): no spatial frequency is sampled more than once (expect from the $\mathbf{0}$ frequency).
- 4. How many objects have the same frequency contents as that of I^F? Remark that any object that has, outside the zero frequency, frequency content only "in the sea", that is, outside the support of $T(\mathbf{u})$ create a totally flat (constant) image. Stated differently, this means that not only O leads to the observed image I^F, but also scaled versions of O plus *any* object that has frequency content outside the support of $T(\mathbf{u})$. This shows another very important point regarding the restoration we can hope to make from I^F: if we make no additional assumption on the geometrical properties of O, we can construct infinitely many different instances of objects that create the image I^F. Using the sole knowledge of I^F, the object remains thus unknown.
- 5. The informative frequency content that has been collected by $T(\mathbf{u})$ is (2.6), and the image that we obtain is the inverse FT of it. How many

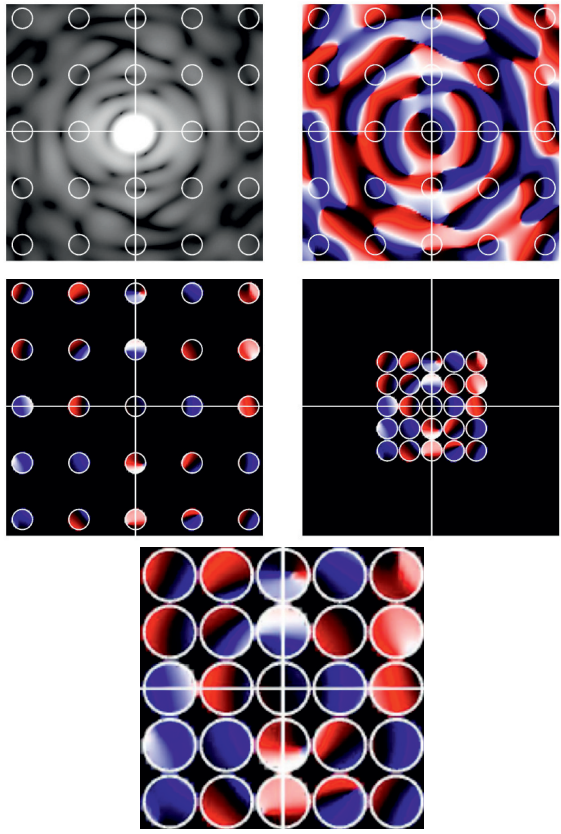


Fig. 2. Illustration of the Fizeau and hypertelescopes imaging properties in Fourier space. *Top left* (resp., right): absolute value (resp. phases) of the Fourier spectrum of the object. The white circles represent the boundaries of the elementary transfer function T_0, *i.e.*, the zones inside which the spectral information is sampled by both systems. *Middle left*: the Fourier content of the Fizeau image. In the black region, no frequency is available. The centers of the circles are located at spatial frequencies $\mathbf{u}_{kl}$ (see text). *Middle right*: the Fourier content of a Michelson image (*i.e.*, densified by a factor γ, full spectral densification is shown). The frequency content of each sampled disk of the Fizeau image has been translated by block, phases and moduli untouched. The centers of the circles are now located at spatial frequencies $\mathbf{u}'_{kl} = \mathbf{u}_{kl}/\gamma$. *Bottom*: the frequency sampling obtained with densification using inverted Galilean telescopes is a dilated (and thus fully equivalent) version of the Michelson mode with periscopic densification.

images contain the same contents of information about O as I^F? To answer this, imagine that the frequency contents of I^F is modified in a reversible way, which preserves hermitian symmetry (thus ensuring that the corresponding

image is real) and which is further such that the resulting image is positive. For instance, appropriately permute the contents of some "islands", or attribute the content of such island to an empty zone, and set the corresponding frequency content of the island to 0. There is a large number of such transformations: all transformations resulting in an OTF which is an autocorrelation function will work. From all the corresponding images (which will be very different from each other), we can take the FT (optically or numerically) and then numerically undo the transformation. We recover by doing so the originally sampled spectrum, and we can reproduce the image I^F. The Michelson densification by a factor γ in periscopic mode is one among such transformations: it leads to an image I^P_γ that is different from I^F, but the informational contents is the same, and the image I^F can be recovered from I^P_γ (at least for values of γ that are not too large). The Appendix shows that the densified image I^G_γ obtained using inverted Galilean telescopes has the same property.

2.2 Michelson configuration: Hypertelescopes

The *Michelson* configuration corresponds to a "densification" of the pupil because the diameters of the subpupils can be increased relatively to their separation. The degree of densification can be quantified by a densification factor called γ (Labeyrie 1996). We consider here the periscopic mode, in which the diameter of the subpupils D is fixed (see the Appendix for the case where the densification is obtained by dilation of the subpupils by a factor γ). The densification factor can in this case be defined as the ratio of the minimal distance between the subapertures before and after the densification: $\gamma = d/d'$. In the extreme case of maximal densification, some subpupils touch each other and are thus separated by a distance of $d' = D$, in which case $\gamma = d/D$.

The image formation model for the *Michelson* stellar interferometer in periscopic mode has been analyzed by Tallon & Tallon-Bosc (1992). The most important difference with the Fizeau configuration is that the image formation model is not a convolution anymore. In the Fourier space, the *Michelson* configuration involves a filtering corresponding to the diluted aperture before densification, followed by a translation of the frequency contents corresponding to the densification. During this translation, the spatial frequencies that are sampled by the input diluted aperture in a disk of width $2D/\lambda$ around frequency $\mathbf{u}_{kl}$ are subsequently carried away, phase and modulus untouched, into a disk of same diameter but centered around the lower center frequencies $\mathbf{u}'_{kl} = \mathbf{u}_{kl}/\gamma$ (see Fig. 2).

The Fourier spectrum of the densified image I^P_γ is now (compare to (2.5) and (2.6))

$$\widehat{I^P_\gamma}(\mathbf{u}) = \widehat{O}(\mathbf{u})\, T_0(\mathbf{u}) + \frac{1}{K}\sum_{l=1}^{K}\sum_{k\neq l}^{K} \widehat{O}(\mathbf{u} - \mathbf{u}'_{kl} + \mathbf{u}_{kl})\, T_0(\mathbf{u} - \mathbf{u}'_{kl}), \tag{2.7}$$

where the term $T_0(\mathbf{u})$ is, as in Section (2.1) the elementary transfer function corresponding to one subaperture and K is the total number of subapertures. Each term of the double sum corresponds to the filtering, by the elementary transfer function centered at frequency $\mathbf{u}'_{kl} = \mathbf{u}_{kl}/\gamma$, of the object's spectrum translated by $\mathbf{u}'_{kl} - \mathbf{u}_{kl}$.

Again, the description in the Fourier space makes some important issues very clear:

- 1. We see that this frequency translation is a perfectly revertible transform[6] as long as the zones around the new frequencies $\mathbf{u}'_{kl}$ do not intersect. Hence, the Fizeau and Michelson images are actually equivalent. One difference arises however in presence of sampling: as visible in Figure 2, the densified image has a lower cut-off frequency than the Fizeau image. Hence, by the Shannon theorem, it may be sampled with less pixels than the Fizeau image.
- 2. When the translated frequency zones intersect, several frequencies melt into a single one at each point of the intersection zone. In this case, the transformation is not invertible, since there not a one-to-one mapping from the initial to the final frequency content. Because the disks corresponding to T_0 around $\mathbf{u}_{kl}$ have width $2D/\lambda$, the lower center frequencies $\mathbf{u}'_{kl}$ cannot be separated less than $2D/\lambda$ to avoid frequency overlap (aliasing). This means that the minimum separation d' between two subapertures in the densified pupil must not be less than $2D$ ($d'/\lambda \geq 2D/\lambda \Leftrightarrow d' \geq 2D$) to avoid information loss. The limiting case $d' = 2D$ is called FSP for Full Spectrum Densification in the literature. The case $d' = D$ (subpupils touching each other) is called FAD for Full Aperture Densification. This case indeed leads to frequency aliasing.

3 Numerical simulations of Fizeau and hypertelescopes interferometric images

3.1 Discretization and periodicity

Sampling and numerical simulations involve discrete approximations of continuous phenomena. We provide here a description of sampling issues which mostly relies on handy notions of Fourier analysis. A rigorous description of sampling theory requires to use distributions and Lebesgue integration, see *e.g.* Chap. 2, 3 and 5 of Mallat (2008).

The physical reference object is considered as constant (or sufficiently slowly varying) in time and as a continuous function of its space (or angle) variables. This object is of infinite resolution (the size of the smallest details in the object is vanishingly small).

[6] Hypertelescopes are an instance of transformations conserving the positivity while changing the image. Studying general properties of such transformations is an interesting point which is left out of the scope of this article.

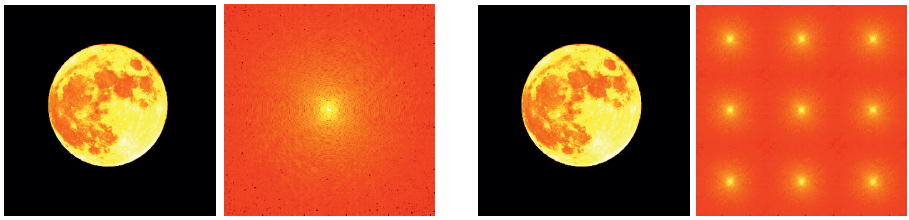
In simulation, the discrete reference object we will consider is an approximation of this ideal reference. We shall assume that the discrete reference object has been obtained by a fine regular sampling from the reference. Let τ_r be the spatial sampling step. The discrete object is the multiplication of the continuous object by a Dirac comb of period τ_r. The numerical representation of the object assigns one number to each sampling cell $\tau_r \times \tau_r$, which is the pixel size. The discrete object is thus often represented as a "staircase" version of the ideal reference object, although a continuous version of the discrete object is indeed possible using other standard interpolation functions. This is illustrated in Figure 3.

a) Moon b) Moon HR c) Moon HR: Zoom

Fig. 3. The discrete reference object (*middle*) is represented as a staircase (pixel) version of the ideal (continuous) reference (*left*). The approximation is visible when zooming (*right*): no detail smaller than the pixel size can be distinguished.

The discretization has indeed very deep implications on the image representations and processing. Let us think of the Fourier spectrum of the continuous reference object. This spectrum possess arbitrarily high frequencies because the smallest spatial structures can be arbitrarily small. Now let us consider the discrete reference object, obtained by multiplication of the continuous object with a Dirac comb. What is the Fourier spectrum of this object? By the convolution theorem, multiplication in one space translates to convolution in the dual space. The spectrum of the sampled object is the convolution of the true (infinite resolution) spectrum by the FT of the spatial sampling comb, that is, a Dirac comb of period $\mathsf{T}_\nu = 1/\tau_r$. The Fourier spectrum of the discrete object is periodic, its frequency period is $\mathsf{T}_\nu = 1/\tau_r$.

At this point we see that a discrete object has a continuous Fourier spectrum that is periodic. But, of course, this continuous spectrum cannot be stored as such in a numerical environment: it must be sampled. Well, the same reasoning as above can be applied to the periodic continuous spectrum. This spectrum is sampled with a step τ_ν in frequency, so it undergoes a multiplication by a Dirac comb with period τ_ν. What is the image corresponding to the resulting discrete, periodic spectrum? The multiplication by a Dirac comb in frequency results in a convolution by a Dirac comb in space, with period $\mathsf{T}_r = 1/\tau_\nu$. Hence, if we consider discretization obtained by regular sampling, we end up with images and spectra that are periodic. Each period is sometimes called the "principal interval" (see Fig. 5). The Discrete FT (DFT) and its inverse compute the representations of



a) Moon: image and Fourier spectrum b) Moon HR: image and Fourier spectrum

Fig. 4. a) Continuous-space reference object and corresponding continuous-frequency spectrum. The spectrum of the reference object is represented only for frequencies $\mathbf{u}(u, v)$: $|u| < \mathsf{T}_\nu/2, |v| < \mathsf{T}_\nu/2$ although of course the spectrum spreads at much higher frequencies. The discretized object and the corresponding (continuous) spectrum are shown in b). The spectrum is periodic: the Fourier space is paved with squares of size $\mathsf{T}_\nu \times \mathsf{T}_\nu$ that replicate the same continuous spectrum. Note that the continuous spectra in each such period are *not* equal to the continuous reference spectrum of the left figure, because the replicas on the right are obtained by superimposition of the spectrum of the left. Since the reference spectrum has no reason to be band-limited in a square of size T_ν (*i.e.* to be zero outside this square), higher frequencies contaminate the replica in the spectrum of the right (this is another instance of aliasing). If T_ν is high enough however, the continuous object will have little frequency content beyond T_ν, so that the replica of the periodic spectrum will be a good approximation of the reference spectrum at frequencies lower than T_ν. In all figures, the origin $(0, 0)$ in space and frequency variables are at the center of the image.

the object in both principal intervals. The DFT assumes that the discrete object and its spectrum are periodic, with a period of N points along each axis. The Fast FT (FFT) allows to compute the DFT in $\mathcal{O}(N \log_2 N)$ instead of N^2 additions and multiplications for monodimensional signals.

3.2 Numerical simulations of Fizeau and hypertelescopes images

The distinction made above between the continuous and discrete cases requires to differentiate the corresponding notation. In the following, a function that is continuous in its variable will be denoted by $f(t)$, and a discrete function by $f[n]$.

The first step is to choose a reference object. As we have seen above, this object must be discrete to be numerically manipulated, and is consequently an approximation of the corresponding continuous object. Anyhow, this discrete object (say, O) will become our reference. This object is composed of $N \times N$ pixels, with $N = 1024$ in all experiments below.

We now turn to the simulation of Equations (2.5) and (2.6) in the Fizeau mode, and of (2.7) in the Michelson (hypertelescope) mode[7]. To do this, we need to choose the subpupils of the array. They will be circular, with some diameter D.

[7]The numerical setting detailed here is essentially the same as that of Aime *et al.* (2012).

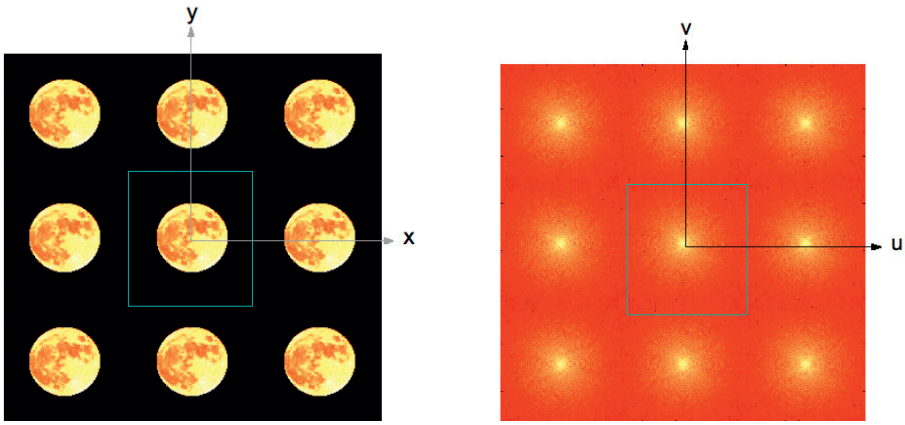


Fig. 5. The principal interval of a discrete reference object in the direct (*left*) and Fourier (*right*) space. Discrete objects of finite spatial and spectral extensions are made periodic when both spaces are related by the DFT. The size of the square cell in the direct space T_r is related to the numerical resolution in frequency (sampling step): $\mathsf{T}_r = 1/\tau_\nu$. The size of the square cell in the Fourier space T_ν is related to the numerical resolution in space (sampling step): $\mathsf{T}_\nu = 1/\tau_r$. The number of points N along each axis is $N = \frac{\mathsf{T}_r}{\tau_r} = \frac{\mathsf{T}_\nu}{\tau_\nu}$.

The center positions of the subpupils in the array are shown in Figure 7, along with their autocorrelation function which gives the central frequencies $\mathbf{u}_{kl}$ defined in Section (2.1). In the Fourier domain, we have seen that the spectra are sampled on a square grid of step-size τ_ν. The transfer function of an elementary circular pupil is the continuous-frequency function $T_0(\mathbf{u})$. When sampled at step-size τ_ν, this function becomes a set of weighted discrete Dirac. In the present case, the support of T_0 spreads essentially over 9 samples (Fig. 7, bottom).

In the considered space-continuous array, the centers of the subpupils are located on the nodes of an integer grid, so the spatial frequencies $\mathbf{u}_{kl}$ defined in Section (2.1) remain, once sampled, on a regular grid. While this may generally not be the case in practice, this setting makes the modeling of the densification easy.

Figure 8, top row, shows the locations where the Fourier samples of images are nonzero in Fizeau mode (left), in Michelson mode with full spectral densification (subpupils' centers are at least separated by $2D$), and in almost full aperture densification (some subpupils are almost touching each other). All samples are located on a 2-dimensional square grid with step-size τ_ν. A white sample means a nonzero sample. A fully white map would indicate that the N^2 samples of the DFT are available, so that the discrete reference object would be perfectly recovered by direct Fourier inversion.

Figure 8 is an illustration of what Equations (2.5)-(2.6) and (2.7) become in our setting. This Figure shows the discrete equivalent of Figure 2. The top left

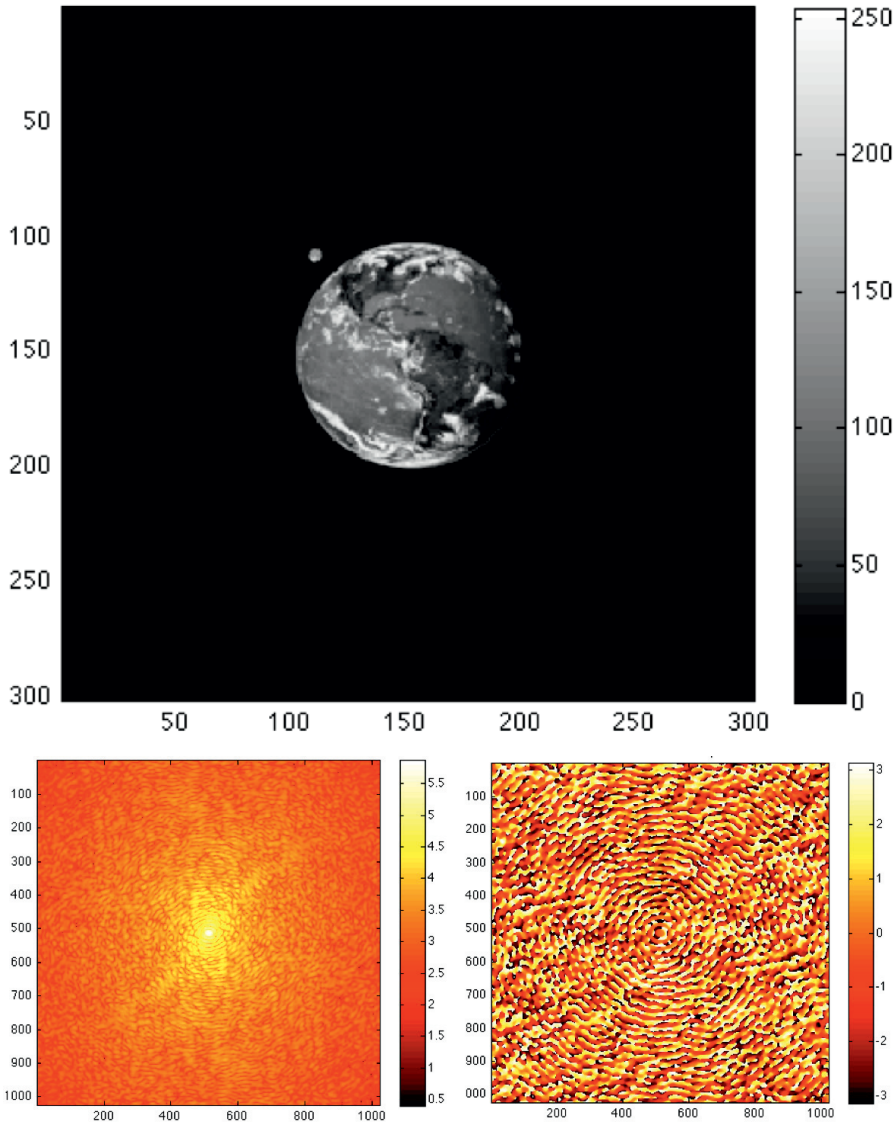


Fig. 6. *Top*: example of a (zoomed) reference object $O[\boldsymbol{\alpha}]$. Bottom, *left*: moduli of $\widehat{O}[\mathbf{u}]$. *Bottom*, *right*: phases of $\widehat{O}[\mathbf{u}]$.

figure shows the support of $T(\mathbf{u})$ in Equation (2.5), once sampled. The sampling "islands" discussed in Section 2.2 for the Fizeau mode are visible in the figures of the second row, as little light squares in the middle of a dark sea of zero samples. These islands correspond to the frequencies sampled around the frequencies $\mathbf{u}_{kl}$ and $\mathbf{u}'_{kl}$ by the elementary OTFs (disks, which once sampled give rise to sets of 9 samples).

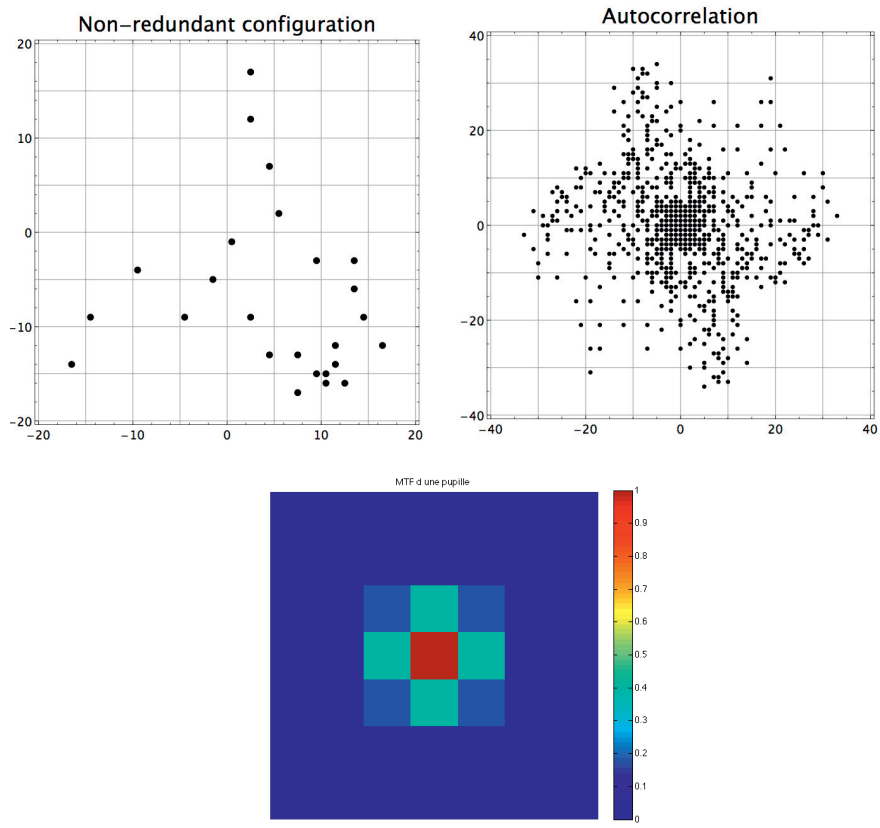


Fig. 7. *Top, left*: position of the centers of the subpupils in the array. These centers are all located at nodes of an integer grid. *Top right*: autocorrelation function of the centers of the subpupils. This function gives the central frequencies $\mathbf{u}_{kl}$ defined in Section 2.1. *Bottom*: sampled OTF $T_0[\mathbf{u}]$ of an elementary aperture.

In the Michelson mode (hypertelescopes), the nonzero frequency contents are moved block-wise towards the frequency origin, and the center frequencies $\mathbf{u}_{kl}$ become $\mathbf{u'}_{kl} = \mathbf{u}_{kl}/\gamma$. In the considered Fizeau mode the minimum separation between two center frequencies $\mathbf{u}_{kl}$ is $7\tau_\nu$. In the FSD mode, for which the translated contributions of the elementary transfer function touch each other, the minimum separation in frequency between the $\mathbf{u'}_{kl}$ is $3\tau_\nu$, so $\gamma = 7/3$. If we increase the densification, the contributions of the elementary transfer function overlap. In the considered (quasi)FAD case, $\gamma = 7/2$. The densification factors can easily been translated in terms of subpupilar distance. In the FSD mode, $d' = 2D$ and in the (quasi)FAD mode $d' = 4/3D$ (some subpupils almost touch each other). Figure 9 summarizes the principle for simulating the formation of hypertelescope images.

We see that starting on integer grid for the Fizeau configuration allows easily to obtain the frequency content for a set of Michelson configurations by simply

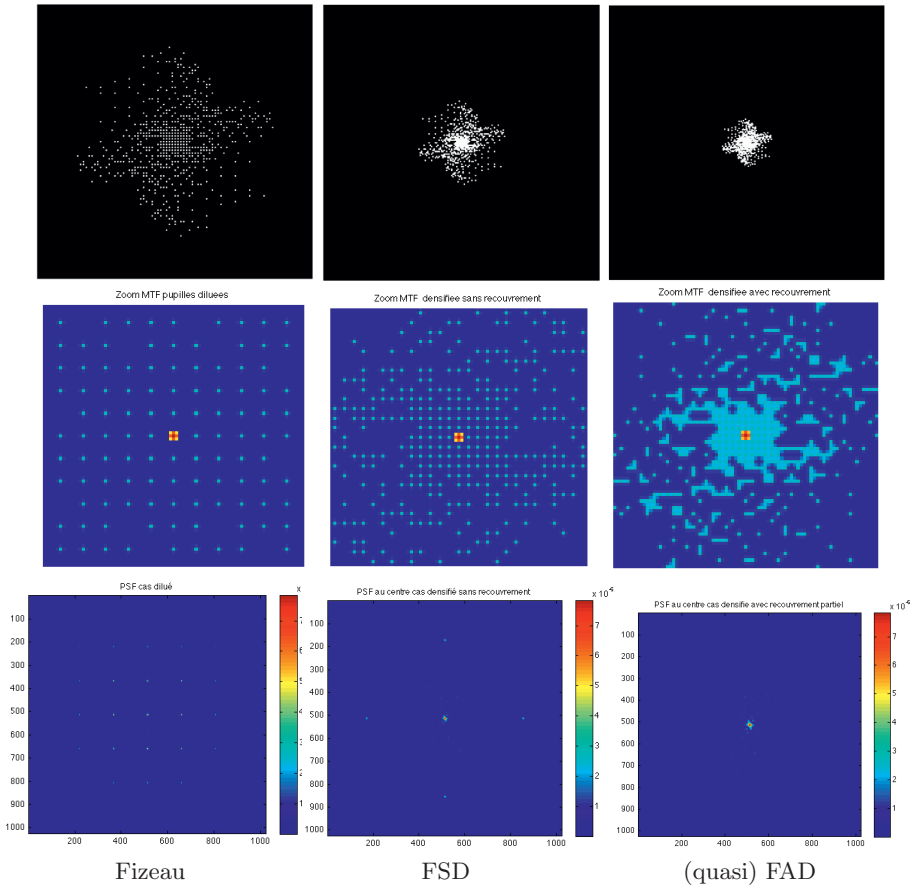


Fig. 8. *Top row*: compared locations of the nonzero Fourier samples that create the observed images in the three settings (Fizeau: *left column*; Michelson FSD: *center column*; Michelson FAD: *right column*). *Middle row*: zooms of the central parts of the top images. *Bottom row*: response to a point source on the optical axis (PSF).

removing the appropriate lines and columns of zeros. In the FSD example, the frequency content in each interval of the FSD image is obtained by removing 3 consecutive lines out of 7, and this process is repeated periodically in the rows $[u]$ and the columns $[v]$ with a period of 7. The outer region of each interval is then zero-padded so that the total number of N^2 samples is conserved. The samples of the densified image are thus again on an integer frequency grid (of same frequency resolution τ_ν).

3.3 Worked-out examples

Examples of images corresponding to the three considered sampling schemes are shown in Figure 10.

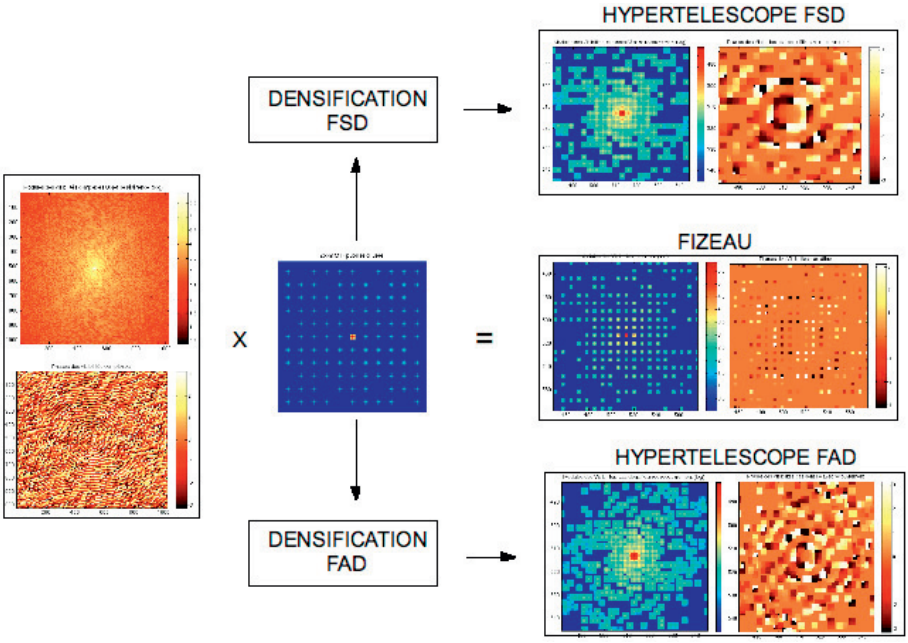


Fig. 9. Simulation of hypertelescope images. The discrete FT of a discrete reference object (*left*: moduli and phases, principal intervals are shown) is multiplied by the transfer function of the pupil array. The densification leads to the frequency contents (moduli and phases) shown on the right. The corresponding images are obtained by inverse DFT.

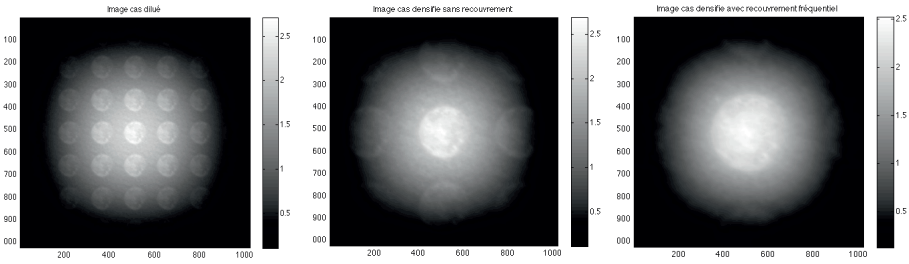


Fig. 10. Examples of images formed in the focal plane for 25 pupils on a grid. *Left*: Fizeau configuration. *Middle*: Michelson FSD. *Right*: Michelson quasi FAD.

There are several important aspects to be noted here.

– 1. First, all images are darker in the outer region. This is caused by the elementary contribution of each subaperture (which are all the same here). Let us consider the Fizeau image. In the Fourier space, the transfer function T is obtained by convolving the autocorrelation function of the pattern

created by Dirac impulsions (Fig. 7, top right) with the elementary transfer function T_0. This convolution acts in the image space as a multiplication by the FT of T_0, that is, by the PSF of the subapertures. What we see here is the diffraction envelope corresponding to the subapertures: imaging is done *inside* the pupils' PSF.

The width of this diffraction envelope defines the total field, which is a disk with diameter about $2.4\lambda/D$. In the Michelson mode this diffraction envelope is obviously visible as well.

In the numerical modeling/sampling, the principal interval should thus have extension $\mathsf{T}_r = 1/\tau_\nu$ that is about this size, $2.4\lambda/D$, since this is the zone we are interested in.

– 2. A striking particularity of these images is the presence of replicated patterns that resemble the reference object. The reason is that the pupils centers are located on a grid. The elementary OTF are in turn located on nodes of the frequency grid, which are here separated by 7 units, that is, by $7\tau_\nu$. If there were not 25 but many more pupils, so that all these nodes would correspond to the center of an elementary OTF, the sampling function would consist in a Dirac comb (of period $7\tau_\nu$) convolved by the elementary OTF T_0. We see that the effect of this system would be, in the image space, to convolve the object intensity distribution with a Dirac comb of period $1/(7\tau_\nu) = \mathsf{T}_r/7$ (and the diffraction envelope would further tapper the result). The effect of this convolution would be to create, in a zone of extension $\mathsf{T}_r \times \mathsf{T}_r$, 7 replicas in both horizontal and vertical dimensions. This is essentially what we see in the Fizeau image of Figure 10. The difference is that for the Fizeau mode there are only 25 pupils. In this case, the Fizeau sampling function can be seen as the previous one multiplied by a function of ones and zeros which kills the frequencies where Fizeau's transfer function has no contribution, and leaves the others unchanged. This function is shown of Figure 8, top left, where the 1 are in white and the 0 in black. It has no regular shape. The Fizeau image is the convolution of the 7 replicas (the image obtained with the full grid) by the inverse FT of this function. The convolution by this "halo function" leads to an irregular and diffuse halo which explains the fuzziness of the Fizeau image. This halo blurs the image and removes some frequency contents (essentially the high frequencies, but also low frequencies in the voids of the sampling function).

– 3. Let us now turn to the Michelson configuration images. Densification translates frequencies block-wise, and thus performs a frequency modulation. This operation is not equivalent to downsampling. Downsampling by a factor k (*i.e.*, keep every k^{th} other sample) contracts the frequency axis uniformly. Each spatial frequency $\mathbf{u}$ is moved to the frequency $\mathbf{u}/k$, so that the object whose spectrum is subsampled appears zoomed (and possibly aliased) by a factor k. In densification, the frequency axis is not uniformly contracted. Only the center frequencies $\mathbf{u}'_{kl} = \mathbf{u}_{kl}/\gamma$ are contracted by a factor γ. For all other frequencies, the result of the translation is only approximately a

contraction by γ. Densification is equivalent to downsampling (followed by lowpass filtering) only in the limit of point pupils: in this case T_0 tends to a Dirac impulsion and only central frequencies $\mathbf{u}'_{kl}$ are sampled. The densified image is not a zoomed version of the Fizeau image. It may be considered as a zoom only in a first approximation (by neglecting the spatial extension of the pupils). This effect nevertheless explains why the replicas appear larger in densified images.

The magnification performed by densification is visible in the FSD image: in the Fourier space, 3 samples out of 7 are kept in the $[u]$ and $[v]$ frequency directions (see Fig. 8), resulting in a magnification of approximately 7/3 ($\approx$2.3) of the size of the replicated pattern. Needless to say, this magnification zoom comes with no gain in resolution at all (the frequency information about the reference objet is the same in both cases, as visible in Figure 2, and in Figure 8, middle left and middle center figures).

The number of replicas is decreased by the same amount $\gamma = 7/3$, because the nodes of the grid on which the frequencies $\mathbf{u'}_{kl} = \mathbf{u}_{kl}/\gamma$ fall are now $\gamma = 7/3$ closer than in the Fizeau case (they are separated by $3\tau_\nu$ instead of $7\tau_\nu$). Thus, the corresponding periodicity in the image space is now $1/(3\tau_\nu) = \mathsf{T}_r/3$: 3 replicas are visible in the vertical and horizontal directions. As in the Fizeau case, the sampling does not yield elementary OTF centered at all the nodes of the frequency grid of step $3\tau_\nu$. Hence, the replicas are convolved by the inverse FT of the 0/1 function shown Figure 8, top middle. This function is not a scaled (contracted) version of the corresponding Fizeau sampling function, unless the subpupils have negligible diameter. So, we see that the densification by a factor γ yields an image which, only in the limit of very small subpupil diameters, corresponds in the diffraction enveloppe to a zoomed (magnified) version of the Fizeau image.

– 4. In the FAD image finally, mainly one replica is visible in the center (actually two halves exist in the borders). This case seems to be a straightforward limiting case of the middle image: increased densification, larger magnification of the center replica, almost total disappearance of off-axis replicas. This is however not the case. The frequency contents in FAD mode has been modified (reduced) because of the too strong spectral densification (Fig. 8, right). During the FAD operation, some frequency cells have collapsed. Overlapping spatial frequencies have been melted, and the overall support size is smaller in the FAD than in the Fizeau/FSD cases. This operation is not invertible. Because of these reasons, it seems better for both purposes of direct imaging and of image restoration to stop the densification at the FSD limit. For direct imaging, FAD and quasi FAD produce images of reduced fidelity with respect to the object. For restoration purpose, FAD increases the difficulty of the inversion, because it adds un-invertibility to the imaging system.
– 5. The left figure in Figure 10 is the convolution of the reference object (Fig. 6) by the PSF shown in Figure 8, bottom left. However, as discussed in Section 2.2, the FSD and quasi FAD images are not the convolution of the

reference object by the PSF shown in Figure 8, bottom center and bottom right. A convolution may be retrieved in the FSD mode only in the limit of vanishingly small diameters (infinite fields of view). This suggests two regimes for densification (hypertelescopes), depending on wether the diameter D is much smaller than the smallest subpupil separation d or not. For hypertelescopes made of very large bases (in the kilometer range) and of many small telescopes (centimeters), $D << d$, and a convolution model may be a good approximation, at least close to the optical axis. For VLTI-like hypertelescopes, made of moderately large bases (in the hundreds of meter) and of a few large telescopes (in the tens of meters), $D \approx d$, and the image formation models strongly departs from convolution.

3.4 Noise

Real images will be affected by several perturbations. Effects caused by perturbations on the phase and by chromaticity are not addressed here. We consider two types of noise: Poisson noise (the number of detected photons in a pixel receiving a constant light flux is Poisson distributed), and Gaussian noise (which usually models the detector read-out noise, or approximates the Poisson distributed in the limit of larges fluxes). The Figure below shows examples of simulated noisy data images that would be obtained for a hypertelescope. The image in FSD mode (left) is sampled on a 1024×1024 pixels CCD detector. The average flux falling on each pixel corresponds to 0.8 photons. The middle image is what the detector sees with Poisson noise (the recorded number of photons in a pixel is the realization of a Poisson process having for mean and variance the noiseless flux on this pixel). In this case the detector noise is negligible with respect to the photon noise. On the other hand, the right image is an instance of what the detector records with a zero mean Gaussian noise having a standard deviation of 1 photon.

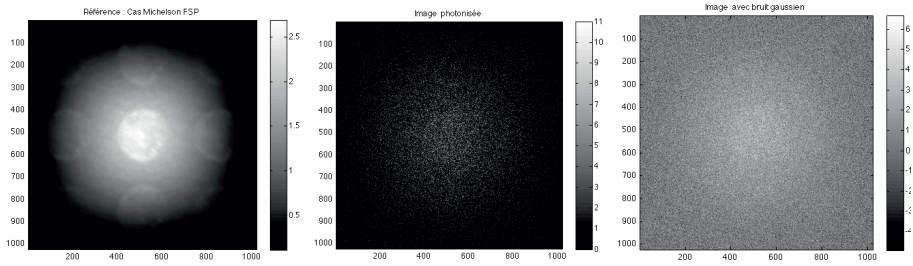


Fig. 11. *Left*: noiseless FSD hypertelescope image. *Middle*: corresponding photonized image. *Right*: noisy image with Gaussian noise.

Clearly, the data images in these cases are quite degraded versions of the noiseless image. However, this visual quality loss is partly illusory, because the noise component in the middle and right images have frequencies in the whole Fourier

space − while we know that our imaging system has measured only the frequencies that belong to the support of the transfer function T. Hence, we can safely (and numerically) filter out all the frequencies outside this support without any degradation of the astrophysical information. This is simply achieved by a DFT (or FFT) of the noisy image, multiplication the result by the corresponding indicator function of Figure 8, top row, and inverse FFT. We obtain the images of Figure 12.

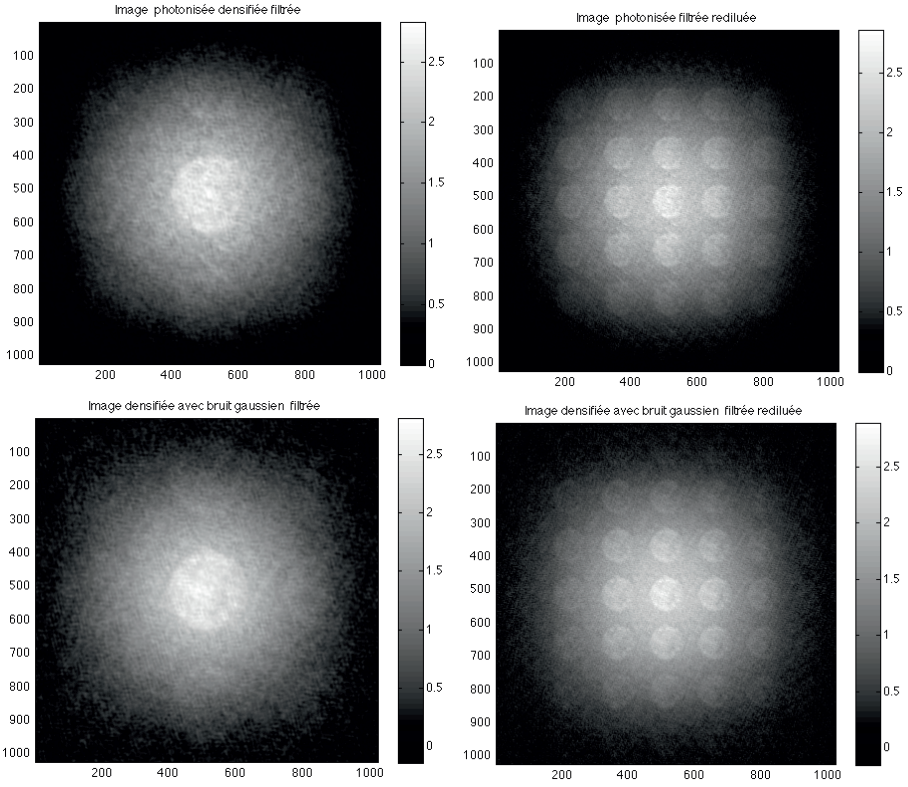


Fig. 12. *Top row*, *left*: filtered version of the photonized FSD hypertelescope image of Figure 11, middle. *Top row*, *right*: corresponding numerically rediluted image. *Bottom row*, *left*: filtered version of the FSD hypertelescope image with Gaussian noise of Figure 11, *right*. *Bottom row*, *right*: corresponding numerically rediluted image.

Clearly, the noise is much less adversarial than suggested by the data image. Note also that a numerical post-processing allows to create, from densified image, images that would have been obtained (up to the noise realization) with the corresponding Fizeau imaging setting. This is simply achieved by computing a FFT, translating back the frequency samples to the location they occupy in the Fizeau sampling, and computing an inverse FFT. This is interesting because now

the model relating the object to the rediluted data is again a convolution, and this allows to use classical deconvolution algorithm even in the densified case.

This example shows that the image quality (*i.e.*, its closeness to the object) can be improved by post-processing. This is the topics of the next Section.

4 Restoration algorithms

This section focuses on methods aimed at improving the estimation of the object from the image. What we want to do here is to infer from the data image, and from the mathematical model of the image-object relation, which object has generated the data. The process of finding back O from I^F, I^P_γ or I^G_γ is sometimes given the catchy name of "inverse crime" in the literature.

We first describe the inversion problem, and then illustrate three classical deconvolution algorithms. The last part proposes a review of a particular class of restoration methods that have received an increasing amount of attention in the last decades. These methods use the notion of sparse representations.

4.1 Inversion

Because restoration algorithms involve digital data and filtering techniques, we will consider from now on a fully discrete model. In this model, let us write the data image and the unknown reference object as vectors $\mathbf{y}$ and $\mathbf{o}$ respectively. These vectors are simply vectorized versions of the discrete arrays containing the intensity values in each pixel of the image and reference object. Considering principal intervals, the reference object and the image have N pixels, and their discrete Fourier spectra have N frequencies. In the case of subpupils placed on an integer grid, the sampled spatial frequencies are also on a grid. The noiseless image formation model in the Fizeau case becomes

$$\mathbf{y}_f = \mathbf{F}^\dagger \mathbf{T} \mathbf{F}\, \mathbf{o} = \mathbf{H}_f\, \mathbf{o}, \tag{4.1}$$

where $\mathbf{F}$ is the matrix form of the DFT, superscript† denotes conjugate transpose, $\mathbf{T} = \mathrm{diag}\{T[1,1] \dots T[N,N]\}$ is the diagonal matrix representing the transfer function of the diluted array, and $\mathbf{H}_f$ is a discrete circular convolution operator (a circulant matrix, because the considered discrete images and spectra are N-periodic). As we have now understood, the inversion is impossible because the solution is not unique: an infinity of objects can lead to the data, because of the zeros of the transfer function. In addition, the transfer function may be nonzero but very small at some frequencies, and the frequency content of the object at some sampled frequencies may be very small relatively to noise. In such cases, the data samples contain essentially noise. Inverting the transfer function at such frequencies leads in the image to fake oscillatory components that can have large amplitudes, a phenomenon called noise amplification. For these reasons, our inversion problem is said to be *ill-posed*, and the illness indeed comes from the instrument, not from the mathematical formulation.

The principal problem of the restoration can be seen as filling "cleverly" the zeros of the transfer function. Cleverly means that the recovered frequencies should, in some sense that remains to be defined, be close to the original frequencies of the object.

The noiseless image formation model in the case of hypertelescopes without spectral overlap (*i.e.*, for densification up to FSD or less) becomes

$$\mathbf{y}_{FSD} = \mathbf{F}^{\dagger}\mathbf{M}_{FSD}\mathbf{T}\mathbf{F}\ \mathbf{o}, \tag{4.2}$$

where $\mathbf{T}$ is the Fizeau transfer function, and $\mathbf{M}_{FSD}$ is the operator implementing the spectral densification (frequency modulation). This operator is a permutation matrix whose 1 specify the positions to which the frequency samples of the input (Fizeau) image are assigned in the Michelson image. This operator is linear, non diagonal (each column and each row of $\mathbf{M}_{FSD}$ have exactly one 1 and $N-1$ zeros) and obviously invertible ($\mathbf{M}_{FSD}^{-1} = \mathbf{M}_{FSD}^{t}$: transposing yields the inverse). The operator $\mathbf{F}^{\dagger}\mathbf{M}_{FSD}\mathbf{T}\mathbf{F}$ does not correspond to a convolution because $\mathbf{M}_{FSD}$ is not diagonal (convolution is diagonalized by the Fourier transform).

If, however, we redilute the densified image

$$\mathbf{y}_{redil} = \mathbf{F}^{\dagger}\mathbf{M}_{FSD}^{-1}\mathbf{F}\ \mathbf{y}_{FSD} = \mathbf{F}^{\dagger}\mathbf{M}_{FSD}^{-1}\underbrace{\mathbf{F}\mathbf{F}^{\dagger}}_{\mathbf{I}}\mathbf{M}_{FSD}\mathbf{T}\mathbf{F}\ \mathbf{o} = \mathbf{H}_f\ \mathbf{o}, \tag{4.3}$$

from which we see that rediluting allows to retrieve the convolution model (4.1).

In the case of a hypertelescope with spectral overlap (or aliasing), that is, in the range of densification from FSD to FAD, the model is

$$\mathbf{y}_{FAD} = \mathbf{F}^{\dagger}\mathbf{M}_{FAD}\mathbf{T}\mathbf{F}\ \mathbf{o} = \mathbf{H}_{FAD}\ \mathbf{o}, \tag{4.4}$$

where $\mathbf{M}_{FAD}$ is not a permutation matrix and is not invertible anymore. $\mathbf{H}_{FAD}$ is not a convolution operator, and we cannot redilute this image because $\mathbf{M}_{FAD}$ is not invertible. In the following, we consider only cases that can be described by a convolution: the Fizeau configuration, and spectrally densified images, without aliasing, and further rediluted. This model will be generically denoted by $\mathbf{y}_0 = \mathbf{H}\ \mathbf{o}$, and we will consider perturbations on $\mathbf{y}_0$ caused by Gaussian and Poisson noises.

We now turn to some standard methods aimed at estimating $\mathbf{o}$ from noisy data $\mathbf{y}$. In estimation theory, the Maximum Likelihood (ML) method is a systematic method aimed at building estimators of parameters considered deterministic. The likelihood of the data is assessed using the model (image-object relationship in our case), and is defined as the probability of observing the data conditioned to the parameters. The (unconstrained) ML method looks for the value of the parameters (the object, in our case) that is the most likely given the data. Unconstrained ML is extremely popular in the context of multiple measurements of the same parameter, because it is often asymptotically (in the number of measurements n) unbiased[8]

[8]This means that if θ denotes the parameter of interest and $\widehat{\theta}_{MV}[n]$ its ML estimate using n measurements, $\lim_{n\to\infty} \mathbb{E}\widehat{\theta}_{MV}[n] = \theta$.

and consistent[9]. In addition, if an estimator that achieves the Cramer-Rao lower bound[10] exists for a finite number of measurements, the ML finds it.

In the framework of images however, the problem is not posed in terms of multiple measurements (if the data image size increases, the number of parameters increases as well). Moreover, the unconstrained ML leads generally, for non invertible operators $\mathbf{H}$, to dramatic noise amplification. Consequently, the sought solution must somehow be constrained for the inversion to be possible. The most obvious constraint is to impose that the object $\mathbf{o}$ is non-negative ($\forall i, \mathbf{o}_i \geq 0$). The two following methods use this constraint, for Poisson and Gaussian data likelihoods respectively. They are relatively popular in the astronomical and optical communities where they have been published, and are synthetically exposed below. A detailed and unified treatment of regularized maximum likelihood methods with non-negativity constraint can be found in Lanteri *et al.* (2002a,b).

4.2 Richardson-Lucy algorithm (1972, 1974)

In the case of Poisson noise, the statistical model is $\mathbf{y} = \mathcal{P}(\mathbf{Ho})$. The likelihood of the data $\mathbf{y}$ is

$$\mathcal{L}(\mathbf{y};\mathbf{o}) = \prod_{i=1}^{N} \frac{([\mathbf{Ho}]_i)^{\mathbf{y}_i}}{\mathbf{y}_i!} \mathrm{e}^{-[\mathbf{Ho}]_i}, \tag{4.5}$$

where the product comes from the independence of the components, guaranteed by that of the noise realization from one pixel to another. By using Stirling's formula, maximizing the likelihood above leads to minimizing

$$J_{Poisson}(\mathbf{o}) = \sum_{i=1}^{N} [\mathbf{Ho}]_i - \mathbf{y}_i \ln[\mathbf{Ho}]_i, \quad \text{subject to} \quad \forall i, \mathbf{o}_i \geq 0, \tag{4.6}$$

which leads to the iterations

$$RL: \qquad \mathbf{o}^{(k+1)} = \mathbf{o}^{(k)} \cdot \mathbf{H}^t \frac{\mathbf{y}}{\mathbf{Ho}^k}. \tag{4.7}$$

In these iterations, the division of $\mathbf{y}$ by $\mathbf{Ho}^k$ is made element-wise. The results is left multiplied by $\mathbf{H}^t$ (which in practice is implemented in the Fourier space using the structure of $\mathbf{H}$ in (4.1)), and the multiplication by the previous estimate $\mathbf{o}^{(k)}$ is again element-wise. The question of deciding when to stop the iterations is a difficult one, as RL (and ISRA) do not possess a natural stopping criterion. Stopping the iterations to some number performs a kind of regularization, although non explicit. Pseudo-codes for RL and ISRA below can be found in Thiébaut (2005).

[9] *i.e.*, $\lim_{n\to\infty} \mathbb{E}\{\widehat{\theta}_{MV}[n] - \theta\}^2 = 0$.

[10] *i.e.*, the smallest Mean Square Error that is achievable by any unbiased estimator, and that is caused the uncertainty inherent to the stochastic perturbations.

4.3 Image space reconstruction algorithm (1986)

When the data are spoiled by a Gaussian noise that is pixel-wise independent but possibly non identically distributed, the model is $\mathbf{y} = \mathbf{H}\mathbf{o} + \mathbf{b}$, $\mathbf{b} \sim \mathcal{N}(0, \mathbf{\Sigma})$, with $\mathbf{\Sigma} = \text{diag}[\sigma_1^2 \dots \sigma_N^2]$. The ISRA algorithm (Daube-Witherspoon & Muehllehner 1986) produces the *non-negative* solution that is the most likely according to the noise model. The likelihood of the data $\mathbf{y}$ is

$$\mathcal{L}(\mathbf{y}; \mathbf{o}) = P([\mathbf{y}_1 \dots \mathbf{y}_N]^t; \mathbf{o}) = \prod_{i=1}^{N} (2\pi\sigma_i^2)^{-\frac{1}{2}} \mathrm{e}^{-\dfrac{(\mathbf{y}_i - [\mathbf{H}\mathbf{o}]_i)^2}{2\sigma_i^2}}. \tag{4.8}$$

Maximising this function on non-negative $\mathbf{o}$ is equivalent to minimize

$$J_{Gauss}(\mathbf{o}) = \frac{1}{2}||\mathbf{y} - \mathbf{H}\mathbf{o}||^2_{\mathbf{\Sigma}^{-1}} = \frac{1}{2}\sum_i \frac{(\mathbf{y}_i - \mathbf{H}\mathbf{o}_i)^2}{\sigma_i^2}, \quad \text{subject to} \quad \forall i, \mathbf{o}_i \geq 0. \tag{4.9}$$

This leads to the iterations

$$ISRA: \qquad \mathbf{o}^{(k+1)} = \frac{\mathbf{o}^{(k)}}{\mathbf{H}^t\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{o}^{(k)}} \cdot \mathbf{H}^t\mathbf{\Sigma}^{-1}\mathbf{y}, \tag{4.10}$$

where the division and multiplications are elemen-twise. In practice however, some data may be negative, in which case the iterations above do not guarantee the non-negativity of $\mathbf{o}^{(k)}$ over the iterations. To overcome this problem, one should work on a data image $\mathbf{y}'$ that is shifted by its minimum value $m = \min_i(\mathbf{y}_i)$ (Lanteri *et al.* 2002b). If we denote by $\mathbf{d}$ the vector with entries $\mathbf{d}_i = -m$, $\forall i$, then $\mathbf{y}' = \mathbf{y} + \mathbf{d}$ is non-negative and the ISRA iterations become

$$\mathbf{o}^{(k+1)} = \mathbf{d} + \frac{\mathbf{o}^{(k)} - \mathbf{d}}{\mathbf{H}^t\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{o}^{(k)}} \cdot \mathbf{H}^t\mathbf{\Sigma}^{-1}\mathbf{y}'. \tag{4.11}$$

When the iterations are stopped at some iteration number k_{stop}, the estimated object is $\widehat{\mathbf{o}} = \mathbf{o}^{(k_{stop})} - \mathbf{d}$.

4.4 Numerical illustrations

The following example (Fig. 13) illustrates what can typically be achieved by such algorithms. This example shows some results of the RL deconvolution algorithm as a function of the iterations. The considered reference object is that of Figure 6, and the data image is a photonized version of the FSD hypertelescope (Fig. 11, middle). Before deconvolution, we first redilute (Fig. 12, top right). We can also, before starting the restoration, filter out the noise. This is not necessary, because the first iteration of the RL algorithm will do it; but seeing the filtered image indicates that the situation is not as bad as visually suggested by the data image (Fig. 11, middle). This helps realizing that some merit should indeed be attributed to the restoration algorithm, but not too much.

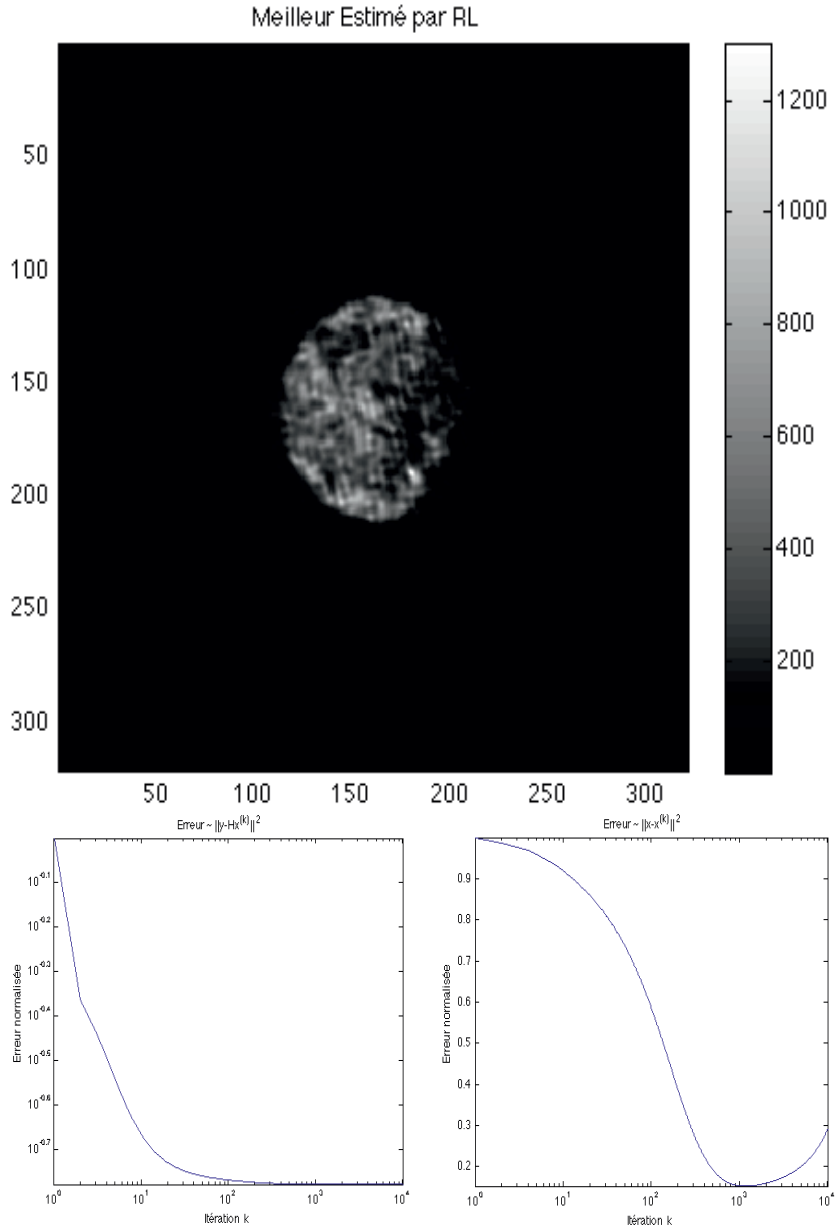


Fig. 13. *Top*: best (in the euclidean distance sense) recovered object for the RL algorithm. The companion is not reproduced. *Bottom*, *left*: data approximation error *vs.* iteration number. *Bottom*, *right*: error with respect to the reference object *vs.* iteration number.

The restored object in Figure 13 is, with respect to the true object, quantitatively and qualitatively inaccurate in the details, but rather fair in the overall shape. The companion is not recovered, and the surface intensity on the main planet has extremely high values. The overall flux of the solution is the same as the flux in the data, because RL iterations structurally preserve the flux. This flux is very close the flux of the reference object, because the 0 frequency is measured by the optical system: the flux is thus known up to the measurement error caused by the noise.

The two lower figures represent the data approximation error ($||\mathbf{y} - \mathbf{H}\mathbf{o}^{(k)}||^2/||\mathbf{y}||^2$) and the distance with respect to the object (measured here as $||\mathbf{o} - \mathbf{o}^{(k)}||^2/||\mathbf{o}||^2$). The data approximation error continuously decreases. This behavior always happens for ISRA (since ISRA iteratively reduces the convex criterion $J_{Gauss}(\mathbf{o}^{(k)})$), and usually for RL as well. On the other hand, the distance of $\mathbf{o}^{(k)}$ to $\mathbf{o}$ decreases and then increases: the best solution in the euclidean distance sense, which is shown in the top figure, is here obtained for an iteration number of $k \approx 1000$. The best iteration number is always impossible to know in practice, because the object is not known. The behavior exhibited by these curves is very general.

This example illustrates important characteristics that share RL and ISRA algorithms. The positivity constraint is easily imposed (and any support constraint could be possibly imposed as well, owing to the multiplicative form of this algorithms), and their implementation is easy. On the other hand, the number of iterations to reach an acceptable solution can be large (slow convergence), and it is difficult to know when to stop the iterations in practice. Finally, the interpolation and extrapolation that are performed in the Fourier space are very difficult to formalize.

An uncountable number of inversion methods can be found in the literature. A large class of those injects an explicit regularization term in the criterion coming from the likelihood. This term reflects *a priori* information about the solution. Explicitly regularized methods allow to put well defined constraints on how the missing frequency content should be added to the data during the iterations, or at least, to impose that the solution exhibits specifically desired properties (*e.g.* smoothness, piecewise constant aspect, etc.). Another set of approaches uses the concept of sparse representations, and a survey is proposed below.

4.5 Deconvolution based on sparse representations

4.5.1 CLEAN and sparsity

Long before the optical interferometry era, radio astronomers had devised various techniques to recover estimates of $\mathbf{o}$ from $\mathbf{y}$ in convolution models. The ancestor of sparsity-based techniques in the radioastronomical community is the CLEAN algorithm, which is used routinely in radioastronomy since almost 40 years now. Numerous variants have been developed (see Cornwell 2008), and in practice CLEAN remains a benchmark method in radioastronomy. It is worth detailing a bit on

this algorithm, as it can directly be applied to the deconvolution of hypertelescope images, and because of its link with sparsity.

CLEAN was proposed by Högbom (1974), and was related to earlier methods in Scharz (1978). See also Cornwell (2009) for a description of the wide impact of CLEAN on Astronomy and beyond. CLEAN can be seen as a method which essentially models $\mathbf{o}$ as a collection of a few (relatively to the number of pixels) point sources. The image $\mathbf{y}$ is thus in turn modeled as a collection of shifted and scaled PSFs. The algorithm subtracts iteratively scaled and shifted PSFs from the image residual, until a stopping criterion is reached. Högbom recommended in its 1974's paper a threshold on the maximum value of the residual. This threshold forces a limited number of iterations, and thus a limited number of detected point sources − this is where sparsity comes in.

This iterative process is very similar to the Matching Pursuit algorithm, which is widely used in the signal processing community (Mallat & Zhang 1993). With a difference, yet: at the end of the iterative deconvolution process of CLEAN, the residual is added back to the synthesized detection map. As a matter of fact, CLEAN works very well when only a few points sources constitute $\mathbf{o}$, and thanks to the residual trick, it remains also relatively efficient even for extended sources, despite the apparent irrelevance of its sparse model for such sources.

The excessive simplicity of CLEAN's "point source" model for general astrophysical sources was nevertheless worked out in the 80 s by several researchers, who looked for more elaborated models of extended sources. Several methods based on multiresolution approaches followed the works by Wakker & Schwarz (1988). An overview of CLEAN's evolutions is given in Rau *et al.* (2009).

4.5.2 Analysis and synthesis sparsity

Sparse representations in dictionaries are ubiquitous in a large number of modern signal processing methods. The reason for this is perhaps double: first, they rely on simple (linear) statistical models; second, they naturally operate a dimension reduction, by focusing on subspaces of reduced dimension where the information of interest actually lies.

Sparse representations can be seen as a generalization of CLEAN's model. In the synthesis model, which is the most intuitive and has historically benefited from more efforts, the object $\mathbf{o}$ is assumed to be well modeled by a linear combination of a few elementary shapes (not just point sources), called *atoms*.

Promoting sparsity relatively to appropriate dictionaries offers a straightforward way to fill the missing frequency contents caused by the zeros of the transfer function. Indeed, sparse methods essentially detect which atoms are present in the data using the measured frequencies: some missing frequencies are then automatically filled with the frequencies of the detected atoms.

Let us now describe in more detail the differences between analysis and synthesis sparsity (Elad *et al.* 2007). We consider the object-image model $\mathbf{y} = \mathbf{H}\mathbf{o} + \mathbf{n}$, with $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$ for simplicity. In the synthesis approach, the unknown intensity distribution $\mathbf{o}$ (of size $N \times 1$, say) is assumed to be sparsely synthesizable by a

few atoms of a given full rank dictionary $\mathbf{S}$ of size $N \times L$. Hence, we write $\mathbf{o}$ as $\mathbf{o} = \mathbf{S}\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ (the synthesis coefficients vector) is sparse. This assumption is widely used in data modeling for denoising, compression, pattern recognition or inpainting applications for instance, because natural signals and images are approximately sparse in appropriate spaces (Mallat 2008).

A synthesis-sparse solution to the deconvolution problem posed by $\mathbf{y} = H\mathbf{o} + \mathbf{n}$ can be obtained by solving:

$$\mathbf{o}_S^* = \mathbf{S} \left\{ \arg\min_{\boldsymbol{\eta}} \frac{1}{2} \parallel \mathbf{HS}\boldsymbol{\eta} - \mathbf{y} \parallel_2^2 + \mu_{\mathrm{p}} \parallel \boldsymbol{\eta} \parallel_{\mathrm{p}}^{\mathrm{p}} \right\}, \tag{4.12}$$

where μ_p is a hyper parameter that tunes the desired sparsity degree, and $\parallel \boldsymbol{\eta} \parallel_p^p = \sum_i |\eta_i|^p$, $0 \leq p \leq 1$, is a function favoring zero values.

The solution $\mathbf{o}_S^*$ is also interpretable in the Bayesian framework as a Maximum A Posteriori (MAP) solution, in which case μ_p is related to the parameters of a Generalized Gaussian prior on $\boldsymbol{\eta}$. The ℓ_0 quasi-norm (which counts the number of nonzero coefficients in $\boldsymbol{\eta}$), obtained for $p = 0$, is the most natural sparsity measure. To ensure the convexity of the resulting cost function, it is often replaced by the ℓ_1 norm $||\cdot||_1$, which still promotes sparsity (and correspond to a Laplacian prior on $\boldsymbol{\eta}$).

In contrast, the analysis approach consists in finding the solution $\mathbf{o}$ that is not correlated with some atoms of a dictionary $\mathbf{A}$ of size $N \times L$: $\mathbf{A}^T\mathbf{o}$ is sparse. An analysis-sparse solution can be obtained by solving:

$$\mathbf{o}_A^* = \arg\min_{\mathbf{o}} \frac{1}{2} \parallel \mathbf{Hx} - \mathbf{y} \parallel_2^2 + \mu_{\mathrm{p}} \parallel \mathbf{A}^{\mathrm{T}}\mathbf{o} \parallel_{\mathrm{p}}^{\mathrm{p}} . \tag{4.13}$$

Note that the synthesis prior is on the synthesis coefficients $\boldsymbol{\eta}$, while the analysis one is on the projection $\mathbf{a} = \mathbf{A}^T\mathbf{o}$ of the signal on the analysis dictionary $\mathbf{A}$.

While both approaches are equivalent when $\mathbf{A}$ and $\mathbf{S}$ are square and invertible, with $\mathbf{A} = \mathbf{S}^{-1}$, they yield in general different solutions for overcomplete dictionaries ($N < L$). Such dictionaries are required for efficient image modeling (see next Subsection). Since natural images can often be approximated by few atomic elements in such dictionaries, the synthesis approach is considered as more intuitive. Its design simplicity (in greedy approaches like CLEAN[11]) has also made it more popular in image processing applications. However, the synthesis solution is restricted to a column subspace of the synthesis dictionary, and the significance of each selected atom is important. On the other hand, the analysis approach may be more robust to "false detections" since the signal is not built from a few number of atoms. Besides, note in Equation (4.12) and (4.13) that the number of unknown in the synthesis case (the number of atoms in the dictionary) can be much larger than in the analysis case (where it remains in the number of pixels).

[11]The greediness of CLEAN is visible in the atom's (shifted PSF) selection rule: select the atom that is the most correlated with the data, that, is, the one that most decreases the norm of the residual.

Thus, analysis-based optimization strategies can be computationally much more efficient for large dictionaries. The comparison of both models is a very active field of research in fundamental signal processing (see the references in Gribonval *et al.* 2009, 2012).

Once a sparsity promoting model is chosen, we still need to choose an appropriate dictionary to perform the deconvolution.

4.5.3 Dictionaries

The sparsity is expressed via dictionaries, which correspond to representation spaces. Dictionaries express geometrical features that are likely to describe the unknown object. In synthesis, the columns of the dictionary are simply the atoms. In analysis, the rows may be atoms as well, or operators (gradient for instance, leading with ℓ_1 norm to the total variation regularization). These dictionaries can be orthonormal transforms (corresponding to orthonormal bases), or more generally redundant (overcomplete) dictionaries. A large variety of representations has been elaborated in the image processing literature, *e.g.*, canonical basis indeed (corresponding to point-like structures), Discrete Cosine Transform (DCT, 2-D plane waves), wavelets (localized patterns in time and frequency), isotropic undecimated wavelets, curvelets, ridgelets, shapelets and many others, see Mallat (2008) and Starck *et al.* (2010) for detailed reviews.

The choice of a dictionary is made with respect to a class of images. In Astronomy, wavelets dictionaries are widely used, but they are known to fail representing well anisotropic structures. In such cases other transforms can be used, that have been designed to capture main features of specific classes of objects. Among them, curvelets sparsify well curved, elongated patterns such as planetary rings or thin galaxy arms for instance; shapelets sparsify well various galaxy morphologies, etc. All these dictionaries have shown empirical efficiency for some specific types of images.

In order to model efficiently complex images with various and different features, several authors have proposed to concatenate dictionaries into a larger dictionary (Chen *et al.* 1998; Gribonval *et al.* 2003). However, the efficiency of a dictionary also critically depends on its size and on the existence of fast operators, without which iterative algorithms cannot run in reasonable time. This is especially true in radiointerferometry and in (possibly polychromatic) optical interferometry where the number of Fourier samples and of pixels can be of the order of hundreds of thousands.

4.6 CS and sparsity in astronomical deconvolution

Since the Compressed Sensing (CS) theory has emerged, providing exciting and beautiful mathematical results about sparse recovery in various cases, the sparsity ideas have benefited from considerable new strengths in the fields of signal and image processing (Donoho 2006; Candès *et al.* 2006).

In the context of the image restoration problem posed by interferometric measurements, the CS theory has provided theoretical proofs that, in idealized

situations, exploiting sparsity is indeed useful. For instance, the CS theory explains why a few point sources may be recovered from random Fourier measurements that are in number far less than specified by the sampling theorem. Of course, this possibility is exploited and implemented in long-existing restoration methods like CLEAN for instance; sparse methods have grown and evolved on their own before CS. They have lead to several types of elaborated sparsity-based algorithms, whose use evidences decades of empirical success.

As in many other applicative fields, references to CS have started hatching in quite a few recent publications about astronomical deconvolution. Yet, one key ingredient in that matter − sparsity − is exploited since at least Högbom's time. Besides, the theoretical CS proofs invoked in the introduction of many such publications turn not to help much in the subsequently proposed restoration methods. This poses the question of the real benefits brought by CS to astronomical deconvolution.

Let us consider the question from an operational point of view, that is, with the concern of better estimating $\mathbf{o}$ from the data $\mathbf{y}$. The benefits from CS with this respect are real but indirect, and they appear to be the following. First, CS clearly drained an increased research effort in fundamental models for sparse representations, like those of Equations (4.12) and (4.13). This in turn lead to improved reconstruction methods, through more elaborated statistical data models. Second, a lot of efficient optimization strategies have been designed to solve problems of the type (4.12) and (4.13), thanks to the new strengths in this field brought by the appealing theoretical results of CS.

In the recent years, sparsity promoting methods were used in interferometry by Wiaux *et al.* (2009a,b), using Basis Pursuit DeNoising (Chen *et al.* 1998) with wavelets dictionaries, and by Vannier *et al.* (2010), with Matching Pursuit algorithms in unions of bases (wavelets/Dirac). Li *et al.* (2011) adopted a synthesis approach with an IUWT (Isotropic Undecimated Wavelet Transform) synthesis dictionary, and solved a Basis Pursuit synthesis criterion through the ISTA minimization algorithm (Iterative Soft-Thresholding Algorithm) and its fast version, FISTA (Fast Iterative Shrinkage-Thresholding Algorithm, Beck & Teboulle 2009). Carillo *et al.* (2012) applied a reweighted ℓ_1 analysis algorithm promoting average signal sparsity over multiple redundant dictionaries, and relying on convex optimization techniques. Dabbech *et al.* (2012) have proposed an hybrid analysis-by-synthesis approach: $\mathbf{o}$ is modeled using sparse synthesis priors as a sum of few objects which, as opposed to classical synthesis-based priors, are unknown. These atoms are iteratively estimated through analysis-based priors, the analysis being based on an IUWT dictionary.

Note that in optical interferometry, the 2012 international Beauty Contest also witnessed an increasing number of sparsity based methods (Baron *et al.* 2012). Finally, in polychromatic optical interferometry, Thiébaut *et al.* (2012) proposed to favor spatial sparsity and spectral grouping of the sources through an alternating direction method of multipliers, a method also issued from the convex optimization literature (see the Article of É. Thiébaut in these proceedings).

We will not include numerical results about restoration algorithms exploiting sparsity in this paper. Examples of results obtained with sparsity promoting methods (comparison of CLEAN, MP and BP) in the case of diluted apertures (Fizeau configuration) can for instance be found in Section 4 of Vannier *et al.* (2010).

To conclude this part, we see that numerous techniques are emerging. They offer sophisticated alternatives to the more traditional and robust constrained ML methods. Indeed, most of these methods comply with the non-negativity constraint. The algorithms described above rely on recent progresses in sparse representations and convex optimization techniques. They allow to solve large scales optimization problems involving complex image models, and they are becoming increasingly popular in interferometry.

We now turn to simulations results aimed at illustrating the effect of densification in the presence of photon noise. We then investigate the possibility of recovering small objects that are far from the optical center.

5 To densify, or not to densify

Of course, the ambition of this section is not to provide a general answer to this question, as many factors should be accounted for (for instance the number of detector pixels, the noise level and its statistical nature, the subupil configurations, etc.). As already emphasized, the FSD densified and the Fizeau images contain the same frequency information. Since the densified image has a lower frequency, it needs less detector pixels to be properly sampled than its Fizeau counterpart. This should be an advantage of hypertelescopes, which will not be illustrated here as this is a straightforward consequence of the sampling theorem. See for instance Lardière *et al.* (2007) for useful insights on these issues.

The question on which we focus here is the following. We are given a reference object $\mathbf{o}$ (the one of Fig. 6) and two sampling schemes (the previously described Fizeau and Michelson FSD configurations, Fig. 8), leading to one-million-pixel images that are contaminated by photon noise (Fig. 11). In these conditions, which scheme leads statistically to the best restored images using a RL algorithm?

We propose to answer this question empirically, by running Monte Carlo simulations. We generated 50 photonized Fizeau images and the same amount of photonized FSD images. The images of the latter set were numerically rediluted so that the object-image model is a convolution. The difference between the two sets of images is in the noise statistics. While it is Poissonian for the first set, this is not the case for the set of rediluted images. These images can (and actually do) exhibit negative values. Thus the RL algorithm, seen as a ML method, is not justified any more because both the image positivity and the Poisson statistics are lost. However, RL can be (as ISRA) simply taken as a deconvolution method which minimizes some loss function between the data and convolved model (the Kullback-Leibler divergence for RL, and the quadratic error loss for ISRA). If non-negative restored objects are expected, care must be taken in this case that the data images are non-negative. This is achieved by setting to 0 the negative values of the rediluted images. We observed for the considered noise level that negative

data values in the rediluted images are very few (typically 5 out of 10^6) and close to 0 (typically less than 0.01 in absolute value). So this non-negativity precaution causes a negligible information loss.

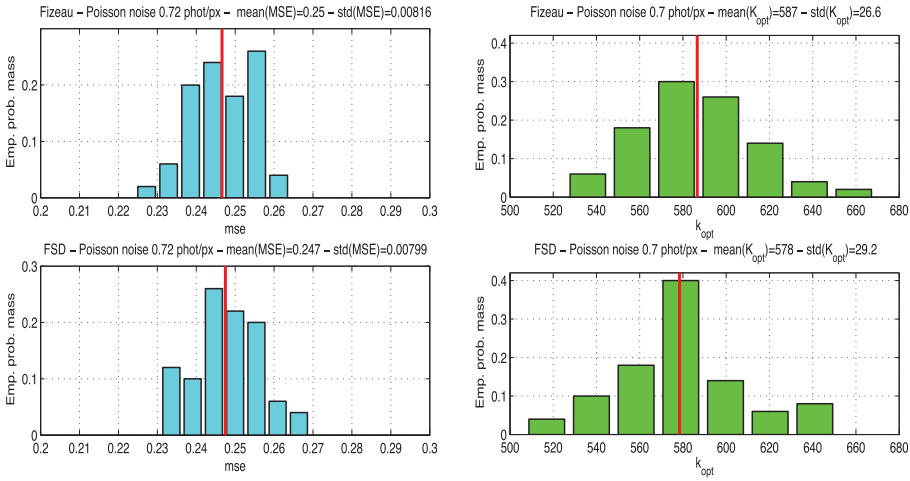


Fig. 14. *Left*: empirical distributions of the best reconstruction errors obtained by RL for photonized Fizeau images (*top*) and rediluted photonized FSD images (*bottom*), for 50 realizations. The average photon noise in the data images corresponds to 0.72 photon/pixel, and the images have 1024×1024 pixels. The vertical line shows the empirical mean of the distribution (which is also indicated in the titles of the figures along with the observed dispersion). *Right*: empirical distributions of the iteration number leading to the best reconstruction (*vertical line*: empirical means; values for means and standard deviations are indicated in the titles).

The results are presented in Figure 14. It is clear from this experiment that the results are essentially equivalent in terms of the quality of the reconstruction error and of the optimal number of iterations. The equivalence in terms of information that holds between Fizeau and FSD configurations appears conserved in images affected by photon noise.

6 Noiseless recovery of a small object outside the "clean" field

6.1 Introduction: Objectives and simulation parameters

We are interested here in the restoration of noiseless images obtained in the Fizeau configuration, or in any densified configuration without spectral aliasing. The objective of this study is to investigate whether a quasi-point source which is located outside the "clean" field (see below) can or not be restored by the RL algorithm. The noise is not considered in order to focus on the effects of the sampling.

The "clean" field (Lardière *et al.* 2007) is the central zone of the image of dimension λ/s, where s is the smallest spatial distance between two subpupils. The global field corresponds to the diffraction envelope of the elementary pupils.

The considered sampling is the same as described in Section 3: 25 non redundant circular apertures on an integer grid. These apertures have the same diameter D, and the principal lobe of the diffraction envelope defines the global field, which has diameter $2 \times 1.22\lambda/D$. For the considered array, the centers of the elementary OTF in the Fizeau sampling are separated by $7\tau_\nu$. Thus, the central part of the Fizeau image is essentially replicated 7 times in each direction, and λ/s corresponds to $1024/7 \approx 146$ pixels (*cf.* Fig. 15, bottom left).

The object we consider is presented in Figure 15. The flux ratio between the central planet and the satellite is $\approx 4.8 \times 10^{-3}$, which corresponds to a difference in magnitude of ≈ 5.8.

6.2 Recovery without spatial aliasing

In this first simulation, the small object is not located on a replica of the main object, but it is quite far from the center (close to the limit set by the global field), and thus highly attenuated. This source is centered around the pixel coordinates $(x = 140, y = 513)$, and is 373 pixels away from the centre $(x = 513, y = 513)$ of the object. This angular distance represents $\approx 2.5 \times \lambda/s$, or $\approx 0.73 \times 1.22\lambda/D$.

Figure 16 illustrates the evolution of the deconvolution along the iterations.

Interestingly, we see that the algorithm first reconstructs a satellite close to the central object, and then transfers the flux from this position to the left by discrete jumps of λ/s (*i.e.*, the clean field), to finally reach the good position:

- Iteration 100: the central planet appears, the replicas and the halo have almost disappeared. No satellite yet.
- Iteration 3500: the central planet starts being fairly well estimated, and a quasi-point source is restored in the vicinity of the planet, at pixel $(x = 432, y = 513)$, *i.e.* at $2\lambda/s$ right of the real position of the satellite $(432 = 140 + 2 \times 146)$.
- Iteration 4500: a second point source appears at pixel $(x = 286, y = 513)$, *i.e.* at λ/s right of the real position. The flux of the first (fake) quasi-point source has decreased with respect to iteration 3500.
- Iteration 5800: The first fake satellite at $(x = 432, y = 513)$ has disappeared.
- Iteration 7800: A third satellite at the right position $(x = 140, y = 513)$ appears. The flux of the second satellite at $(x = 286, y = 513)$ decreases.
- Iteration 30 000: The second fake satellite $(x = 286, y = 513)$ has almost disappeared. The algorithm has (almost) converged to a correct reconstruction of the central planet and of its satellite.

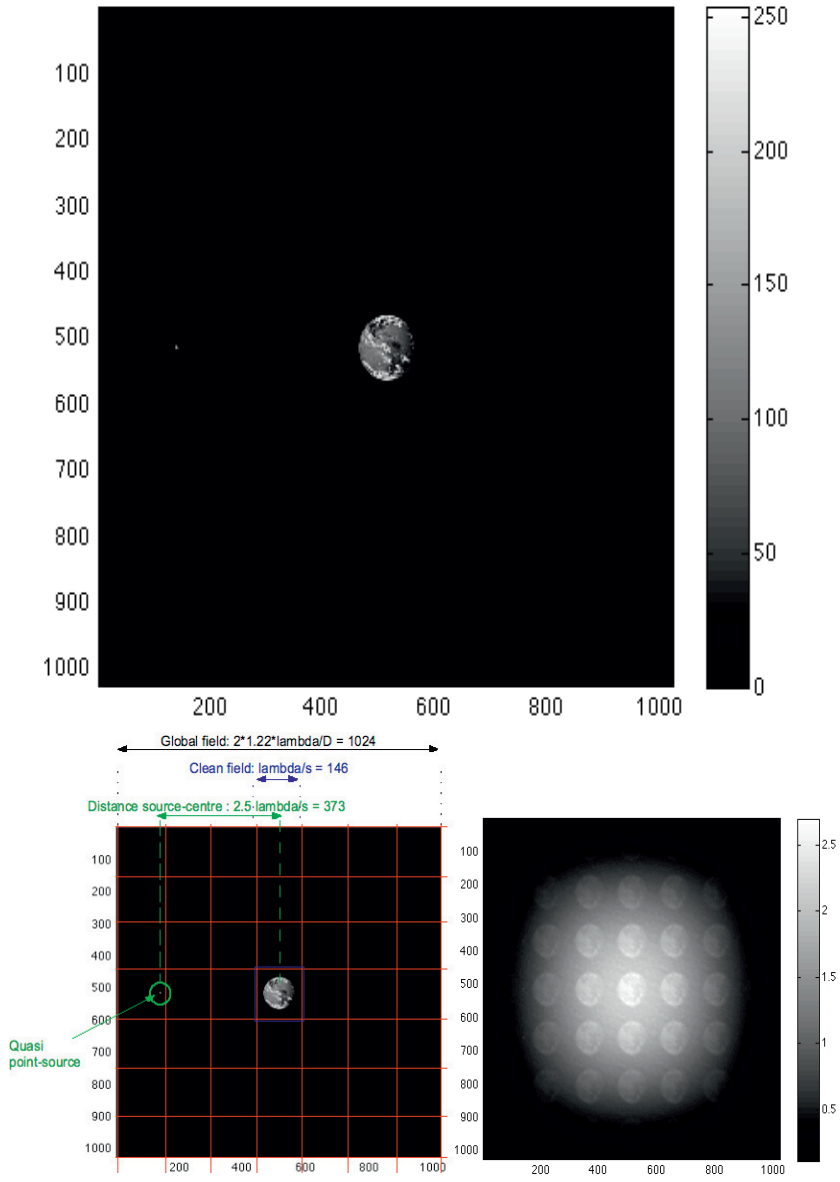


Fig. 15. *Top*: reference object (planet-satellite). *Bottom*, *left*: same object with the fields shown. The global field represents 1024 pixels in each direction, and the clean field 146 pixels. The quasi-punctual source contributes flux in 28 pixels. It is 373 pixels away from the centre, which corresponds to $\approx 2.5\lambda/s$. *Bottom right*: image obtained with the considered diluted pupil.

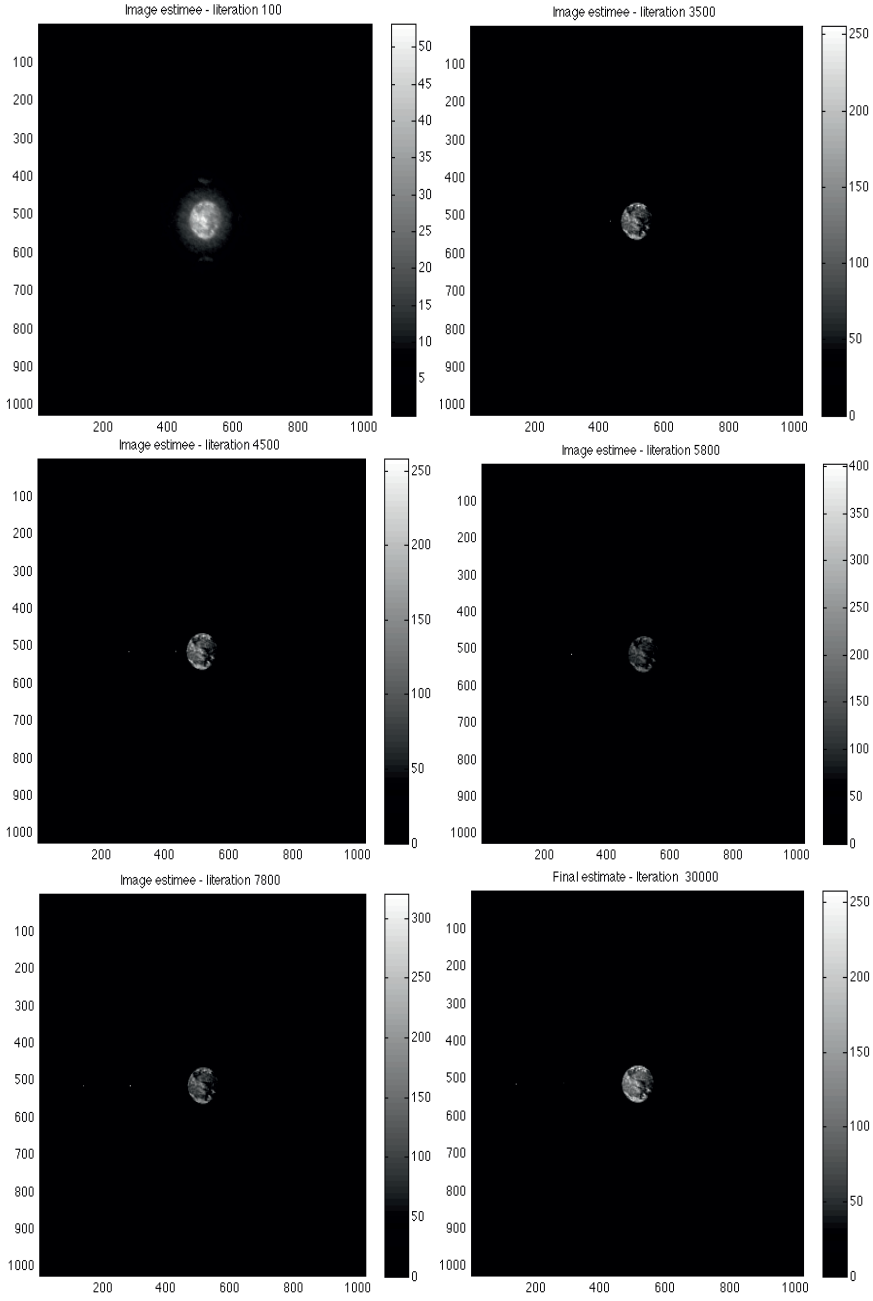


Fig. 16. Snapshots of restored object for some iterations of the RL algorithm. In lexicographic order $k = 100$, 3500, 4500, 5800, 7800, 30 000.

The satellite reconstruction by jumps of extension λ/s can be followed in the Fourier space. As illustrated below, the RL algorithm fills the Fourier space by progressive interpolation of the spectrum around the available samples. The moduli of the frequency samples where information is available are represented in Figure 17, left, and the total spectral information to be recovered is in Figure 17, right. Figure 18 shows the same for the phases. The satellite information appears essentially as a modulation on the moduli and on the phases of the central planet's spectrum.

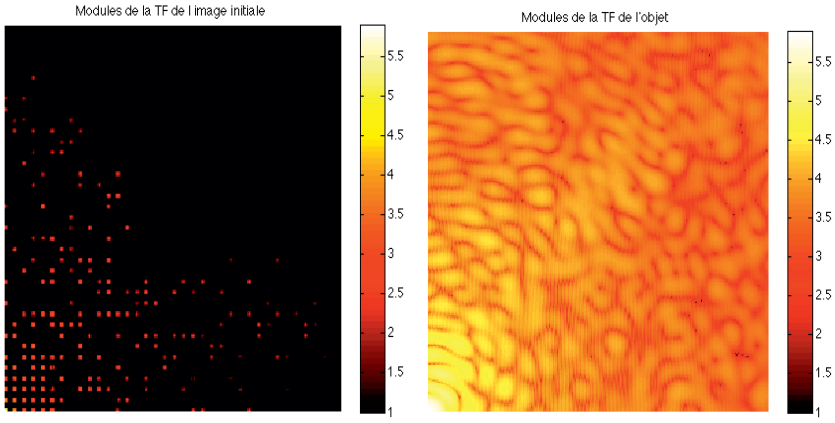


Fig. 17. Zoom on the moduli of the Fourier spectra. *Left*: available moduli (the missing samples are in black). *Right*: moduli of the spectrum of the considered planet-satellite system.

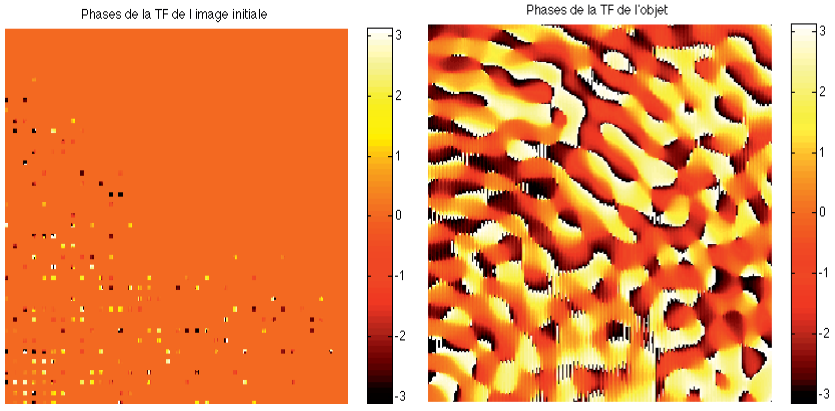


Fig. 18. Same as Figure 17, but for the phases.

As the iterations go, the "holes" at low frequencies are progressively filled, and the high frequencies are then estimated, as illustrated in Figure 19.

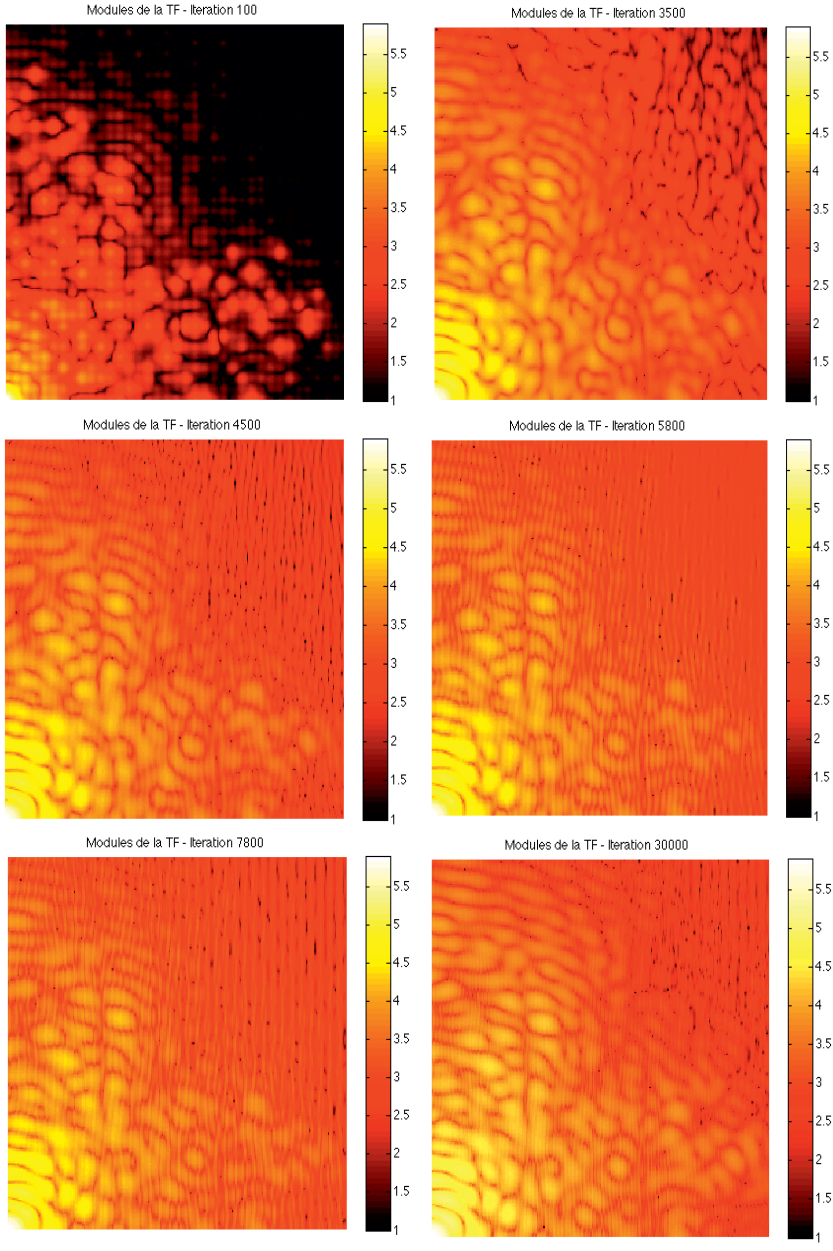


Fig. 19. Zoom on the moduli of the estimated objects for the iterations of the RL algorithm shown in Figure 16 (in lexicographic order $k = 100$, 3500, 4500, 5800, 7800, 30 000).

After 30 000 iterations, the object restored by the algorithm is relatively close to the original, at least as far as the satellite position is concerned. The Figure 20 presents a zoom on the central planet (left column) for the reference object (top, left) and on the deconvolved object (middle, left). Similarly, the right column shows a zoom on the satellite of the reference object (top, right) and deconvolved (middle, right). The total flux of the deconvolved satellite is about 60% of the total flux of the reference satellite. Of course the deconvolution is not perfect (and in cannot be, as too many frequency are lost by the sampling). But the result is comparable to the direct image that would be produced by a monolithic Extremely Large Telescope having the same diameter as the largest base of the hypertelescope[12] (Fig. 20, bottom row).

The important thing is that the point source is fairly recovered, a point which was not obvious considering the ambiguity posed by the sampling scheme. This result is encouraging, efforts for the quest of high angular resolution do not seem to be vain. We may pause here to think that some day in the future, the detection of such a faint little point in the dark corner of a real hypertelescope image might be the origin of a great discovery for the Human civilization.

Enough dreams for now, *sine experientia nihil sufficienter sciri potest*: let us come back to the prosaic reality of simulations and try a more difficult recovery.

6.3 Recovery with spatial aliasing

The considered object is still of the planet-satellite type but the satellite replicas are superimposed on the replicas of the central object in the data image, see Figure 21.

The point source is now at coordinates $(x = 213, y = 513)$, which is 300 pixels away from the centre $(x = 513, y = 513)$. This represents $\approx 2\lambda/s$. As visible in Figure 21 right, this is a clear case of spatial aliasing. The results of the deconvolution for some iterations are presented in Figure 22.

- Iteration 100: the planet starts being well restored. No satellite in the vicinity. Note that a bright spot, of about the satellite size, is created on the central planet. This spot comes from the spatial aliasing (replica of the satellite superimposed on the planet).
- Iteration 3000: the reconstruction seems to stabilize on an object without satellite, with a surface spot at the place of the satellite. It is unclear whether the algorithm will be able to distinguish between a satellite to be placed further away, and a bright surface spot.
- Iteration 5000: a satellite appears around pixel $(x = 359, y = 513)$, that is, at λ/s right of the real position. In the same time, the bright artifact at the center is less visible than in the previous iterations.

[12] This ideal telescope is called "Metatelescope" in Aime *et al.* (2012).

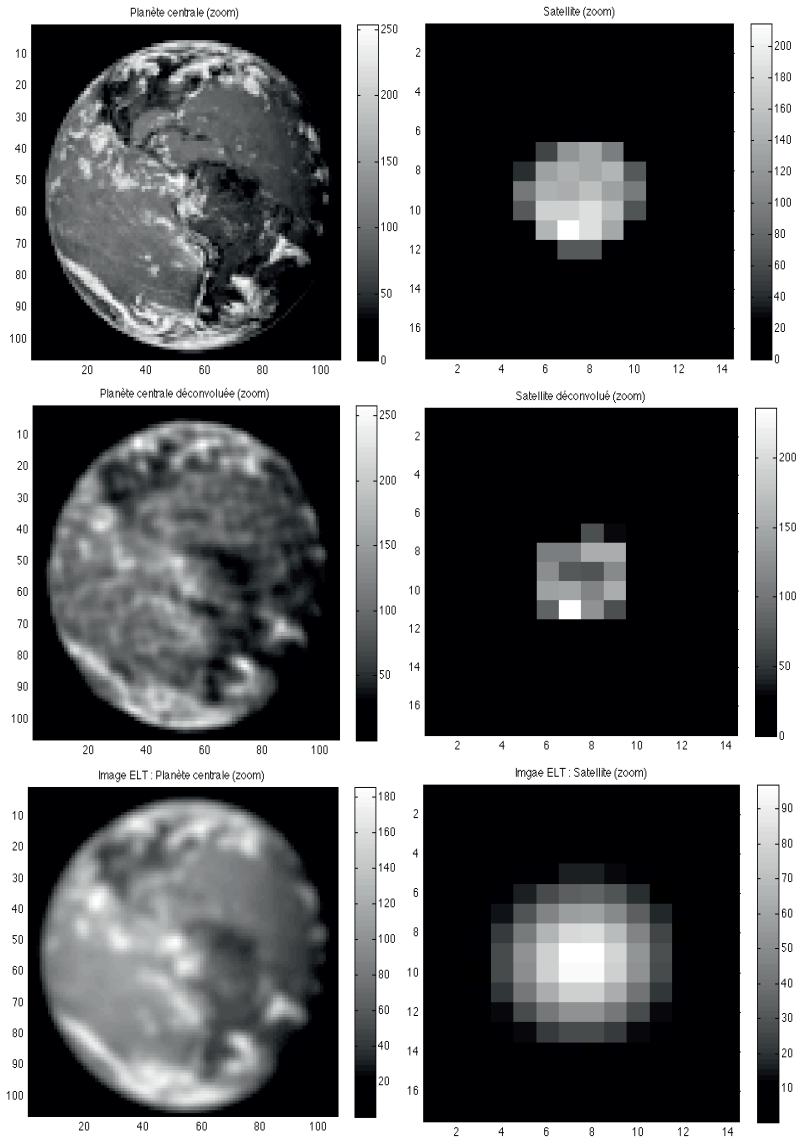


Fig. 20. Zoom on the central planet (*left colum*) and on the satellite (*right column*). *Top row*: reference object. *Middle row*: deconvolved object after 300 000 RL iterations. *Bottom row*: Image of a monolithic ELT having the same high frequency cut-off as the diluted array.

– Iteration 40 000: a second satellite appears at the real position. The flux attributed to the first fake satellite decreases, and the artifact at the centre of the planet is almost not visible anymore.

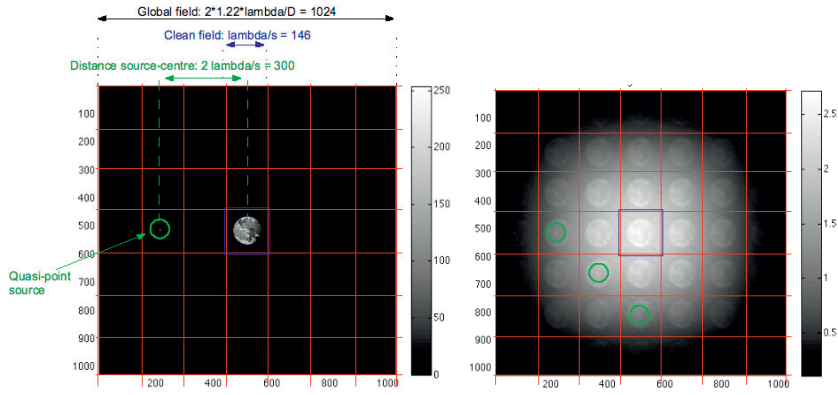


Fig. 21. Case where the satellite "falls" in a replica of the central planet. *Left*: considered system planet-satellite. *Right*: image produced by the diluted array. The light circles evidence the spatial aliasing on some replicas (it is the case for all of them).

– Iterations 80 000 et 200 000: the flux is progressively transferred from the first to the second satellite, but the convergence is very slow. Figure 23 zooms on the object deconvolved at iteration 200 000: zoom on the central planet (left) and on the satellite position (right). The flux estimated for the satellite at the real position is still insufficient (compare to Fig. 22).

We see that although the convergence is slow, the algorithm is on the way to find the right configuration. Figure 24 presents, for each estimated object $\mathbf{o}^{(k)}$ at iteration k, $k = 1, \ldots, 200\,000$, the normalized error in approximating the data (left), and the normalized error with respect to the true object (right).

We see that the convergence is very slow. Note also that the error with respect to the object is not constantly decreasing. An intermediate solution corresponding to a local maximum ($k \approx 8000$) corresponds to an estimated object with one satellite that is too close to the planet. But this solution does not perfectly explain the data. Some flux then starts being injected at the right position, so that the error decreases again. These results suggest that the right configuration can be recovered, even if the flux is not perfectly estimated, at least with negligible noise.

How is the algorithm able to find out, from data where the satellite is everywhere superimposed on the surface of the central object, that there is satellite, and that the surface has no bright spot? The reason is that the if the central planet had a bright surface spot, this spot should be less bright in replicas that are further away from the center (because of the diffraction envelope). But this is not the case for the replicas of the satellite: the brightest replicas are the ones that are close to the true position of the satellite. This discrepancy makes the algorithm to eventually inject the flux at the right position. In other words, it is the diffraction envelope which saves the reconstruction here.

A last remark. To see to which solution the algorithm would eventually converge, and how accurate the recovery would be in this case, there should be several

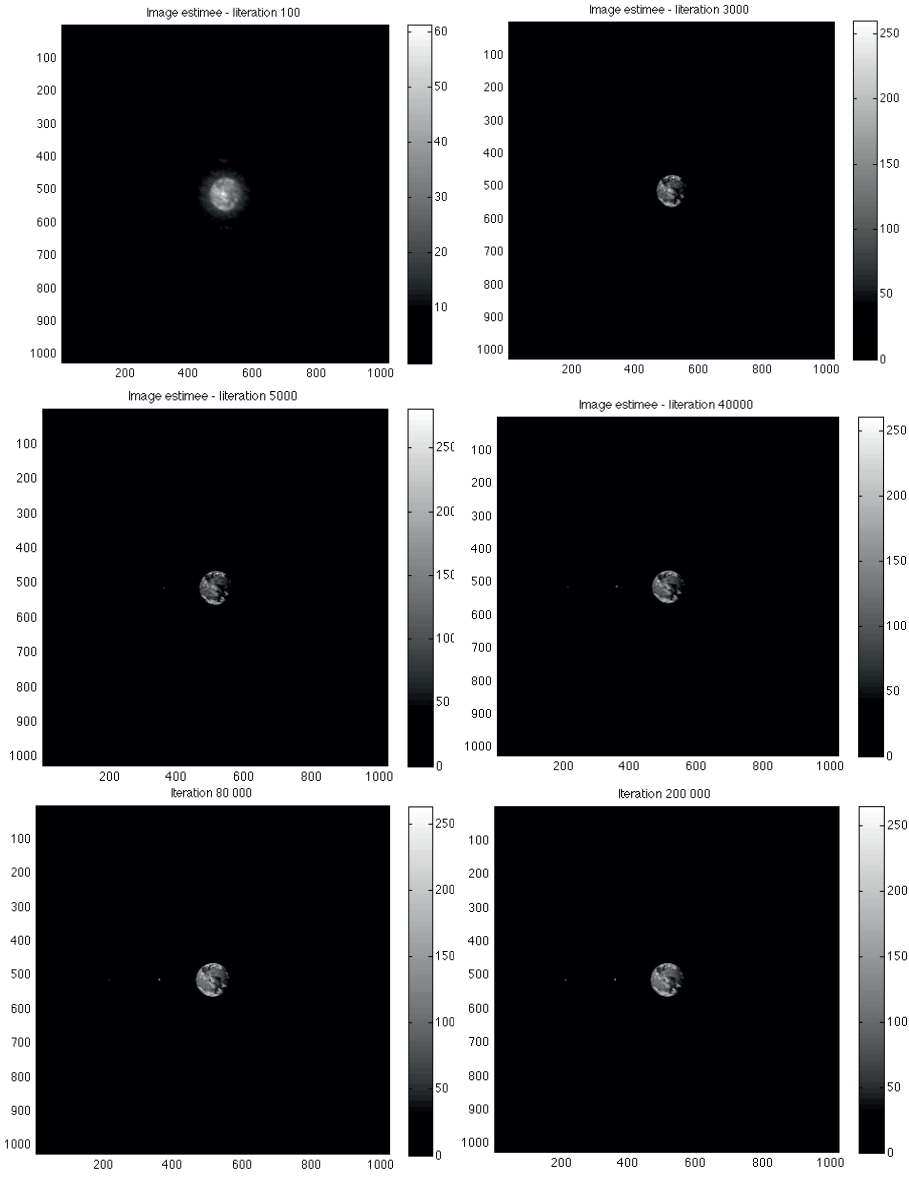


Fig. 22. Estimated objects for some iterations of the RL algorithm. By lexicographic order: $k = 100$, 3000, 5000, 40 000, 80 000, 200 000. At iteration $k = 40\,000$ a source starts being visible at the right position.

hundred thousands iterations more. This is very time consuming: 200 000 RL iterations on 1024×1024 images represent ≈80 h on a standard laptop. This clearly illustrates the importance of designing fast algorithms for image restoration.

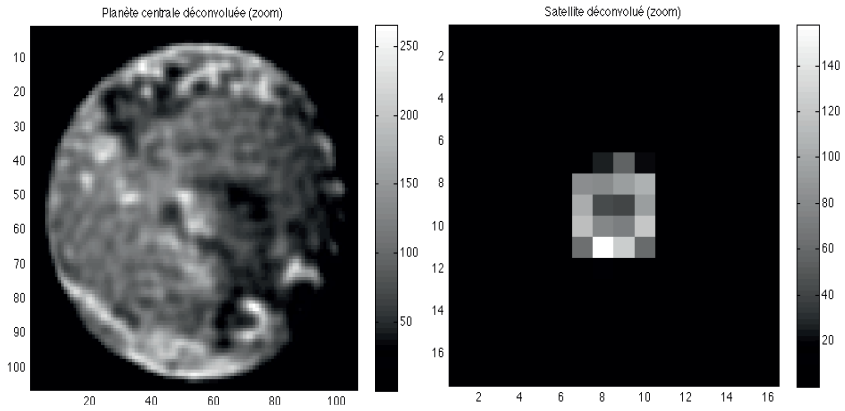


Fig. 23. Deconvolved object in the case of spatial aliasing by RL after 200 000 iterations: zoom on the central planet (*left*) and on the satellite position (*right*).

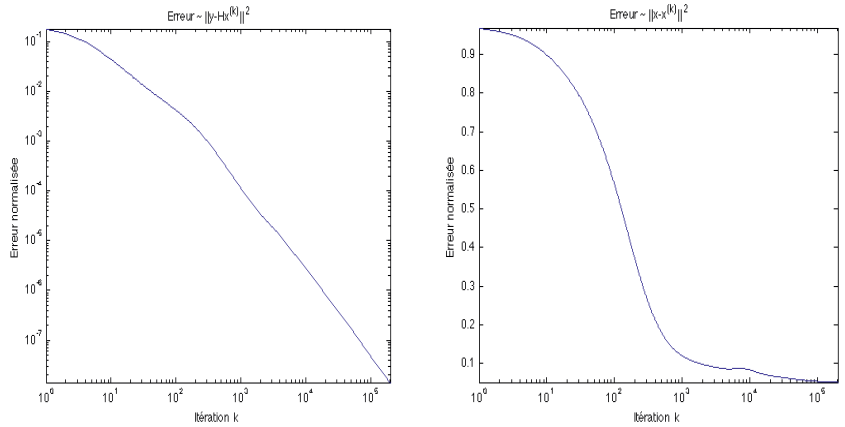


Fig. 24. Normalised error in approximating the data $\frac{||\mathbf{y}-H\mathbf{o}^{(k)}||^2}{||\mathbf{y}||^2}$ (*left*), and normalized error with respect to the true object $\frac{||\mathbf{o}-\mathbf{o}^{(k)}||^2}{||\mathbf{o}||^2}$ (*right*) as a function of the iteration number.

7 Summary and conclusions

This article tried to provide a detailed introduction to the description of the image formation models for diluted pupils array and their densified versions called hypertelescopes. These optical systems represent one of the main promises for the next generation of high angular resolution instruments.

The introduction underlined using historical elements how essential have been high angular resolution observations, transmission of knowledge, and reliance on long term research projects to our current representation of the Universe.

A substantial part of the paper was devoted to the explanations of sampling issues, of their effects on the observed images, and of possible settings that can be used to simulate hypertelescopes images.

In the Fizeau mode (no densification), the image model is a convolution. The densified mode corresponds to hypertelescopes and can be done using either a periscopic setting or inverted Galilean telescopes. We showed in the Appendix that both settings are fully equivalent. In the densified case, convolution generally disappears because frequencies are modulated (translated block-wise), and FAD yields information loss while FSD does not. A convolution may be retrieved in the FSD mode only in the limit of vanishingly small subapertures (infinite fields of view).

This suggests two modeling regimes for densification (hypertelescopes), depending on whether the diameter D is much smaller than the pupil separation d or not. For hypertelescopes made of very large bases (in the kilometer range) and of many small telescopes (centimeters), $D << d$, and a convolution model may be a good approximation, at least close to the optical axis. For VLTI-like hypertelescopes, made of moderately large bases (in the hundreds of meter) and of a few large telescopes (in the tens of meters), $D \approx d$ and the image formation models strongly departs from convolution.

We also addressed the issue of restoring such images, and presented classical methods of constrained ML for Gaussian and Poisson noises (RL and ISRA). Faster and regularized deconvolution algorithms should be preferred to RL and ISRA. We provided a detailed survey of such recent methods based on sparse representations.

The two last sections of the paper were dedicated to original studies.

The first study showed that the restoration quality achieved by constrained ML from photon limited images obtained from a diluted array on a grid, or from a densified (but free from spectral aliasing) array are essentially equivalent. We still expect a gain of densified w.r.t. Fizeau images because of the relatively lower cutoff frequency of the former, although we did not provide results supporting this assertion.

The second study (last section) showed that it is possible to recover or at least to "detect" in hypertelescopes (or more generally, interferometric) images quasi point sources that are not only far outside the clean field, but also superimposed on the replicas of other objects. This is true at least for the considered pupil array and in the limit of no noise. Further studies should investigate the effect of noise on the recovery, and of the magnitude difference for the satellite to be recoverable.

Appendix: Densification using Galilean inverted telescopes and recovery of former periscopic expressions

The densification of hypertelescopes can be operated in two ways: using a periscope as in Michelson's stellar interferometer, or using Galilean inverted telescopes. In the first case, the distance between the subapertures is reduced in the output pupil with respect to the input pupil, while their diameter remains fixed. In

the second case, the relative distances between the subapertures is conserved but their diameter is magnified. In practice of course, the images are rescaled in both settings. We show here, by treating in detail the image formation model of Galilean inverted telescopes, that these two settings are equivalent.

Let us consider a monochromatic plane wave of amplitude $A(\boldsymbol{\beta})$ coming from an angular direction $\boldsymbol{\beta}$ (β_x, β_y) on the sky, and emitted by an object of intensity $O(\boldsymbol{\beta}) = |A(\boldsymbol{\beta})|^2$. This wave produces at position $\mathbf{r}$ (r_x, r_y) in the plane of the input pupil of a telescope an amplitude $A(\boldsymbol{\beta}) \exp(2i\pi\boldsymbol{\beta}.\mathbf{r}/\lambda)$, where the phase factor accounts for the tilt of the wavefront and the bold dot means scalar product.

The wave $\Psi_1(\mathbf{r}, \boldsymbol{\beta})$ in the input pupil plane of an interferometer made of an array of K cophased identical apertures $P_0(\mathbf{r})$ centered at spatial positions $\mathbf{r}_k$ can thus be written as

$$\begin{aligned}\Psi_1(\mathbf{r}, \boldsymbol{\beta}) &= A(\boldsymbol{\beta}) \sum_{k=1}^{K} P_0(\mathbf{r} - \mathbf{r}_k) \exp\left(2i\pi \frac{\boldsymbol{\beta}.\mathbf{r}}{\lambda}\right) \\ &= A(\boldsymbol{\beta}) \sum_{k=1}^{K} P_0(\mathbf{r}) \exp\left(2i\pi \frac{\boldsymbol{\beta}.\mathbf{r}}{\lambda}\right) \star \delta(\mathbf{r} - \mathbf{r}_k) \exp\left(2i\pi \frac{\boldsymbol{\beta}.\mathbf{r}_k}{\lambda}\right), \qquad (7.1)\end{aligned}$$

where the last form was first used by Tallon & Tallon-Bosc (1992) to treat the effect of the periscopic transformation in Michelson interferometry. This form explicits the separation between the positions and the geometry of the elementary apertures.

For a hypertelescope, the densification using the periscopic mode basically consists of translating the apertures images from the positions $\mathbf{r}_k$ to the new positions $\mathbf{r}'_k = \mathbf{r}_k/\gamma$, where γ is called the densification factor (Labeyrie 1996). These aspects have been presented in several papers (Tallon & Tallon-Bosc 1992; Labeyrie 1996; Lardiere *et al.* 2007; Aime 2008; Aime *et al.* 2012) and will not be further detailed here.

In contrast to these papers, we present here the formalism for the densification using inverted Galilean telescopes, and show that it leads to results that are identical to the periscopic technique. From a physical point of view this is expected since the two images of the resulting apertures are identical, up to an irrelevant magnifying factor. Nevertheless, the presentation of the theory for the inverted Galilean telescope approach is of interest, at least from a pedagogic point of view.

Using inverted Galilean telescopes for densification amounts to applying a magnification by a real factor $\gamma > 1$ of the wave on each elementary aperture, leaving unchanged the center positions $\mathbf{r}_k$. In this operation the amplitude of the light is divided by the factor γ, to keep unchanged the energy. In Equation (7.1) this aperture reshaping consists in applying the dilation factor γ to the first term of the convolution. Let us denotes $\Psi_\gamma(\mathbf{r}, \boldsymbol{\beta})$ this amplitude:

$$\Psi_\gamma(\mathbf{r}, \boldsymbol{\beta}) = \frac{A(\boldsymbol{\beta})}{\gamma} \sum_{k=1}^{K} P_0\left(\frac{\mathbf{r}}{\gamma}\right) \exp\left(2i\pi \frac{\boldsymbol{\beta}.\mathbf{r}}{\gamma\lambda}\right) \star \delta(\mathbf{r} - \mathbf{r}_k) \exp\left(2i\pi \frac{\boldsymbol{\beta}.\mathbf{r}_k}{\lambda}\right). \qquad (7.2)$$

For $\gamma = 1$ we obviously recover the original wavefront $\Psi_1(\mathbf{r}, \boldsymbol{\beta})$ of Equation (7.1).

Now let $A_\gamma(\boldsymbol{\alpha},\boldsymbol{\beta})$ denote the complex amplitude of the wave in the focal plane of the telescope, at angular position $\boldsymbol{\alpha}(\alpha_x,\alpha_y)$ in this plane. This wave is obtained by a scaled Fourier transform of $\Psi_\gamma(\mathbf{r},\boldsymbol{\beta})$ (see Aime *et al.* in these proceedings):

$$\begin{aligned} A_\gamma(\boldsymbol{\alpha},\boldsymbol{\beta}) &= \frac{1}{i\lambda}\iint \Psi_\gamma(\mathbf{r},\boldsymbol{\beta})\exp\left(-2i\pi\frac{\mathbf{r}.\boldsymbol{\alpha}}{\lambda}\right)\mathrm{d}\mathbf{r} \\ &= \frac{\gamma A(\boldsymbol{\beta})}{i\lambda}\sum_{k=1}^{K}\widehat{P_0}\left(\frac{\gamma\boldsymbol{\alpha}-\boldsymbol{\beta}}{\lambda}\right)\exp\left(-2i\pi\mathbf{r}_k.\left(\frac{\boldsymbol{\alpha}-\boldsymbol{\beta}}{\lambda}\right)\right). \end{aligned} \tag{7.3}$$

The factor γ appears now at the numerator after a change of variable in the 2D integral. The elementary intensity in the case of inverted Galilean telescopes $I_\gamma^G(\boldsymbol{\alpha},\boldsymbol{\beta})$ produced by the point source coming from the direction $\boldsymbol{\beta}$ at position $\boldsymbol{\alpha}$ in the focal plane can therefore be written as:

$$\begin{aligned} I_\gamma^G(\boldsymbol{\alpha},\boldsymbol{\beta}) &= |A_\gamma(\boldsymbol{\alpha},\boldsymbol{\beta})|^2 = A_\gamma(\boldsymbol{\alpha},\boldsymbol{\beta})A_\gamma^*(\boldsymbol{\alpha},\boldsymbol{\beta}) \\ &= O(\boldsymbol{\beta})\frac{\gamma^2}{\lambda^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\left|\widehat{P_0}\left(\frac{\gamma\boldsymbol{\alpha}-\boldsymbol{\beta}}{\lambda}\right)\right|^2\exp\left(2i\pi(\mathbf{r}_k-\mathbf{r}_l).\left(\frac{\boldsymbol{\alpha}-\boldsymbol{\beta}}{\lambda}\right)\right), \end{aligned} \tag{7.4}$$

where superscript * denotes complex conjugate. $I_\gamma^G(\boldsymbol{\alpha},\boldsymbol{\beta})$ is indeed real (the imaginary parts of the complex exponentials involving $\mathbf{r}_k-\mathbf{r}_l$ and $\mathbf{r}_l-\mathbf{r}_k$ cancel by pairs); the notation with complex exponentials will later evidence a Fourier transform that will be used in Equation (7.7).

The image in the focal plane $I_\gamma^G(\boldsymbol{\alpha})$ is obtained by summing all contributions coming from the object:

$$I_\gamma^G(\boldsymbol{\alpha}) = \iint I_\gamma^G(\boldsymbol{\alpha},\boldsymbol{\beta})\mathrm{d}\boldsymbol{\beta}. \tag{7.5}$$

Taking the Fourier transform of $I_\gamma^G(\boldsymbol{\alpha})$, we also have:

$$\widehat{I_\gamma^G}(\mathbf{u}) = \iint \widehat{I_\gamma^G}(\mathbf{u},\boldsymbol{\beta})\mathrm{d}\boldsymbol{\beta}, \tag{7.6}$$

where $\mathbf{u}$ is the angular frequency associated to $\boldsymbol{\alpha}$.

Using the notation $\mathbf{u}_{kl} = (\mathbf{r}_k-\mathbf{r}_l)/\lambda$ in Equation (7.4), the expression of $\widehat{I_\gamma^G}(\mathbf{u},\boldsymbol{\beta})$ can be written as:

$$\begin{aligned} \widehat{I_\gamma^G}(\mathbf{u},\boldsymbol{\beta}) &= \iint I_\gamma^G(\boldsymbol{\alpha},\boldsymbol{\beta})\exp(-2i\pi\mathbf{u}.\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\alpha} \\ &= O(\boldsymbol{\beta})\sum_{k=1}^{K}\sum_{l=1}^{K}\exp(-2i\pi\frac{\boldsymbol{\beta}}{\gamma}.(\mathbf{u}+(\gamma-1)\mathbf{u}_{kl})\iint|\widehat{P_0}(\boldsymbol{\xi})|^2\exp(-2i\pi\boldsymbol{\xi}.\frac{1}{\gamma}(\mathbf{u}-\mathbf{u}_{kl}))\mathrm{d}\boldsymbol{\xi} \end{aligned} \tag{7.7}$$

If we denote by S the telescope area and by $T_0(\mathbf{u})$ the normalized optical transfer function (OTF) defined by

$$T_0(\mathbf{u}) = \frac{1}{S}\iint P(\mathbf{r})P^*(\mathbf{r}-\lambda\mathbf{u})\mathrm{d}\mathbf{r}, \tag{7.8}$$

Equation (7.7) becomes

$$\widehat{I_\gamma^G}(\mathbf{u},\boldsymbol{\beta}) = O(\boldsymbol{\beta}) \sum_{k=1}^{K}\sum_{l=1}^{K} \exp\left(-2i\pi\frac{\boldsymbol{\beta}}{\gamma}.(\mathbf{u}+(\gamma-1)\mathbf{u}_{kl}\right)\ ST_0\left(\frac{1}{\gamma}(\mathbf{u}-\mathbf{u}_{kl})\right). \tag{7.9}$$

Substituting this expression into Equation (7.6), we obtain:

$$\widehat{I_\gamma^G}(\mathbf{u}) = \sum_{k=1}^{K}\sum_{l=1}^{K} \widehat{O}\left(\frac{1}{\gamma}(\mathbf{u}+(\gamma-1)\mathbf{u}_{kl})\right)\ ST_0\left(\frac{1}{\gamma}(\mathbf{u}-\mathbf{u}_{kl})\right). \tag{7.10}$$

This expression depends on the collecting surface of the telescope. We can get rid of this surface by dividing by KS, which finally leads to:

$$\begin{aligned}\widehat{\mathrm{I}_\gamma^G}(\mathbf{u}) &= \frac{1}{KS}\,\widehat{I_\gamma^G}(\mathbf{u}) \\ &= \frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{K} \widehat{O}\left(\frac{1}{\gamma}(\mathbf{u}+(\gamma-1)\mathbf{u}_{kl})\right)\ T_0\left(\frac{1}{\gamma}(\mathbf{u}-\mathbf{u}_{kl})\right) \\ \widehat{\mathrm{I}_\gamma^G}(\mathbf{u}) &= \widehat{O}\left(\frac{\mathbf{u}}{\gamma}\right)\ T_0\left(\frac{\mathbf{u}}{\gamma}\right) + \frac{1}{K}\sum_{k=1}^{K}\sum_{l\neq k}^{K} \widehat{O}\left(\frac{1}{\gamma}(\mathbf{u}+(\gamma-1)\mathbf{u}_{kl})\right)\ T_0\left(\frac{1}{\gamma}(\mathbf{u}-\mathbf{u}_{kl})\right).\end{aligned} \tag{7.11}$$

We see that the sampling in this case operates on a dilated version of the spectrum $\widehat{O}(\frac{\mathbf{u}}{\gamma})$ using transfer functions that are dilated as well. Performing the change of variable $\boldsymbol{\nu} = \mathbf{u}/\gamma$, we recover the periscopic mode:

$$\widehat{I_\gamma^P}(\boldsymbol{\nu}) = \widehat{O}(\boldsymbol{\nu})\ T_0(\boldsymbol{\nu}) + \frac{1}{K}\sum_{k=1}^{K}\sum_{l\neq k}^{K} \widehat{O}\left(\boldsymbol{\nu}+\mathbf{u}_{kl}-\frac{\mathbf{u}_{kl}}{\gamma}\right) T_0\left(\boldsymbol{\nu}-\frac{\mathbf{u}_{kl}}{\gamma}\right), \tag{7.12}$$

which shows that images obtained by densification in periscopic mode or using inverted Galilean telescopes are the same (compare to Eq. (2.7) and see Fig. 2).

In both cases, if we take $\gamma = 1$ we indeed recover the Fizeau mode of Equation (2.6):

$$\widehat{I_1^G}(\mathbf{u}) = \widehat{I_1^P}(\mathbf{u}) = \widehat{I^F}(\mathbf{u}) = \sum_{k=1}^{K}\sum_{l=1}^{K} \widehat{O}(\mathbf{u})T_0(\mathbf{u}-\mathbf{u}_{kl}). \tag{7.13}$$

References

Aime, C., 2008, A&A, 483, 361

Aime, C., Lantéri, H., Diet, M., & Carlotti, A., 2012, A&A, 543, A42

Anderson, J.A., 1920, ApJ, 51, 263

Baron, F., *et al.*, 2012, The 2012 Interferometric Imaging Beauty Contest, Proc. SPIE: Astronomical Telescopes and Instrumentation Conference (Amsterdam)

Beck, A., & Teboulle, M., 2009, Siam J. Imaging Sciences, 2, 183

Candès, E. J., Romberg, J., & Tao, T., 2006, IEEE Trans. Inf. Theory, 52, 489

Carrillo, R.E., McEwen, J.D., & Wiaux, Y., 2012 [arXiv:1205.3123]

Chen, S.S., Donoho, D.L., & Saunders, M.A., 1998, SIAM J. Scientific Computing, 20, 33

Cornwell, T.J., 2009, A&A, Special issue, 500, 65

Cornwell, T.J., 2008, IEEE J. Selected Topics Signal Proc., 2, 793

Dabbech, A., Mary, D., & Ferrari, C., 2012, Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 3665

Daube-Witherspoon, M.E., & Muehllehner, G., 1986, IEEE Trans. Med. Imaging, 5, 61

Donoho, D.L., 2006, IEEE Trans. Inf. Theory, 52, 1289

Elad, M., Milanfar, P., & Rubinstein, R., 2007, Inverse Probl., 23, 947

Fizeau, H., 1868, C. R. Hebd. Seanc. Acad. Sci. Paris, 66, 934

Fornassier, M., 2010, Theoretical Foundations and Numerical Methods for Sparse Recovery, De Gruyter; 1 edition (2010)

Giovannelli, J.-F., & Coulais, A., 2005, A&A, 439, 401

Gribonval, R., 2009-2012, http://small-project.eu/publications

Gribonval, R., & Nielsen, M., 2003, IEEE Trans. Inf. Theory, 49, 3320

Edmond Halley, 1715, A short History of several New-Stars. Phil. Trans, XXIX, 354 (1715) Available online

Högbom, J.A., 1974, A&AS, 15, 417

Labeyrie, A., *et al.*, 2012, Optical and Infrared Interferometry III, Proceedings of the SPIE, Vol. 8445

Labeyrie, A., 1996, A&AS, 118, 517

Labeyrie, A., 1975, ApJ, 196, L71

Lantéri, H., Roche, M., & Aime, C., 2002, Inverse Probl., 18, 1397

Lantéri, H., *et al.*, 2002, Signal Proc., 82, 1481

Lardière, O., Martinache, F., & Patru, F., 2007, MNRAS, 375, 977

Li, F., Cornwell, T.J., & de Hoog, F., 2011, A&A, 528, 31

Lucy, L.B., 1974, AJ, 79, 745

Magain, P., Courbin, F., & Sohy, S., 1998, ApJ, 494, 472

Mallat, S., 2008, A wavelet tour of signal processing: the sparse way, 3rd edition (Academic Press)

Mallat, S., & Zhang, Z., 1993, IEEE Trans. Sig. Proc., 41, 3397

McEwen, J.D., & Wiaux, Y., 2011, MNRAS, 413, 1318

Michelson, A.A., 1891, Nature, 45, 160

Michelson, A.A., 1920, ApJ, 51, 257

Michelson, A.A., & Pease, F.G., 1921, ApJ, 53, 249

Mignard, F., & Martin, C., 1997, Pour la Science, 235

Pirzkal, N., Hook, R.N., & Lucy, L.B., 2000, In ASP Conference Series, Astronomical Data Analysis, Software and Systems IX, Paris, ed. N. Manset, C. Veuillet & D. Crabtree, 216, 657

Rau, U., Bhatnagar, S., Voronkov, M.A., &,Cornwell, J.T., 2009, Proc., 97, 1472

Richardson, W.H., 1972, J. Opt. Soc. Am., 62, 55

Schwarz, U.J., 1978, A&A, 65, 417

Starck, J.L, Pantin, E., & Murtagh, F., 2002, PASP, 114, 1051

Starck, J.L, Murtagh, F., & Fadili, M.-J., 2010, Sparse Image and Signal Processing - Wavelets, Curvelets, Morphological Diversity (Cambridge University Press)

Stephan, E., 1873, C. R. Hebd. Seanc. Acad. Sci. Paris, 76, 1008

Stephan, E., 1873, C. R. Hebd. Seanc. Acad. Sci. Paris, 78, 1008

Kopilovich, L.E., & Sodin, L.G., 2001, Multielement System Design in Astronomy and Radio Science, Astrophysics and Space Science Library (Kluwer, 2001)

Tallon, M., & Tallon-Bosc, I., 1992, A&A, 253, 641

Thiébaut, E., 2005, NATO ASIB Proc. 198: Optics Astrophys., 397

Thiébaut, É., Soulez, F., & Denis, L., 2012, J. Opt. Soc. A, submitted

Vannier, M., *et al.*, 2010, Spectral regularization and sparse representation bases for interferometric imaging, Proc. SPIE: Astronomical Telescopes and Instrumentation (Conference, San Diego)

Wakker, B.P., & Schwarz, U.J., 1988, A&A, 200, 312

Wenger, S., Darabi, S., Sen, P., Glassmeier, K.H., & Magnor, M., 2010, Proc. IEEE Int. Conf., Image Process., IEEE Signal Process. Soc., 1381

Wiaux, Y., Jacques, L., Puy, G., Scaife, A.M.M., & Vandergheynst, P., 2009a, MNRAS, 395, 1733

Wiaux, Y., Puy, G., Boursier, Y., & Vandergheynst, P., 2009b, MNRAS, 400, 1029

Statistical Models in Signal and Image Processing

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 265–301*

INTRODUCTION TO THE RESTORATION OF ASTROPHYSICAL IMAGES BY MULTISCALE TRANSFORMS AND BAYESIAN METHODS

A. Bijaoui[1]

Abstract. The image restoration is today an important part of the astrophysical data analysis. The denoising and the deblurring can be efficiently performed using multiscale transforms. The multiresolution analysis constitutes the fundamental pillar for these transforms. The discrete wavelet transform is introduced from the theory of the approximation by translated functions. The continuous wavelet transform carries out a generalization of multiscale representations from translated and dilated wavelets. The *à trous* algorithm furnishes its discrete redundant transform. The image denoising is first considered without any hypothesis on the signal distribution, on the basis of the *a contrario* detection. Different softening functions are introduced. The introduction of a regularization constraint may improve the results. The application of Bayesian methods leads to an automated adaptation of the softening function to the signal distribution. The MAP principle leads to the *basis pursuit*, a sparse decomposition on redundant dictionaries. Nevertheless the posterior expectation minimizes, scale per scale, the quadratic error. The proposed deconvolution algorithm is based on a coupling of the wavelet denoising with an iterative inversion algorithm. The different methods are illustrated by numerical experiments on a simulated image similar to images of the deep sky. A white Gaussian stationary noise was added with three levels. In the conclusion different important connected problems are tackled.

1 Introduction

The astrophysical images observed by modern instruments are today currently enhanced by digital processing. In particular, many efforts are done for their denoising and their deblurring. These operations are often coupled in a global

[1] University of Nice Sophia Antipolis, UMR CNRS 6202, OCA, BP. 4229, 06304 Nice Cedex 04, France

DOI: 10.1051/eas/1359013

image restoration. Generally the problem is written as:

$$\mathbf{Y} = \mathbf{AX} + \mathbf{N}; \tag{1.1}$$

where $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{N}$ are respectively the observed image, the image to be restored, the blurring linear operator and the noise image.

The inversion is conditioned by two features: the matrix singularities and the noise level. In the case of a regular blurring matrix and a signal without noise, the inversion is unique, the problem being only to minimize the number of operations. In general, the noise level and the matrix singularities lead to inconsistency and instability. Statistical rules are needed to define the correct solution. A regularity constraint is also required to select the best one from a given criterion.

In this context, the image representation plays an important part. In the general case of a space invariant *Point Spread Function* (PSF), the matrix product corresponds to a convolution in the direct space and a filtering in the Fourier one. Thus, the matrix singularities are associated to the frequency holes. Thus, a method based on the Fourier transform can not fill these holes without a constraint. In the case of a representation different from a Fourier series, this filling becomes possible. This is the case for the CLEAN algorithm (Högbom 1974) based on the consideration that the information consists in sparse Dirac peaks.

The representation plays an important part for the denoising. Its quality depends on the efficiency to concentrate the information into the minimum number of coefficients; these coefficients being obtained by a suitable transform.

Multiscale transforms were early developed and applied for the image processing (Starck *et al.* 1998; Mallat 1998). The *Multiresolution Theory* developed in 80's is a beautiful framework to get multiscale representations (Mallat 1989). It leads to the *Discrete Wavelet Transform* (DWT). Closely related redundant transforms, connected to the *continuous wavelet transform* (CWT) (Morlet *et al.* 1982) carried out better results (Raphan & Simoncelli 2008). Nevertheless a correct CWT development needs the multiresolution theory.

The present paper constitutes an introduction to this large topic. In Section 2, the multiresolution theory is developed in the context of the approximation theory from translated scaling functions. In Section 3, the CWT is then described. It is shown that the use of scaling functions unifies DWT and CWT. The denoising is examined in Section 4 from different thresholding rules. In Section 5, a first Bayesian approach, derived from the Maximum a Posteriori (MAP), leads to the Basis Pursuit (BP). The Bayesian posterior mean is applied in Section 6. In Section 7, an application to a deconvolution problem is developed. Finally, in the conclusion, different uncovered problems are scanned.

2 The multiresolution theory and the DWT

2.1 The approximation by translated functions

The Shannon interpolation. The Shannon interpolation theorem was a milestone in the signal processing progress (Shannon 1948). A function $f(x) \in L^2(\mathbb{R})$ is

interpolated from regularly spaced samples thanks to the relation:

$$f_0(x) = \sum_{k=-\infty}^{+\infty} f(kh)\mathrm{sinc}\left(\frac{x}{h} - k\right); \tag{2.1}$$

where $\mathrm{sinc}(x)$ is the sine cardinal function ($\frac{\sin \pi x}{\pi x}$) and h the sampling step. The interpolation is perfect ($f_0(x) \equiv f(x)$) if $h \leq h_0$, h_0 being the Nyquist-Shannon step:

$$h_0 = \frac{1}{2\nu_0}; \tag{2.2}$$

where ν_0 is the cut-off frequency of the function $f(x)$. In practice, this theorem is not directly applied due to the slow convergence of the sine cardinal function. It introduced the idea of interpolations based on translated functions that played a fundamental part for the building of the multiresolution theory.

The L^2 approximation by translated functions. We set now (Schoenberg 1946; Strang & Fix 1971):

$$f_0(x) = \sum_{k=-\infty}^{+\infty} a(k)\varphi(x-k). \tag{2.3}$$

Compared to Equation (2.1), the sampling step is set to 1, the sine cardinal function is changed to the φ one, the values at the interpolation mesh $f(nh)$ are changed to the $a(k)$ coefficients. The goal is not to get $f_0(k) \equiv f(k)$. Here, we search the coefficients $a(k)$ such that the distance between the functions $f(x)$ and $f_0(x)$ is minimum in the $L^2(\mathbb{R})$ space, *i.e.*:

$$R = \int_{-\infty}^{+\infty} |f(x) - f_0(x)|^2 dx \tag{2.4}$$

is minimum. Taking into account Equation (2.3) we get:

$$\int_{-\infty}^{+\infty} \varphi(x-k)[f(x) - \sum_{l=-\infty}^{+\infty} a(k)\varphi(x-l)]dx = 0. \tag{2.5}$$

The following equation is derived:

$$c(k) = \sum_{l=-\infty}^{+\infty} a(l)A(k-l); \tag{2.6}$$

with:

$$A(k-l) = \int_{-\infty}^{+\infty} \varphi(x-k)\varphi(x-l)dx; \tag{2.7}$$

and

$$c(k) = \int_{-\infty}^{+\infty} f(x)\varphi(x-k)dx \equiv < f(x), \varphi(x-k) > . \tag{2.8}$$

Equation (2.6) is a discrete convolution which can be solved by the application of the Fourier transform:

$$\hat{c}(\nu) = \hat{a}(\nu) \sum_{n=-\infty}^{+\infty} \hat{A}(\nu + n); \tag{2.9}$$

with $\hat{A}(\nu) = |\hat{\varphi}(\nu)|^2$. The inversion is possible if:

$$\hat{S}(\nu) \equiv \sum_{n=-\infty}^{+\infty} |\hat{\varphi}(\nu + n)|^2 \neq 0. \tag{2.10}$$

$\varphi(x)$ is called the scaling function. $c(k)$ is a weighted mean of $f(x)$ around k. $f_0(x)$ is the projection of $f(x)$ into a subspace V_0 of $L^2(R)$.

Duality and orthogonal scaling functions. If $\hat{S}(\nu) = 1$ $a(k) = c(k)$, the approximation is easily computed from the scalar products $c(k)$.

If, more generally, Relation 2.10 is satisfied for all frequencies, we can derive a new scaling function $\bar{\varphi}(x)$ from the Fourier transform of the initial one (Daubechies *et al.* 1986):

$$\hat{\bar{\varphi}}(\nu) = \frac{\hat{\varphi}(\nu)}{\sqrt{\hat{S}(\nu)}}. \tag{2.11}$$

The set $\{\bar{\varphi}(x - k)\}$ is an orthonormal basis of the V_0 subspace. Here the $a(k)$ coefficients are identical to the $c(k)$ ones. The same scaling function is used for the analysis ($c(k)$) and the synthesis ($a(k)$). This is the case for the Shannon interpolation, the sine cardinal function being an orthogonal scaling function.

In the case of a non orthogonal scaling function, it is also convenient to introduce the dual scaling function:

$$\hat{\tilde{\varphi}}(\nu) = \frac{\hat{\varphi}(\nu)}{\hat{S}(\nu)}. \tag{2.12}$$

In this framework, it results that:

$$f_0(x) = \sum_{k=-\infty}^{+\infty} c(k)\tilde{\varphi}(x - k). \tag{2.13}$$

Here the $c(k)$ coefficients are also identical to the $a(k)$ ones. But it is not the same scaling function used for the analysis ($c(k)$) and the synthesis ($a(k)$).

Normalization of the scaling function. In Equation (2.3) the scaling function is considered without normalization. The orthonormal scaling functions associated to Equation (2.11) have, by construction, their square integral equal to 1. This is the general setting in the framework of the multiresolution theory. Nevertheless, it could be also convenient to choose the integral equal to 1. In this case the approximation coefficients are local means, weighted by the scaling function.

The Shannon scaling function. It is easy to show that the shifted sine cardinal functions with integers are orthogonal. The scaling function is here:

$$\varphi(x) = \frac{\sin \pi x}{\pi x}. \tag{2.14}$$

The corresponding V_0 subspace is the one of the functions having a frequency support in $[-0.5, 0, 5]$. The approximation for a function belonging to V_0 corresponds to the Shannon interpolation. It can be noted that this approximation is invariant by translation.

The Haar scaling function. The characteristic function, $H(x) = 1$ for $x \in [0, 1]$ and null outside this interval, is called the Haar scaling function. The functions shifted with integers are orthogonal. The corresponding V_0 subspace is the space of the staircase functions. Note that this approximation is only invariant by an integer shift.

2.2 The pyramid of resolution

Scale modification and the dilation equation. The scaling function is dilated by a factor a, the approximation coefficients become:

$$c(a, k) =< f(x), \frac{1}{a}\varphi\left(\frac{x}{a} - k\right) > . \tag{2.15}$$

The factor $\frac{1}{a}$ is introduced to keep constant the integral of the dilated scaling function. In the case of an orthonormal scaling function, the factor becomes $\frac{1}{\sqrt{a}}$ in order to keep the square integral equal to 1.

There is a linear relation between $c(k)$ and $c(a, k)$ if the scaling function satisfies to the dilation equation (Strang 1989):

$$\frac{1}{a}\varphi(\frac{x}{a}) = \sum_{n=-\infty}^{+\infty} h_a(n)\varphi(x - n). \tag{2.16}$$

It results that

$$c(a, k) = \sum_{n=-\infty}^{+\infty} h_a(n)c(ak + n). \tag{2.17}$$

In this framework, the function $f(x)$ has to be known only by its approximation coefficients $c(k)$. Note that the number of coefficients (for a finite set) is reduced by a factor a. Most often $a = 2$, this leads to the so-called dyadic analysis. The resulting approximation $f_a(x)$ belongs to a subset V_a which is embedded in V_0.

The resolution pyramid. The dilation of the scaling function may be iterated, leading to the approximations $f_0(x)$, $f_a(x)$, $f_{a^2}(x)$, ... These functions constitute the pyramid of resolution associated to this analysis. The functions belong to the subsets $V_0 \supset V_a \supset V_{a^2} \ldots$. For a finite initial number of approximation coefficients, their number is divided by a at each iteration step.

Examples of scaling functions. It is easy to show that the sine cardinal function obeys to the dilation equation, whatever the integer a. The generated subspace corresponds to the functions with a frequency bandwidth 1, $\frac{1}{a}$, $\frac{1}{a^2}$, ...

The Haar scaling function obeys also to the dilation equation:

$$H_a(x) = \frac{1}{a}[H(x) + H(x-1) + \ldots + H(x-a+1)]. \tag{2.18}$$

The B-spline functions (Hou & Andrews 1978) generalize the Haar one. Its centered version is defined by its Fourier transform:

$$\hat{B}_l(\nu) = \mathrm{sinc}^{l+1}(\nu). \tag{2.19}$$

The Fourier transform of its dilated version is:

$$\hat{B}_{l,a} = \mathrm{sinc}^{l+1}(\mathrm{a}\nu). \tag{2.20}$$

Its quotient with $B_l(\nu)$ is:

$$\hat{h}_{l,a}(\nu) = \frac{\mathrm{sinc}^{l+1}(\mathrm{a}\nu)}{\mathrm{sinc}^{l+1}(\nu)}. \tag{2.21}$$

It is easy to show that it is a 1-periodic function. In particular, for $a = 2$ we get:

$$\hat{h}_l = \cos^{l+1}(\nu); \tag{2.22}$$

which leads to:

$$h_l(n) = \frac{1}{2^{l+1}} C_{l+1}^{\frac{l+1}{2}-n}. \tag{2.23}$$

The cubic B-spline is often used (Starck *et al.* 1998). Its coefficients are:

$$h_3(n) = \frac{1}{16} C_4^{2-n}. \tag{2.24}$$

Case of an orthonormal scaling function. From Equation (2.12) we derive:

$$\sum_{n=-\infty}^{+\infty} |\hat{\varphi}(\nu+n)|^2 = 1. \tag{2.25}$$

That leads directly to:

$$\sum_{n=-\infty}^{+\infty} |\hat{\varphi}(2\nu+n)|^2 = 1. \tag{2.26}$$

The dilation equation in the Fourier space is written as:

$$\hat{\varphi}(2\nu) = \hat{h}(\nu)\hat{\varphi}(\nu). \tag{2.27}$$

We have:

$$\sum_{n=-\infty}^{+\infty} |\hat{\varphi}(2\nu+n)|^2 = \sum_{n=-\infty}^{+\infty} |\hat{\varphi}(2\nu+2n)|^2 + \sum_{n=-\infty}^{+\infty} |\hat{\varphi}(2\nu+2n+1)|^2 = 1 \quad (2.28)$$

Relation 2.27 is applied:

$$\sum_{n=-\infty}^{+\infty} |\hat{h}(\nu+n)|^2 |\hat{\varphi}(\nu+n)|^2 + \left|\hat{h}\left(\nu+n+\frac{1}{2}\right)\right|^2 \left|\hat{\varphi}\left(\nu+n+\frac{1}{2}\right)\right|^2 = 1 \quad (2.29)$$

Taking Relation 2.25 and taking into account the periodicity of the function $\hat{h}(\nu)$ it results finally:

$$|\hat{h}(\nu)|^2 + \left|\hat{h}\left(\nu+\frac{1}{2}\right)\right|^2 = 1. \quad (2.30)$$

2.3 The 1D multiresolution analysis

The complementary subspace. As $V_1 \subset V_0$, we can write:

$$f_0(x) = f_1(x) + g_1(x); \quad (2.31)$$

where $f_0 \in V_0$ and $f_1 \in V_1$. g_1 is a function of the complementary subspace W_1 of V_1 in V_0, *i.e.* $V_0 = V_1 + W_1$.

The wavelet basis. $g_1(x)$ can be written as:

$$g_1(x) = \sum_{k=-\infty}^{\infty} w(1,k)\tilde{\psi}\left(\frac{x}{2}-k\right). \quad (2.32)$$

The detail coefficients $w(1,k)$ are obtained by projection on a translated set:

$$w(1,k) = < f, \frac{1}{2}\psi\left(\frac{x}{2}-k\right) > . \quad (2.33)$$

The $w(1,k)$ computation from the $c(k)$ ones, requires that:

$$\frac{1}{2}\psi\left(\frac{x}{2}\right) = \sum_n g(n)\varphi(x-n); \quad (2.34)$$

i.e. that $\frac{1}{2}\psi(\frac{x}{2})$ belongs to V_0. That leads to the relation:

$$w(1,k) = \sum_n g(n)c(2k+n). \quad (2.35)$$

Orthogonal wavelets. In this case we have:

$$\sum_{n=-\infty}^{+\infty} |\hat{\psi}(\nu+n)|^2 = 1. \quad (2.36)$$

Equation (2.34) is equivalent to:

$$\hat{\psi}(2\nu) = \hat{g}(\nu)\hat{\psi}(\nu). \tag{2.37}$$

Thus, it is derived that:

$$|\hat{g}(\nu)|^2 + |\hat{g}\left(\nu + \frac{1}{2}\right)|^2 = 1. \tag{2.38}$$

The subspace V_1 and W_1 being orthogonal we have:

$$\sum_n \hat{\varphi}(\nu + n)\hat{\psi}^*(\nu + n) = 0. \tag{2.39}$$

The following relation is derived:

$$\hat{h}(\nu)\hat{g}^*(\nu) + \hat{h}\left(\nu + \frac{1}{2}\right)\hat{g}^*\left(\nu + \frac{1}{2}\right) = 0. \tag{2.40}$$

Filters h and g satisfying Relations 2.30, 2.38 and 2.40 generate conjugate orthogonal scaling and wavelet functions. For a given filter h obeying to 2.30, we can associate the filter g given by the relation:

$$\hat{g}(\nu) = e^{-2i\pi\nu}\hat{h}^*\left(\nu + \frac{1}{2}\right) \tag{2.41}$$

h and g are called *Quadrature Mirror Filters* (QMF) (Esteban & Galland 1977).

Reconstruction. Any V_0 basis function $\varphi(x-k)$ can be written as a sum of V_1 and W_1 base functions:

$$\varphi(x-k) = 2[\sum_l h(k-2l)\varphi_1(x-2l) + g(k-2l)\psi_1(x-2l)]. \tag{2.42}$$

That leads to get the approximation coefficients by projection:

$$c(k) = 2[\sum_l h(k-2l)c(1,l) + g(k-2l)w(1,l)]. \tag{2.43}$$

The multiresolution analysis. From the approximation coefficients $c(1,k)$, it is possible to iterate by a new dilation of the scaling and of the wavelet functions. By iteration we obtained a set of details $w(i,k)$ such that the function $f(x)$ can be written as:

$$f(x) = \sum_{i,k=-\infty}^{\infty} w(i,k)\tilde{\psi}\left(\frac{x}{2^i} - k\right) \tag{2.44}$$

for any function of the $L^2(\mathbb{R})$ space (Mallat 1989).

The recurrence formulae and the filter bank. The previous developments are summarized as:

Approximation $c(i,k) = \sum_n h(n)c(i-1, 2k+n)$;

Wavelets $w(i,k) = \sum_n g(n)c(i-1, 2k+n)$;

Reconstruction $c(i,k) = 2[\sum_l \tilde{h}(k+2l)c(i+1,l) + \tilde{g}(k+2l)c(i+1,l)].]$

In the case of orthogonal scaling and wavelet functions, $\tilde{h}(n) = h(-n)$ and $\tilde{g}(n) = g(-n)$. The resulting algorithm flow-chart is drawn in Figure 1. This algorithm is known as the filter bank one (Vitterli 1986). The data vector inputs at the top left. It is convolved with the two filters H (low passband) and G (high passband). The resulting vectors are decimated, by removing every other point. The smoothed values are convolved again, and so on, up to get one value. The restoration consists to introduce a 0 between two approximation or two detail coefficients. We start from the bottom right, and progressively the signal is restored from the largest scale to the smallest one. The convolutions are done with the filter $\tilde{H}$ and $\tilde{G}$.

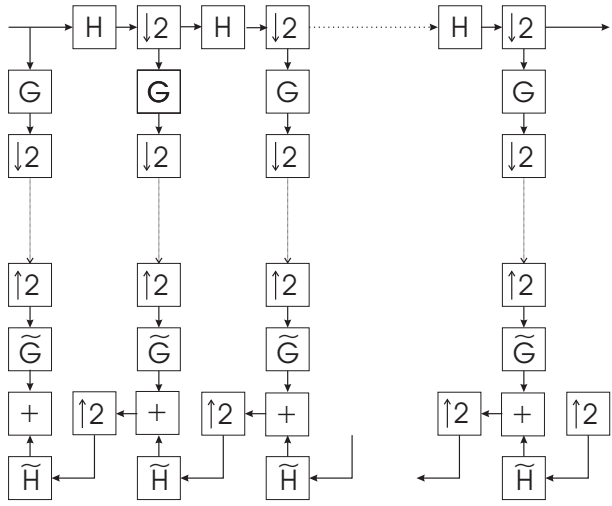


Fig. 1. Flow-chart of the filter bank algorithm.

The algorithm was developed from the multiresolution theory. But it is more general. The restoration is perfect only if the filters h, $\tilde{h}$, g and $\tilde{g}$ satisfied the following conditions called perfect reconstruction and antialiasing conditions:

$$\hat{h}(\nu)\hat{\tilde{h}}(\nu) + \hat{g}(\nu)\hat{\tilde{g}}(\nu) = 1; \tag{2.45}$$

$$\hat{h}(\nu)\hat{\tilde{g}}(\nu) + \hat{h}\left(\nu + \frac{1}{2}\right)\hat{\tilde{g}}\left(\nu + \frac{1}{2}\right) = 0. \tag{2.46}$$

The Daubechies wavelets. The Haar transform is associated to the filters:

$$h(0) = h(1) = \frac{1}{\sqrt{2}} \qquad h(n) = 0 \quad n \neq (0,1). \tag{2.47}$$

The corresponding high pass filter is

$$g(0) = \frac{1}{\sqrt{2}} \qquad g(1) = -\frac{1}{\sqrt{2}} \qquad h(n) = 0 \quad n \neq (0,1). \tag{2.48}$$

The algorithm is very fast. K operations are required for the transform and its inverse for a signal with K elements.

Daubechies (1988) generalized the Haar transform with compact filters. They are widely applied in modern signal processing. Later on in this paper, Daubechies' filters of length 8 are applied.

2.4 The 2D multiresolution

The 2D approximation by translated scaling functions. The concept of approximation by translated scaling functions in $L^2(\mathbb{R})$ is easily extended to 2 (and more) dimensions. If $f(x,y)$ is the function to be approximated and $\varphi(x,y)$ the scaling function, the approximation coefficients are:

$$c(0,k,l) =< f(x,y), \varphi(x-k, y-l) > . \tag{2.49}$$

The corresponding approximation is:

$$f_0(x,y) = \sum_{k,l} c(0,k,l)\tilde{\varphi}(x-k, y-l) \tag{2.50}$$

where $\tilde{\varphi}(x,y)$ is the dual scaling function. This function exists if:

$$\sum_{n,m} |\hat{\varphi}(u+n, v+m)|^2 \neq 0. \tag{2.51}$$

The approximation is a $f(x,y)$ projection on the V_0 subspace of $L^2(\mathbb{R}^2)$.

The 2D dilation equation. The approximation for a scaling function dilated by a factor a in each direction can be computed from the $c(0,k,l)$ coefficients if:

$$\frac{1}{a^2}\varphi(\frac{x}{a}, \frac{y}{a}) = \sum_{n,m} h(n,m)\varphi(x-n, y-m). \tag{2.52}$$

Most often the variables are separated:

$$\varphi(x,y) \equiv \varphi(x)\varphi(y) \tag{2.53}$$

where $\varphi(x)$ satisfies the 1D dilation equation.

The wavelets. Taking into account the variable separation the V_0 subspace is divided in four subspaces:

V_1 the subspace corresponding to the approximation at scale 2. It is computed with the filter $h(n)h(m)$. The scaling function is $\frac{1}{4}\varphi(\frac{x}{2})\varphi(\frac{y}{2})$.

$W_{1,h}$ associated to the horizontal details. The wavelet coefficients are computed with the filter $g(n)h(m)$. The wavelet function is $\frac{1}{4}\psi(\frac{x}{2})\varphi(\frac{y}{2})$.

$W_{1,v}$ associated to the vertical details. The wavelet coefficients are computed with the filter $h(n)g(m)$. The wavelet function is $\frac{1}{4}\varphi(\frac{x}{2})\psi(\frac{y}{2})$.

$W_{1,d}$ associated to the diagonal details. The wavelet coefficients are computed with the filter $g(n)g(m)$. The wavelet function is $\frac{1}{4}\psi(\frac{x}{2})\psi(\frac{y}{2})$.

In this framework, the 2D filter bank algorithm is easily deduced from the 1D one.

3 The continuous wavelet transform

3.1 Generalities

Definition and main properties. The Morlet-Grossmann definition of the continuous wavelet transform (Grossmann & Morlet 1984) for a $1D$ signal $f(x) \in L^2(\mathbb{R})$ is:

$$W(a,b) = N(a) \int_{-\infty}^{+\infty} f(x)\psi^*\left(\frac{x-b}{a}\right) dx; \tag{3.1}$$

where z^* notes the complex conjugate of z, $\psi^*(x)$ is the analyzing wavelet, a (> 0) is the scale parameter and b is the position parameter. Grossmann & Morlet set $N(a) = \frac{1}{\sqrt{a}}$, but it is often convenient to set $N(a) = \frac{1}{a}$. The transformation is linear, covariant under translations and under dilations. The last property makes the wavelet transform very suitable for analyzing hierarchical structures. It is like a mathematical microscope with properties that do not depend on the magnification.

Inversion. Consider now a function $W(a,b)$ which is the wavelet transform of a given function $f(x)$. $f(x)$ can be restored by using the formula (Grossmann & Morlet 1984):

$$f(x) = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{a}} W(a,b)\psi\left(\frac{x-b}{a}\right) \frac{da.db}{a^2}; \tag{3.2}$$

where:

$$C_\psi = \int_0^{+\infty} \frac{|\hat{\psi}(\nu)|^2}{\nu} d\nu. \tag{3.3}$$

The reconstruction is only available if C_ψ is defined (admissibility condition). This condition is generally true if $\hat{\psi}(0) = 0$, *i.e.* the mean of the wavelet function is 0.

3.2 The discrete wavelet transform

The transform sampling. The image sampling is generally made according to the Shannon theorem. The discrete wavelet transform (DWT) can be derived from

this theorem. If the wavelet function has no cut-off frequency, the transform cut-off frequency is the signal one, whatever the scale. So if K is the number of signal elements and I the number of scales the transform has KI elements.

In the case of a wavelet function having the cut-off frequency $\frac{1}{2}$, at each dyadic scale the cut-off frequency is divided by two. Thus, the sampling step can be multiplied by a factor 2. The total transform length becomes about $2K$.

Generally the scales are sampled according to a 2^i law. Nevertheless it is not guaranteed that all the information on the CWT is kept by this sampling.

Direct transformations. The DWT can be obtained directly by convolution, using a compact wavelet function. As the scale increases, the CPU time increases in proportion. So, in practice this method is not easy to use.

It is possible to work in the Fourier space, computing the transform scale by scale. The number of elements for a scale can be reduced for a wavelet having a cut-off frequency. Here, the CPU time is proportional to $K\log(K)$.

The transform from the filter bank. In fact, in the previous section, beyond the multiresolution theory, we examined a fast DWT algorithm based on the filter bank. The transform size is K and the computing time is proportional to K.

3.3 The wavelet approximation and the à trous algorithm

The sampled wavelet function at scale 1. Let us consider a real wavelet function which can be written as an approximation from translated scaling functions (Eq. (2.34)). Let us admit that we know the sampled approximation coefficients $c(0,k) =< f(x), \varphi(x-k) >$. The sampled continuous wavelet function at scale $a = 2$ can written as ($N(a) = 1/a$):

$$w(1,k) = \sum_n g(n)c(0,k+n). \tag{3.4}$$

This expression is similar to 2.3, but the array is not decimated, the factor 2 being not present. Note that in the following, the scale of the wavelet transform will design the exponent i of the true scale $a = 2^i$.

The recurrence expressions. We want to compute $w(2,k)$ using a similar formula. The scaling function $\varphi(x)$ is chosen to satisfy the dilation Equation (2.16). In this framework, we have for the approximation coefficients:

$$c(i+1,k) =< f(x), \frac{1}{2^{i+1}}\varphi\left(\frac{x}{2^{i+1}} - k\right) > =< f(x), \frac{1}{2^i}\sum_n h(n)\varphi\left(\frac{x}{2^i} - k\right) >; \tag{3.5}$$

which leads to:

$$c(i+1,k) = \sum_n h(n)c(i,k+2^i n). \tag{3.6}$$

Similarly we get:

$$w(i+1,k) = \sum_n g(n)c(i,k+2^i n). \tag{3.7}$$

The interpolation and the wavelet coefficients are computed with a linear operation similar to a convolution but we jump a set of $2^i - 1$ points. For that reason, this algorithm is called the *à trous* algorithm (algorithm with holes) (Holschneider *et al.* 1989). The flow-chart of this algorithm is drawn in Figure 2. A set of $K \log_2 K$ values of the wavelet transform is obtained, with a number of operations proportional to $K \log_2 K$.

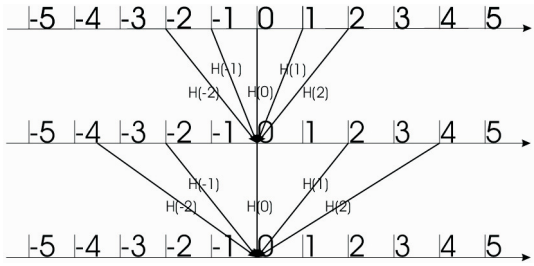


Fig. 2. Flow-chart of the à trous algorithm.

The inversion. Here, the transform is undecimated. Thus we get a transform size larger than the input signal one. The inverse system is over-determined. If the $\{w(i,k)\}$ set is a wavelet transform, it is easy to restore the approximation scale by scale using dual filters $\tilde{h}$ and $\tilde{g}$ which have to satisfy the relation:

$$\hat{h}(\nu)\hat{\tilde{h}}(\nu) + \hat{g}(\nu)\hat{\tilde{g}}(\nu) = 1. \tag{3.8}$$

The filters choice is large. A simple algorithm consists in making the difference between two approximations (Bijaoui *et al.* 1994):

$$w(i+1,k) = c(i,k) - c(i+1,k). \tag{3.9}$$

Here the inversion is obvious.

The inversion of a modified wavelet transform. If the $\{w(i,k)\}$ set is not the wavelet transform of a given signal, nevertheless a signal $\{c(0,k)\}$ will be obtained by inversion. But its wavelet transform $\{w_s(i,k)\}$ can be different from $\{w(i,k)\}$. This point is important for image restoration. In this framework most often a softening rule is applied on the wavelet coefficients. There is a duality between the wavelet transform and the signal for the orthogonal DWT. But this duality vanishes for the redundant undecimated wavelet transform (UDWT) associated to the *à trous* algorithm. Some cautions have to be taken for the inversion. A classical solving method consists to obtain the orthogonal projection of $\{w(i,k)\}$ in the space generated by the wavelet transforms of all the signals. That corresponds to get the set $\{c(0,k)\}$ such that its wavelet transform $\{w_s(i,k)\}$ has the minimum distance to the input set $\{w(i,k)\}$. Obviously the inversion algorithm is slowed.

The shift-invariant wavelet transform. Coifman & Donoho (1995) introduced a variant of the *à trous* algorithm based on the orthogonal wavelet transform. The

transform becomes shift invariant by removing the decimation of the coefficients (approximation and wavelet). The inversion is done by taking into account the coefficient interleaving generated by the lack of decimation. A same pixel is thus many times reconstructed, the mean is done in the proposed algorithm. This algorithm has the advantage to inverse from the filter bank.

The pyramidal transform. The undecimated wavelet transform may correspond to a too important data array for large images. At each step of the algorithm the approximation coefficients can be removed without decimating the wavelet array. We get a pyramidal set of values. The number of data is now $2N$ and the number of operations is proportional to N. The inversion is based on an orthogonal projection, obtained by an iterative algorithm.

3.4 The two-dimensional continuous wavelet transform

General definition. The wavelet dilation is not necessarily isotropic, *i.e.* identical whatever the direction. But it can be seen as a dilation in two orthogonal directions. The reference frame can be also rotated with a θ angle. That leads to the coordinate transform:

$$R(x, y, a_x, a_y, \theta) = \begin{pmatrix} \frac{1}{a_x} & 0 \\ 0 & \frac{1}{a_y} \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \tag{3.10}$$

The wavelet transform becomes:

$$w(b_x, b_y, a_x, a_y, \theta) = N(ab) < f(x, y), \psi^*(|R(x - b_x, y - b_y, a_x, a_y, \theta)|) >; \tag{3.11}$$

where $|R|$ designs the module of the vector R. The resulting transform is thus a 5D function. Its sampling rules are not evident, especially for the angular variable.

The isotropic 2D CWT. Most often the 2D CWT is applied in its simplified isotropic version:

$$w(b_x, b_y, a) = N(a^2) < f(x, y), \psi^* \left(\frac{x}{a}, \frac{y}{a}\right) > . \tag{3.12}$$

The two dimensional à trous algorithm. The *à trous* and the pyramidal algorithms can be easily transposed in two dimensions taking into account the use of a scaling function which satisfies to the 2D dilation equation. The algorithms are simplified in the case of a variable separation ($\varphi(x, y) \equiv \varphi(x)\varphi(y)$). The *à trous* computation of the approximation coefficients is obtained with the expression:

$$c(i + 1, k, l) = \sum_{n,m} h(n)h(m)c(i, k + 2^i n, l + 2^i m); \tag{3.13}$$

while the wavelet coefficients are computed with:

$$w(i + 1, k, l) = \sum_{n,m} g(n, m)c(i, k + 2^i n, l + 2^i m). \tag{3.14}$$

The wavelet function is not necessarily separated in the two directions. For example, the wavelet associated to the difference between two successive approximations is not separable.

Quasi isotropic scaling functions. For the coherence of the method it would be convenient to process the data with an isotropic scaling function. The Gaussian is the single function which is separable and isotropic. But this function does not satisfy the dilation equation.

The centered B-splines tends for an increasing index to the Gaussian. Thus, its use allows a quasi isotropic analysis with fast computations, for the *à trous* and for the pyramidal algorithms.

4 Image denoising from significant coefficients

4.1 The significant coefficients

The quality criterion. Let us consider a discrete noisy signal $\mathbf{Y} = \mathbf{X} + \mathbf{N}$. $\mathbf{X}$ is the true signal vector and $\mathbf{N}$ its associated noise. The signal denoising consists into the operation $O(\mathbf{Y}) \rightarrow \bar{\mathbf{X}}$ such that this vector is the closest to $\mathbf{X}$. The distance criterion depends on the noise statistics. The case of a stationary white Gaussian noise is only examined in the present paper. Some methods adapted to other noise statistics are indicated in the conclusion. For this statistical distribution, the Euclidian distance is the universal criterion. It is converted into the Signal to Noise Ratio (SNR) defined as:

$$SNR = 10 \log_{10} \frac{|\bar{\mathbf{X}} - \mathbf{X}|^2}{|\mathbf{X}|^2}. \tag{4.1}$$

The distribution of the transform coefficients. Let us admit that an orthonormal transform is applied on $\mathbf{Y}$. That corresponds to apply a rotation in the signal space. Thus, the noise is still Gaussian, stationary and white. This operation seems to be useless. But the transform can deeply modify the signal statistics. For example, let us consider a signal which is spatially quite uniformly distributed. Even if the pixel distribution law seems to do not depend on the position, its Fourier transform at the lowest frequencies correspond generally to the highest coefficient values; while the values at the highest frequencies can appear very faint compared to the noise deviation. The Wiener denoising (Wiener 1949) is a filtering based on the ratio between the signal and the noise at each frequency. This filter takes thus into account the information content.

This separation between the low and the high frequencies comes at the cost of a space delocalization. At the contrary of the Fourier transform, the DWT allows both a space and a frequency (scale) representation. Even if there is no global information detected at small scales, few coefficients could be significant. A wavelet denoising would restore this information while the Fourier filtering would remove it.

The best transform for the denoising is the one which optimizes the separation between the signal and the noise. That depends on the considered images; but many experiments showed that the DWT is well-adapted for the astronomical images.

Mean coefficient property. Let us consider a DWT coefficient, it can be written as:

$$w(i,k,l) = \sum_{n,m} g_i(n,m)c(k+n,l+m); \tag{4.2}$$

where $\{c(k,l)\}$ is the discrete signal and $g_i(n,m)$ is the discrete wavelet filter at scale i. The filter is a pass-band one, thus we have:

$$\sum_{n} g_i(n,m) = 0. \tag{4.3}$$

By consequence the mean DWT coefficient is also equal to 0 whatever the image background. The distribution of the wavelet coefficients is centered at each scale.

Now, if the signal is constant on the support of the filter (admitted to be compact), the wavelet coefficient is null. Due to the noise, the distribution of the observed coefficients would be a centered Gaussian with a deviation equal to the noise deviation σ for an orthonormal transform.

Significant coefficients. Let us consider a coefficient $w(i,k,l)$. If its value is positive, we consider the probability $p = Prob[W > w(i,k,l)]$, where W is the stochastic variable associated to the noise distribution of the wavelet coefficient. $p < \epsilon$ means that the probability of getting the value from a constant signal is fainter that the significance level ϵ. That leads to introduce a threshold $T(\epsilon)$ such that:

$$Prob(w(i,k) > T) < \epsilon \qquad \text{or} \qquad Prob(w(i,k) < -T) < \epsilon. \tag{4.4}$$

For a Gaussian distribution, the significance ϵ is translated into a factor of the noise deviation ($T = \kappa\sigma$). Note that, if the image has 1000×1000 pixels and $\epsilon = 0.001$, statistically 1000 positive coefficients (false alarms) appear significant for a noisy uniform image. 1000 negative coefficients also appear significant. The false alarm rate is identical to the significance threshold. Here this threshold is equal to 3.09σ, σ being the standard deviation of the Gaussian noise distribution.

4.2 Denoising from thresholdings

The material for the experiments. The restoration tests were done on a simulated image (Mel1) composed as a sum of $2D$ Gaussian functions. Their amplitudes are distributed according to a power law, in order to get an image like astronomical ones. A white Gaussian noise at the levels 0.007, 0.07 and 0.7 was added. In Figure 3 the simulated 256×256 images are displayed. Their SNRs are respectively 14.73, -5.27 and -25.27 dB.

Hard thresholding (HT). The basic method consists into the image reconstruction from only the significant coefficients, according to the threshold T (Starck & Bijaoui 1994).

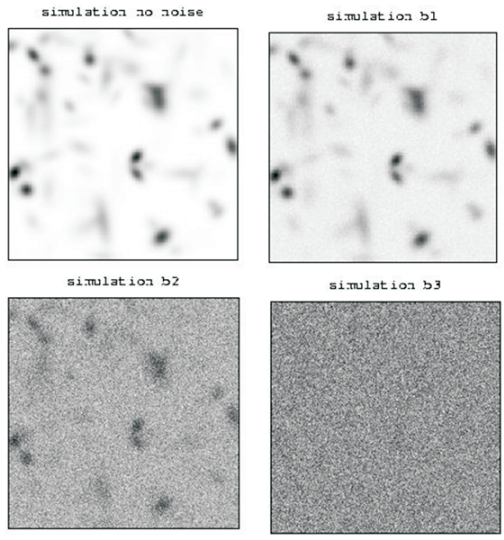


Fig. 3. The simulated images used for the tests by lexicographic order: reference and noisy images with increasing noise level b1, b2 and b3.

Soft thresholding (ST). The images restored with the previous method display some punctual defaults due to the discontinuities introduced by the hard thresholding. Donoho (1995) proposed to soften them with the following rules

$$w(i,k,l) > T \qquad \tilde{w}(i,k,l) = w(i,k,l) - T; \tag{4.5}$$

$$w(i,k,l) < -T \qquad \tilde{w}(i,k,l) = w(i,k,l) + T; \tag{4.6}$$

$$|w(i,k,l)| < T \qquad \tilde{w}(i,k,l) = 0. \tag{4.7}$$

Modified soft thresholding (MST). In previous rules the coefficients are modified even if they are largely significant. In a modified softening we proposed rules with two thresholds to keep them (Bijaoui *et al.* 1997).

$$|w(i,k,l)| \geq T_2 \qquad \tilde{w}(i,k,l) = w(i,k,l); \tag{4.8}$$

$$|w(i,k,l)| \leq T_1 \qquad \tilde{w}(i,k,l) = 0; \tag{4.9}$$

$$T_1 < |w(i,k,l)| < T_2 \qquad \tilde{w}(i,k,l) = w(i,k,l)\frac{|w(i,k,l)| - T_1}{T_2 - T_1}. \tag{4.10}$$

The thresholds. The denoising depends on the chosen κ parameter. In the presented experiments, we set different values, often the same for the whole scales. In the two thresholds case, we set $\kappa_1 = 3.5$ and $\kappa_2 = 4.5$. The corresponding false alarm rates are respectively 4.710^{-4} and $6.7.10^{-6}$, taking into account the two signs.

Donoho & Johnstone (1994) introduced a thresholding rule (DST) based on the minimum risk leading to a threshold depending on the number K of independent

wavelet coefficients (which is the case for an orthogonal DWT):

$$\kappa = \sqrt{2\log_2(K)}. \tag{4.11}$$

As K decreases with the scale by a factor 4 in 2D, the threshold decreases with it. A similar rule was also applied to the à trous algorithm with MST (SMST).

The experiments. In Figure 4 The Haar transform of the Mel1 and the b1 images are plotted. That shows the effect of the noise with the scales. At right of the figure, the denoised b1 image seems quite good. A faint block effect can be identified on this image. On Figure 5 the denoised images obtained for b2 (left) and b3 (right) are displayed. The best images were selected on the different thresholding methods. The block effect largely increases with the noise. This is due to the fact that the noise increasing more and more coefficients become insignificant. The images are thus reconstructed by less and less coefficients, displaying the staircases associated to the Haar scaling function.

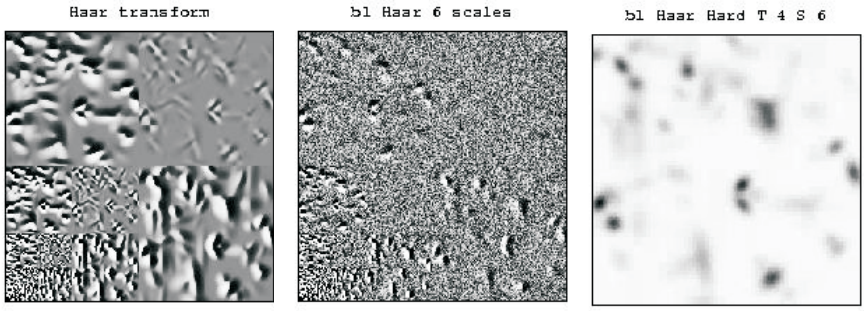


Fig. 4. The Haar transform of the Mel1 and b1 images. At right, the denoised image with the Haar transform and a hard thresholding at 3σ.

In Figure 6 the best denoised images obtained for b1 (left), b2 (middle) and b3 (right) with the Dauchechies 8 transform are displayed. There is no block effect, but a ringing appears around the bright objets. Due the reduction of the number of coefficients, these objects are characterized by peaks in the wavelet transform. Their reconstruction corresponds to the wavy wavelet pattern.

In Figure 7 the wavelet transform obtained with the *à trous* algorithm on b1 is displayed on 6 scales. The noise is clearly identified at the first scale. In Figure 8 the best denoised images obtained for b1 (left), b2 (middle) and b3 (right) with the *à trous* algorithm are displayed. There is no block effect, neither ringing. Some faint holes appeared around bright objects.

In Table 1 the SNRs obtained from different experiments are given. We can note that the denoisings obtained from the Haar transform are generally the worst ones. The effect of the thresholdings depends on the transform, the threshold and the initial SNR. The application of the redundant *à trous* algorithm seems to improve the denoising; but for a low SNR the Dauchechies 8 transform carries out the best result. Thus, the analysis shows that the choice of the best method

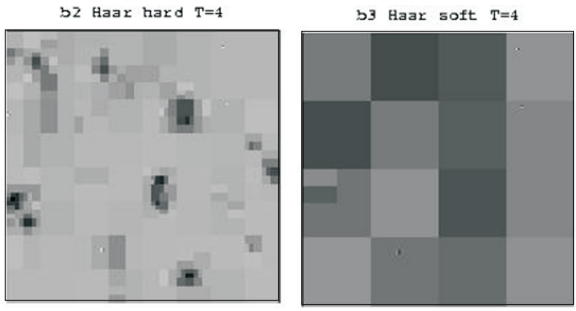


Fig. 5. The best denoised images for b2 and b3 with the Haar transform.

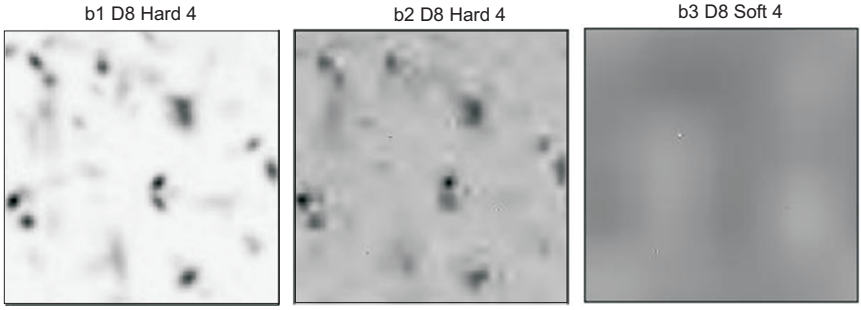


Fig. 6. The best denoised images with the Daubechies 8 transform.

Table 1. SNR obtained on the three different images with different algorithms.

Method	b1	b2	b3
Haar HT $\kappa = 3$	19.49	6.06	−11.18
Haar HT $\kappa = 4$	18.77	7.82	−0.60
Haar ST $\kappa = 3$	17.25	7.06	0.37
Haar ST $\kappa = 4$	15.89	5.94	1.09
Daubechies 8 HT $\kappa = 3$	24.26	7.12	−10.78
Daubechies 8 HT $\kappa = 4$	24.46	9.33	−0.19
Daubechies 8 ST $\kappa = 3$	21.28	8.03	0.52
Daubechies 8 ST $\kappa = 4$	19.72	6.69	1.23
AT MST	26.46	12.25	−0.53
AT SMST	28.22	13.83	−2.96

depends on the input SNR. The adaptation needs to implement a thresholding algorithm taking into account the prior signal distribution, *i.e.* an algorithm based on a Bayesian statistics.

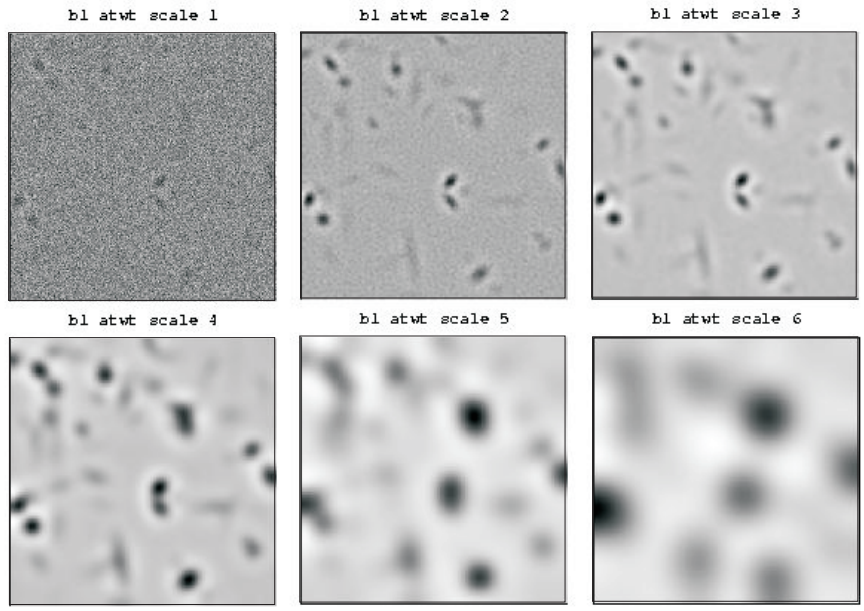


Fig. 7. The *à trous* wavelet transform of the b1 image.

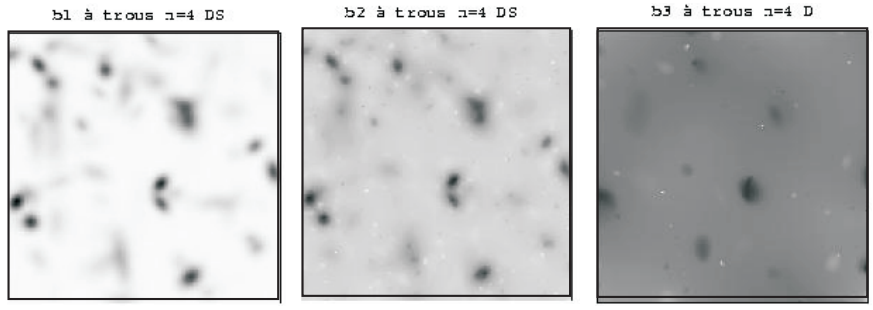


Fig. 8. The best denoised images with the *à trous* algorithm.

4.3 The regularization and the thresholding

Relation between close coefficients. In the previous methods the denoising was based on a local approach, the pixels were thresholded independently of each other. The environment played a role only in the computation of the wavelet coefficient. There are different ways to exploit the correlation between close wavelet coefficients. Here, a simple method based on the regularization is presented.

Values of non significant coefficients. In the HT case the wavelet coefficients are separated in two classes according to their values compared to the threshold. The raw reconstruction consists into the application of the inverse transform with

setting 0 for the non significant coefficients. Some artifacts, like block effects for the Haar transform, appeared in the restoration. In reality, the true value of a non significant coefficient is not null but faint.

Instead of inversing with null values we search to get a function $f(k,l)$ which minimizes a given objective function $C(f)$. The problem can be written as: *determine $f(k,l)$ such that $C(f)$ is minimum and $w_f(i,k,l) = w(i,k,l)$ for each significant coefficient.*

The application of the Tikhonov objective criterion. We set (Bobichon & Bijaoui 1997):

$$C(f) \equiv |D_k(f)|^2 + |D_l(f)|^2; \tag{4.12}$$

where $D_k(f)$ and $D_l(f)$ are respectively the derivatives on the k and l directions. The $C(f)$ minimization is equivalent to:

$$L(f) = 0 \tag{4.13}$$

where L is the image Laplacian. The Van-Cittert algorithm (see Sect. 7.2) leads to an iterated solution:

$$f^{(n+1)} = f^{(n)} + \alpha[0 - L(f^(n))] = f^{(n)} - \alpha L(f^{(n)}). \tag{4.14}$$

where α is an adapted factor. For significant wavelet coefficients, we set:

$$w_{f^{(n+1)}}(i,k,l) = w(i,k,l). \tag{4.15}$$

This operation allows the reduction of the block effects for the Haar transform. The restoration is also improved for the other DWT. Few iterations are generally needed.

Case of a softening function. Here, the wavelet coefficients are softened by a relation:

$$\tilde{w}(i,k,l) = \varpi w(i,k,l) \quad \text{with} \quad \varpi = S(w(i,k,l)). \tag{4.16}$$

Here $S(w)$ is called the softening function. It takes values in the interval $[0,1]$. The application of the regularization can be done by considering ϖ as a weight. So after applying 4.14, we set (Jammal & Bijaoui 2004):

$$\tilde{w}_{f^{n+1}}(i,k,l) = \varpi w(i,k,l) + (1-\varpi))w_{f^{n+1}}(i,k,l). \tag{4.17}$$

For the highly significant coefficients, $\varpi \simeq 1$, no modification is done. While, for non significant ones $\varpi \simeq 0$, the algorithm furnishes the values given from the regularization.

The experiments. In Figure 9 the best denoised images obtained for b1 (left), b2 (middle) and b3 (right) with the *à trous* algorithm with regularization are displayed. There is no block effect, neither ringing. Some faint holes appeared around bright objects.

In Table 2 the results obtained with the application of the regularization, with the *à trous* algorithm are given. 10 iterations were applied. The regularization

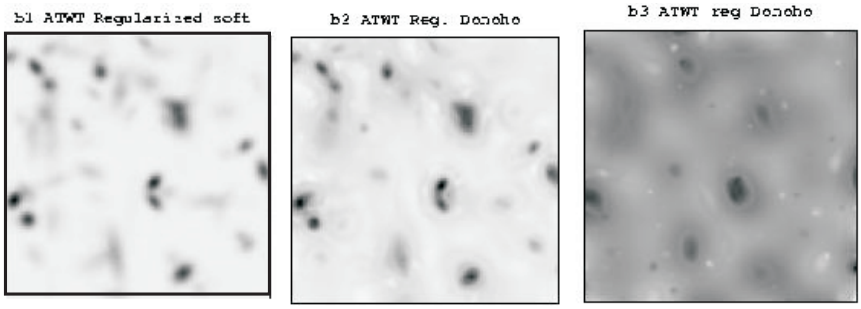


Fig. 9. The best denoised images with the *à trous* algorithm and Tikhonov regularization of the image.

Table 2. The SNRs obtained after regularization with the Tikhonov constraint.

Method	b1	b2	b3
AT-HT	28.29	11.67	−5.48
AT-ST	27.27	10.85	−7.70
AT-DST	24.67	12.55	1.2

allowed a gain in the SNR for the hard thresholding at high SNR level. The gain is less clear at low SNR.

The method was applied to a deconvolution in order to examine the possibility to restore images compressed with the Haar transform (Bobichon & Bijaoui 1997; Jammal & Bijaoui 2004; Dollet *et al.* 2004).

5 Image denoising from the maximum *a posteriori*

5.1 The Bayesian estimations

The posed problem. Let us consider a statistical variable x. It is observed with a noise having a dispersion law $q(y|x)$. The denoising needs to answer the question of what we can say about x knowing y. In the previous sections, only significant coefficients were kept, possibly after a softening. The signal distribution was not taken into account. It was noted that the method carrying out best results depends on the signal-to-noise ratio. Thus, it is necessary to take into account the prior distribution $p_x(x)$. That leads to apply the Bayes rule to get the conditional posterior PDF:

$$p_{x|y} = \frac{p_x(x)q(y|x)}{p_y(y)}; \qquad \text{with} \qquad p_y(y) = \int_{-\infty}^{+\infty} p_x(x)q(y|x)dx. \tag{5.1}$$

From the observation y, knowing the dispersion and the x prior PDF, we get the x posterior PDF. As it is irrelevant to furnish this law for each observed value,

a simple estimate is derived. The maximum of the posterior PDF (MAP) is the most often furnished estimate.

The MAP and the regularization theory. The MAP estimate can be written as:

$$\hat{x} = \text{Arg} \quad \min_x[-\log(q(y|x)) - \log(p_x(x))]. \tag{5.2}$$

This expression is similar to the one introduced in the regularization theory. The first term corresponds to the data attachment $J_2(y, x)$; the second one to the objective function $J_1(x) = -\log(p_x(x))$. In the general case of the regularization theory, the objective function can be formalized independently of a prior distribution.

The case of a Gaussian white noise. From Equation (5.2) we get directly:

$$\hat{x} = y + \sigma^2 \frac{\partial \log(p_x(\hat{x}))}{\partial x}; \tag{5.3}$$

where σ is the standard deviation of the Gaussian noise distribution. Thus, $\hat{x}$ is obtained by solving an equation which may have many roots. That depends on the prior signal law. This law could be determined from the observed y PDF, by solving a deconvolution equation. This operation is delicate due to the histogram fluctuations.

The generalized Gaussian function and the L_q regularization. Experiments on natural images have been carried to model the out PDFs of their wavelet coefficients with extended tails. The prior PDFs were fitted with generalized Gaussian (Moulin & Liu 1998):

$$p_x(x) = ae^{-\frac{|x|^q}{b}}. \tag{5.4}$$

Coupled to Equation (5.2) that leads to the relation:

$$\hat{x} = \text{Arg} \quad \min_x \left[\frac{(y-x)^2}{2\sigma^2} + \lambda |x|^q \right]. \tag{5.5}$$

In this relation $\lambda = 1/b$ In the Gaussian case ($q = 2$), with a signal with a variance s, we get the Wiener filter (Wiener 1949):

$$\hat{x} = \frac{s^2}{s^2 + \sigma^2} y. \tag{5.6}$$

In the case of a Laplacian distribution ($q = 1$) of parameter b, the solution is the soft thresholding with the threshold $\frac{\sigma^2}{b}$. For a lower exponent, the filter tends to become a hard thresholding.

5.2 The basis pursuit

Principles. Chen *et al.* (1998) posed the restoration problem as:

$$\mathbf{Y} = \mathbf{A}\mathbf{\Psi}\mathbf{Z} + \mathbf{N} \tag{5.7}$$

where $\mathbf{\Psi}$ is a set (dictionary) of atoms and $\mathbf{Z}$ the related coefficients allowing the restoration of the image $\mathbf{X} = \mathbf{\Psi Z}$. In this section, only the case $\mathbf{A} = \mathbf{I}$ is examined. An image atom is a given function, generally, but not necessarily, with a null mean. It can be obtained by translation and dilation of a generative function, like for the wavelet transform. A dictionary may be the union of primary dictionaries. A dictionary can be redundant, in which case the Gram matrix computed with these atoms is singular. In this case, a selection has to be done in order to restore the image with few atoms. There are many atom combinations leading to restore the same signal. The basis pursuit consists into the application of the MAP principle. That leads to search the atoms such that:

$$|\mathbf{Y} - \mathbf{\Psi Z}|^2 + \lambda |\mathbf{Z}|_q \tag{5.8}$$

is minimum. $q = 0$ corresponds to minimize the number of atoms (ℓ_0). Chen *et al.* (1998) proposed $q = 1$ which furnishes also a sparse representation. λ is the Lagrangian parameter. Its value results from the respect of the data attachment constraint.

The matching pursuit algorithm. In the ℓ_0 case, Equation (5.8) can be approximatively solved through a matching pursuit algorithm (Mallat & Zhang 1993). It is a greedy algorithm which progressively identifies the different atoms. Whatever the algorithm used to solved 5.8, generally it can not be proved that the minimum number of components is reached. Nevertheless, the matching pursuit algorithms, and specifically the orthogonal matching pursuit (OMP) (Pati *et al.* 1993) algorithms are very popular (Tropp 2004).

The ℓ_1 case. Let us consider the case of an orthogonal wavelet dictionary. Equation (5.8) leads to:

$$\mathbf{\Psi}^T Y - (\mathbf{\Psi}^T \mathbf{\Psi})^{-1} \mathbf{Z} - \lambda sign(\mathbf{Z}) = 0. \tag{5.9}$$

$\mathbf{\Psi}_T \mathbf{Y}$ is the image wavelet transform $\mathbf{W}$. $\mathbf{\Psi}^T \mathbf{\Psi}$ is the identity matrix. Thus we write:

$$\mathbf{Z} = \mathbf{W} - \lambda sign(\mathbf{Z}). \tag{5.10}$$

The algorithm corresponds to a soft thresholding; λ being chosen to satisfy the data attachment condition.

In the case of redundant dictionaries, the basis pursuit algorithm allows the minimization with a sparse representation. Different optimization algorithms were proposed. In particular the Block-Coordinate Relaxation (BCR) method (Bruce *et al.* 1998) leads to fast computations (Starck *et al.* 2004).

The Morphological Component Analysis (MCA). (Starck *et al.* 2004) This method was developed from the use of different transforms, such as the wavelet, the ridgelet (Candès 1998), the curvelet (Candès & Donoho 1999) and DCT (Ahmed *et al.* 1974) ones. BCR (Bruce *et al.* 1998) was used for obtaining the optimal representation. MCA was developed both with the ℓ_0 and the ℓ_1 norms.

The matching pursuit with the à trous algorithm. Instead of using dictionaries of orthonormal bases, it is possible to develop a sparse representation from the *à*

trous or the pyramidal algorithms (Bijaoui 2008). The corresponding atoms are progressively identified taking into account a threshold which decreases at each iteration. It is not guaranteed that the minimum number of atoms is reached with this greedy algorithm.

In Figure 10 the denoised images obtained for b1 (left), b2 (middle) and b3 (right) with this algorithm are displayed. The atoms are identified with the *à trous* wavelet transform using a matching pursuit algorithm A hard thresholding is performed with a threshold equal to 4. Thanks to the representation, the images appear very clean, without noise. Nevertheless the SNRs are not the best ones, 26.17 for b1, 12.07 for b2 and −1.09 for b3. The number of pyrels are respectively 702, 149 and 25. Compared to the number of pixels (65736) that corresponds to a high compression factor (94, 441 and 2629).

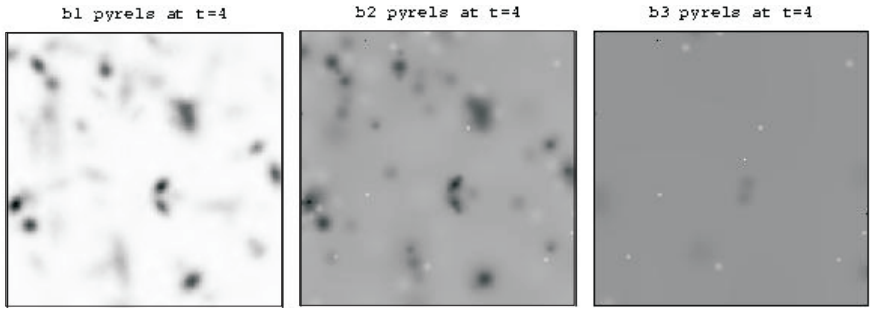


Fig. 10. The images resulting from the matching pursuit algorithm.

6 Image denoising from the posterior mean

6.1 The posterior mean

The minimum mean square estimator. MAP takes into account only the posterior distribution around its maximum which does not characterize the whole distribution. The expectation minimizes the mean square error (MMSE). Its value is:

$$\hat{x} = \int_{-\infty}^{+\infty} x \frac{p_x(x)q(y|x)dx}{\int_{-\infty}^{+\infty} p_x(x)q(y|x)dx}. \tag{6.1}$$

The MMSE evaluation. From Equation (6.1) we note that the evaluation of $\hat{x}$ needs to know the dispersion law $q(y|x)$ and the prior one $p_x(x)$. In this paper, it is admitted that the noise is white and Gaussian. As for the MAP estimation we have to evaluate the prior distribution.

Since Robbins' seminal work (Robbins 1956), it is known that the MMSE can be determined directly from the observed posterior distribution $p_y(y)$ for different dispersion PDFs. This is the case for the Gaussian one.

The Miyasawa relation. Taking into account that:

$$q(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-x)^2}{2\sigma^2}}; \tag{6.2}$$

we get:

$$\hat{x} = y + \frac{1}{\sqrt{2\pi}\sigma} \frac{\int_{-\infty}^{+\infty} (x-y) p_x(x) e^{-\frac{(y-x)^2}{2\sigma^2}} dx}{p_y(y)}. \tag{6.3}$$

As:

$$\frac{\partial p_y(y)}{\partial y} = \frac{1}{\sigma^2} \int_{-\infty}^{+\infty} (x-y) p_x(x) q(y|x) dx; \tag{6.4}$$

we get the Miyasawa relation (Miyasawa 1961):

$$\hat{x} = y + \sigma^2 \frac{\partial \log p_y(y)}{\partial y}. \tag{6.5}$$

Note that the estimate depends only on the distribution of the observed variable y. We can also note the similarity between the Equations (5.3) and (6.5). But in the first case (MAP) the estimate is the solution of an equation which requires the knowledge of the prior PDF, while for the MMSE the estimate is directly furnished by a relation which takes into account the PDF of the observed variable.

6.2 Application of the Miyasawa relation

Application to the denoising with the DWT. Here, it is set that the variable y is the wavelet coefficient at a given scale. Thus, each wavelet plane is analyzed separately. It exists some correlation between the coefficients at the different scales, even for an orthogonal DWT. Thus, even if the MMSE is obtained at each scale, that does not guarantee that it is globally reached.

The estimation of the coefficients distribution. Relation 6.5 appears at the first glance very easy to exploit. However, the estimate depends on the derivative of the logarithm of the PDF, which is hard to correctly estimate. It is posed that the image is the realization of a stationary process. Thus, the coefficient histogram is the empirical $p(y)$ statistics. Its noise results from a Bernoulli distribution. That leads to very bad estimations for the distribution tails. Different approaches were proposed to improve the estimation (Bijaoui 2006, 2009):

- The Parzen method based on a sum of shifted windows (Parzen 1962). This method gives bad estimation for the tails. The window size has to be adapted to the event frequency.
- A denoising based on the wavelet transform (Bijaoui 2006). The Bernoulli noise increases the difficulty.
- A PDF model based on truncated distributions. Raphan & Simoncelli (2007) proposed exponentials. The use of Gaussians leads to easier computations (Bijaoui 2009).

- In the case of astronomical images, the PDF of the observed wavelet coefficients may be fitted by a Voigt function (García 2006), a convolution of a Gaussian with a Lorentzian function (Laplace PDF).

- The PDF can be approximated by a sum of Gaussians (Bijaoui 2002). The EM algorithm can be applied to determine the parameters (Bijaoui 2011). A simple mixture model furnishes a nice approximation to compute the softening filter (Bijaoui 2002).

A simple model for a MMSE estimation. First experiments showed on astronomical images that two Gaussian functions were extracted from the histogram of the wavelet transform at small scales: i/the Gaussian corresponding to the noise, ii/ a second one larger which corresponds to the sum of a signal with the noise. Thus, the PDF can be written as:

$$p(y) = (1-a)G(y,N) + aG(y,S+N); \tag{6.6}$$

where $G(y,V)$ is a centered Gaussian with a variance V.

The noise variance N is supposed to be known. Only a and S have to be determined at each scale. These parameters can be estimated from the variance M_2 and the 4th-order moment M_4:

$$M_2 = (1-a)N + a(S+N); \tag{6.7}$$
$$M_4 = 3(1-a)N^2 + 3a(S+N)^2. \tag{6.8}$$

Thus:

$$S = \frac{\frac{M_4}{3} - N^2}{M_2 - N}; \tag{6.9}$$
$$a = \frac{(M_2-N)^2}{\frac{M_4}{3} - N^2}. \tag{6.10}$$

If $M_2 - N < 0$ or if $M_4 < 3N^2$, S and a are set to 0. If $a > 1$, a is set to 1 and S is estimated only from the variance. We used these simple rules for estimating the model parameters at each wavelet scales.

Experimentations. In Figure 11 the denoised images obtained for b1 (left), b2 (middle) and b3 (right) with this algorithm (FONDW) are displayed. The SNRs are respectively 28.38, 13.15 and 1.84. The softening function is scale/scale determined by an algorithm free of tuning parameter. It is automatically adapted to the SNR.

At this experimentation level, the redundant à trous algorithm with MMSE carried out the best denoisings.

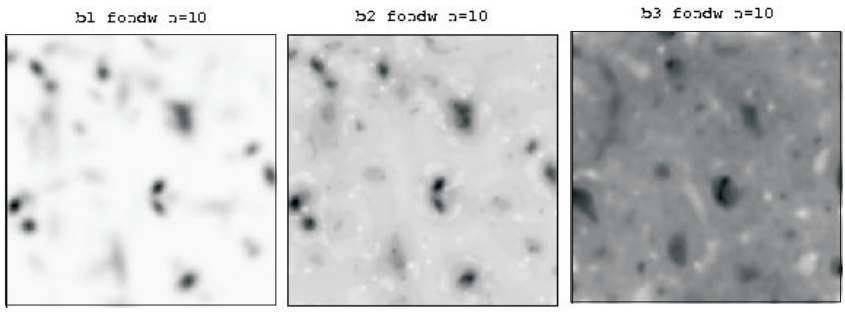


Fig. 11. The images resulting from the algorithm based on a scale/scale posterior expectation.

7 Application to astronomical image deconvolution

7.1 Image deconvolution using the wavelet transform

The direct inversion. The trivial image restoration would consist to denoise the image and to inverse the result (Starck & Bijaoui 1994). This method furnishes correct solutions for a regular blurring operator; but, it is not adapted to the case of a PSF with frequency holes.

The application of iterative inversions. Instead to inverse using the FFT, it is possible to apply an iterative scheme. Necessarily, the number of iterations is limited. This number plays a regularization role. For example, two classical iterative algorithms, the Van-Cittert and the Landweber ones are further presented. The information is not similarly restored for each frequency. If the modulation transfer function (MTF) is high the convergence is very fast; in contrast, the convergence is very low for the smallest MTF values. The number of iterations allows the obtention of a frequency filter depending of the MTF.

Many other iterative algorithms were proposed, such that the Richardson-Lucy one (Lucy 1974); which is very popular in the astronomical laboratories, due to its tendency to carry out images with point-like structures.

The wavelet-vaguelette decomposition. The multiresolution analysis carries out a linear representation with wavelet functions. The inversion is a linear operation so that the previous wavelet representation can be directly translated into a vaguelette decomposition; the vaguelettes being the wavelet functions deconvolved with the PSF (Donoho 1992). This is also convenient for a regular operator, but not adapted to a singular one.

The mirror wavelet. Among the different proposed methods, Kalifa *et al.* (2003) proposed first to deconvolve the image and then to denoise taking account a soft thresholding with mirror wavelet bases. This decomposition is a specific wave packet (Coifman & Wickerhauser 1992) for which the frequency band is decomposed like a mirror compared to the wavelet decomposition. This property allowed them to improve the signal localization both spatially and in frequency.

The multi resolution CLEAN. Wakker *et al.* (1988) proposed a restoration method for aperture synthesis images based on a two-stages CLEAN algorithm, with identification of Dirac peaks and extended Gaussians. Starck *et al.* (1994) developed this idea in the framework of the wavelet transform. The algorithm is similar to the matching pursuit one, with specific wavelet functions.

The application of the basis pursuit. In the previous method, the goal was to represent the image with few wavelet patterns. The basis pursuit algorithm allows the obtention of a solution, with an ℓ_1 minimization (Daubechies *et al.* 2004).

7.2 The Van Cittert and the Landweber iterative inversions

The basic relation. Let us consider the classical inverse problem without noise $\mathbf{Y} = \mathbf{AX}$. Van Cittert (1931) introduced a simple iterative inversion algorithm. The idea consists into writing $\mathbf{B} = \mathbf{I} - \mathbf{A}$; where $\mathbf{I}$ is the identity matrix. Thus, the solution is written as:

$$\mathbf{X} = [\mathbf{I} - \mathbf{B}]^{-1}\mathbf{Y}. \tag{7.1}$$

If all the $\mathbf{B}$ eigenvalues are in the open interval $]-1, +1[$, it can be derived that:

$$\mathbf{X} = [\mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \ldots]\mathbf{Y}. \tag{7.2}$$

The development is limited at the order n:

$$\mathbf{X}^{(n)} = [\mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^{(n)}]\mathbf{Y}. \tag{7.3}$$

That leads to:

$$\mathbf{X}^{(n)} = \mathbf{Y} + \mathbf{B}[\mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^{(n-1)}]\mathbf{Y}; \tag{7.4}$$

Or:

$$\mathbf{X}^{(n)} = \mathbf{Y} + \mathbf{B}\mathbf{X}^{(n-1)} = \mathbf{Y} + [\mathbf{I} - \mathbf{A}]\mathbf{X}^{(n-1)}. \tag{7.5}$$

Thus, finally:

$$\mathbf{X}^{(n)} = \mathbf{X}^{(n-1)} + [\mathbf{Y} - \mathbf{A}\mathbf{X}^{(n-1)}]. \tag{7.6}$$

The convergence factor. It is also possible to accelerate the convergence by the introduction of a convergence factor α_n:

$$\mathbf{X}^{(n)} = \mathbf{X}^{(n-1)} + \alpha_n[\mathbf{Y} - \mathbf{A}\mathbf{X}^{(n-1)}]. \tag{7.7}$$

Its best value is reached for the minimum norm of $\mathbf{R}^{(n)} = \mathbf{Y} - \mathbf{A}\mathbf{X}^{(n)}$:

$$\alpha_n = \frac{\mathbf{R}^{(n-1)}\mathbf{A}\mathbf{R}^{(n-1)}}{|\mathbf{A}\mathbf{R}^{(n-1)}|^2}. \tag{7.8}$$

The distance minimization. With the previous algorithm, the algorithm does not converge if $\mathbf{A}$ is singular, *i.e.* it exists null eigenvalues ($\lambda_p = 0$). In the direction of the corresponding eigenvectors the equation is written as $Y_p = \lambda_p X_p = 0$. If

$Y_p \neq 0$ the equation is not consistent. Thus, the problem is modified by writing $r = |\mathbf{Y}-\mathbf{AX}|^2$ is minimum. That leads to:

$$\mathbf{A}^T[\mathbf{Y} - \mathbf{AX}] = 0; \tag{7.9}$$

which is written as:

$$\mathbf{A}^T\mathbf{Y} = \mathbf{A}^T\mathbf{AX}. \tag{7.10}$$

For $\lambda_p = 0$, we get $0 = 0$. The system is now underdetermined but consistent. A unique solution can be obtained by regularization. A classical constraint consists into setting $X_p = 0$ for $\lambda_p = 0$ (minimum energy).

The Landweber algorithm. The Van-Cittert algorithm is applied to Equation (7.10). That leads to:

$$\mathbf{X}^{(n)} = \mathbf{X}^{(n-1)} + \alpha_n\mathbf{A}^T[\mathbf{Y} - \mathbf{AX}^{(n-1)}] = \mathbf{X}^{(n-1)} + \alpha_n\mathbf{A}^T\mathbf{R}^{(n-1)}. \tag{7.11}$$

This relation is known as the Landweber algorithm, which can be directly developed from a gradient descent (Landweber 1951). α_n is computed as it is upper indicated for the Van Cittert The previous approach allows the comparison between the two algorithms for the convergence. In Relation 7.10 the data are smoothed by the joint matrix. For an eigendirection p, the left value is $\lambda_p Y_p$. If $\lambda_p = 0$, its value becomes also null in this direction. The eigenvalue of $\mathbf{A}^T\mathbf{A}$ is λ_p^2, always positive or null. If we consider a non null eigenvalue, the convergence is always assumed. For a null eigenvalue the algorithm keeps the initial value. If we set $\mathbf{X}^{(0)} = 0$, no new information is added. Thus, the algorithm furnishes the solution which minimizes the energy.

The convergence speed depends on the matrix conditioning, *i.e.* the ratio between the highest eigenvalue and its lowest (and different from 0) one of the matrix $\mathbf{A}^T\mathbf{A}$. Many methods were proposed to accelerate the convergence. The conjugate gradient is the most popular one (Hestenes & Stiefel 1952).

7.3 Deconvolution from the significant residual

The significant residuals. The Landweber algorithm consists into the addition of the residual $\mathbf{R}^{(n-1)}$ smoothed with the adjoint matrix $\mathbf{A}^T$ to the previous approximation. Even if the solution at step $(n-1)$ is not noisy we add a noisy residual. The smoothing by the joint matrix removes only a part of the noise. Thus it is needed to denoise the residual in order to avoid adding noise to the solution. It is considered that it is the same noise for the successive residuals than for the signal $\mathbf{Y}$. In (Murtagh *et al.* 1995) different iterative inversion algorithms were examined. The Van-Cittert and the Richardson-Lucy ones are also available for that purpose. I prefer to present the algorithm using the Landweber algorithm, more stable than the other ones. In the previous sections different denoising methods were examined. The best ones have to be applied. In (Murtagh *et al.* 1995) a hard thresholding was applied. Upper, it was clear that the Bayesian posterior mean leads to the best results for the different SNRs. It is thus chosen for the deconvolution algorithm.

The deconvolution algorihm. It is the following:

1. Set $\mathbf{X}^{(0)} = 0$ and $n = 1$.
2. $\mathbf{Z} = \mathbf{A}\mathbf{X}^{(n-1)}$.
3. $\mathbf{R} = \mathbf{Y} - \mathbf{Z}$.
4. $\mathbf{R}$ is denoised to $\tilde{\mathbf{R}}$.
5. $\mathbf{S} = \mathbf{A}^T\tilde{\mathbf{R}}$.
6. $\alpha_n = \mathbf{S}.\tilde{\mathbf{R}}/|\mathbf{S}|^2$.
7. $\mathbf{X}^{(n)} = \mathbf{X}^{(n-1)} + \alpha_n\mathbf{S}$ and $n = n + 1$.
8. According to the chosen convergence criterion (number of iterations, residual energy ...) the algorithm stops or comes back to step 2.

The positivity constraint. Generally the astrophysical sources are positive functions. A background is preliminary subtracted in order to get a positive image function. The positivity constraint can be easily satisfied by a thresholding to positive values at each iteration. The application of this constraint to a deconvolution with a PSF having frequency holes may lead to a significant gain in resolution.

The deconvolution experiments. The simulated image was smoothed with a Gaussian PSF having the size 1 (Msg1), 2 (Msg2) and 4 (Msg4). A white Gaussian noise was added with the deviations 0.007, 0.07 and 0.7. The applied deconvolution program was based on the FONDW denoising. The residuals are scale by scale examined. The wavelet coefficients are softened with a filter derived from the Miyasawa relation. Their histograms are fitted with the Gaussian mixture associated to the simple model. After deconvolution, the SNR for the deconvolved image compared to the initial one (Mel1) is computed. The SNR for the denoised image compared to the blurred image, free of noise, is also determined.

The results. In Figure 12 the denoised and restored images obtained with this algorithm (dgondw) are displayed for the Mel1 image blurred with a Gaussian PSF with $\sigma = 1$. The lines correspond respectively to the b1, b2 and b3 images. At left the smoothed noisy images, at middle the smoothed denoised ones and at right the Mel1 restoration. In Table 3 the resulting SNRs are given. It can be noted that the deconvolved images for b1 and b2 lead to a better SNR than the images obtained without blurring. That is probably due to the a regularization effect introduced by the Landweber iterations. In Table 3 the SNRs for the restored images with a blurring at $\sigma = 2$ and $\sigma = 4$ are also indicated. The positive constraint brings a significant gain, especially for a low SNR. For a large PSF, the number of iterations inside the Landweber algorithm was also increased to improve the results for a high SNR. On Figure 13 the denoised and restored images are displayed for $\sigma = 4$.

The denoising from deconvolution taking into account an hypothetic PSF. In the previous paragraph the denoising derived from a deconvolution with a known PSF.

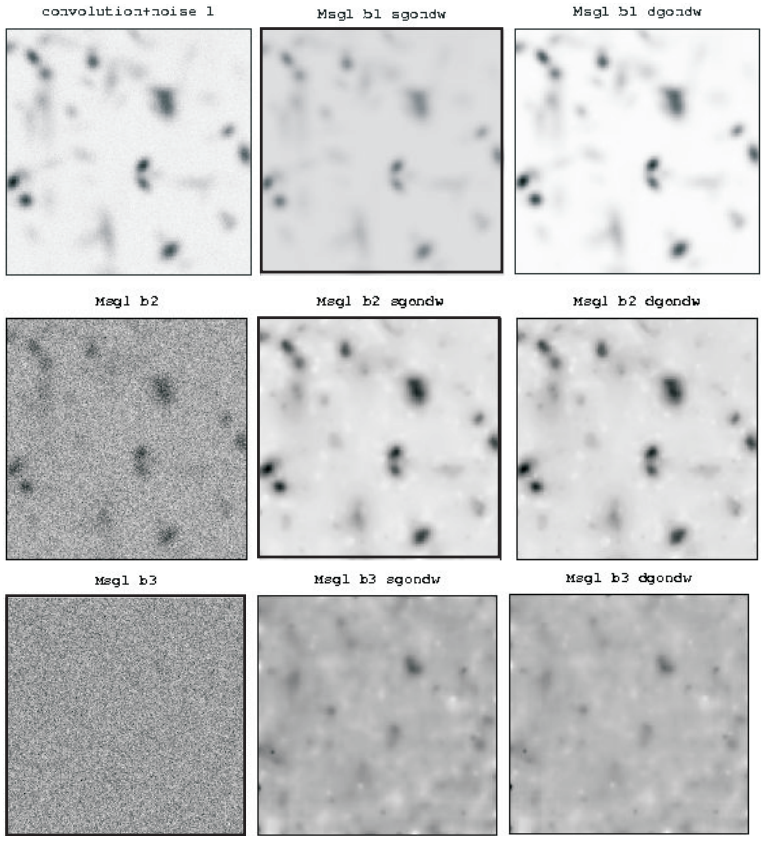


Fig. 12. The blurred noisy images at $\sigma = 1$ (*left*), its resulting denoised (*middle*) and restored images (*right*).

Table 3. SNR resulting from the deconvolution experiments. p: the positivity constraint is applied; x: 100 iteration steps were done instead to 10. (s) means smoothing and (d) deconvolution.

Image	(s)b1	(d)b1	(s)b2	(d)b2	(s)b3	(d)b3
Msg1	29.90	28.89	14.47	14.03	−1.12	−1.27
Msg2	30.02	27.25	13.87	12.78	−0.16	−0.42
Msg2 p	30.67	28.29	15.54		1.79	1.52
Msg4 p	30.56	14.07	15.32	11.48	0.82	0.08
Msg4 px	30.90	22.35	14.97	11.30	−0.35	−2.12

It can be noted that the SNRs were very fine. But the comparison was done to the blurred images without noise (Msg1, Msg2, Msg4) and not to the initial one (Mel1).

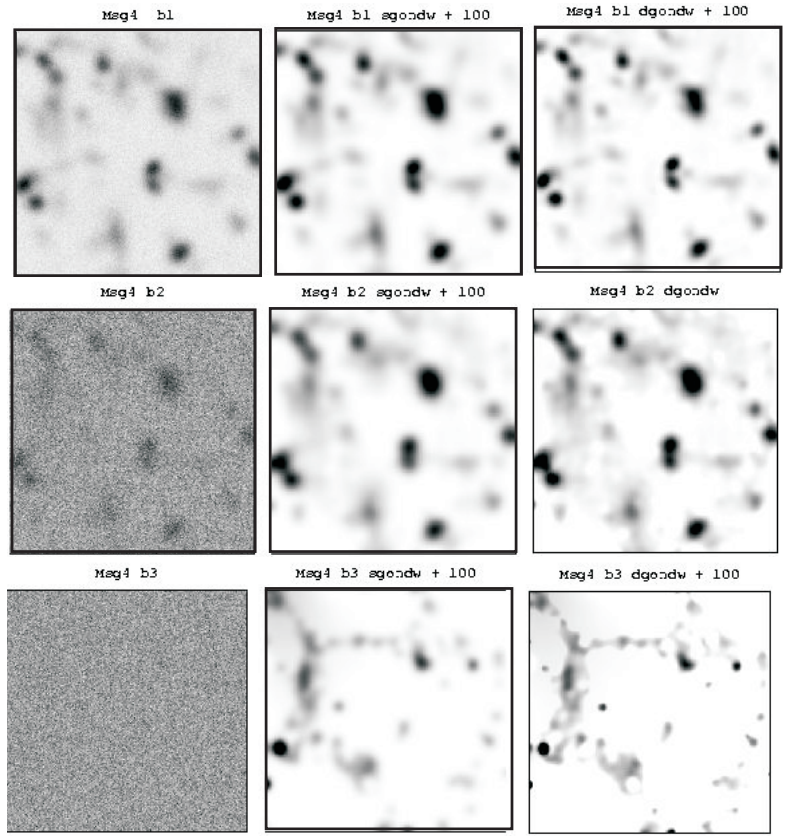


Fig. 13. The blurred noisy images at $\sigma = 4$ (*left*), its resulting denoised (*middle*) and restored images (*right*).

Table 4. The SNRs obtained after denoising taking into account an hypothetic PSF. The first number is the size of the hypothetic Gaussian PSF. x means that 100 iteration steps were done instead of 10.

PSF size	b1	b2	b3
1	29.99	15.55	1.65
2	30.11	15.24	1.72
4 x	30.11	15.80	2.69

It is posed that the observed image (for example b1) is the noisy blurred version of an image with a Gaussian PSF but of unknown width. Different widths are tested. The denoised image obtained after deconvolution is compared to Mel1. If the chosen PSF is close to a Dirac peak, no gain can be obtained compared to a direct denoising. For a too extended chosen PSF, information is lost on the details

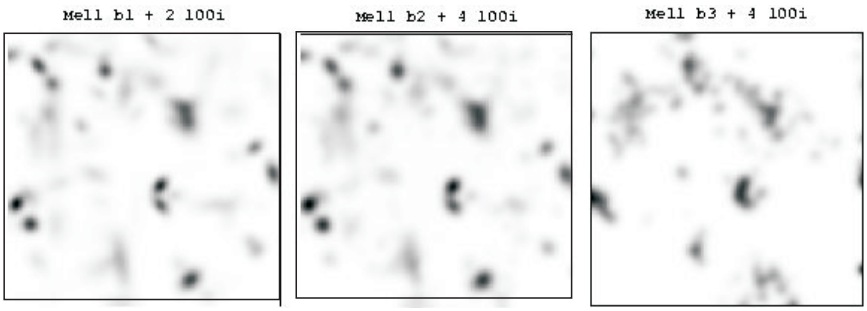


Fig. 14. The denoisings obtained from a deconvolution from an hypothetic PSF at $\sigma = 4$.

at small scales. In Table 4 the resulting SNRs are given for different PSF. In the whole cases the positivity constraint was applied. On Figure 14 the denoised Mel1 images are displayed for $\sigma = 4$. They correspond to the best denoised images.

It can be noted that the best SNRs were obtained with the present method.

8 Conclusion

The study limits. In the present paper, an introduction to the image restoration using multiscale methods was given. A large number of papers were published since two decades on this topic. The paper was centered on the use of a peculiar algorithm, the *à trous* one, for the restoration of astrophysical images corrupted by a white Gaussian noise. In the literature a large panel of transforms and different noises were examined.

The Poisson case. Many modern astronomical detectors furnish images for which the noise statistics is dominated by the photon noise. That led to introduce specific methods to restore these images with a Poisson noise (Murtagh *et al.* 1995). Different methods were proposed to quantify the significance of a wavelet coefficient. A simple way consists into the transformation of the initial pixel values such its variance becomes constant. The Anscombe transform (1948) was first proposed. The results are fine for a mean number of photons greater than about 10. Below this value, other strategies were proposed (Fadda *et al.* 1998; Bijaoui & Jammal 2001; Jammal & Bijaoui 2004; Zhang *et al.* 2008).

Other cases. Different other noise processes were also considered. A generalization of the Anscombe transform was proposed to process images with a mixing of a Gaussian and a Poisson noise (Murtagh *et al.* 1995). More generally the noise properties can be provided by a table. That leads to laborious computations to estimate for each wavelet coefficient its distribution.

Oriented wavelets. Orthogonal DWT carries out simple directional information. Discrete CWT allows the application of oriented wavelets using the FFT (Antoine *et al.* 1996). The steerable pyramids were introduced in order to overcome the limitations of orthogonal separable wavelet decompositions that do not represent

oblique orientations well (Simoncelli *et al.* 1992). This transform is shift invariant and also rotation invariant. Kingsbury (2002) generalized the use of complex DWT for the image processing. This transform allows an approximate shift invariance and some oriented information. More recently an oriented transform based on wavelets and the lifting scheme was introduced by Chapellier & Guillemot (2006) for image compression. This work brings a general framework for oriented DWTs.

Other representations. This paper was oriented to the application of the discrete transforms derived from the continuous one. Other multiscale representations were mentioned such that the ridgelets, the curvelets or the wave packets. Many other multiscale geometric representations were developed with a directional and frequency selectivity (Jacques *et al.* 2011).

A quality criterion for a given image. Beyond the proposed representations, the main problem resides in their capability to best restore the images. But, what is "a best restored image"? Here, the SNR was the alone considered criterion. The search of sparse representations leads to images which a clean appearance, fully denoised. But that does not mean that the results are better than the ones obtained with other algorithms leading to a faint noisy appearance. The posterior expectation leads to a better estimation than the MAP in term of quadratic error, even if residual fluctuations can be identified.

In fact, the main question concerns the use of the restored images. The image restoration is only a step in the image analysis. The astrophysicists are interested by the detection and the characterization of cosmic sources. They apply on the restored images programs for the source identification. Thus, more accurate quality criteria may derive from the detection rate, the false-alarm rate and the quality of the resulting measurements. This feature is fully outside this present introduction. Nevertheless the readers can get useful information on these questions inside the reference papers.

I thanks Pr. D. Mary and Pr. D. Nuzillard for their helpful comments on the draft version.

References

Ahmed, N., Natarajan, T., & Rao, K.R., IEEE Comp., 32, 90

Anscombe, F.J., 1948, Biometrika, 15, 246

Antoine, J.-P., Vandergheynst, P., & Murenzi, R., 1996, Int. J. Im. Sys. Tech., 152

Bijaoui, A., 2002, Signal Proc., 82, 709

Bijaoui, A., 2008, Astronomical Data Analysis V, ed. F., Murtagh & J.L. Starck, `http://www.ics.forth.gr/ada5/pdf_files/Bijaoui_talk.pdf`, 2008

Bijaoui, A., 2009, GRETSI Conference, ed. D. Ginhac (Dijon University)

Bijaoui, A., 2011, GRETSI Conference, ed. Y. Berthoumieu & M. Najim (Bordeaux University)

Bijaoui, A., Bobichon, Y., Fang, Y., & Rué, F., 1997, Traitement du Signal, 14, 179

Bijaoui, A., & Jammal, G., 2001, Signal Proc., 81, 1789

Bijaoui, A., Nowak, R., & Martin, A., 2006, Astronomical Data Analysis IV, ed. C. Surace, http://www.oamp.fr/conf/ada4/pub/19/DSD/ADAIV_Bijaoui.pdf

Bijaoui, A., Starck, J.L., & Murtagh, F., 1994, Traitement du Signal, 11, 229

Bobichon, Y., & Bijaoui, A., 1997, Exp. Astr., 7, 239

Bruce, A., Sardy, S., & Tseng, P., 1998, Proc. SPIE, 3391, 75

Candès, E., 1998, Ridgelets: theory and applications, Ph.D. Thesis, Statistics (Standford University)

Candès, E.J., & Donoho, D.L., 1999, in Curves and Surfaces, ed. Schumaker *et al.* (Vanderbilt University Press, Nashville, TN)

Chappellier V., & Guillemot, C., 2006, IEEE IP, 15, 2892

Chen, S.S., Donoho, D.L., & Saunders, M.A., 1988, SIAM J., 20, 33

Coifman, R.R., & Donoho, D.L., 1995, Technical report 475, Dpt. of Statistics (Stanford University)

Coifman, R.R., & Wickerhauser, M.V., 1992, IEEE IT, 38, 876

Daubechies, I., 1988, Com. Pure Appl. Math., 41, 909

Daubechies, I., Grossmann, A., & Meyer, Y., 1986, J. Math. Phys., 27, 1271

Daubechies, I., Defrise, M., & De Mol, C., 2004, Com. Pure Appl. Math., 57, 1413

Dollet, C., Bijaoui, A., & Mignard, F., 2004, A&A, 2004, 426

Donoho, D.L., 1992, Appl. Comput. Harmonic Anal., 2, 101

Donoho, D.L., 1995, IEEE IT, 41, 613

Donoho, D.L., & Johnstone, J.M., 1994, Biometrika, 81, 425

Fadda, D., Slezak, E., & Bijaoui, A., 1998, A&ASS, 127, 335

García, T.T., 2006, MNRAS, 369, 2025

Grossmann, A., & Morlet, J., 1984, SIAM J., 15, 723

Hestenes, M.R., & Stiefel, E., 1952, J. Res. National Bureau Stand., 49, 409

Högbom, J., 1974, A&ASS, 15, 417

Holschneider, M., Kronland-Martinet, R., Morlet, J., & Tchamichian, P., 1989, in Wavelets Combes, ed. J.M. *et al.* (Springer-Verlag), 286

Hou, H., & Andrews, H., 1978, IEEE ASSP, 26, 508

Jammal, G., & Bijaoui, A., 2004, Signal Proc., 84, 1049

Esteban, D., & Galland, C., 1977, in Proc. ICASSP, 191

Kalifa, J., Mallat, S., & Rougé, B., 2003, IEEE IT, 12, 446

Kingsbury, N.G., 2001, J. Appl. Comp. Harmonic Anal., 10, 234

Landweber, L., 1951, Am. J. Math., 73, 615

Jacques, L., Duval, L., Chaux, C., & Peyré, G., 2011, Signal Proc., 91, 2699

Lucy, L.B., 1974, AJ, 79, 745

Mallat, S., 1989, IEEE PAMI, 11, 674

Mallat, S., 1998, A wavelet tour of signal processing (Academic Press, San Diego)

Mallat, S., & Zhang, Z., 1993, IEEE SP, 41, 3397

Miyasawa, K., 1961, Bull. Inst. Internat. Statist., 38, 181

Morlet, J., Arens, G., Fourgeau, E., & Giard, D., 1982, Geophysics, 47, 203

Moulin, P., & Liu, J., 1998, IEEE, IT, 45, 909

Murtagh F., Starck J.L., & Bijaoui A., 1995, A&ASS, 112, 179

Parzen, E., 1962, Ann. Math. Stat., 33, 1065

Pati, Y., Rezaiifar, R., & Krishnaprasad, P., 1993, in Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers

Raphan, M., & Simoncelli, E.P., 2007, Comp. Sc. Tech. Report TR2007-900, Courant Institute of Math. Sci. (New York University)

Raphan, M., & Simoncelli, E.P., 2008, IEEE IP, 17, 1342

Robbins, H., 1956, Proc. Third Berkley Symp. on Math. Stat., 1, 157

Shannon, C.E., 1948, Bell Syst. Tech. J., 27, 379

Schoenberg, I.J., 1946, Quart. Appl. Math., 4, 45, 112

Simoncelli, E.P., Freeman, W.T., Adelson, E.H., & Heeger, D.J., 1992, IEEE IT, Special Issue on Wavelets, 38, 587

Starck, J.L., & Bijaoui A., 1994, Signal Proc., 35, 195

Starck, J.L., Bijaoui, A., Lopez, B., & Perrier, C., 1994, A&A, 283, 349

Starck, J.L., Murtagh, F., & Bijaoui, A., 1998, Image Processing and Data Analysis in the Physical Sciences. The Multiscale Approach (Cambridge University Press)

Starck, J.L., Elad, M., & Donoho, D.L., 2004, Adv. Imaging Electr. Phys., 132

Strang, G., 1989, SIAM Review, 31, 614

Strang, G., & Fix, G., 1971, in Constructive Aspects of Functional Analysis, Edizioni Cremonese, Rome, 796

Tropp, J.A., 2004, IEEE IT, 50, 2231

Van Cittert, P.H., 1931, Z. Phys., 69, 298

Vitterli, M., 1986, Signal Proc., 10, 219

Wakker, B.P., & Schwarz, U.J., 1988, A&A, 200, 312

Wiener, N., 1949, Extrapolation, Interpolation, and Smoothing of Stationary Time Series (Wiley, New York)

Zhang, B., Fadili, J.M., & Starck J.L., 2008, IEEE IT, 17, 1093

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 303–324*

CONSTRAINED MINIMIZATION ALGORITHMS

H. Lantéri[1], C. Theys[1] and C. Richard[1]

Abstract. In this paper, we consider the inverse problem of restoring an unknown signal or image, knowing the transformation suffered by the unknowns. More specifically we deal with transformations described by a linear model linking the unknown signal to an unnoisy version of the data. The measured data are generally corrupted by noise. This aspect of the problem is presented in the introduction for general models. In Section 2, we introduce the linear models, and some examples of linear inverse problems are presented. The specificities of the inverse problems are briefly mentionned and shown on a simple example. In Section 3, we give some information on classical distances or divergences. Indeed, an inverse problem is generally solved by minimizing a discrepancy function (divergence or distance) between the measured data and the model (here linear) of such data. Section 4 deals with the likelihood maximization and with their links with divergences minimization. The physical constraints on the solution are indicated and the Split Gradient Method (SGM) is detailed in Section 5. A constraint on the inferior bound of the solution is introduced at first; the positivity constraint is a particular case of such a constraint. We show how to obtain strictly, the multiplicative form of the algorithms. In a second step, the so-called flux constraint is introduced, and a complete algorithmic form is given. In Section 6 we give some brief information on acceleration method of such algorithms. A conclusion is given in Section 7.

1 Introduction

Inverse problems arise in a variety of important applications in science and industry, such as optical and geophysical imaging, medical diagnostic, remote sensing. More generaly such problem occurs when the measured quantities are not directly the quantities of interest (parameters). In such applications, the goal is to estimate the unknown parameters, given the data. More precisely, denoting y the measured

[1] Laboratoire Lagrange, UMR 7293, Université de Nice Sophia Antipolis, CNRS, Observatoire de la Côte d'Azur, Campus Valrose, 06108 Nice Cedex 2, France

DOI: 10.1051/eas/1359014

data (output of a physical system, generaly corrupted by noise), x the input of the system, $m(a,x)$ the model and a the internal parameters of the model ($m(.,.)$ is a known function), four cases must be considered:

- when y, $m(.,.)$ and x are known, the goal is to identify the optimal values of the internal parameters a of the model $m(a,x)$;
- when y, $m(.,.)$ and a are known, the goal is to found the optimal value of x;
- when x, $m(.,.)$ and a are known, y is easy to compute; it is the direct problem;
- when y, $m(.,.)$ only are known, we can say that we have a "blind inverse problem" which is much more difficult to solve than the previous ones (see for example NMF and blind deconvolution).

To solve such inverse problems we are generally faced with the problem of minimization of a discrepancy function between the noisy data y and the (unnoisy) model $m(a,x)$. The discrepancy function must deal with significant properties from the physical point of view, and must leads to a mathematically tractable minimization problem.

Moreover, to be physically acceptable, the solution is subjected to some specific (physical) constraints that have to be taken into account during the minimization process.

In this paper we are mainly concerned with physical processes described by a linear model. An algorithmic method allowing to deal with the minimization of any stricly convex differentiable discrepancy function is proposed; classical constraints such as positivity and fixed sum (integral) are taken into account.

2 Inverse problems with linear models (Bertero 1989; Bertero *et al.* 1998)

2.1 Linear models

In this case, the model $m\,(a,x)$ is simply described by a linear relation between the unknown (input) signal x and the unnoisy transformed signal $\widetilde{y}$ (output), we simply write:

$$\widetilde{y} = m\,(a,x) = Hx. \tag{2.1}$$

More generally, if H (the parameters of the system) is known, for a given x, we can compute $\widetilde{y}$.

On the other hand we have the experimental data y, that is a noisy version of $\widetilde{y}$. The problem is to find a solution x such that Hx is as close as possible to y. This is generally performed by minimizing a discrepancy function between y and Hx, eventually subject to constraints.

This brief presentation shows that we are typically dealing with an inverse problem (Bertero *et al.* 1998) whose difficulties will be briefly indicated in the following sections. We first give some examples of problems in which the model is described by a linear relation.

2.2 Some examples of linear models

2.2.1 Linear unmixing (Heinz & Chang 2001)

In such problems, the model is described by the relation $\widetilde{y} = Hx$. The experimental data y is a one dimensional optical spectrum sampled at various (equispaced) wavelenghts; these (noisy) observations are obtained for example by the spectroscopic analysis of the light contained in a given pixel of an image. The matrix $[H]$ is formed by the juxtaposition of columns containing the (known) spectra of basis possible component (the endmembers, that is, the elements of a dictionnary), sampled at the same wavelenght as the data. The unknown vector x contains the weights (abundances) corresponding to the endmembers, so that the data vector is described as a weighted sum of the endmembers. The constraints in this problem are the following: the weights must be positive or zero, moreover their sum must be 1 (that is, they express a percentage).

One can think that in order to solve this problem in full generality, a supplementary condition must be that the sum of the components of the data, and the sum of the components of the endmembers must be equal.....

2.2.2 Non negative Matrix Factorization N.M.F. - Hyperspectral data (Lee & Seung 2001; Cichocki *et al.* 2009)

Extending first the previous problem, the model can be described by a matrix equation $\left[\widetilde{Y}\right] = [H]\,[X]$. The matrix $[H]$ is the one described in the previous problem, it contains the "endmembers". The unknowns are organized in a matrix $[X]$, each column of this matrix contains the weights (abundances), so that the column "n" of $\left[\widetilde{Y}\right]$ is modeled as the sum of the endmembers (columns of $[H]$) weighted by the elements of the column "n" of $[X]$.

The experimental data are organized in a matrix $[Y]$; each column of $[Y]$ is an optical spectrum analogous to those considered in the previous problem, they correspond to all the pixels of an image. If the matrix $[H]$ is known, the problem will be a simple succession of "linear unmixing" problems.

The NMF problem becomes much more complicated because the endmembers are not known, so that the matrix $[H]$ is unknown as well as $[X]$.

Roughly speaking, the problem is then: knowing the (noisy) data matrix $[Y]$ described as the product of two matrix $[H]$ and $[X]$, found such two matrices subject to some constraints.

2.2.3 Deconvolution (Andrews & Hunt 1977; Demoment 1989; Bertero *et al.* 2008)

Let us consider the case of images. In the space of continuous functions, the model is described by a first kind Fredholm integral with space invariant kernel. After discretization, the data (noisy blurred image), the point spread function (PSF) and the unknown object are obtained as tables of dimensions (N*N) and the model is described by a discrete convolution between the PSF and the object (that can be easily performed using FFT). However for sake of generality, we adopt a matrix notation, so that the columns (or rows) of the data and of the unknown object tables are organized in stack vectors y and x respectively (lenght N^2), the transformation matrix H is then $(N^2 * N^2)$, moreover, if the kernel of the Fredholm equation is space invariant, H is Block-Toeplitz; note that this is not the case for example in medical imaging where, while we have a linear model, the kernel is no more space invariant and corellatively, the matrix H does not have any specific property.

Let us focus more specifically on the deconvolution problem for astrophysical imagery. In such a case, the kernel of the integral equation is not only space invariant, but also positive and moreover, its integral is equal to 1, so that the convolution (blurring operation) of a positive object of known integral gives a positive image with the same integral; such a convolution acts as a low pass spatial filtering operation. The intensities in the image pixels have been redistributed, while the total intensity in the image is equal to the total intensity in the object.

For the discretized problem, this is analogous to say that each column of the matrix H is of sum 1.

The first constraint of our problem is then: the "solution must be positive or zero", while a second constraint will be "the flux must be maintained". Frequently, this last constraint is not clearly taken into account. One can consider that the deconvolution problem is closely related to the "linear unmixing" problem with however some specific difficulties due to the low pass filtering effect of such convolution.

2.2.4 Blind deconvolution (Ayers & Dainty 1988; Lane 1992)

The blind deconvolution can be considered with respect to the classic deconvolution as an analogous of the NMF problem with respect to the linear unmixing problem. Indeed, the data model boils down to the convolution product of two unknown functions, then, the number of unknown is (two times) higher than the number of data values; the convolution is however commutative, while for NMF, the matrix product is not, moreover the specific problems appearing in classic deconvolution obviously remains. Then, this problem is very hard to solve and it is out of the scope of the present analysis.

2.3 Some generalities on inverse problems

Inverse problems are generally ill-posed problems in the sense of Hadamard; the conditions of Hadamard (1923) for well-posed problems are:

- the solution must "exist"
- the solution must be "unique"
- the solution must be "stable with respect to the measurement errors" (the noise).

If any of these conditions is not fulfilled, the problem is "ill-posed".

While in finite dimensional spaces, the difficulties linked to the existence and uniqueness of the solution could be circumvented, the problem of stability remains because it is a consequence of the ill-conditionning of the matrix H, that is the condition number K of the matrix H (ratio of the maximum singular value to the minimun singular value $K = \frac{\lambda_{Max}}{\lambda_{min}}$) is high.

To clarify this point in a very simple way, let us consider a simple system of two linear equations with two unknowns, illustrated in Figures 1 and 2.

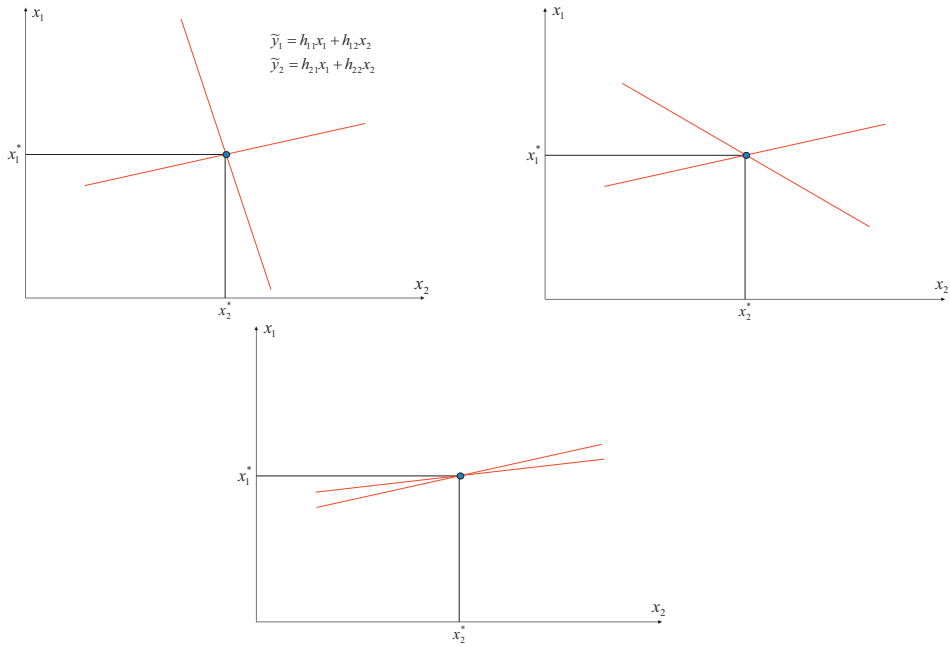


Fig. 1. For a system of two equations with two unknown, three cases are examined, depending on the condition number of the matrix H.

The Figure 1 correspond to the case of an unperturbed system (no noise added). In Figure 1 (upper left), the two lines are almost orthogonal ($K \approx 1$), the system is very well conditionned. If we think for example, to solve the system with an

iterative method operating by successive orthogonal projections on the two lines (1 iteration = 2 projections), it is clear that only a very small number of iterations is necessary to reach an acceptable point (close enough to the solution).

In Figure 1 (upper right), the condition number K has been increased, the two lines are no more orthogonal. Using the iterative method previously described, the iteration number allowing to reach the solution has been increased, but we can expect to reach an acceptable point.

In Figure 1 (lower), the value of K has been strongly increased, the problem is now ill-conditionned, clearly, the solution is always unique, but the necessary number of iterations heavily increases.

To summarize, the only difference between the three cases is an increase of the iteration number and then of computing time when the problem become ill-conditionned.

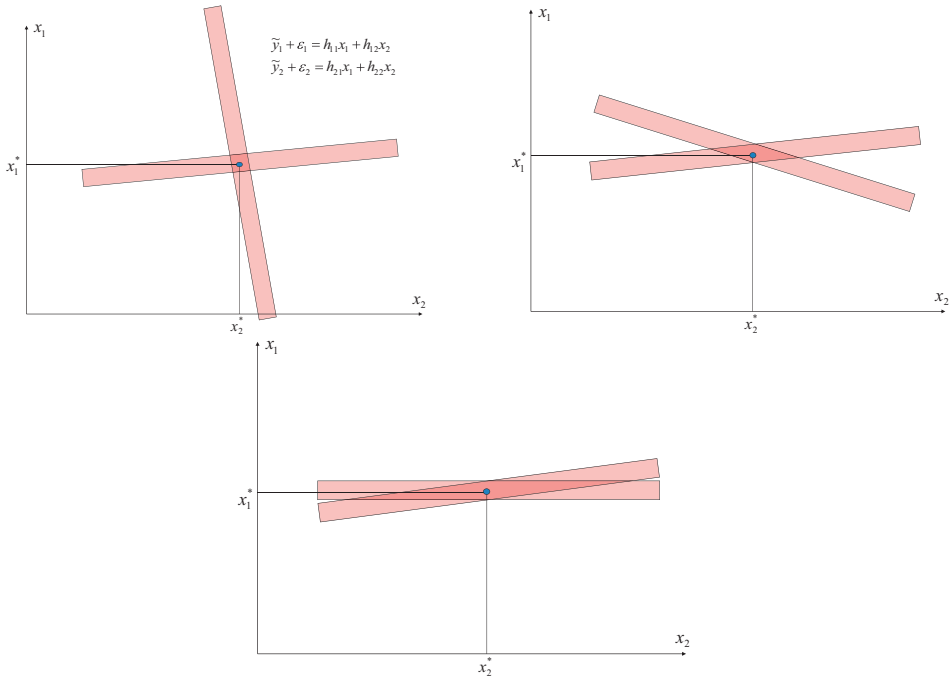


Fig. 2. For a system of two equations with two unknown, three cases are examined, depending on the condition number of the matrix H. A small amount of noise ϵ has been added to the unnoisy data $\widetilde{y}$.

In Figure 2, a small amount of error (noise) has been added to the data.

Depending on the value of the error, the lines remains parallel to themselves, but moves in their respective shaded areas.

Clearly, there will be always one and only one solution that will be located somewhere in the intersection of the two shaded areas.

In Figure 2 (upper left), the solution is close to the one of the initial noiseless problem Figure 1 (upper left) then, a small error on the data will corresponds to a small error on the solution, this behavior is typical of the well-posed problems. In Figure 2 (upper right), the condition number K increases as in Figure 1 (upper right). The solution is always unique and located in the intersection of the shaded areas, but it can be in some cases far from the solution of the initial noiseless problem. In Figure 2 (lower), K has a very large value, the problem is now ill-conditionned and the solution can be very far from the true solution Figure 1 (lower).

This is a simplebut explicit illustration if the difficulties related to the stability of the solution with respect to the measurement errors in ill-posed problems.

3 Distances and divergences (Basseville 1996; Taneja 2005)

To solve the inverse problem, *i.e.* to recover the solution x such that the model $m(a, x)$ is as close as possible to the noisy data y, we must minimize a scalar discrepancy function between y and $m(a, x)$ quantifying the gap between them.

Let p_i and q_i the elements of two data fields p and q, the discrepancy function $D(p, q)$ between the two fields must have the following properties:

1. $D(p, q)$ must be positive if $p \neq q$
2. $D(p, p) = 0$
3. $D(p, q)$ must be convex (strictly) with respect to p and q (at least w.r.t. the field corresponding to the model).

With these properties, $D(p, q)$ is a "divergence". If, moreover the triangular inequality is fullfilled, then $D(p, q)$ is a distance. This last point is not necessary for our purpose. Finally, we consider that generally, such quantity allowing to deal with the whole data fields is the sum of analogous distances (divergences) between corresponding elements of the two fields.

$$D\left(p, q\right) = \sum_i D\left(p_i, q_i\right) \tag{3.1}$$

$$D\left(y, m\left(a, x\right)\right) = \sum_i D\left(y_i, \left\{m\left(a, x\right)\right\}_i\right). \tag{3.2}$$

3.1 Csiszar divergences (Csiszar 1991)

Let $f\left(x\right)$ be a strictly convex function, with $f\left(1\right) = 0$, and for our specific use $f^{'}\left(1\right) = 0$; this last property is very important in our case.

The general class of Csiszar divergences is defined as:

$$C_f\left(p, q\right) = \sum_i q_i f\left(\frac{p_i}{q_i}\right). \tag{3.3}$$

Generally $C_f(p,q) \neq C_f(q,p)$.
This divergence is jointly convex w.r.t. p and q.

3.2 *Divergences founded on convexity measures*

3.2.1 Jensen or Burbea-Rao divergences (Burbea & Rao 1982)

This class of divergences is founded on the classical definition of the convex functions that can be expressed as: let $f(x)$, a strictly convex function, and let p and q two values of the variable, the secant between the points $\{p, f(p)\}$ and $\{q, f(q)\}$ is always superior to the curve between the same points. This is represented in Figure 3 and expressed by the relation (3.4)

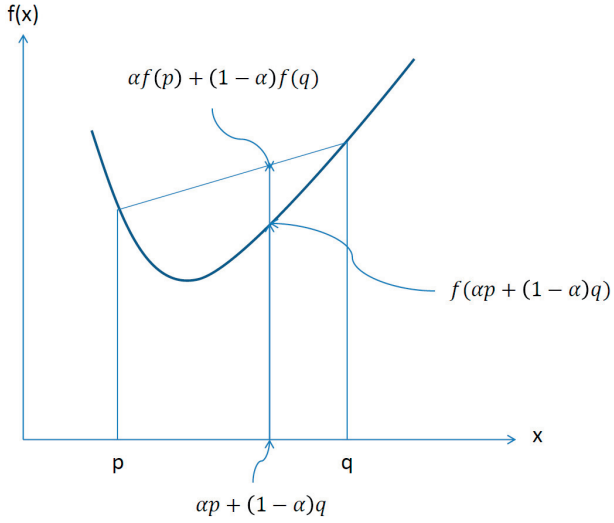


Fig. 3. Strictly convex function.

$$\alpha f(p) + (1-\alpha) f(q) - f[\alpha p + (1-\alpha) q] \geq 0. \tag{3.4}$$

The divergence is then:

$$J_f(p,q) = \sum_i \{\alpha f(p_i) + (1-\alpha) f(q_i) - f[\alpha p_i + (1-\alpha) q_i]\}. \tag{3.5}$$

Note that the convexity of the basis function $f(x)$ does not ensure the conxexity of the corresponding divergence.

3.2.2 Bregman divergences (Bregman 1967)

These divergences are founded on another property of convex functions:
a (strictly) convex function is always greater than any tangent line, that is to the

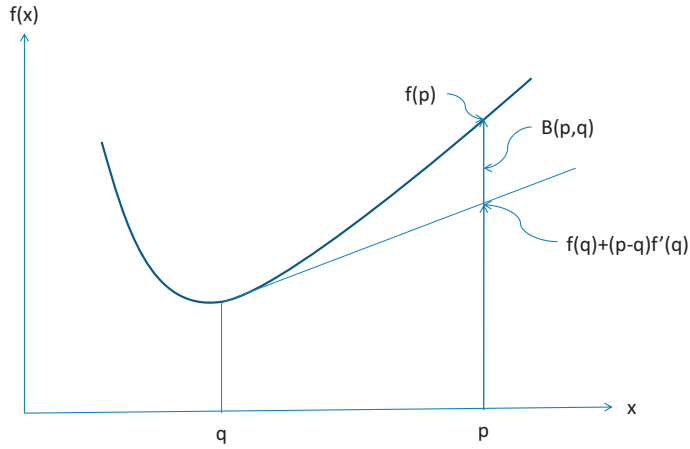


Fig. 4. First order Taylor expansion of a strictly convex function.

first order Taylor expansion of the function; this is represented in Figure 4 and expressed by the relation (3.6)

$$f(p) - f(q) - (p-q) f'(q) \geq 0. \tag{3.6}$$

The Bregman divergence is then:

$$B_f(p,q) = \sum_i \left\{ f(p_i) - f(q_i) - (p_i - q_i) f'(q_i) \right\}. \tag{3.7}$$

Note that this divergence is always convex w.r.t. p, but its convexity w.r.t. q depends on the function f.

This classification of divergences is artificial because it is founded on their constructive method only. A Jensen or Bregman divergence can also be a Csiszar divergence. Moreover, in this brief presentation, we do not consider the extensions or generalization of these divergences, but it is important to know that they exist and could be used as well. Then, at this point it is clear that there are many ways to quantify the discrepancy between two data fields; the question is then: how to choose a "good", that is a "significant" divergence or distance. A partial answer is given by the Maximum Likelihood principle.

4 Maximum likelihood solutions (Taupin 1988)

In this case, we take into account the statistical properties of the noise corrupting the data. We consider that we know the analytical expression of the likelihood that is of the conditional probability law $p(y/x)$, and we want to obtain the value of x corresponding to the maximum of this law.

In each pixel the noisy data y_i depends on the model $[m(a,x)]_i$ which is the mean value; moreover we assume that the noise realizations in the different pixels

are independent. In what follows, the internal parameters a of the model are known, so they are omitted in our notations, $[m(a,x)]_i = [m(x)]_i$, then we have:

$$p(y|m(x)) = \prod_i p(y_i|\{m(x)\}_i) \tag{4.1}$$

and

$$\max_x [p(y|m(x)] \equiv \min_x \left[-\ln \prod_i p(y_i|\{m(x)\}_i) \right]. \tag{4.2}$$

The solution x is obtained as:

$$x = \arg\min \sum_i -\ln [p(y_i|\{m(x)\}_i)]. \tag{4.3}$$

Two cases are generally exhibited in the literature corresponding to physical situations, the zero mean Gaussian additive noise and the Poisson process. We now examine these two cases and we show the relations with the divergences minimization problem.

4.1 Gaussian additive noise case

The likelihood is given by:

$$p(y|x) = p(y|m(x)) \approx \prod_i \exp - \frac{[y_i - \{m(x)\}_i]^2}{\sigma_i^2} \tag{4.4}$$

where σ_i^2 is the noise variance in the pixel i. This leads to an objective function which is the Euclidean distance between y and $m(x)$ in a space weighted by the variances:

$$J(x) = -\ln[p(y|m(x))] \approx \frac{1}{2} \sum_i \frac{[y_i - \{m(x)\}_i]^2}{\sigma_i^2}. \tag{4.5}$$

If the variance is not known or if the variance is identical for all the pixels, we obtain the pure Euclidean distance:

$$J(x) \approx \frac{1}{2} \sum_i [y_i - \{m(x)\}_i]^2. \tag{4.6}$$

One can observe that such a distance is defined for any value of x even if $m(x) \leq 0$.

4.2 Poisson noise case

The likelihood is given by:

$$p(y|x) = p(y|m(x)) = \prod_i \frac{[\{m(x)\}_i]^{y_i}}{y_i!} \exp[-\{m(x)\}_i] \tag{4.7}$$

$$J(x) = -\ln[p(y|m(x))] = \sum_i \{m(x)\}_i + \ln y_i! + y_i \ln \frac{y_i}{\{m(x)\}_i}. \tag{4.8}$$

Which is equivalent to:

$$J(x) = \sum_i \{m(x)\}_i - y_i + y_i \ln \frac{y_i}{\{m(x)\}_i}. \tag{4.9}$$

This expression is the Kullback-Leibler divergence (Kullback & Leibler 1951), adapted to data fields that are not necessarily probability laws. On the contrary to the Euclidean distance, one can note that the K.L. divergence is not defined if $m(x) \leq 0$. In the case of our linear model $x > 0 \Rightarrow m(x) = Hx > 0$. When the positivity constraint is required, the constraints domain is entirely contained in the domain of the objective function $J(x)$. Then if the solution is searched for in the constraints domain, the minimization can be performed. It is one of the reasons which leads to use an interior points algorithmic method. In such an iterative method, the successive estimates are feasibles solutions *i.e.* they fulfill all the constraints. We propose now a minimization method dealing with strictly convex differentiable functionnals, subject to a constraint on the inferior bound of the solution. The positivity constraint will appear as a particular case.

5 The Split Gradient Method (SGM)

This iterative method has been developped initially in the context of the deconvolution problem with non negativity constraint (Lanteri *et al.* 2001, 2002a,b). The multiplicative form of the algorithms is an obvious byproduct. It can be easily extended to regularized functionnals. The method is founded on the Karush-Kuhn-Tucker (KKT) conditions (Bertsekas 1995). We first recall these conditions. A simple example with only one variable clarifies this point.

5.1 Karush-Kuhn-Tucker conditions for inequality constraints

We denote $J_1(x)$, the "data consistency" term, $J_2(x)$, the "regularization" term and $\gamma \geq 0$, the regularization factor. The problem is to minimize w.r.t. x, the strictly convex differentiable functionnal: $J(x, \gamma) = J_1(x) + \gamma J_2(x)$.

The constraints are: $x_i \geq m_i \geq 0 \ \ \forall i \quad \Rightarrow \quad x_i - m_i \geq 0 \ \ \forall i$.

In what follows, the parameter γ will be omitted for sake of clarity.

Let us denote λ the Lagrange multiplier vector, and $\langle .,. \rangle$ the classical inner product.

The Lagrange function writes:

$$L(x, \lambda) = J(x) - \langle \lambda, (x - m \rangle). \tag{5.1}$$

The KKT conditions writes: at the solution (x^*, λ^*)

$$\nabla_x L(x^*, \lambda^*) = 0 \quad \Rightarrow \quad \lambda^* = \nabla_x J(x^*) \tag{5.2}$$

$$\lambda^* \geq 0 \quad \Rightarrow \quad \nabla_x J\left(x^*\right) \geq 0 \tag{5.3}$$

$$\left\langle \lambda^*, \left(x^* - m\right)\right\rangle = 0 \;\Rightarrow\; \left\langle \nabla_x J\left(x^*\right), \left(x^* - m\right)\right\rangle = 0. \tag{5.4}$$

This last condition must be understood as:

$$\left[\nabla_x J\left(x^*\right)\right]_i \left(x_i^* - m_i\right) = 0 \quad \forall i. \tag{5.5}$$

Indeed, because $x_i^* - m_i \geq 0\ \forall i$, and $\left[\nabla_x J\left(x^*\right)\right]_i \geq 0\ \forall i$, the inner product will be zero if and only if all the terms of the inner product are separately 0.

The Split Gradient Method is founded precisely on this condition.

The KKT conditions for inequality constraints can be understood easily on a simple example for a function of one variable $f(x)$ with an inferior bound constraint $x \geq m$.

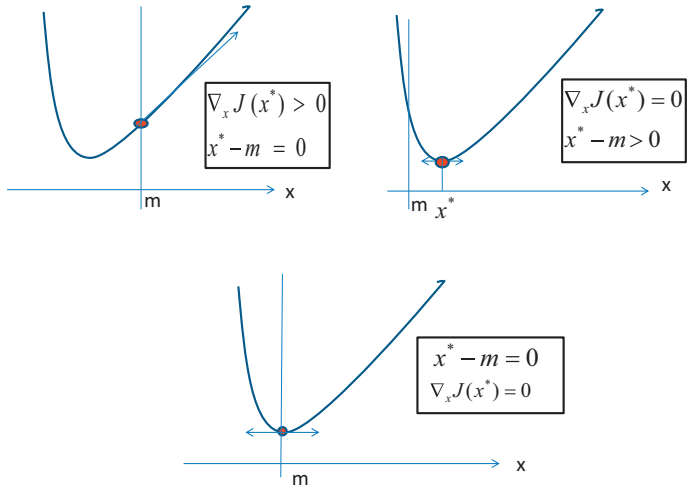


Fig. 5. Illustration of KKT conditions for non negativity constraint in the one dimensional case.

Let us consider the case represented in Figure 5 (upper right).

The minimum of the function is clearly reached for x such that $f'(x) = 0$. The solution is the same to the one of the unconstrained problem; in the case of a function of several variables, the solution will be reached when $\left[\nabla_x f\left(x\right)\right]_i = 0$, for the corresponding components i.

Then, for such indexes, $\left(x_i - m_i\right)\left[\nabla_x f\left(x\right)\right]_i = 0$ because $\left[\nabla_x f\left(x\right)\right]_i = 0$.

In Figure 5 (upper left), the solution is on the constraint $x - m = 0$. At this point, we have $\lambda = f'(x) > 0$. In a multi variables case the equivalent condition will be: if a component i is on the constraint $x_i - m_i = 0$, we will have $\lambda_i = \left[\nabla_x f\left(x\right)\right]_i > 0$, so that **for each component on the constraint**, we will have $\left(x_i - m_i\right)\left[\nabla f\left(x\right)\right]_i = 0$.

In Figure 5 (lower), the minimum of the function is exactly on the constraint, so that we have simultaneously $x - m = 0$ and $f^{'}(x) = 0$, then obviously, their product is zero. In a multi variables case the equivalent condition will be: if a component i is on the constraint $x_i - m_i = 0$, and if moreover for the same components, we have $\lambda_i = [\nabla_x f(x)]_i = 0$, for such components we will have $(x_i - m_i)[\nabla_x f(x)]_i = 0$.

Then for each component of the solution, the KKT condition expresses as: $(x_i - m_i)[\nabla_x f(x)]_i = 0$.

5.2 Principle of the Split Gradient Method

The problem is set as: let $\gamma \geq 0$ and y the noisy data, found

$$x = \arg\ \min\ J(x, \gamma) = J_1(x) + \gamma J_2(x). \tag{5.6}$$

Subject to the constraint

$$0 \leq m_i \leq x_i \quad \forall i. \tag{5.7}$$

Moreover, in the particular case $m_i = 0\ \forall i$, we will introduce a supplementary equality constraint:

$$\sum_i x_i = \sum_i y_i. \tag{5.8}$$

In a first step the equality constraint is not considered; it will be introduced later. Considering now that for convex differentiable functionnals such as $J(x, \gamma) \equiv J(x)$, the negative gradient is a descent direction, we want to solve w.r.t. x an equation of the form:

$$[-\nabla_x J(x^*)]_i (x_i^* - m_i) = 0 \quad \forall i. \tag{5.9}$$

Note that the multiplication of this equation by a positive term do not change solution.

Then, an iterative algorithm can be writen in the form:

$$x_i^{k+1} = x_i^k + \alpha_i^k \left(x_i^k - m_i\right) \left[-\nabla_x J\left(x^k\right)\right]_i. \tag{5.10}$$

In this algorithm, α_i^k is a positive descent step that must be computed to ensure the convergence of the algorithm. Moreover the form of the algorithm implies that at each iterative step, we must ensure that $x_i^k - m_i \geq 0$. This last point is of major importance in SGM.

The negative gradient is now written as the difference between two positive quantities $U\left(x^k\right)$ and $V\left(x^k\right)$:

$$-\nabla_x J\left(x^k\right) = U\left(x^k\right) - V\left(x^k\right). \tag{5.11}$$

Obviously, such a decomposition is not unique, indeed a constant term can be added and subtracted to the gradient, leading to shifted values of U and V, with the only condition that the shifted values remains positive.

We then propose to modify the algorithm as follows:

$$x_i^{k+1} = x_i^k + \alpha_i^k \left(x_i^k - m_i\right) \frac{1}{\left[V\left(x^k\right)\right]_i} \left[\underbrace{U\left(x^k\right) - V\left(x^k\right)}_{-\nabla J(x^k)}\right]_i \cdot \tag{5.12}$$

In the rest of the paper, we will use for sake of clarity, the notations: $\left[U\left(x^k\right)\right]_i = U_i^k$ and $\left[V\left(x^k\right)\right]_i = V_i^k$.

We can observe that the descent property is maintained even if the descent direction is changed.

The starting iterate will be $x_i^0 \geq m_i, \quad \forall i$.

The first step of the method is to compute for each component of the solution vector, the maximal step size ensuring $x_i^{k+1} \geq m_i \quad \forall i$, knowing that $x_i^k \geq m_i \quad \forall i$.

Obviously, such restriction on the step size is only necessary for the indexes i for which $\left[\nabla J\left(x\right)\right]_i \geq 0$.

This leads to:

$$\alpha_i^k \leq \frac{V_i^k}{V_i^k - U_i^k} \cdot \tag{5.13}$$

Then, at the iteration "k" the maximal step size allowing to fulfill the inferior bound constraint for all components, will be:

$$\alpha_{Max}^k = \min_i \left[\alpha_i^k\right]. \tag{5.14}$$

We note that $\alpha_{Max}^k \geq 1$.

As a consequence, with a stepsize equal to 1 the inferior bound constraint is always fulfilled. The proposed algorithm can then be written in matrix form:

$$x^{k+1} = x^k + \alpha_c^k \, diag\left[x_i^k - m_i\right] diag\left[\frac{1}{V_i^k}\right] \underbrace{\left(U^k - V^k\right)}_{-\nabla J^k} \cdot \tag{5.15}$$

It is a descent algorithm of scaled gradient type, that is of the general form:

$$x^{k+1} = x^k + \alpha_c^k d^k. \tag{5.16}$$

The descent direction is:

$$d^k = diag\left[x_i^k - m_i\right] diag\left[\frac{1}{V_i^k}\right] \underbrace{\left(U^k - V^k\right)}_{-\nabla J^k} \cdot \tag{5.17}$$

The descent stepsize α_c^k must be computed on the range $\left[0, \alpha_{Max}^k\right]$ to ensure the convergence of the algorithm. However, if we use a stepsize equal to 1 $\forall k$, we obtain a very attractive simple "quasi multiplicative" form, whose convergence is not demonstrated in full generality, but only in some specific cases:

$$x^{k+1} = m + diag\left[x^k - m\right] \frac{U^k}{V^k} \cdot \tag{5.18}$$

For a non negativity constraint ($m_i = 0\ \forall i$), the classical multiplicative form is immediately obtained:

$$x^{k+1} = diag\left[x^k\right] \frac{U^k}{V^k}. \tag{5.19}$$

In the two last equations, the ratio $\frac{U^k}{V^k}$ of the vectors U^k and V^k is performed component wise. With this simplified form, we can recover two classical algorithms: ISRA (Daube-Witherspoon *et al.* 1986) and RLA (Richardson 1972; Lucy 1974) corresponding respectively to the hypothesis of a Gaussian, zero mean additive noise, and to a Poisson noise process.

5.3 Examples with non negativity constraint

5.3.1 Gaussian additive noise case - Least squares

As previously indicated in Equation (4.6), the objective function writes:

$$J(x) = \frac{1}{2}\left\|y - Hx\right\|^2 \tag{5.20}$$

$$-\nabla J(x) = H^T y - H^T H x. \tag{5.21}$$

A decomposition can be:

$$U = H^T y; \quad V = H^T H x \tag{5.22}$$

Then the algorithm with non negativity constraint, in the non-relaxed form writes:

$$x_i^{k+1} = x_i^k \frac{\left(H^T y\right)_i}{\left(H^T H x^k\right)_i}. \tag{5.23}$$

This is the classical Image Space Reconstruction Algorithm (ISRA) whose convergence has been demonstrated by De Pierro (1987).

If some of the components of $V = H^T y$ is negative, we can add to all the components of U and V, the quantity $-\min\left(H^T y\right) + \epsilon$, so that the shifted values become positive.

5.3.2 Poisson noise case - Kullback-Leibler divergence

As previously indicated in Equation (4.9), the objective function writes:

$$J(x) = \sum_i y_i \ln \frac{y_i}{(Hx)_i} + (Hx)_i - y_i \tag{5.24}$$

$$-\nabla J(x) = H^T\left(\frac{y}{Hx} - \mathbf{1}\right). \tag{5.25}$$

In this equation the ratio of two vectors is performed component wise; the result of the operation is a vector. A decomposition can be:

$$U = H^T \frac{y}{Hx}; \quad V = H^T \mathbf{1}. \tag{5.26}$$

Then the algorithm with non negativity constraint, in the non-relaxed form writes:

$$x_i^{k+1} = x_i^k \frac{\left(H^T \frac{y}{Hx^k}\right)_i}{\left(H^T \mathbf{1}\right)_i} = x_i^k \frac{\left(H^T \frac{y}{Hx^k}\right)_i}{a_i}. \tag{5.27}$$

This is the classical E.M. (Dempster *et al.* 1977), Richardson-Lucy algorithm.

Some remarks then occur:

1. In the previous equation we have introduced the notation: $\left(H^T \mathbf{1}\right)_i = a_i$, howe-ver, in many cases for example in deconvolution problem with a convolution kernel normalized to "1", all the columns of H are of sum 1, that is $a_i = 1 \ \forall i$. Unfortunately, an oversimplified expression of the algorithm in which a_i is omitted, frequently appears; this can be a source of errors.

2. The algorithm of Richardson-Lucy with a kernel normalized to "1", have the "magic" and unwanted property to be flux conservative, that is $\sum_i x_i^k = \sum_i y_i \ \forall k$; this property does not exist with ISRA.

An interesting question is: why?

The answer lies in the particular expression of the K.L. divergence and in the associated properties.

5.4 Flux (intensity) conservation constraint (Lanteri et al. 2009, 2010)

We propose now to introduce a supplementary equality constraint in order to take into account the so called flux constraint or fixed sum constraint. While the method can be applied to the problem adressed in the previous section with a constant inferior bound constraint (which is typical of deconvolution problems), for sake of simplicity, we restrict the presentation to the case of a non negativity constraint.

The equality constraint writes:

$$\sum_i x_i = \sum_i y_i. \tag{5.28}$$

Moreover, because we want to remain in the class of interior points methods, such a constraint must be fulfilled at each iteration, that is:

$$\sum_i x_i^k = \sum_i y_i \quad \forall k. \tag{5.29}$$

The two previous relations expressing a sum constraint are typical of the deconvolution problem. In problems such as the linear unmixing one, we have to simply replace $\sum_i y_i$ by 1, without changing anything else in what follows.

The basic idea to take into account this constraint is to use the following procedure:

- Introduce the variable change:

$$x_i = \frac{u_i}{\sum_m u_m} \sum_m y_m. \tag{5.30}$$

- Proceed to a minimization w.r.t. the new variable u, subject to non negativity constraint only.
- Go back (correctly) to the initial variables x.

To minimize w.r.t. the new variable u, subject to non negativity constraint, we use the SGM previously described.

However, a fundamental question arises first: if $J(x)$ is convex w.r.t. x, did the function $\widetilde{J}(u)$ transformed function of $J(x)$ is still convex w.r.t. u?

The answer may be as follows: if during the iterative minimization process w.r.t. u, we are able to maintain $\sum_i u_i^k = Cst \ \ \forall k$, then the convexity w.r.t. u is ensured.

Moreover, we show that this property will allow us to go back "correctly" to the initial variables x.

To apply SGM, we compute the gradient of $\widetilde{J}(u)$ w.r.t. u

$$\frac{\partial \widetilde{J}(u)}{\partial u_j} = \sum_i \frac{\partial J}{\partial x_i} \frac{\partial x_i}{\partial u_j}. \tag{5.31}$$

We then obtain after some simple but tedious algebra:

$$-\frac{\partial \widetilde{J}(u)}{\partial u_j} \approx \left(-\frac{\partial J}{\partial x_j}\right) - \sum_i \frac{u_i}{\sum_m u_m} \left(-\frac{\partial J}{\partial x_i}\right). \tag{5.32}$$

We can now use SGM to minimize $\widetilde{J}(u)$ w.r.t. u with the non negativity constraint only, but we want also that $\sum_i u_i^{k+1} = \sum_i u_i^k \ \forall k$.

To reach such an objective, at first sight, we can choose:

$$\begin{aligned} U_j &= -\frac{\partial J}{\partial x_j} = \left(-\frac{\partial J}{\partial x}\right)_j \\ V_j &= \sum_i \frac{u_i}{\sum_m u_m} \left(-\frac{\partial J}{\partial x_i}\right) = \sum_i \frac{u_i}{\sum_m u_m} \left(-\frac{\partial J}{\partial x}\right)_i. \end{aligned} \tag{5.33}$$

However, with such a choice, we cannot ensure that U_j and V_j are positive.

To have this property which is necessary in S.G.M., we will choose:

$$U_j = \left(-\frac{\partial J}{\partial x}\right)_j - \min\left(-\frac{\partial J}{\partial x}\right) + \epsilon \tag{5.34}$$

$$V_j = \sum_i \frac{u_i}{\sum_m u_m} \left(-\frac{\partial J}{\partial x}\right)_i - \min\left(-\frac{\partial J}{\partial x}\right) + \epsilon \tag{5.35}$$

One can also write:

$$V_j = \sum_i \frac{u_i}{\sum_m u_m} \left[\left(-\frac{\partial J}{\partial x}\right)_i - \min\left(-\frac{\partial J}{\partial x}\right) + \epsilon\right]. \tag{5.36}$$

Obviously, the shift $-\min\left(-\frac{\partial J}{\partial x}\right) + \epsilon$ does not change the gradient, but now, we are sure that U_j and V_j are positive. Let us note that V_j is in fact constant and independent of the index j.

We can now apply SGM to obtain the relaxed algorithm:

$$u_j^{k+1} = u_j^k + \alpha^k u_j^k \left(\frac{\left(-\frac{\partial J}{\partial x^k}\right)_j - \min\left(-\frac{\partial J}{\partial x^k}\right) + \epsilon}{\sum_i \frac{u_i^k}{\sum_m u_m^k}\left[\left(-\frac{\partial J}{\partial x^k}\right)_i - \min\left(-\frac{\partial J}{\partial x^k}\right) + \epsilon\right]} - 1 \right). \tag{5.37}$$

The step size α^k is obviously computed as indicated in Section 5.2.

In the non relaxed case, that is , with $\alpha^k = 1\ \forall k$, we have:

$$u_j^{k+1} = u_j^k \frac{\left(-\frac{\partial J}{\partial x^k}\right)_j - \min\left(-\frac{\partial J}{\partial x^k}\right) + \epsilon}{\sum_i \frac{u_i^k}{\sum_m u_m^k}\left[\left(-\frac{\partial J}{\partial x^k}\right)_i - \min\left(-\frac{\partial J}{\partial x^k}\right) + \epsilon\right]}. \tag{5.38}$$

Clearly, with such form of the algorithm, relaxed or non-relaxed, we will have:

$$\sum_j u_j^{k+1} = \sum_j u_j^k. \tag{5.39}$$

Then, during the iterative process, the solution u^k is positive and remains in the convexity domain of the objective function $\widetilde{J}(u)$. Moreover the flux conservation property of the previous algorithms (5-37, 5-38) allows us to turn back "correctly" to the initial variables x. Indeed, multiplying the two members of these algorithms by $\frac{\sum_m y_m}{\sum_j u_j^{k+1}} = \frac{\sum_m y_m}{\sum_j u_j^k}$, and taking into account the change of variables (5-30), the final algorithm is obtained in the relaxed case as:

Let $x^0 = Cst \geq 0$ such that $\sum_i x_i^0 = \sum_i y_i$,

$$x_j^{k+1} = x_j^k + \alpha^k x_j^k \left(\frac{\left(-\frac{\partial J}{\partial x^k}\right)_j - \min\left(-\frac{\partial J}{\partial x^k}\right) + \epsilon}{\sum_i \frac{x_i^k}{\sum_m y_m}\left[\left(-\frac{\partial J}{\partial x^k}\right)_i - \min\left(-\frac{\partial J}{\partial x^k}\right) + \epsilon\right]} - 1 \right). \tag{5.40}$$

Let us observe that with such a relaxed algorithm, we obtain:

$$\sum_i x_i^{k+1} = (1 - \alpha^k) \sum_i x_i^k + \alpha^k \sum_i y_i. \tag{5.41}$$

So that, the flux conservation is related to the properties of the initial estimate, that is $\sum_i x_i^0 = \sum_i y_i$.

In the non relaxed case, that is, with $\alpha^k = 1 \forall k$, we obtain:

$$x_j^{k+1} = x_j^k \frac{\left(-\frac{\partial J}{\partial x^k}\right)_j - \min\left(-\frac{\partial J}{\partial x^k}\right) + \epsilon}{\sum_i x_i^k \left[\left(-\frac{\partial J}{\partial x^k}\right)_i - \min\left(-\frac{\partial J}{\partial x^k}\right) + \epsilon\right]} \sum_m y_m. \tag{5.42}$$

One can easily check that $x_i^{k+1} \geq 0$ if $x_i^k \geq 0 \ \ \forall k$, and that $\sum_i x_i^{k+1} = \sum_i y_i \ \forall k$ even if $\sum_i x_i^k \neq \sum_i y_i$.

This is basically different of the property of the non-relaxed algorithm.

Unfortunately, to our experience, such beautifull non relaxed algorithm does not converge, and the relaxed version must always be used. The corollary remark is that the only effective property concerning the flux constraint will be:

$$\sum_i x_i^k = \sum_i x_i^0. \tag{5.43}$$

All the algorithms founded on SGM are sometimes considered as having a slow convergence rate. In the relaxed form, the stepsize computation allows to ensure the convergence and moreover to (slightly) modify the convergence speed. Then we briefly indicate in the following section the general rules of the acceleration methods proposed in the literature.

6 Acceleration methods (Biggs *et al.* 1997; Nesterov 1983; Beck *et al.* 2010)

6.1 Principle of the method

Considering that we have a basis convergent algorithm analogous to (5.40), written in the form:

$$x^{k+1} = F(x^k). \tag{6.1}$$

Remember that in such an algorithm, the solution x^{k+1} is at each step non negative and of fixed sum if x^k is non negative and of fixed sum.

The general form of the acceleration methods proposed in the litterature could be summarized as follows:

1. Given the initial estimate x^0 fulfilling all the constraint, compute x^1 (which obviously fulfill all the constraints).

2. Knowing x^k and x^{k-1}, proceed to a linear extrapolation step to obtain the prediction $\widehat{x}^{k+1}$ as:

$$\widehat{x}^{k+1} = x^k + \delta^k \left(x^k - x^{k-1}\right) \tag{6.2}$$

   where the extrapolation step size δ^k is positive or zero $\forall k$.

   Two expressions allowing to obtain this stepsize are given in (Biggs *et al.* 1997; Nesterov 1983; Beck *et al.* 2010); however some supplementary restrictions on this stepsize are necessary as indicated in the comments.

3. Proceed to an iteration of the basis algorithm:

$$x^{k+1} = F(\widehat{x}^{k+1}). \tag{6.3}$$

6.2 Comments

All the difficulties are in the choice of the extrapolation step size, indeed:

- The extrapolated solution $\widehat{x}^{k+1}$ must be a non negative solution.

  Depending on the choice of δ^k, some components of $\widehat{x}^{k+1}$ can become negative, this is not allowed; if one think to project orthogonally $\widehat{x}^{k+1}$ on the space of non negative vectors, then, the flux constraint is not fulfilled; as a conclusion, the extrapolation step, must lead to $\widehat{x}^{k+1} \geq 0$. Then, due to the linearity of the extrapolation step, $\widehat{x}^{k+1}$ will fulfill the flux constraint.

  To fulfill the non negativity constraint on $\widehat{x}^{k+1}$, some restrictions of the extrapolation step size must be introduced.

- Even if such restrictions are taken into account, the algorithm can be non-monotonic, that is, the objective function can increase locally. This could be a source of problems.

  The solution generally proposed is simply to remove the extrapolation step when this happens.

- If the extrapolation is too strong, the accelerated algorithm may even diverge.

Then, clearly, the main problem is in the value of the extrapolation step size. Even if several methods are proposed in the literature to compute such a step size, as far we know, the convergence of accelerated algorithms is not clearly demonstrated and remain an open problem.

7 Conclusion

In the present work, we analyze mainly the inverse problems in which the overall effect of the physical system corresponds to a linear transformation of the input signal. The discrepancy between the experimental noisy data and the linear model must be quantified. Several classes of divergences or distances are then proposed as discrepancy functions. The problem is then to recover the unknown signal by minimization of the adequate divergence, subject to physical constraints.

The main point of this presentation is the Split Gradient Method. When this method has been elaborated, the objective was to recover, using classical optimization ideas, several algorithms that have been proposed in the field of image restoration or deconvolution. The main constraint introduced in these problems was the non negativity constraint. More generally such constraint has been extended to an inferior bound constraint.

In a second step, we have taken into account explicitely the flux conservation or the fixed sum constraint. The corresponding algorithms have been exhibited in the context of the SGM. These algorithms have been applied successfully in the fields of linear unmixing, NMF and deconvolution. Finally, the acceleration method of such algorithms is considered and briefly discussed at the end of the paper.

References

Bertero, M., 1989, Adv. Electr. Elect. Phys., 75, 1

Bertero, M., & Boccacci, P., 1998, Introduction to inverse problems in imaging (IOP Publishing)

Heinz, D.C., & Chang, C.I., 2001, IEEE. Trans. G.R.S, 39, 529

Lee, D.D., & Seung, H.S., 2000, NIPS

Cichocki, A., Zdunek, R., Phan, A.H., & Amari, S.I., 2009, Non negative matrix and tensor factorization (J. Wiley)

Andrews, H.C., & Hunt, B.R., 1977, Digital Image Restoration (Prentice Hall)

Demoment, G., 1989, IEEE Trans. ASSP, 12, 2024

Bertero, M., Lanteri, H., & Zanni, L., 2008, in Mathematical methods in Biomedical imaging and IMRT (Edizioni della normale, Pisa)

Ayers, G.R., & Dainty, J.C., 1988, Opt. Lett., 13, 428

Lane, R.G., 1992, J. Opt. Soc. Am. A, 9, 1508

Hadamard, J., 1923, Lectures on the Cauchy problem in linear partial differential equations (Yale University Press, New Haven)

Basseville, M., 1996, Information: entropies, divergences et moyennes, Technical Report, 1020, IRISA

Taneja, I.J., 2005, On mean divergences measures, Math. ST

Csiszar, I., 1991, Ann. Statist., 19, 2032

Burbea, J., & Rao, C.R., 1982, IEEE Trans. IT, 28, 489

Bregman, L.M., 1967, URSS Comput. Math. Math. Phys., 7, 200

Taupin, D., 1988, Probabilities, data reduction and error analysis in the physical sciences (Les Editions de Physique)

Kullback, S., & Leibler, R.A., 1951, Annals Math. Statistics, 22, 79

Lanteri, H., Roche, M., Cuevas, O., & Aime, C., 2001, Sig. Proc., 54, 945

Lanteri, H., Roche, M., & Aime, C., 2002, Inv. Probl., 18, 1397

Lanteri, H., Roche, M., Gaucherel, P., & Aime, C., 2002, Sig. Proc., 82, 1481

Bertsekas, D., 1995, Non Linear Programming (Athena Scientific)

Daube-Witherspoon, M.E., & Muehlehnner, 1986, IEEE Trans. Medical Imaging, 5, 61

Richardson, W.H., 1972, J. Opt. Soc. Am., 1, 55

Lucy, L.B., 1974, AJ, 79, 745

De Pierro, A.R., 1985, IEEE Trans. Medical Imaging, 6, 124

Dempster, A.D., Laird, N.M., & Rubin, D.B., 1977, J. Royal Stat. Soc. B, 39, 1

Lanteri, H., Theys, C., Benvenuto, F., & Mary, D., 2009, Gretsi

Lanteri, H., Theys, C., Fevotte, C., & Richard, C., 2010, Eusipco

Biggs, D.S.C., & Andrews, M., 1997, Appl. Optics, 36, 1766

Nesterov, Yu., E., 1983, Soviet Math. Dokl, 27, 372

Beck, A., & Teboulle, M., 2010, in Convex Optimization in Signal Processing and Communications, ed. D. Palomar & Y. Eldar (Cambridge University Press), 33

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 325–356*

SCALED GRADIENT PROJECTION METHODS FOR ASTRONOMICAL IMAGING

M. Bertero[1], P. Boccacci[1], M. Prato[2] and L. Zanni[2]

Abstract. We describe recently proposed algorithms, denoted *scaled gradient projection* (SGP) methods, which provide efficient and accurate reconstructions of astronomical images. We restrict the presentation to the case of data affected by Poisson noise and of nonnegative solutions; both maximum likelihood and Bayesian approaches are considered. Numerical results are presented for discussing the practical behaviour of the SGP methods.

1 Introduction

Image deconvolution is an important tool for reducing the effects of noise and blurring in astronomical imaging. In this paper we assume that blurring is described by a space invariant *point spread function* (PSF) and that a model of the PSF is available, accounting for both telescope diffraction and adaptive optics (AO) correction of the atmospheric blur. Therefore we do not consider topics such as space-variant deblurring or blind deconvolution.

Since the deconvolution problem is ill-posed, it should be formulated by using all the information available on the image formation process: not only the PSF is required but also a knowledge of the statistical properties of the noise affecting the data. These properties are used, for instance, for reformulating deconvolution as a *maximum likelihood* (ML) problem, which is also ill-posed in many instances (even if, presumably, with a lower degree of ill-posedness). Then prior information on the unknown astronomical target is required and this, if available, can be taken into account by extending the ML approach to a *Bayesian* approach. In both cases one reformulates deconvolution as a discrete variational problem and therefore the use of methods derived from numerical optimization becomes essential.

As concerns noise modeling, a crucial point is that astronomical images are typically detected by charged coupled device (CCD) cameras so that one can use,

[1] DIBRIS, Università di Genova, via Dodecaneso 45, 16146 Genova, Italy
[2] DISFIM, Università di Modena e Reggio Emilia, via Campi 213/b, 41125 Modena, Italy

DOI: 10.1051/eas/1359015

for instance, the accurate model described by Snyder *et al.* (1993). According to this model, if we denote by y_i the value of the image y detected at pixel i, then (after correction for flat field, bad pixels etc.) y_i is given by

$$y_i = y_i^{(\mathrm{obj})} + y_i^{(\mathrm{back})} + y_i^{(\mathrm{ron})}, \tag{1.1}$$

where $y_i^{(\mathrm{obj})}$ is the number of photoelectrons due to radiation from the object, $y_i^{(\mathrm{back})}$ is the number of photoelectrons due to internal and external background, dark current, etc., and $y_i^{(\mathrm{ron})}$ is the contribution of the *read-out noise* (RON) due to the amplifier. The first two terms are realizations of Poisson random variables (r.v.) while the third is a realization of an additive Gaussian r.v.. Therefore the noise affecting the data is a mixture of Poisson (due to photon counting) and additive Gaussian noise, due to RON. However, a refined model taking into account this particular structure of the noise does not provide significant improvement with respect to a simplified model also proposed by Snyder *et al.* (1993) (for a comparison see, for instance, Benvenuto *et al.* 2008, 2012). Indeed, Snyder *et al.* propose that, after the substitution $y_i \to y_i + \sigma^2$, where σ^2 is the variance of the RON, the RON can be treated as the realization of a Poisson r.v. with mean and variance being the same as σ^2. In this paper we use this approximation, which is quite accurate in the case of *near infrared* (NIR) observations, characterized by a large background emission. In conclusion we assume that the data y_i, shifted by σ^2, are realizations of suitable Poisson r.v.s.

In the framework of this model several iterative methods have been proposed for solving the ML or the Bayes problem. These methods are, in general, easy to implement but very slow: they require a large number of iterations, so that the computational cost can become excessive for the present and future large telescopes, able to acquire images of several mega-pixels so that the problem of image deconvolution in Astronomy becomes a large scale one.

An interesting property of some of the proposed algorithms is that they are first-order optimization methods using as a descent direction a suitable (diagonal) scaling of the negative gradient of the objective function. As a consequence, using these scalings, it is possible to apply a recently proposed approach denoted as *scaled gradient projection* (SGP) method and described in its general form by Bonettini *et al.* (2009). As shown by several numerical experiments this approach can provide a considerable speed-up of the standard algorithms.

In this paper SGP is not only considered for single-image deconvolution, the typical problem arising in the improvement of images provided by telescopes consisting of a monolithic mirror, but also for multiple-image deconvolution, a problem arising when different images of the same astronomical target are available. A significant application of this approach is the deconvolution of the images of the future Fizeau interferometer of the Large Binocular Telescope (LBT) called LINC-NIRVANA (Herbst *et al.* 2003).

LBT (`http://www.lbto.org`) is the world's largest optical and infrared telescope since it consists of two 8.4 m primary mirrors with the total light-gathering power of a single 11.8 m telescope. The two mirrors have an elevation over an

azimuth mounting and the elevation optical support structure moves on two large C-shaped rings (see Fig. 1). They are mounted with a 14.4 m centre separation, hence with an edge-to-edge distance of 22.8 m. This particular structure makes possible Fizeau interferometry, with a maximum baseline of 22.8 m, corresponding to a theoretical resolution of a 22.8 m mirror in the direction of the line joining the two centres.

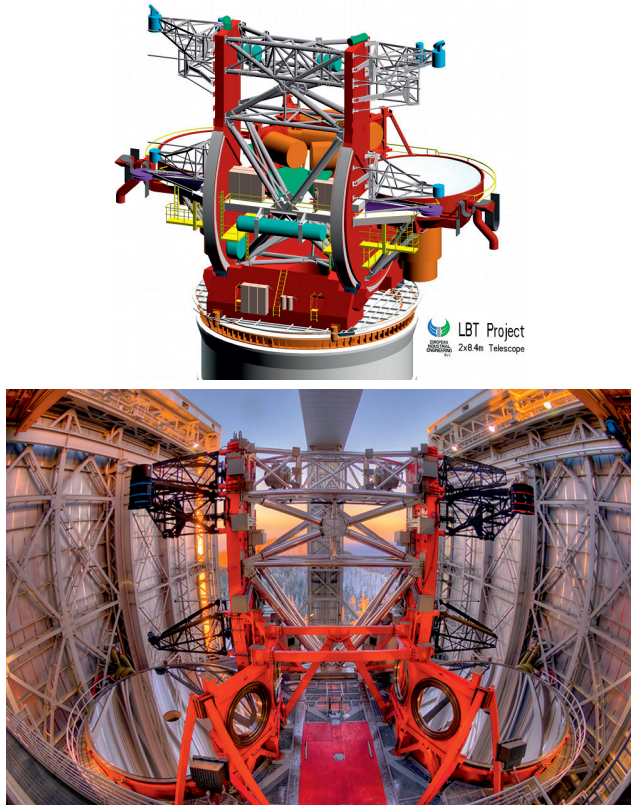


Fig. 1. A design view of LBT (*upper panel*), and a fish-eye image of the opposite side of LBT inside the enclosure (*lower panel*), as it appears to the visitors of the observatory (photo courtesy of W. Ruyopakam and the Large Binocular Telescope Observatory).

LINC-NIRVANA (LN for short) will operate as a true imager. Indeed, in the Fizeau mode, the two beams from the primary mirrors are combined in a common focal plane (not in the pupil plane as with essentially all the existing interferometers). LN is in an advanced realization phase by a consortium of German and Italian institutions, leaded by the Max Planck Institute for Astronomy in Heidelberg (`http://www.mpia.de/LINC/`). When completed, the instrument will be mounted in the centre of the platform of LBT (clearly visible in the lower panel of Fig. 1). It will be fully commissioned and available for scientific studies in 2014.

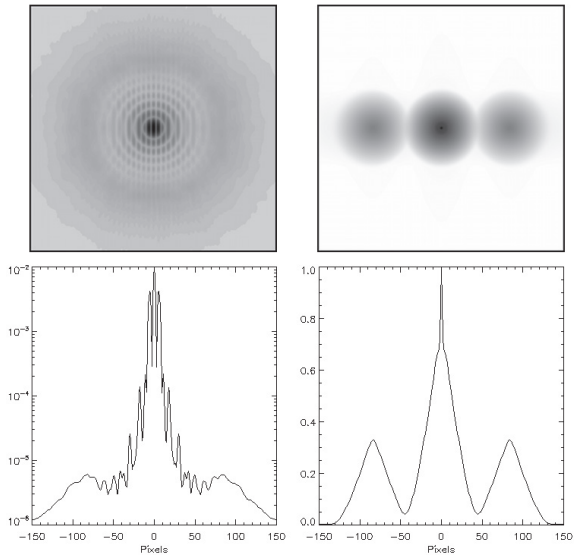


Fig. 2. Simulated PSF of LINC-NIRVANA with SR = 70% (*upper-left panel*), and the corresponding MTF (*upper-right panel*), both represented with reversed gray scale. The fringes are orthogonal to the baseline. In the lower panels we show the cut of the PSF along the baseline (*left*) and the cut of the MTF along the same direction (*right*).

In Figure 2 we show a simulated point spread function (PSF) with SR = 70%, together with the corresponding modular transfer function (MTF), *i.e.* the modulus of the Fourier transform of the PSF. This PSF, as well as others used in this paper, has been obtained with the code LOST (Arcidiacono *et al.* 2004). It is monochromatic ($\lambda = 2.2$ μm, *i.e.* K band), and, as clearly appears from this figure, it is the PSF of a 8.4 m telescope modulated by the interferometric fringes; accordingly the central disc of the MTF corresponds to the band of a 8.4 m mirror while the two side disks are replicas, due to interferometry, with a weaker intensity than the central one. These disks contain the precious additional information on the target due to interferometry.

As follows from this analysis, LN images will be characterized by an anisotropic resolution: that of a 22.8 m telescope in the direction of the baseline, and that of a 8.4 m in the orthogonal direction. Therefore, in order to get the maximum resolution in all directions, it will be necessary to acquire different images of the same astronomical target with different orientations of the baseline and to combine these images into a unique high-resolution image by means of suitable image reconstruction methods. In other words LN will routinely require multiple-image deconvolution.

The paper is organized as follows. In Section 2 we outline the mathematical model based on the approximation of the RON mentioned above and we describe the main algorithms introduced for solving the ML and the Bayesian problems both

for single and multiple-image deconvolution. Moreover we recall an approach, proposed in Bertero & Boccacci (2005), for boundary effect correction. In Section 3 we describe the algorithm SGP in the particular case of the nonnegativity constraint and the optimization of its parameters. In Section 4 we demonstrate its efficiency by several numerical experiments and finally in Section 5 we derive some conclusions.

2 Mathematical modeling

As outlined in the Introduction we assume that the value y_i of an astronomical image y detected at pixel i is the realization of a Poisson r.v. Y_i with unknown expected value λ_i. A further assumption is that the r.v.s. associated with different pixels are statistically independent. As a consequence their joint probability distribution is given by

$$P_Y(y|\lambda) = \prod_{i \in S} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, \tag{2.1}$$

the data being assumed to be integer numbers and S being the set of the index values.

In the case of a linear model for image formation, with the imaging system described by a space-invariant PSF, the unknown expected value is given by

$$\lambda_i = (Hx)_i + b_i, \quad Hx = K * x, \tag{2.2}$$

where: x is the unknown astronomical target; b the background emission, including the σ^2 term due to the RON; H the imaging matrix and K the PSF of the system satisfying the conditions

$$K_i \ \geq \ 0, \quad \sum_{i \in S} K_i = 1. \tag{2.3}$$

Assuming that b and K are known, the image restoration problem requires the development of methods for providing an estimate of x, given y.

2.1 Maximum likelihood approach

In the ML approach, given the detected image y, as well as b and K, one introduces the likelihood function, which is the function of x defined by

$$\mathcal{L}_y(x) = P_Y(y|Hx + b) \tag{2.4}$$

and obtained by inserting the image y and the model (2.2) in Equation (2.1). Then, a ML estimate of the unknown object is any image x^* which maximizes the likelihood function. However, since the likelihood is the product of a large number of functions, it is more convenient to take the negative logarithm of the likelihood

and minimize the resulting function. It is easy to see that, by rearranging terms independent of x, the negative logarithm of $\mathcal{L}_y(x)$ is given by

$$f_0(x;y) = \sum_{i \in S} \left\{ y_i \ln \frac{y_i}{(Hx+b)_i} + (Hx+b)_i - y_i \right\}, \tag{2.5}$$

which is the so-called generalized *Kullback-Leibler* (KL) *divergence* of the computed data $Hx + b$ from the detected data y. This function is nonnegative and is zero iff $Hx + b = y$; it is also convex and coercive, *i.e.* $f_0(x;y) \to +\infty$ if $||x||_2 \to +\infty$. The KL-divergence is not a metric distance, because it is not symmetric in the two terms and does not satisfy the triangle inequality. However it can be taken as a measure of the discrepancy between $Hx+b$ and y; it will be called the *data fidelity function*. The properties of $f_0(x;y)$ imply the existence of global minima of this function on the nonnegative orthant and therefore the existence of nonnegative ML estimates of the unknown. If all data are strictly positive and the imaging matrix is nonsingular, then $f_0(x;y)$ is strictly convex, a sufficient condition for the uniqueness of the solution.

As shown in Barrett & Meyers (2003), the nonnegative minimizers of $f_0(x;y)$ are sparse objects, *i.e.* they consist of bright spots over a black background (sometimes are called *star-night solutions*). Therefore they can be reliable solutions in the case of simple astronomical objects, such as binaries or open star clusters, but they are not in the case of more complex objects, such as nebulae, galaxies etc..

The standard algorithm for the minimization of $f_0(x;y)$ is the so-called *Richardson-Lucy* (RL) algorithm (Richardson 1972; Lucy 1974), defined by

$$x^{(k+1)} = x^{(k)} \cdot H^T \frac{y}{Hx^{(k)} + b}, \tag{2.6}$$

where the $\cdot$ denotes Hadamard product of two vectors and similarly the fraction symbol indicates component-wise division of two vectors. In the case $b = 0$ convergence of the iteration to the minimizers of $f_0(x;y)$ has been proved, but it is important to remark that the algorithm has also well-known "regularization" properties: in the case of complex objects sensible solutions can be obtained by a suitable early stopping of the iterations, even if this approach may not provide satisfactory results in some specific cases, for instance in the case of objects with sharp structures. Then a more refined regularization can be obtained by the use of prior information on the solution in a Bayesian framework (see, the next subsection).

An extension of the previous approach is required when different images of the same object are available. This problem, as discussed in the Introduction, is fundamental for the future Fizeau interferometer of LBT or for the "co-adding" method of images with different PSFs proposed by Lucy & Hook (1992).

Let p be the number of detected images $y^{(j)}$, j=1,..,p, with corresponding PSFs $K^{(j)}$, all normalized to unit volume, $H^{(j)}x = K^{(j)} * x$, and backgrounds $b^{(j)}$ (including the term σ^2 due to RON). It is quite natural to assume that the p images are statistically independent, so that the likelihood of the problem is the

product of the likelihoods associated to the different images. If we assume again Poisson statistics, and we take the negative logarithm of the likelihood, then the ML estimates are the minimizers of a data-fidelity function which is the sum of KL divergences, one for each image, *i.e.*

$$f_0(x;y) = \sum_{j=1}^{p} \sum_{i \in S} \left\{ y_i^{(j)} \ln \frac{y_i^{(j)}}{(H^{(j)}x + b^{(j)})_i} + (H^{(j)}x + b^{(j)})_i - y_i^{(j)} \right\}. \tag{2.7}$$

If we apply the standard expectation maximization method (Shepp & Vardi 1982) to this problem, we obtain the iterative algorithm

$$x^{(k+1)} = \frac{1}{p} x^{(k)} \cdot \sum_{j=1}^{p} (H^{(j)})^T \frac{y^{(j)}}{H^{(j)}x^{(k)} + b^{(j)}}, \tag{2.8}$$

which we call the *multiple-image* RL method (multiple RL, for short).

For the reconstruction of LN images an acceleration of this algorithm is proposed in Bertero & Boccacci (2000) by exploiting an analogy between the images of the interferometer and the projections in tomography. In this approach called OSEM (ordered subset expectation maximization; Hudson & Larkin 1994), the sum over the p images in Equation (2.8) is replaced by a cycle over the same images. To avoid oscillations of the reconstructions within the cycle, a preliminary step is the normalization of the different images to the same flux, if different integration times are used in the acquisition process. The method OSEM is summarized in Algorithm 1.

Algorithm 1 Ordered subset expectation maximization (OSEM) method

Choose the starting point $x^{(0)} > 0$.

For $k = 0, 1, 2, ...$ do the following steps:

Step 1. Set $h^{(0)} = x^{(k)}$;

Step 2. For $j = 1, ..., p$ compute

$$h^{(j)} = h^{(j-1)} \cdot (H^{(j)})^T \frac{y^{(j)}}{H^{(j)}h^{(j-1)} + b^{(j)}}; \tag{2.9}$$

Step 3. Set $x^{(k+1)} = h^{(p)}$.

End

As follows from practice and theoretical remarks, this approach reduces the number of iterations by a factor p. However, the computational cost of one multiple RL iteration is lower than that of one OSEM iteration: we need $3p + 1$ FFTs in the first case and $4p$ FFTs in the second. In conclusion, the increase in efficiency provided by OSEM is roughly given by $(3p + 1)/4$. When $p = 3$ (the number of images provided by the interferometer will presumably be small), the efficiency

is higher by a factor of 2.5, and a factor of 4.7 when $p = 6$. These results must be taken into account when evaluating the efficiency of SGP with respect to that of multiple RL. We can add that the convergence of SGP is proved while that of OSEM is not, even if it has always been verified in our numerical experiments.

2.2 Bayesian approach

As already remarked, the regularization of the ML estimates obtained by an early stopping of the previous algorithms may not be satisfactory in some cases. A more general kind of regularization can be obtained with the so-called Bayesian approach. In this approach one assumes that the unknown object is also a realization of a suitable r.v. X whose probability distribution expresses information available on its properties, such as smoothness, sharp details etc..

A frequently used probability distribution has the following form, which is typical in statistical mechanics

$$P_X(x) = \frac{1}{Z} e^{-\beta\ f_1(x)}, \tag{2.10}$$

where Z (also called the partition function) is a normalization constant, β is a hyper-parameter, playing the role of a regularization parameter in our application, and $f_1(x)$ is a potential function characterizing the known properties of the unknown object, called in the following *regularization function* or also *regularizer*. $P_X(x)$ is called the *prior*.

If the probability distribution $P_Y(y|Hx + b)$, obtained by combining Equations (2.1) and (2.2), is interpreted as the conditional probability of Y for a given value of X, then, from Bayes formulas we obtain that the conditional probability of X for a given value of Y is given by

$$P_X(x|y) = \frac{P_Y(y|Hx + b)P_X(x)}{P_Y(y)}, \tag{2.11}$$

where $P_Y(y)$ is the marginal probability distribution of Y.

If in this equation we insert the detected image y, we obtain a function of x which is called the *posterior probability* of x and is essentially the product of the likelihood and the prior (the value of the marginal distribution of Y computed in y is a constant which can be neglected). The maximizers of this function are the *maximum a posteriori* (MAP) estimates of the unknown object. By taking again the negative log of this function we find that they are the nonnegative minimizers of the function

$$f_\beta(x; y) = f_0(x; y) + \beta f_1(x), \tag{2.12}$$

where the second term is the negative log of the prior. If the function $f_1(x)$ is convex and nonnegative, then $f_\beta(x; y)$ is also convex and nonnegative; moreover it is also coercive, thanks to the coercivity of $f_0(x; y)$, so that MAP estimates of the unknown object exist. Given the regularizer, a crucial point in this approach is the choice of the regularization parameter β. This point will be briefly discussed in the following.

For our purposes an interesting algorithm for the minimization of $f_\beta(x;y)$ is the so-called *split-gradient method* (SGM) proposed by Lantéri *et al.* (2002), which consists in a simple modification of the RL algorithm. If $f_1(x)$ is differentiable and $U_1(x), V_1(x)$ is a pair of nonnegative functions such that

$$-\nabla_x f_1(x) = U_1(x) - V_1(x), \tag{2.13}$$

then the algorithm is as follows

$$x^{(k+1)} = \frac{x^{(k)}}{\hat{1} + \beta\ V_1(x^{(k)})} \cdot \left\{ H^T \frac{y}{Hx^{(k)} + b} + \beta\ U_1(x^{(k)}) \right\}, \tag{2.14}$$

where $\hat{1} = (1, \ldots, 1)^T$. The choice of the pair $U_1(x), V_1(x)$ is not unique but, for each one of the standard regularizers, one can find a quite natural choice (Lantéri *et al.* 2002). As concerns the extension to the case of multiple image deconvolution (Bertero *et al.* 2011), the updating rule of SGM becomes

$$x^{(k+1)} = \frac{x^{(k)}}{p\hat{1} + \beta V_1(x^{(k)})} \cdot \left\{ \sum_{j=1}^{p} (H^{(j)})^T \frac{y^{(j)}}{H^{(j)}x^{(k)} + b^{(j)}} + \beta U_1(x^{(k)}) \right\}, \tag{2.15}$$

while the OSEM algorithm, with regularization, is given by Algorithm 1 where Equation (2.9) is replaced by

$$h^{(j)} = \frac{h^{(j-1)}}{\hat{1} + \frac{\beta}{p} V_1(h^{(j-1)})} \cdot \left\{ (H^{(j)})^T \frac{y^{(j)}}{H^{(j)}h^{(j-1)} + b^{(j)}} + \frac{\beta}{p} U_1(h^{(j-1)}) \right\}. \tag{2.16}$$

2.3 Boundary effect corrections

If the target x is not completely contained in the image domain, then the previous deconvolution methods produce annoying boundary artifacts. It is not the purpose of this paper to discuss the different methods for solving this problem. We focus on an approach proposed in Bertero & Boccacci (2005) for single-image and in Anconelli *et al.* (2006) for multiple-image deconvolution. Here we present the approach in the case of multiple images (single image corresponds to $p = 1$).

The idea is to reconstruct the object x over a domain broader than that of the detected images and to merge, by zero padding, the arrays of the images and the object into arrays of dimensions that enable their Fourier transform to be computed effectively by means of FFT. We denote by $\bar{S}$ the set of values of the index labeling the pixels of the broader arrays containing S, and by R that of the object array contributing to S, so that $S \subset R \subset \bar{S}$. It is obvious that also the PSFs must be defined over $\bar{S}$ and that this can be done in different ways, depending on the specific problem one is considering. We point out that they must be normalized to unit volume over $\bar{S}$. We also note that R corresponds to the part of the object contributing to the detected images and that it depends on the extent of the PSFs. The reconstruction of x outside S is unreliable in most

cases, but its reconstruction inside S is practically free of boundary artifacts, as shown in the papers cited above and in the experiments of Section 4.

In order to estimate the reconstruction domain R we can proceed as follows. Let M_S be the characteristic function (mask) of S in $\bar{S}$, *i.e.* the array which is 1 inside S and 0 outside; moreover, let us introduce the following arrays, defined over $\bar{S}$, which appear in the computation of the gradient of $f_0(x;y)$ as defined below

$$\gamma^{(j)} = K_-^{(j)} * M_S \ , \ \gamma = \sum_{j=1}^{p} \gamma^{(j)}, \tag{2.17}$$

where $(K_-^{(j)})_i = (K^{(j)})_{-i}$. These arrays are essentially the images of M_S in $\bar{S}$ and are computable by FFT. Their extent outside S (they can be either very small or zero in pixels of $\bar{S}$ outside S) depends on the extent of the PSF and therefore they can be used for defining the reconstruction domain R. Given a thresholding value ϵ, we define R as follows

$$R = \{l \in \bar{S} \mid \gamma_l^{(j)} \geq \epsilon \ ; j = 1, .., p\}; \tag{2.18}$$

Next, if M_R is the characteristic function of R, we introduce the following matrices $H^{(j)}$ and $(H^{(j)})^T$

$$H^{(j)}x = M_S \cdot K^{(j)} * (M_R \cdot x), \tag{2.19}$$

$$(H^{(j)})^T h = M_R \cdot K_-^{(j)} * (M_S \cdot h), \tag{2.20}$$

where, in the second equation, h denotes a generic array defined over $\bar{S}$. Again, both matrices can be computed by means of FFT. We point out that, in the case of a regularization function containing the discrete gradient of x, it could be convenient to slightly modify the definition of M_R: not use exactly the characteristic function of R, but an array which is 1 over R and tends smoothly to 0 outside R (obtained, for instance, by convolving the characteristic function of R with a suitable Gaussian). In this way one can avoid discontinuities at the boundary of R in $\bar{S}$.

With the previous definitions, the data fidelity function is given again by Equation (2.7), with S replaced by $\bar{S}$ and the matrices $H^{(j)}$ defined as in the previous equation. Then the multiple RL algorithm, with regularization and boundary effect correction, is given by

$$x^{(k+1)} = \frac{M_R \cdot x^{(k)}}{\gamma + \beta V_1(x^{(k)})} \cdot \left\{ \sum_{j=1}^{p} (H^{(j)})^T \frac{y^{(j)}}{H^{(j)}x^{(k)} + b^{(j)}} + \beta U_1(x^{(k)}) \right\}, \tag{2.21}$$

the quotient being zero in the pixels outside R. Similarly, the OSEM algorithm, with regularization and boundary effect correction, is given by Algorithm 1 where Equation (2.9) is replaced by

$$h^{(j)} = \frac{M_R \cdot h^{(j-1)}}{\gamma^{(j)} + \frac{\beta}{p} V_1(h^{(j-1)})} \cdot \left\{ (H^{(j)})^T \frac{y^{(j)}}{H^{(j)}h^{(j-1)} + b^{(j)}} + \frac{\beta}{p} U_1(h^{(j-1)}) \right\}. \tag{2.22}$$

We stress again that the convergence of OSEM is not proved in the case of noisy data but that it has been always verified numerically in our applications to astronomical imaging.

3 The scaled gradient projection method

Let us consider, for generality, the case of multiple images with boundary effect correction and regularization. It is easy to verify that the gradient of $f_\beta(x;y)$, with x restricted to R, is given by

$$\nabla_x f_\beta(x;y) = M_R \cdot \left(\gamma - \sum_{j=1}^{p} (H^{(j)})^T \frac{y^{(j)}}{H^{(j)}x + b^{(j)}} \right) + \beta \nabla f_1(x), \tag{3.1}$$

where the definitions and notations introduced in the previous sections are used. If x is an admissible image, $x \geq 0$, then it is also easy to verify that, for each $\alpha \in (0,1]$ the image

$$x_\alpha = x - \alpha \frac{x}{\gamma + \beta\ V_1(x)} \nabla_x f_\beta(x;y), \tag{3.2}$$

where $V_1(x)$ is the array related to the gradient of $f_1(x)$ (see Eq. (2.13)), is also an admissible image. If we do the substitutions $x_\alpha = x^{(k+1)}$, $x = x^{(k)}$ and $\alpha = 1$, we re-obtain the algorithm of Equation (2.21).

Since all the algorithms in the previous section can be obtained as particular cases of this one, we can conclude that all these algorithms are scaled gradient method, with a descent direction which is also feasible just thanks to the scaling of the gradient which has been introduced. This property may suggest that all the scalings previously considered may be very useful for designing efficient first order methods and this is just what is obtained thanks to the SGP method proposed in Bonettini *et al.* (2009).

3.1 The algorithm

In many astronomical applications both ML and Bayes problems are particular cases of the following general convex optimization problem

$$\min \quad f(x), \quad \text{sub.to} \quad x \geq 0, \tag{3.3}$$

where f is a continuously differentiable, nonnegative, convex and coercive function. In the following we denote as P_+ the projection onto the nonnegative orthant, *i.e.* the operator setting to zero the negative component of a vector. Moreover, we introduce the set $\mathcal{D}$ of the diagonal positive definite matrices, whose diagonal elements have values between L_1 and L_2, for given thresholds $0 < L_1 < L_2$. Then the general SGP can be stated as in Algorithm 2.

In practice, at iteration k, given the step-length α_k and the scaling matrix $D_k \in \mathcal{D}$, a descent direction $d^{(k)}$ is obtained as difference between the projection of the

Algorithm 2 Scaled gradient projection (SGP) method

Choose the starting point $x^{(0)} \geq 0$ and set the parameters $\eta, \theta \in (0,1)$, $0 < \alpha_{min} < \alpha_{max}$.

FOR $k = 0, 1, 2, \ldots$ DO THE FOLLOWING STEPS:

STEP 1. Choose the parameter $\alpha_k \in [\alpha_{min}, \alpha_{max}]$ and the scaling matrix $D_k \in \mathcal{D}$;

STEP 2. Projection:
$$z^{(k)} = P_+(x^{(k)} - \alpha_k D_k \nabla f(x^{(k)}));$$

STEP 3. Descent direction: $d^{(k)} = z^{(k)} - x^{(k)}$;

STEP 4. Set $\lambda_k = 1$;

STEP 5. Backtracking loop:
let $f_{new} = f(x^{(k)} + \lambda_k d^{(k)})$;
IF
$$f_{new} \leq f(x^{(k)}) + \eta \lambda_k \nabla f(x^{(k)})^T d^{(k)}$$
THEN
go to step 6;
ELSE
set $\lambda_k = \theta \lambda_k$ and go to step 5.
ENDIF

STEP 6. Set $x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$.

END

vector $x^{(k)} - \alpha_k D_k \nabla f(x^{(k)})$ and the current iteration $x^{(k)}$. The descent direction is then used to define the new approximation $x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$, where the line-search parameter λ_k is defined by a standard Armijo line-search procedure that ensures the monotone reduction of the objective function at each iteration. The global convergence can be obtained by following Birgin *et al.* (2000, 2003) and Bonettini *et al.* (2009), where the more general case based on non-monotone line-search procedures is also considered. We emphasize that any choice of the step-length $\alpha_k \in [\alpha_{min}, \alpha_{max}]$ and the scaling matrix $D_k \in \mathcal{D}$ are allowed; this freedom of choice can then be fruitfully exploited for introducing performance improvements, as discussed in the next section.

3.2 Scaling matrix and step-length

The choice of the scaling matrix has to be addressed with the goal of improving the convergence rate of the image reconstruction process while avoiding to increase excessively the computational cost of the single iteration. In the case of twice continuously differentiable objective function, a possible choice is to use a diagonal scaling matrix whose nontrivial elements approximate the diagonal entries of the

inverse of the Hessian matrix $\nabla^2 f(x)$, for example by choosing

$$(D_k)_{ii} = \min\left\{L_2, \max\left\{L_1, \left(\left(\nabla^2 f(x^{(k)})\right)_{ii}\right)^{-1}\right\}\right\}. \tag{3.4}$$

However, since the computation of the diagonal entries of the Hessian might represent an expensive task, the commonly used choice for the scaling matrix is the one suggested by the RL algorithm and its regularized versions, namely

$$D_k = \mathrm{diag}\left(\min\left\{L_2, \max\left\{L_1, \frac{x^{(k)}}{\gamma + \eta V_1(x^{(k)})}\right\}\right\}\right), \tag{3.5}$$

where only the indexes in R are considered in the case of boundary effect correction. In several applications of SGP to image deblurring the above scaling matrix has been shown to be very successful in accelerating the approximation of suited reconstructions, in comparison with gradient projection based approaches that avoid the use of scaling matrices (Bonettini *et al.* 2009, 2012).

As concerns the step-length parameter, an effective selection strategy is obtained by adapting to the context of the scaled gradient projection methods the two Barzilai & Borwein (1988) rules (hereafter denoted by BB), which are widely used in standard non-scaled gradient methods for unconstrained minimization problems. For the non-scaled case, the recent literature suggests effective alternation strategies of two BB step-length updating rules, derived by a careful analysis of their properties in the case of unconstrained minimization of quadratic functions. In particular, their ability in approximating the eigenvalues of the objective Hessian is exploited to design adaptive alternation strategies able to improve significantly the convergence rate of the gradient scheme (Zhou *et al.* 2006; Frassoldati *et al.* 2008). Numerical evidence is available that confirms the efficiency of these alternated BB rules also in case of nonlinear constrained minimization problems (Serafini *et al.* 2005; Loris *et al.* 2009).

When the scaled direction $D_k \nabla f(x^{(k)})$ is exploited within a step of the form $x^{(k)} - \alpha_k D_k \nabla f(x^{(k)})$, the standard BB step-length rules can be generalized as follows:

$$\alpha_k^{(BB1)} = \frac{(s^{(k-1)})^T D_k^{-1} D_k^{-1} s^{(k-1)}}{(s^{(k-1)})^T D_k^{-1} t^{(k-1)}}, \tag{3.6}$$

$$\alpha_k^{(BB2)} = \frac{(s^{(k-1)})^T D_k t^{(k-1)}}{(t^{(k-1)})^T D_k D_k t^{(k-1)}}, \tag{3.7}$$

where $s^{(k-1)} = x^{(k)} - x^{(k-1)}$ and $t^{(k-1)} = \nabla f(x^{(k)}) - \nabla f(x^{(k-1)})$; when $D_k = I$ the above formulas lead to the standard BB rules.

In SGP, the values produced by these rules are constrained into the interval $[\alpha_{min}, \alpha_{max}]$ in the following way:

IF $(s^{(k-1)})^T D_k^{-1} t^{(k-1)} \leq 0$ THEN
  $\alpha_k^{(1)} = \min\{10 \cdot \alpha_{k-1},\ \alpha_{max}\}$;
ELSE

$$\alpha_k^{(1)} = \min\left\{\alpha_{max},\ \max\left\{\alpha_{min},\ \alpha_k^{(BB1)}\right\}\right\};$$

ENDIF

IF $(s^{(k-1)})^T D_k t^{(k-1)} \leq 0$ THEN

$$\alpha_k^{(2)} = \min\left\{10 \cdot \alpha_{k-1},\ \alpha_{max}\right\};$$

ELSE

$$\alpha_k^{(2)} = \min\left\{\alpha_{max},\ \max\left\{\alpha_{min},\ \alpha_k^{(BB2)}\right\}\right\};$$

ENDIF

The criterion adopted in SGP for alternating between the above step-lengths is derived from that proposed in Frassoldati *et al.* (2008) and can be stated as follows:

IF $\alpha_k^{(2)}/\alpha_k^{(1)} \leq \tau_k$ THEN

$$\alpha_k = \min_{j=\max\{1,k+1-M_\alpha\},\ldots,k} \alpha_j^{(2)}; \tag{3.8}$$

$$\tau_{k+1} = 0.9 \cdot \tau_k;$$

ELSE

$$\alpha_k = \alpha_k^{(1)}; \qquad \tau_{k+1} = 1.1 \cdot \tau_k;$$

ENDIF

where M_α is a prefixed positive integer and $\tau_1 \in (0, 1)$. In contrast to the criterion proposed in Frassoldati *et al.* (2008), that is thought for the non-scaled case ($D_k = I$) and uses a constant threshold $\tau_k = \tau \in (0, 1)$ in the switching condition, here a variable threshold is exploited with the aim of avoiding the selection of the same rule for a too large number of iterations. A wide computational study suggests that this alternation criterion is more suitable in terms of convergence rate than the strategy proposed by Zhou *et al.* (2006) and the use of a single BB rule (Bonettini *et al.* 2009; Favati *et al.* 2010; Zanella *et al.* 2009). Furthermore, in our experience, the use of the BB values provided by Equation (3.8) (that are generally lower than those provided by $\alpha_k^{(1)}$) in the first iterations slightly improves the reconstruction accuracy and, consequently, in the proposed SGP version we start the step-length alternation only after the first 20 iterations.

3.3 Choice of the parameters and implementation

Even if the number of SGP parameters is certainly higher than those of the RL and OSEM approaches, the huge amount of tests carried out in several applications has led to an optimization of these values, which allows the user to have at his disposal a robust approach without the need of an expensive problem-dependent parameter tuning. In the following we provide some comments on each of these parameters:

- $x^{(0)}$: although any array can be used as starting point of the algorithm, the two commonly used images are either the detected one (or one of the detected images in the case of multiple deconvolution) or a constant image with pixel values equal to the background-subtracted flux (or mean flux in the case of multiple deconvolution) of the noisy data divided by the number

of pixels. If the boundary effect correction is considered, only the pixels in the object array R become equal to this constant, while the remaining values of $\bar{S}$ are set to zero. Our experience showed no clear preference of the former choice with respect to the latter, that is typically used in the standard RL approach;

- η, θ: the sufficient decrease parameter η and the step-reduction parameter θ control, respectively, the severity of the objective function decrease condition and the number of backtracking reductions. The parameter η has been set to 10^{-4} as usually done in literature (see, for example, Birgin *et al.* 2000), while the value $\theta = 0.4$ resulted to be a good compromise to get a sufficiently large step size calculated with a low number of reductions;

- $\alpha_{min}, \alpha_{max}, \alpha_0$: the bounds $\alpha_{min}, \alpha_{max}$ of the step-length parameter α_k are safeguard values that have to be considered for the algorithm to ensure the theoretical convergence. Usually, a very large range $(\alpha_{min}, \alpha_{max})$ is exploited in combination with BB-like step-length selections (Birgin *et al.* 2000, set such values to 10^{-30} and 10^{30}); we found that the interval $(10^{-5}, 10^{5})$ is suited both for working with the rules (3.6)-(3.7) and for avoiding extreme step-length values. As far as the starting parameter α_0 concerns, the value 1.3 has been chosen to have an initial step slightly longer than the RL one;

- initial value for τ_k: as previously observed, the switching condition between the step-length (3.8) and the value $\alpha_k^{(1)}$ works after the first 20 iterations and we choose the value 0.5 as first value for the switching parameter τ_k. In our experience, in the considered imaging applications, the values provided by Equation (3.7) are generally lower than those given by (3.6) and the starting value chosen for τ_k seems well suited to activate the alternation between the two step-length rules (remember that in the non-scaled case, if $(s^{(k-1)})^T t^{(k-1)} > 0$, the inequality $\alpha_k^{(BB2)} \leq \alpha_k^{(BB1)}$ holds).

- M_α: in case of non-scaled gradient schemes for unconstrained quadratic minimization, the use of the minimum of the step-lengths $\alpha_{k-j}^{(BB2)}, j = 0, \ldots, M_\alpha$ increased the ability of the first BB rule to approximate, in the subsequent iterations, the inverse of the Hessian's smallest eigenvalues, with interesting convergence rate improvements (Frassoldati *et al.* 2008). In Bonettini *et al.* (2009), by using the setting $M_\alpha = 3$, the importance of this trick is numerically confirmed also on more general minimization problems and in case of scaled gradient projection methods; for this reason we adopted the same setting also for our SGP version.

- L_1, L_2: while in the original paper of Bonettini *et al.* (2009) the choice of the bounds (L_1, L_2) for the scaling matrices was a couple of fixed values $(10^{-10}, 10^{10})$, independent of the data, we prefer to make automatically these bounds suitable for images of any scale. In details, one step of the RL method is performed and the parameters (L_1, L_2) are tuned according to

the min/max positive values $y_{\min}/y_{\max}$ of the resulting image; moreover, for avoiding too close bounds, the following rule is implemented

IF $y_{\max}/y_{\min} < 50$ THEN
  $L_1 = y_{\min}/10$;
  $L_2 = y_{\max} \cdot 10$;
ELSE
  $L_1 = y_{\min}$;
  $L_2 = y_{\max}$;
ENDIF

The above parameter settings are at the basis of the SGP versions currently available for ML deconvolution of astronomical images, which we briefly describe.

- IDL implementation: an Interactive Data Language (IDL) package for the single and multiple deconvolution of 2D images corrupted by Poisson noise, with the optional inclusion of the boundary effect correction.
- IDL-GPU implementation: an extended version of the above IDL implementation able to exploit the resources available on Graphics Processing Units (GPUs). This SGP version is obtained by means of the CUDA (Compute Unified Device Architecture) technology, developed by NVIDIA for programming their GPUs. The CUDA framework is available within an IDL implementation through the GPUlib, a software library developed by Tech-X Corporation, that enables GPU-accelerated computation.
- Matlab implementation: a Matlab package for the deconvolution of 2D and 3D images through the minimization of the function (2.5) and the early stopping of the iterations.

These implementations and the relative documentation can be downloaded from the URL `http://www.unife.it/prin/software`. A complete C++ and C++/CUDA library collecting all the described SGP versions is in progress and will be soon available by request.

4 Numerical experiments

The application of SGP to ML problems described in Section 3 is presented, discussed and illustrated with several numerical examples in Prato *et al.* (2012). In this section we show the SGP behaviour by discussing a few of the numerical experiments presented in Prato *et al.* (2012) as well as a numerical experiment of regularized deconvolution described in Staglianò *et al.* (2011).

In the case of methods for ML problems a crucial point is the choice of the number of iterations, *i.e.* the introduction of sensible stopping rules providing sensible solutions. On the other hand, in the case of regularization methods, the crucial point is the choice of the regularization parameter. We first discuss the stopping of the iterative methods for ML reconstructions.

In the case of the reconstruction of stellar objects such as binaries, clusters etc., SGP can be pushed to convergence; in other words, iteration can be stopped when the following condition is satisfied

$$|f_0(x^{(k)}; y) - f_0(x^{(k-1)}; y)| \;\leq\; tol\; f_0(x^{(k-1)}; y), \tag{4.1}$$

where *tol* is a parameter selected by the user (in most cases we use $tol = 10^{-7}$, but a larger value can be selected to reduce the number of iterations if a poorer accuracy of the result is sufficient). We remark that the application of this criterion does not require an additional cost because $f_0(x^{(k)}; y)$ is already computed within the algorithm.

In the case of early stopping the choice of the stopping rule is a difficult task. In numerical simulations the reference object is known, let us denote it as $\tilde{x}$, and therefore at each iteration one can compute (with a small additional cost) some "distance" between $\tilde{x}$ and $x^{(k)}$. A frequently used indicator is the relative r.m.s. error defined by

$$\rho^{(k)} = \frac{||x^{(k)} - \tilde{x}||_2}{||\tilde{x}||_2}, \tag{4.2}$$

or other indicators in terms of ℓ_1-norm, KL divergence etc.. Iterations can be stopped when $\rho^{(k)}$ reaches a minimum value, thus defining a reconstruction which is "optimal" according to this criterion.

Obviously such a strategy can not be applied in the case of real data. In the vein of a discrepancy principle used for Tikhonov regularization, one can introduce the following quantity, which must be computed at each iteration and can be called a "discrepancy function"

$$D_y^{(k)} = \frac{1}{\#S} \left\| \frac{Hx^{(k)} + b - y}{\sqrt{Hx^{(k)} + b}} \right\|^2. \tag{4.3}$$

It is derived from Bardsley & Goldes (2009) while in Staglianò *et al.* (2011) it is shown that this quantity is a decreasing function of k; moreover, in the latter paper, it is proposed, on the basis of statistical considerations, that iterations could be stopped when $D_y^{(k)} \leq 1$. Another criterion, also based on a statistical analysis, is proposed in Bertero *et al.* (2010). In this case the "discrepancy function" is defined by

$$D_y^{(k)} = \frac{2}{\#S} f_0(x^{(k)}; y), \tag{4.4}$$

and its computation does not require any additional cost. It is proved that it is a decreasing function of k and again iterations can be stopped when $D_y^{(k)} \leq 1$. Examples of the application of this criterion are given in Bertero *et al.* (2010).

In the case of regularized solutions, for a given value of the regularization parameter, the iterations must be pushed to convergence using, for instance, a criterion similar to (4.1), with $f_0(x^{(k)}; y)$ replaced by $f_\beta(x^{(k)}; y)$. The problem is to select a value of β. Again, one must use different strategies in the case of simulated and real data.

In the first case, if we denote as x^*_β the minimizer of $f_\beta(x;y)$ (in practice, its approximation computed by means of an iterative method), then one can introduce again a relative r.m.s. error using a "distance" between x^*_β and $\tilde{x}$, for instance in terms of the ℓ_2-norm,

$$\rho(\beta) = \frac{||x^*_\beta - \tilde{x}||_2}{||\tilde{x}||_2}, \tag{4.5}$$

(or another indicator) and searching for the value of β minimizing this quantity. This approach obviously requires the computation of x^*_β for several values of β and can be computationally expensive.

In the case of real data one can use the discrepancy function introduced by Bardsley & Goldes (2009)

$$D_y(\beta) = \frac{1}{\#S} \left\| \frac{Hx^*_\beta + b - y}{\sqrt{Hx^*_\beta + b}} \right\|^2, \tag{4.6}$$

or that introduced by Bertero *et al.* (2010)

$$D_y(\beta) = \frac{2}{\#S} f_0(x^*_\beta; y). \tag{4.7}$$

In both cases one must search for the value of β satisfying the equation $D(\beta) = 1$. A secant-like method can be used for solving this equation; if a tolerance 10^{-3} is used, in general only 4-5 iterations are required. This approach can be useful also in the case of simulations because the value of β minimizing the error (4.5) can be searched in a neighborhood of the value provided by the discrepancy principle.

4.1 Acceleration of the RL method

In this section we show the effectiveness of SGP with respect to the RL and OSEM approaches, highlighting the speedups achievable thanks to both the algorithmic acceleration provided by SGP and the parallel implementation of the codes on GPU. We consider 256×256 HST images of the planetary nebula NGC 7027 and the galaxy NGC 6946, with two different integrated magnitudes (m) of 10 and 15, not corresponding to the effective magnitudes of these objects but introduced for obtaining simulated images with different noise levels. Such images have been convolved with an ideal PSF, simulated assuming a telescope of diameter 8.25 m, a wavelength of 2.2 μm, and a pixel size of 50 mas. A constant background term of about 13.5 mag arcsec^{-2}, corresponding to observations in K-band, is added and the resulting images are perturbed with Poisson noise and additive Gaussian noise with $\sigma = 10$ e$^-$/px. Original objects and the corresponding blurred and noisy images are shown in Figure 3. As suggested in Snyder *et al.* (1994), compensation for RON is obtained in the deconvolution algorithms by adding the constant $\sigma^2 = 100$ to the images and the background. We obtained test problems of larger size (up to 2048×2048) by means of a Fourier-based rebinning, preserving

the same background and the same noise level. The results are reported in Tables 1 and 2, where we highlight both the speedup observed between GPU and serial implementations (labeled "Par") and the one provided by the use of SGP instead of RL (labeled "Alg").

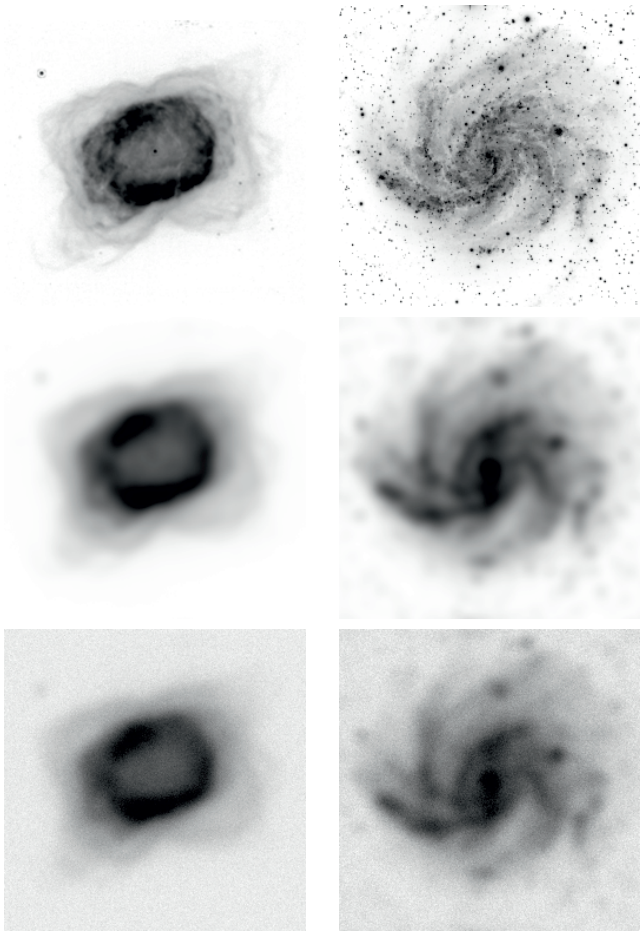


Fig. 3. Original images (*top panels*) and blurred noisy images with $m = 10$ (*middle panels*) and $m = 15$ (*bottom panels*).

As concerns the multiple-image deconvolution problem, we test the efficiency of multiple RL, OSEM, and SGP (applied to multiple RL), by means of synthetic images of LN. In particular, we simulate a model of an open star cluster based on an image of the Pleiades, by selecting the nine brightest stars characterized by the following name, position and magnitude.

Table 1. Relative r.m.s. errors, computational times, and speedups obtained by the accelerating features of SGP with respect to RL ("Alg") and by the GPU implementations ("Par"), for the nebula NGC 7027 with different image sizes. Iterations are stopped at a minimum relative r.m.s. error in the serial algorithms.

$m = 10$					
Algorithm	Size	Err	Sec	SpUp (Par)	SpUp (Alg)
RL	256^2	0.051	783.9	-	-
It = 10000*	512^2	0.051	4527	-	-
	1024^2	0.051	17610	-	-
	2048^2	0.051	80026	-	-
RL_CUDA	256^2	0.051	35.63	22.0	-
It = 10000*	512^2	0.051	69.77	64.9	-
	1024^2	0.051	149.5	118	-
	2048^2	0.051	469.1	171	-
SGP	256^2	0.052	26.14	-	30.0
It = 272	512^2	0.051	143.6	-	31.5
	1024^2	0.051	554.0	-	31.8
	2048^2	0.051	2493	-	32.1
SGP_CUDA	256^2	0.052	1.797	14.5	19.8
It = 272	512^2	0.052	3.469	41.4	20.1
	1024^2	0.052	8.016	69.1	18.7
	2048^2	0.052	25.66	97.2	18.3
$m = 15$					
Algorithm	Size	Err	Sec	SpUp (Par)	SpUp (Alg)
RL	256^2	0.068	48.27	-	-
It = 612	512^2	0.064	278.7	-	-
	1024^2	0.062	1068	-	-
	2048^2	0.062	4897	-	-
RL_CUDA	256^2	0.068	2.219	21.8	-
It = 612	512^2	0.064	4.109	67.8	-
	1024^2	0.062	9.250	115	-
	2048^2	0.062	29.13	168	-
SGP	256^2	0.068	3.016	-	16.0
It = 31	512^2	0.064	16.95	-	16.4
	1024^2	0.062	65.22	-	16.4
	2048^2	0.061	290.8	-	16.8
SGP_CUDA	256^2	0.068	0.218	13.8	10.2
It = 31	512^2	0.064	0.421	40.3	9.76
	1024^2	0.062	1.063	61.4	8.70
	2048^2	0.061	3.406	85.4	8.55

Table 2. Relative r.m.s. errors, computational times, and speedups obtained by the accelerating features of SGP with respect to RL ("Alg") and by the GPU implementations ("Par"), for the galaxy NGC 6946 with different image sizes. Iterations are stopped at a minimum relative r.m.s. error in the serial algorithms.

$m = 10$					
Algorithm	Size	Err	Sec	SpUp (Par)	SpUp (Alg)
RL	256^2	0.293	786.0	-	-
It = 10000*	512^2	0.293	4545	-	-
	1024^2	0.293	17402	-	-
	2048^2	0.293	80022	-	-
RL_CUDA	256^2	0.293	36.64	21.5	-
It = 10000*	512^2	0.293	67.94	66.9	-
	1024^2	0.293	146.7	119	-
	2048^2	0.293	463.9	172	-
SGP	256^2	0.292	88.72	-	8.86
It = 928	512^2	0.291	484.3	-	9.38
	1024^2	0.291	1854	-	9.19
	2048^2	0.291	8386	-	9.54
SGP_CUDA	256^2	0.293	7.219	12.3	5.08
It = 928	512^2	0.293	11.14	43.5	6.10
	1024^2	0.293	25.86	71.7	5.67
	2048^2	0.293	81.02	104	5.73
$m = 15$					
Algorithm	Size	Err	Sec	SpUp (Par)	SpUp (Alg)
RL	256^2	0.311	114.9	-	-
It = 1461	512^2	0.307	644.3	-	-
	1024^2	0.306	2574	-	-
	2048^2	0.306	11689	-	-
RL_CUDA	256^2	0.311	5.375	21.4	-
It = 1461	512^2	0.307	9.656	66.7	-
	1024^2	0.306	22.41	115	-
	2048^2	0.306	68.44	171	-
SGP	256^2	0.311	3.672	-	31.3
It = 38	512^2	0.308	20.36	-	31.6
	1024^2	0.307	78.20	-	32.9
	2048^2	0.306	354.0	-	33.0
SGP_CUDA	256^2	0.311	0.266	13.8	20.2
It = 38	512^2	0.307	0.531	38.3	18.2
	1024^2	0.307	1.344	58.2	16.7
	2048^2	0.306	4.188	84.5	16.3

Star Name	X	Y	m
ALCYONE	228	246	12.86
ATLAS	156	237	13.62
ELECTRA	340	247	13.70
MAIA	299	295	13.86
MEROPE	277	216	14.17
TAYGETA	326	313	14.29
PLEIONE	155	253	15.09
CELAENO	343	280	15.44
ASTEROPE	296	330	15.64

The coordinate values are deduced from the picture found in the Wikipedia page (`http://en.wikipedia.org/wiki/Pleiades`), resized to a 256×256 pixels image, and immersed in a 512×512 pixels image. In this way we generated a relatively compact cluster in the center of the image. These objects are convolved with three PSFs corresponding to three equispaced orientations of the baseline, $0°$, $60°$, and $120°$, obtained by rotating the PSF described in the Introduction and shown in Figure 2. Background emission in K band (13.5 mag/arcsec^2) is added to the results, which are also perturbed with Poisson and Gaussian ($\sigma = 10$ e$^-$/px) noise. The object and one of the corresponding blurred and noisy images are shown in Figure 4.

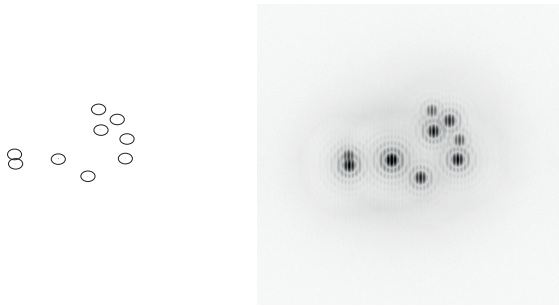


Fig. 4. Star cluster data: simulated object (*left panel*, stars are marked by circles) and corresponding blurred and noisy image (*right panel*).

In this case, iterations are pushed to convergence and therefore the stopping rule is given by the condition (4.1). We use different values of *tol*, specifically 10^{-3}, 10^{-5}, and 10^{-7}. In order to measure the quality of the reconstruction, we introduce an average relative error of the magnitudes defined by

$$\text{av\_rel\_er} = \frac{1}{q} \sum_{j=1}^{q} \frac{|m_j - \widetilde{m}_j|}{\widetilde{m}_j}, \tag{4.8}$$

where q is the number of stars (in our case $q = 9$) and $\widetilde{m}_j$ and m_j are respectively the true and the reconstructed magnitudes. The results are reported in Table 3.

Table 3. Reconstruction of the star cluster with three 512×512 equispaced images. The error is the average relative error in the magnitudes defined in Equation (4.8).

	tol = 1e-3			
Algorithm	It	Err	Sec	SpUp
RL	319	2.39e-4	393.4	-
RL_CUDA	319	2.38e-4	4.641	84.8
OSEM	151	1.63e-4	220.8	-
OSEM_CUDA	151	1.62e-4	2.421	91.2
SGP	71	1.35e-3	97.80	-
SGP_CUDA	71	1.29e-3	1.641	59.6
	tol = 1e-5			
Algorithm	It	Err	Sec	SpUp
RL	1385	6.65e-5	1703	-
RL_CUDA	1385	6.64e-5	19.38	87.9
OSEM	675	5.64e-5	980.6	-
OSEM_CUDA	675	5.64e-5	10.75	91.2
SGP	337	5.89e-4	455.2	-
SGP_CUDA	337	1.79e-4	7.187	63.3
	tol = 1e-7			
Algorithm	It	Err	Sec	SpUp
RL	7472	5.64e-5	9180	-
RL_CUDA	7472	5.98e-5	104.8	87.6
OSEM	3750	6.13e-5	5442	-
OSEM_CUDA	3750	5.98e-5	59.52	91.4
SGP	572	7.37e-5	772.6	-
SGP_CUDA	572	7.05e-5	12.20	63.3

4.2 Boundary effect correction

We show now the effectiveness of the boundary effect correction described in Section 2.3 on the RL, OSEM and SGP algorithms. The numerical experiments are designed according to the following procedure: we select a 256×256 HST image of the Crab nebula NGC 19521, and we build the blurred and noisy image by means of the same procedure (and the same parameters) adopted in the previous tests, but using the AO-corrected PSF[3] shown in Figure 5.

The parameters of this PSF (pixel size, diameter of the telescope, etc.) are not provided. However, it has approximately the same width as the ideal PSF described in Section 4.1. We apply RL and SGP first to the full image, and then to four 160×160 partly overlapping sub-domains with the addition of the boundary effect correction. The full deconvolved image is obtained as a mosaic of the central parts (see Fig. 6). The same comparison is performed in the multiple-image case

[3]Downloaded from `http://www.mathcs.emory.edu/~nagy/RestoreTools/index.html`

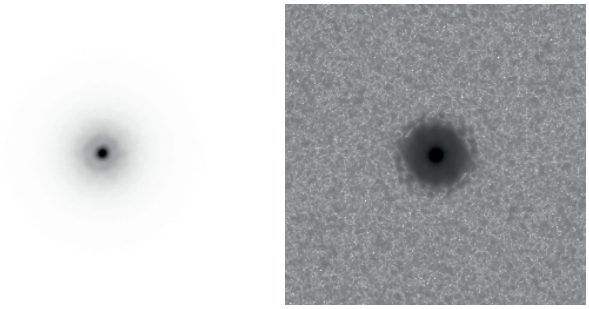


Fig. 5. The PSF used for the single deconvolution experiments with boundary effect correction (*left panel*) and the corresponding MTF (*right panel*). Both are represented in reversed gray scale.

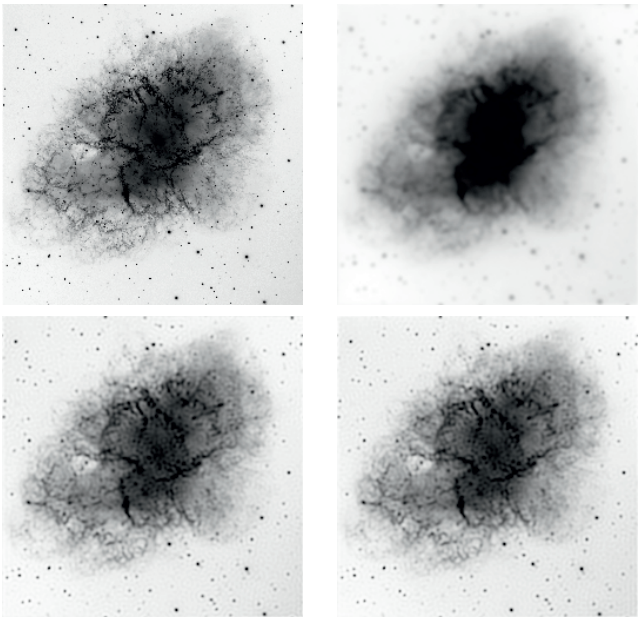


Fig. 6. Crab nebula test: the object (*top left*), its blurred and noisy image in the case $m = 10$ (*top right*), the reconstructions of the full image with SGP (*bottom left*) and as a mosaic of four reconstructions of partially overlapping subdomains, using SGP with boundary effect correction (*bottom right*).

by using three 512×512 images of the nebula NGC 7027 obtained by means of the LN PSFs described in the previous section (in this test, 320×320 sub-domains are extracted). In Tables 4 and 5 we report the serial and parallel performances of RL, OSEM (when multiple images were available) and SGP in both cases of

Table 4. Reconstruction of the 256×256 Crab object with a standard deconvolution and as a mosaic of the reconstructions of four subimages with boundary effect correction.

	Standard deconvolution				Boundary effects correction			
				$m = 10$				
Algorithm	It	Err	Sec	SpUp	It	Err	Sec	SpUp
RL	5353	0.128	419.8	-	4070	0.129	1146	-
RL_CUDA	5353	0.128	19.45	21.6	4070	0.129	61.55	18.6
SGP	151	0.129	14.28	-	129	0.129	46.42	-
SGP_CUDA	151	0.129	1.219	11.7	129	0.133	4.342	10.7
				$m = 12$				
Algorithm	It	Err	Sec	SpUp	It	Err	Sec	SpUp
RL	954	0.136	74.83	-	696	0.137	196.5	-
RL_CUDA	954	0.136	3.516	21.3	696	0.137	10.99	17.9
SGP	52	0.137	4.984	-	53	0.137	19.41	-
SGP_CUDA	52	0.137	0.406	12.3	53	0.137	1.922	10.1
				$m = 15$				
Algorithm	It	Err	Sec	SpUp	It	Err	Sec	SpUp
RL	128	0.172	10.09	-	99	0.172	28.08	-
RL_CUDA	128	0.172	0.483	20.9	99	0.172	1.704	16.5
SGP	10	0.172	1.093	-	9	0.172	3.859	-
SGP_CUDA	10	0.172	0.093	11.8	9	0.172	0.360	10.7

Table 5. Reconstruction of the nebula using three equispaced 512×512 images, in the cases of standard deconvolution and as a mosaic of four reconstructed subimages with boundary effect correction.

	Standard deconvolution				Boundary effects correction			
				$m = 10$				
Algorithm	It	Err	Sec	SpUp	It	Err	Sec	SpUp
RL	3401	0.032	4364	-	2899	0.034	13978	-
RL_CUDA	3401	0.032	48.00	90.9	2899	0.034	174.2	80.2
OSEM	1133	0.032	1602	-	950	0.034	5447	-
OSEM_CUDA	1133	0.032	18.59	86.2	950	0.034	64.03	85.1
SGP	144	0.033	220.7	-	160	0.034	873.3	-
SGP_CUDA	144	0.033	3.563	61.9	160	0.034	15.45	56.5
	Standard deconvolution				Boundary effects correction			
				$m = 15$				
Algorithm	It	Err	Sec	SpUp	It	Err	Sec	SpUp
RL	353	0.091	441.5	-	243	0.094	1174	-
RL_CUDA	353	0.091	4.937	89.4	243	0.094	15.28	76.8
OSEM	117	0.091	165.7	-	81	0.094	479.1	-
OSEM_CUDA	117	0.091	2.062	80.4	81	0.094	5.939	80.7
SGP	16	0.087	26.14	-	11	0.087	69.88	-
SGP_CUDA	16	0.087	0.546	47.9	11	0.086	1.532	45.6

full and splitted deconvolution. The computational times reported in the case of boundary effect correction refer to the reconstruction of all the four sub-domains.

4.3 Edge-preserving regularization

As an example of regularized reconstruction we consider the case of an edge-preserving prior, called hypersurface (HS) regularization (Charbonnier *et al.* 1997). It is defined by

$$f_1(x) = \sum_{j_1,j_2=1}^{n} \psi_\delta(D^2_{j_1,j_2}), \quad \delta \neq 0, \tag{4.9}$$

where

$$\psi_\delta(t) = \sqrt{t+\delta^2}, \quad D^2_{j_1,j_2} = (x_{j_1+1,j_2} - x_{j_1,j_2})^2 + (x_{j_1,j_2+1} - x_{j_1,j_2})^2. \tag{4.10}$$

For δ small this regularization is used as a smoothed approximation to total variation (TV) (see, for instance, Vogel 2002; Bardsley & Luttman 2009; Zanella *et al.* 2009; Defrise *et al.* 2011; Staglianò *et al.* 2011; for TV regularization, see Dey *et al.* 2006; Le *et al.* 2007; Brune *et al.* 2010; Setzer *et al.* 2010; Bonettini & Ruggiero 2011).

By computing the gradient of $f_1(x)$ one finds the following natural choice for the function $V_1(x)$ to be inserted in the scaling of the gradient of the complete objective function (see Eq. (3.5))

$$[V_1(x)]_{j_1,j_2} = [2\psi'_\delta(D^2_{j_1,j_2}) + \psi'_\delta(D^2_{j_1,j_2-1}) + \psi'_\delta(D^2_{j_1-1,j_2})], \tag{4.11}$$

where $\psi'_\delta(t)$ is the derivative of $\psi_\delta(t)$.

We consider as reference object the frequently used spacecraft image characterized by sharp details (Fig. 7). The size is 256×256 (in Fig. 7 we show only the central part), and the maximum value is 255; it is superimposed to a background $b = 1$. Moreover, for generating images with different noise levels, we consider three other versions with maximum values 2550, 25 500 and 25 5000, respectively (and backgrounds 10, 100, 1000), obtained by scaling the original object. Next, the four versions are convolved with a PSF and then perturbed with Poisson noise (we did not add Gaussian noise). The PSF used is the one already described in the previous section and shown in Figure 5. For each image we generate 25 different realizations of noise so that we have a total of 200 noisy images.

We first consider unregularized reconstructions. Early stopping of the iteration is based on two stopping rules. The first consists in computing at each iteration the relative r.m.s. error $\rho^{(k)}$, defined in Equation (4.2) and stopping the iteration when this parameter reaches its minimum value. The second consists in computing the discrepancy $D^{(k)}$, introduced by Bardsley & Goldes (2009), defined in Equation (4.3), and stopping the iteration when it crosses 1. Iteration is initialized with $x^{(0)} = y_{am} - b$ (where y_{am} is the arithmetic mean of the image values). In all cases, $D^{(0)} > 1$ and $D^{(k)}$ is decreasing for increasing k, providing a solution of the equation $D^{(k)} = 1$.

The results are given in Table 6 for the four images of the spacecraft with different noise levels. For each image we report average value and standard deviation both of the number of iterations and of the reconstruction error, computed using

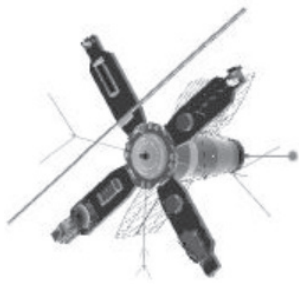


Fig. 7. The spacecraft image, represented in reversed gray scale.

the 25 realizations of noise. The reconstruction error is weakly dependent on the noise realization even if the number of iterations is strongly varying. Stopping based on Bardsley & Goldes criterion works better at the highest noise level.

Table 6. Unregularized reconstructions of the spacecraft: errors and iterations.

	Minimum error		**Discrepancy**	
	iter	**error (%)**	**iter**	**error (%)**
255	73 ± 19	40.1 ± 0.6	33 ± 14	43.9 ± 9.6
2550	186 ± 58	33.5 ± 0.4	117 ± 84	30.9 ± 11.5
25 500	465 ± 198	29.3 ± 0.3	593 ± 322	30.0 ± 1.1
255 000	1449 ± 376	26.9 ± 0.2	1788 ± 553	27.3 ± 0.5

In column (a) of Figure 8 we show the four images with different noise levels; in columns (b) and (c) the reconstructions corresponding to the minimum r.m.s. error and to the criterion of Bardsley & Goldes, respectively; finally, in the last column, we show the normalized residuals defined by

$$R^{(k)} = \frac{Hx^{(k)} + b - y}{\sqrt{Hx^{(k)} + b}}, \tag{4.12}$$

and computed in the case of the reconstructions of column (b). Artifacts are present at the lowest noise levels, due to the reconstruction method.

The previous numerical test is performed for investigating possible improvements of the reconstructions due to the use of edge-preserving regularization, as provided by the penalty function of Equation (4.9), with $\delta = 10^{-4}$, and we use the SGP algorithm with the scaling defined in terms of the function (4.11). This scaling has been already successfully used in the case of denoising of Poisson data (Zanella *et al.* 2009) and we use the same parameters of the algorithm described in that paper. For a given β, iteration is stopped when $|f_\beta(x^k; y) - f_\beta(x^{k-1}; y)| \leq 10^{-7} f_\beta(x^{k-1}; y)$. The choice of β is performed by computing x^*_β and using a secant-like method for satisfying the criterion of Bardsley & Goldes, with a tolerance of 10^{-3}. Next, the value of β providing the minimum r.m.s. error is obtained by

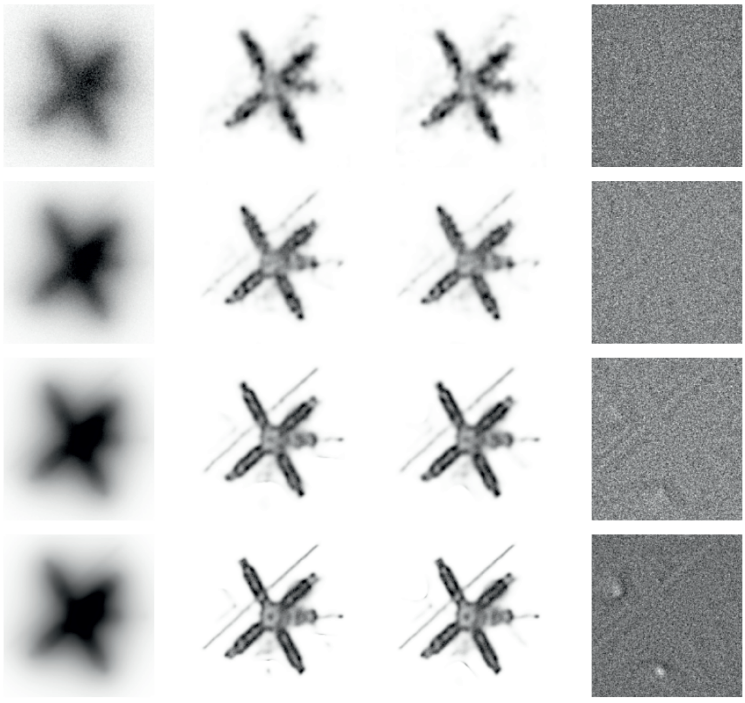


Fig. 8. Unregularized reconstructions of the spacecraft: (a) the blurred images; (b) the reconstructions with minimum r.m.s. error; (c) the reconstructions satisfying the criterion of Bardsey & Goldes; (d) the normalized residuals in the case of the reconstructions of column (b).

searching in an interval around the value provided by the discrepancy equation. Also in this experiment we considered 25 different realization of noise for each test image.

The reconstruction errors and the number of required iterations are reported in Table 7. The average reconstruction errors are smaller than those obtained in the unregularized case, with comparable standard deviations. As concerns the use the discrepancy criterion, it provides acceptable results except at the highest noise level. The reconstructions and the normalized residuals are shown in Figure 9. The residuals are still affected by strong artifacts, at least in the case of the lowest noise levels.

5 Concluding remarks and perspectives

We briefly discuss the main points of this paper by considering first the case of the ML problems.

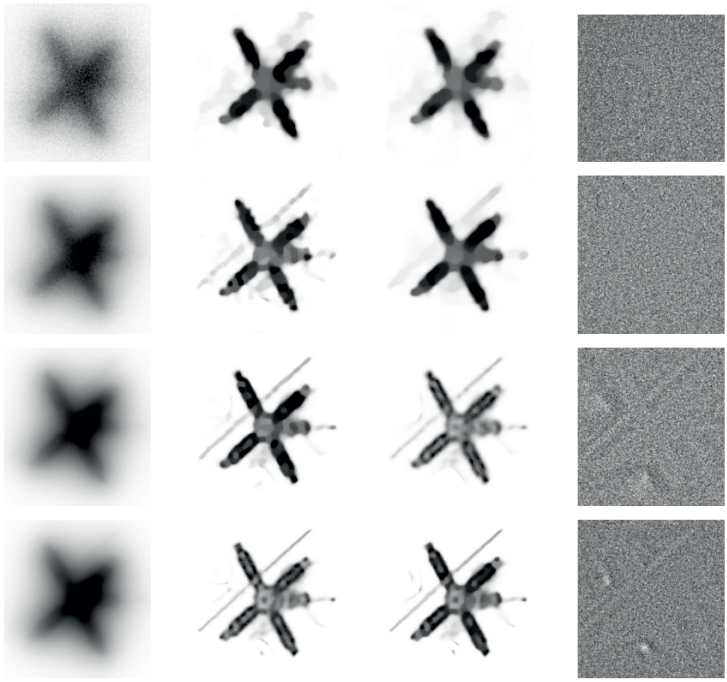


Fig. 9. Regularized reconstructions of the spacecraft: (a) the blurred images; (b) the reconstruction with the minimum r.m.s. error; (c) the reconstructions satisfying the criterion of Bardsley & Goldes; (d) the normalized residuals in the case of the reconstructions of column (b).

Table 7. Regularized reconstructions of the spacecraft: iterations and errors.

	Minimum error		**Discrepancy**	
	iter	**error (%)**	**iter**	**error (%)**
255	247 ± 54	36.4 ± 0.6	367 ± 198	40.7 ± 4.5
2550	458 ± 138	30.7 ± 0.3	462 ± 210	32.2 ± 1.0
25 500	1308 ± 124	26.1 ± 0.2	933 ± 221	26.9 ± 0.7
255 000	2190 ± 409	24.3 ± 0.8	1700 ± 462	24.9 ± 1.0

Both RL and SGP (with the scaling suggested by RL) converge to minimizers of the data fidelity function defined in terms of the generalized KL divergence, in particular to the unique minimizer if the function is strictly convex. In the case of the reconstruction of binaries or star clusters, the algorithms must be pushed to convergence and, of course, they provide the same result. However, as follows for instance from Table 4, the convergence of SGP is much faster than that of RL with a speed-up increasing from 4 to about 12 for the serial implementation, and from 3 to 9 for the parallel implementation, if the required accuracy is increased.

On the other hand, in the case of complex objects such as nebulae, galaxies or similar, it is well known that an early stopping of the iterations is required in the case of RL. Indeed the algorithm has the so-called *semi-convergence* property, in the sense that the iterations first approach the true object (we are talking about simulations) and then go away. Therefore it is interesting to remark that the iterations of SGP have a similar behaviour. The trajectories formed by the iterations of the two algorithms are different even when the starting point is the same, but the two points of minimal distance from the true object are very close (in general, visually indistinguishable), and SGP reaches the point with a number of steps much smaller than RL. The gain in computational time is considerable in spite of the fact that the cost of one SGP iteration is about 30% higher than that of one RL iteration (Bonettini *et al.* 2009). If we look at Tables 1 and 2, we find a speed-up ranging from 10 to 30 in the serial implementation, and from 6 to 20 in the parallel implementation. The speed-up depends on the specific object and, in general, it is higher when a higher number of iterations is required. We conclude these brief remarks by pointing out that, in the case of faint objects, SGP implemented on GPU is able to process a 2048×2048 image in a few seconds.

In the case of a Bayesian approach we do not still have estimates of the speed-up provided by SGP algorithms (with the scaling suggested by SGM) with respect to other algorithms and, in particular, SGM (with or without line-search in terms, for instance, of Armijo rule). In this paper we give only a few preliminary results obtained in the case of SGP deconvolution with edge-preserving regularization. The speed-up provided by GPU implementation of SGP edge-preserving denoising of Poisson data is estimated in Serafini *et al.* (2010) (see also Ruggiero *et al.* 2010, for GPU implementation of SGP deconvolution without regularization). A speed-up of the order of 20 is observed.

We expect that also in the case of regularized deconvolution SGP can provide very fast algorithms, reducing the computational time required for the estimation of the value of the regularization parameter with one of the methods described in Section 4 or other proposed methods. These topics are under investigation by our group. The goal is to provide a library of algorithms for different regularization functions.

We conclude by remarking that the SGP approach has been already applied to other problems, in particular to the computation of nonnegative least-square solutions (Benvenuto *et al.* 2010), to the nonnegative reconstruction of astronomical data from sparse Fourier data (Bonettini & Prato 2010) and to the least-squares problem with a sparsity regularization (Loris *et al.* 2009).

References

Anconelli, B., Bertero, M., Boccacci, P., Carbillet, M., & Lanteri, H., 2006, A&A, 448, 1217

Arcidiacono, C., Diolaiti, E., Tordi, M., Ragazzoni, R., Farinato, J., Vernet, E., & Marchetti, E., 2004, Appl. Opt., 43, 4288

Bardsley, J.M., & Goldes, J., 2009, Inverse Probl., 25, 095005

Bardsley, J.M., & Luttman, A., 2009, Adv. Comput. Math., 31, 35

Barrett, H.H., & Meyers, K.J., 2003, Foundations of Image Science (Wiley and Sons, New York), 1047

Barzilai, J., & Borwein, J.M., 1988, IMA J. Numer. Anal., 8, 141

Benvenuto, F., La Camera, A., Theys, C., Ferrari, A., Lantéri, H., & Bertero, M., 2008, Inverse Probl., 24, 035016

Benvenuto, F., La Camera, A., Theys, C., Ferrari, A., Lantéri, H., & Bertero, M., 2012, Inverse Probl., 28, 069502

Benvenuto, F., Zanella, R., Zanni, L., & Bertero, M., 2010, Inverse Probl., 26, 025004

Bertero, M., & Boccacci, P., 2000, A&AS, 144, 181

Bertero, M., & Boccacci, P., 2005, A&A, 437, 369

Bertero, M., Boccacci, P., Talenti, G., Zanella, R., & Zanni, L., 2010, Inverse Probl., 26, 10500

Bertero, M., Boccacci, P., La Camera, A., Olivieri, C., & Carbillet, M., 2011, Inverse Probl., 27, 113001

Birgin, E.G., Martínez, J.M., & Raydan, M., 2000, SIAM J. Optimiz., 10, 1196

Birgin, E.G., Martínez, J.M., & Raydan, M., 2003, IMA J. Numer. Anal., 23, 539

Bonettini, S., Zanella, R., & Zanni, L., 2009, Inverse Probl., 25, 015002

Bonettini, S., & Prato, M., 2010, Inverse Probl., 26, 095001

Bonettini, S., & Ruggiero, V., 2011, Inverse Probl., 27, 095001

Bonettini, S., Landi, G., Loli Piccolomini, E., & Zanni, L., 2012, Intern. J. Comp. Math., in press, DOI: 10.1080/00207160.2012.716513

Brune, C., Sawatzky, A., & Burger, M., 2010, J. Computer Vision, 92, 211

Charbonnier, P., Blanc-Féraud, L., Aubert, G., & Barlaud, A., 1997, IEEE T. Image Process., 6, 298

Defrise, M., Vanhove, C., & Liu, X., 2011, Inverse Probl., 27, 065002

Dey, N., Blanc-Féraud, L., Zimmer, C., *et al.*, 2006, Micros. Res. Techniq., 69, 260

Favati, P., Lotti, G., Menchi, O., & Romani, F., 2010, Inverse Probl., 26, 085013

Frassoldati, G., Zanni, L., & Zanghirati, G., 2008, J. Indust. Manag. Optim., 4, 299

Herbst, T.M., Ragazzoni, R., Andersen, D., *et al.*, 2003, Proc. SPIE, 4838, 456

Hudson, H.M., & Larkin, R.S., 1994, IEEE T. Med. Imaging, 13, 601

Lantéri, H., Roche, M., & Aime, C., 2002, Inverse Probl., 18, 1397

Le, T., Chartran, R., & Asaki, T.J., 2007, J. Math. Imaging Vis., 27, 257

Loris, I., Bertero, M., De Mol, C., Zanella, R., & Zanni, L., 2009, Appl. Comput. Harmon. A., 27, 247

Lucy, L.B., 1974, AJ, 79, 745

Lucy, L.B., & Hook, R.N., 1992, ASP Conf. Series, 25, 277

Prato, M., Cavicchioli, R., Zanni, L., Boccacci, P., & Bertero, M., 2012, A&A, 539, A133

Richardson, W.H., 1972, J. Opt. Soc. Am., 62, 55

Ruggiero, V., Serafini, T., Zanella, R., & Zanni, L., 2010, J. Global Optim., 48, 145

Serafini, T., Zanghirati, G., & Zanni, L., 2005, Optim. Method Softw., 20, 353

Serafini, T., Zanella, R., & Zanni, L., 2010, Adv. Parallel Comput., 19, 59

Setzer, S., Steidl, G., & Teuber, T., 2010, J. Vis. Commun. Image R., 21, 193

Shepp, L.A., & Vardi, Y., 1982, IEEE T. Med. Imaging, 1, 113
Snyder, D.L., Hammoud, A.M., & White, R.L., 1993, J. Opt. Soc. Am., A10, 1014
Staglianò, A., Boccacci, P., & Bertero, M., 2001, Inverse Probl., 27, 125003
Vogel, C.R., 2002, Computational Methods for Inverse Problems (SIAM, Philadelphia)
Zanella, R., Boccacci, P., Zanni, L., & Bertero, M., 2009, Inverse Probl., 25, 045010
Zhou, B., Gao, L., & Dai, Y.H., 2006, Comput. Optim. Appl., 35, 69

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 357–379*

SGM TO SOLVE NMF – APPLICATION TO HYPERSPECTRAL DATA

C. Theys[1], H. Lantéri[1] and C. Richard[1]

Abstract. This article deals with the problem of minimization of a general cost function under non-negativity and flux conservation constraints. The proposed algorithm is founded on the Split Gradient Method (SGM) adapted here to solve the Non Negative Matrix Factorization (NMF). We show that SGM can be easily regularized, allowing to introduce some physical constraints. Finally, to validate the algorithm, we propose an example of application to hyperspectral data unmixing.

1 Introduction

In the field of image reconstruction or deconvolution, the minimization of a cost function between noisy measurements and a linear model is usually performed, subject to positivity and flux constraints. The well known, in astrophysical area, are the Iterative Space Reconstruction Algorithm (ISRA) (Daube-Witherspoon 1986), and the Expectation Minimization (EM) (Dempster *et al.* 1977) or Richardson Lucy (RL) (Lucy 1974; Richardson 1972) algorithm. In the last ten years, a general algorithmic method, called Split Gradient Method (SGM) (Lantéri *et al.* 2001, 2002), has been developed to derive multiplicative algorithms for minimizing any convex criterion under positivity constraints. It leads to ISRA and EM-RL algorithm as particular cases. SGM has recently been extended to take into account a flux conservation constraint (Lantéri *et al.* 2009).

During the last few years, many papers have been published in the field of Nonnegative Matrix Factorization (NMF) with multiplicative algorithms (Lee & Seung 2001; Cichoki *et al.* 2006; Févotte *et al.* 2009). This problem is closely related to the blind deconvolution one (Desidera *et al.* 2006; Lantéri *et al.* 1994) and consists in estimating $\mathbf{W}$ and $\mathbf{H}$, nonnegative, such that $\mathbf{V} \approx \mathbf{WH}$. The aim

[1] Laboratoire Lagrange, Université de Nice Sophia-Antipolis, Observatoire de la Côte d'Azur, CNRS, Nice, France

DOI: 10.1051/eas/1359016

of this paper is to propose a unified framework based on SGM, an interior-point algorithm, to derive algorithms for NMF, in a multiplicative form or not.

To illustrate the general interest of SGM for NMF, we also show how to regularize the problem by introducing smoothness or sparsity constraints on the columns of $\mathbf{W}$ and $\mathbf{H}$ respectively, Lantéri *et al.* (2011), Lantéri *et al.* (2011). The choice of these different regularization terms are motivated by the application on hyperspectral imagery, Theys *et al.* (2009). The paper is organized as follows. In Section 2, we describe the problem at hand and notations for non-negative matrix factorization. In Section 3, we describe the Split Gradient Method (SGM). In Section 4, we show how to add a sum-to-one constraint in the SGM algorithm. In Section 5, we briefly discuss the choice of the step size. Section 6 introduces the physical context and some simulation results are given in Section 7. The regularized SGM is developed in section 8 with a smoothness constraint on the columns of $\mathbf{W}$ and then a sparsity constraint on the columns of $\mathbf{H}$, with typical numerical examples in Section 9. Section 10 concludes the paper.

2 Nonnegative matrix factorization

We consider here the problem of nonnegative matrix factorization (NMF), which is now a popular dimension reduction technique, employed for non-subtractive, part-based representation of nonnegative data. Given a nonnegative data matrix $\mathbf{V}$ of dimension $F \times N$, the NMF consists of seeking a factorization of the form

$$\mathbf{V} \approx \mathbf{WH} \tag{2.1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are nonnegative matrices of dimensions $F \times K$ and $K \times N$, respectively. Dimension K is usually chosen such that $FK + KN \ll FN$, that is, much more equations than unknowns. For example with $F = N = 3$ and $K = 1$:

$$\begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \mathbf{V}_{23} \\ \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{V}_{33} \end{bmatrix} = \begin{bmatrix} \mathbf{W}_{11} \\ \mathbf{W}_{21} \\ \mathbf{W}_{31} \end{bmatrix} \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \mathbf{H}_{13} \end{bmatrix}. \tag{2.2}$$

This problem is encountered at each time we want to find both the basis and the coefficients of projection. The factorization (2.1) is usually sought through the minimization problem

$$\min_{\mathbf{W},\mathbf{H}} \mathcal{D}(\mathbf{V}, \mathbf{WH}) \quad \text{s.t.} \quad [\mathbf{W}]_{ij} \geq 0, [\mathbf{H}]_{ij} \geq 0 \tag{2.3}$$

with $[\mathbf{V}]_{ij}$ and $[\mathbf{WH}]_{ij}$ the (i,j)-th entries of $\mathbf{V}$ and $\mathbf{WH}$, respectively. In the above expression, $\mathcal{D}(\mathbf{V}, \mathbf{WH})$ is a cost function defined by

$$\mathcal{D}(\mathbf{V}, \mathbf{WH}) = \sum_{ij} d([\mathbf{V}]_{ij}, [\mathbf{WH}]_{ij}) = \sum_{ij} d_{ij}. \tag{2.4}$$

In the general case, $d(u, v)$ is a positive convex function that is equal to zero if $u = v$.

2.1 Unicity

The solution of (2.3) is, obviously, not unique. One way to overcome this problem is to normalize the columns of $\mathbf{W}$ or $\mathbf{H}$. We propose, here, to normalize to *one* the columns of $\mathbf{W}$. As a direct consequence of (2.1), this implies a constraint-sum condition on the columns of $\mathbf{H}$.

The minimization problem (2.3) becomes:

$$\min_{\mathbf{W},\mathbf{H}} \mathcal{D}(\mathbf{V}, \mathbf{WH}) \quad \text{s.t.} \quad [\mathbf{W}]_{ij} \geq 0, \quad [\mathbf{H}]_{ij} \geq 0, \\ \sum_i [\mathbf{W}]_{ij} = 1, \quad \sum_i [\mathbf{H}]_{ij} = \sum_i [\mathbf{V}]_{ij}. \quad (2.5)$$

This constant-sum constraint is motivated by applications such as, for example, hyperspectral data unmixing. In this case, $\mathbf{W}$ is the matrix of basis spectra that are supposed to be normalized to *one*. Another source of indetermination is that the solutions are given up to a permutation on rows and columns of $\mathbf{W}$ and $\mathbf{H}$. The problem established by (2.3) is a convex optimization problem under inequality constraint and problem (2.5) is a convex optimization problem under both equality and inequality constraints. We propose to consider first the problem (2.3), the inequality constraint is treated by solving the Karush-Kuhn-Tucker conditions. Second, we consider the problem (2.5) and the equality constraint is added by introducing normalized variables. Once the conditions satisfying the constraints have been established, an iterative algorithm should be applied alternatively on $\mathbf{W}$ and $\mathbf{H}$. The proposed iterative algorithm founded on the Split Gradient Method (SGM), a scaled gradient descent algorithm. The way to obtain it is detailed in the following section.

3 Minimization under non-negativity constraints: The SGM

The SGM was initially formulated and developed to solve the minimization of a positive convex function under non-negativity constraint of the solution, problem (2.3).

3.1 The Lagrangian function

The non-negativity constraint is expressed by the Lagrangian function associated to (2.3), given by:

$$\mathcal{L}(\mathbf{V}, \mathbf{WH}; \mathbf{\Lambda}, \mathbf{\Omega}) = \mathcal{D}(\mathbf{V}, \mathbf{WH}) - \langle \mathbf{\Lambda}, \mathbf{W} \rangle - \langle \mathbf{\Omega}, \mathbf{H} \rangle \quad (3.1)$$

where $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ are the matrices of positive Lagrange multipliers, and $\langle \cdot, \cdot \rangle$ is the inner product defined by:

$$\langle \mathbf{U}, \mathbf{V} \rangle = \sum_{ij} [\mathbf{U}]_{ij} [\mathbf{V}]_{ij}. \quad (3.2)$$

The Lagrange multipliers method allows to find an optimum of a function under some constraints.

3.2 Minimization with respect to $\mathbf{W}$

Minimization of (3.1) with respect to $\mathbf{W}$ leads to the following Karush-Kuhn-Tucker conditions for all i, j at the solution $\mathbf{W}^*$, $\mathbf{\Lambda}^*$:

$$[\nabla_W \mathcal{L}(\mathbf{V}, \mathbf{W}^*\mathbf{H}; \mathbf{\Lambda}^*, \mathbf{\Omega})]_{ij} = 0, \tag{3.3}$$
$$[\mathbf{\Lambda}^*]_{ij} \geq 0, \tag{3.4}$$
$$[\mathbf{W}^*]_{ij} \geq 0, \tag{3.5}$$
$$\langle \Lambda^*, \mathbf{W}^* \rangle = 0 \Leftrightarrow [\Lambda^*]_{ij}[\mathbf{W}^*]_{ij} = 0. \tag{3.6}$$

Condition (3.3) immediately leads to

$$[\Lambda^*]_{ij} = [\nabla_W \mathcal{D}(\mathbf{V}, \mathbf{W}^*\mathbf{H})]_{ij}. \tag{3.7}$$

Condition (3.6) then becomes

$$\begin{aligned}[\mathbf{W}^*]_{ij}[\nabla_W \mathcal{D}(\mathbf{V}, \mathbf{W}^*\mathbf{H})]_{ij} = 0 \\ \Leftrightarrow [\mathbf{W}^*]_{ij}[-\nabla_W \mathcal{D}(\mathbf{V}, \mathbf{W}^*\mathbf{H})]_{ij} = 0\end{aligned} \tag{3.8}$$

where the extra minus sign in the last expression is just used to make apparent the negative gradient descent direction of $\mathcal{D}(\mathbf{V}, \mathbf{WH})$.

The expression (3.6) gives the condition that must be satisfied for any optimization problem under non-negativity constraint. At the solution, the inner product between the gradient of the cost function and the variables must be equal to zero. The interpretation is the following: either our solution is the one that minimizes the cost function and the minimizer is positive, either the minimizer of the cost function is negative or zero and the constrained solution is zero.

This condition is non linear w.r.t. the unknowns, an analytical solution does not exist.

3.2.1 Gradient descent method

Since the gradient of the functional has an analytical form, a natural choice for the iterative algorithm is a gradient descent method.

If we consider first the minimization problem without non-negativity constraint:

$$\min_{\mathbf{W},\mathbf{H}} \mathcal{D}(\mathbf{V}, \mathbf{WH}), \tag{3.9}$$

we use the negative gradient as a descent direction and we write:

$$[\mathbf{W}^{k+1}]_{ij} = [\mathbf{W}^k]_{ij} + \alpha_{ij}^k [-\nabla_W \mathcal{D}(\mathbf{V}, \mathbf{W}^*\mathbf{H})]_{ij} \tag{3.10}$$

with α_{ij}^k a positive step size that allows to control convergence of the algorithm.

If now, we consider the minimization problem with non-negativity constraint, Equation (2.3), the descent direction becomes $[\mathbf{W}^*]_{ij}[-\nabla_W \mathcal{D}(\mathbf{V}, \mathbf{W}^*\mathbf{H})]_{ij}$, Equation (3.8) and the descent algorithm is:

$$[\mathbf{W}^{k+1}]_{ij} = [\mathbf{W}^k]_{ij} + \alpha_{ij}^k [\mathbf{W}^*]_{ij}[-\nabla_W \mathcal{D}(\mathbf{V}, \mathbf{W}^*\mathbf{H})]_{ij}. \tag{3.11}$$

More generally:

$$\mathbf{M} \cdot \mathbf{W} \cdot [-\nabla_W \mathcal{D}(\mathbf{V}, \mathbf{W}^*\mathbf{H})] \tag{3.12}$$

is a scaled gradient descent direction of $\mathcal{D}$ if $\mathbf{M}$ is a matrix with positive entries, where $\cdot$ denotes the Hadamard product. A particular choice for $\mathbf{M}$ with an adequate particular decomposition of $[-\nabla_W \mathcal{D}(\mathbf{V}, \mathbf{W}^*\mathbf{H})]$ leads to the SGM algorithm.

3.2.2 Split Gradient Method (SGM)

The SGM algorithm is a descent algorithm whose direction is constructed in such a way that, for a step size equal to one, we obtain a multiplicative algorithm. To obtain it, an additional point is that $[-\nabla_W \mathcal{D}]_{ij}$ can always be decomposed as $[\mathbf{P}]_{ij} - [\mathbf{Q}]_{ij}$, where $[\mathbf{P}]_{ij}$ and $[\mathbf{Q}]_{ij}$ are positive entries, let us note that this decomposition is obviously not unique. If we take for $\mathbf{M}$, Equation (3.12):

$$[\mathbf{M}]_{ij} = \frac{1}{[\mathbf{Q}]_{ij}} \tag{3.13}$$

we obtain the following gradient-descent algorithm:

$$[\mathbf{W}^{k+1}]_{ij} = [\mathbf{W}^k]_{ij} + \alpha_{ij}^k \frac{[\mathbf{W}^k]_{ij}}{[\mathbf{Q}]_{ij}^k} [-\nabla_W \mathcal{D}(\mathbf{V}, \mathbf{W}^k\mathbf{H})]_{ij} \tag{3.14}$$

with α_{ij}^k a positive step size that allows to control convergence of the algorithm. If we write explicitly the decomposition of the gradient, Equation (3.11) becomes:

$$[\mathbf{W}^{k+1}]_{ij} = [\mathbf{W}^k]_{ij} + \alpha_{ij}^k \frac{[\mathbf{W}^k]_{ij}}{[\mathbf{Q}^k]_{ij}} \left([\mathbf{P}^k]_{ij} - [\mathbf{Q}^k]_{ij}\right) \tag{3.15}$$

or

$$[\mathbf{W}^{k+1}]_{ij} = [\mathbf{W}^k]_{ij} + \alpha_{ij}^k [\mathbf{W}^k]_{ij} \left(\frac{[\mathbf{P}^k]_{ij}}{[\mathbf{Q}^k]_{ij}} - 1 \right). \tag{3.16}$$

Once we have the gradient type descent algorithm, we must determine the maximum value for the step size in order that $[\mathbf{W}^{k+1}]_{ij} \geq 0$, given $[\mathbf{W}^k]_{ij} \geq 0$. Note that, according to (3.15) or (3.16), a restriction must only apply if

$$[\mathbf{P}^k]_{ij} - [\mathbf{Q}^k]_{ij} < 0 \tag{3.17}$$

since the other terms are positive. The maximum step size which ensures the positivity of $[\mathbf{W}^{k+1}]_{ij}$ is given by

$$(\alpha_{ij}^k)_{\max} = \frac{1}{1 - \frac{[\mathbf{P}^k]_{ij}}{[\mathbf{Q}^k]_{ij}}} \tag{3.18}$$

which is strictly greater than 1. Finally, the maximum step size over all the components must satisfy

$$(\alpha^k)_{\max} \leq \min\{(\alpha_{ij}^k)_{\max}\}. \tag{3.19}$$

This choice ensures the non-negativity of all the components of $\mathbf{W}^k$ from iteration to iteration. Then, convergence of the algorithm is guaranteed by computing an appropriate step size, at each iteration, over the range $[0, (\alpha^k)_{\max}]$ by means of a simplified line search such as the Armijo rule for example. Finally, it is important to notice that the use of a step size equal to 1 leads to the very simple and well-known multiplicative form:

$$[\mathbf{W}^{k+1}]_{ij} = [\mathbf{W}^k]_{ij} \frac{[\mathbf{P}^k]_{ij}}{[\mathbf{Q}^k]_{ij}}. \tag{3.20}$$

This form is used because it is very easy to implement and it guarantees the non-negativity of successive iterates for an initial non-negative value $[\mathbf{W}^0]_{ij} \geq 0$. The main and important drawback is that the convergence of the algorithm is not assured in the general case, but only for specific cases of $[\mathbf{P}]$ and $[\mathbf{Q}]$.

3.3 Minimization with respect to $\mathbf{H}$

Minimization of (3.1) with respect to $\mathbf{H}$ leads to the following Karush-Kuhn-Tucker conditions for all i, j at the solution $\mathbf{W}^*$, $\mathbf{\Lambda}^*$:

$$[\nabla_H \mathcal{L}(\mathbf{V}, \mathbf{W}^*\mathbf{H}; \mathbf{\Lambda}, \mathbf{\Omega}^*)]_{ij} = 0, \tag{3.21}$$
$$[\mathbf{\Omega}^*]_{ij} \geq 0, \tag{3.22}$$
$$[\mathbf{H}^*]_{ij} \geq 0, \tag{3.23}$$
$$\langle \mathbf{\Omega}^*, \mathbf{H}^* \rangle = 0 \Leftrightarrow [\mathbf{\Omega}^*]_{ij}[\mathbf{H}^*]_{ij} = 0. \tag{3.24}$$

Condition (3.21) immediately leads to

$$[\mathbf{\Omega}^*]_{ij} = [\nabla_H \mathcal{D}(\mathbf{V}, \mathbf{W}\mathbf{H}^*)]_{ij}. \tag{3.25}$$

Condition (3.24) then becomes

$$\begin{aligned} [\mathbf{H}^*]_{ij}[\nabla_H \mathcal{D}(\mathbf{V}, \mathbf{W}\mathbf{H}^*)]_{ij} &= 0 \\ \Leftrightarrow [\mathbf{H}^*]_{ij}[-\nabla_H \mathcal{D}(\mathbf{V}, \mathbf{W}\mathbf{H}^*)]_{ij} &= 0. \end{aligned} \tag{3.26}$$

where the extra minus sign in the last expression is just used to make the negative gradient descent direction of $\mathcal{D}(\mathbf{V}, \mathbf{W}\mathbf{H})$ apparent.

The expression (3.24) gives the condition that must be satisfied for any optimization problem under non-negativity constraint. At the solution, the inner product between the gradient of the cost function and the variables must be equal to zero. The interpretation is the following: either our solution is the one that minimizes the cost function and the minimizer is positive, either the minimizer of the cost function is negative or zero and the constrained solution is zero.

This condition is non linear w.r.t. the unknowns, an analytical solution does not exist.

3.3.1 Gradient descent method

Since the gradient of the functional is computable, a natural choice for the iterative algorithm is a gradient descent method.

If we consider first the minimization problem without non-negativity constraint:

$$\min_{\mathbf{W},\mathbf{H}} \mathcal{D}(\mathbf{V}, \mathbf{WH}), \tag{3.27}$$

we use the negative gradient as a descent direction and we write:

$$[\mathbf{H}^{k+1}]_{ij} = [\mathbf{H}^k]_{ij} + \beta_{ij}^k [-\nabla_H \mathcal{D}(\mathbf{V}, \mathbf{WH}^*)]_{ij} \tag{3.28}$$

with β_{ij}^k a positive step size that allows to control convergence of the algorithm.

If now, we consider the minimization problem with non-negativity constraint, Equation (2.3), the descent direction becomes $[\mathbf{H}^*]_{ij}[-\nabla_H \mathcal{D}(\mathbf{V}, \mathbf{WH}^*)]_{ij}$, Equation (3.26) and the descent algorithm is:

$$[\mathbf{H}^{k+1}]_{ij} = [\mathbf{H}^k]_{ij} + \beta_{ij}^k [\mathbf{H}^k]_{ij} [-\nabla_H \mathcal{D}(\mathbf{V}, \mathbf{WH}^*)]_{ij}. \tag{3.29}$$

More generally:

$$\mathbf{N} \cdot \mathbf{H} \cdot [-\nabla_H \mathcal{D}(\mathbf{V}, \mathbf{WH}^*)] \tag{3.30}$$

is a gradient descent direction of $\mathcal{D}$ if $\mathbf{N}$ is a matrix with positive entries, where $\cdot$ denotes the Hadamard product. A particular choice for $\mathbf{N}$ with a specific decomposition of $[-\nabla_H \mathcal{D}(\mathbf{V}, \mathbf{WH}^*)]$ leads to the SGM algorithm.

3.3.2 Split Gradient Method (SGM)

The SGM algorithm is a descent algorithm whose direction is constructed in such a way that, for a step size equal to one, we obtain a multiplicative algorithm. To obtain it, an additional point is that $[-\nabla_H \mathcal{D}]_{ij}$ can always be decomposed as $[\mathbf{R}]_{ij} - [\mathbf{S}]_{ij}$, where $[\mathbf{R}]_{ij}$ and $[\mathbf{S}]_{ij}$ are positive entries, let us note that this decomposition is obviously not unique. If we take for $\mathbf{N}$, Equation (3.30):

$$[\mathbf{N}]_{ij} = \frac{1}{[\mathbf{S}]_{ij}}, \tag{3.31}$$

we obtain the following gradient-descent algorithm:

$$[\mathbf{H}^{k+1}]_{ij} = [\mathbf{H}^k]_{ij} + \beta_{ij}^k \frac{[\mathbf{R}^k]_{ij}}{[\mathbf{S}^k]_{ij}} [-\nabla_H \mathcal{D}(\mathbf{V}, \mathbf{WH}^k)]_{ij} \tag{3.32}$$

with β_{ij}^k a positive step size that allows to control convergence of the algorithm. If we write explicitly the decomposition of the gradient, Equation (3.32) becomes:

$$[\mathbf{H}^{k+1}]_{ij} = [\mathbf{H}^k]_{ij} + \beta_{ij}^k \frac{[\mathbf{H}^k]_{ij}}{[\mathbf{R}^k]_{ij}} \left([\mathbf{R}^k]_{ij} - [\mathbf{S}^k]_{ij} \right) \tag{3.33}$$

or

$$[\mathbf{H}^{k+1}]_{ij} = [\mathbf{H}^k]_{ij} + \beta_{ij}^k [\mathbf{H}^k]_{ij} \left(\frac{[\mathbf{R}^k]_{ij}}{[\mathbf{S}^k]_{ij}} - 1 \right). \tag{3.34}$$

Once we have the gradient type descent algorithm, we must determine the maximum value for the step size in order that $[\mathbf{H}^{k+1}]_{ij} \geq 0$, given $[\mathbf{H}^k]_{ij} \geq 0$. Note that, according to (3.33), a restriction must only apply if

$$[\mathbf{R}^k]_{ij} - [\mathbf{S}^k]_{ij} < 0 \tag{3.35}$$

since the other terms are positive. The maximum step size which ensures the positivity of $[\mathbf{H}^{k+1}]_{ij}$ is given by

$$(\beta_{ij}^k)_{\max} = \frac{1}{1 - \frac{[\mathbf{R}^k]_{ij}}{[\mathbf{S}^k]_{ij}}} \tag{3.36}$$

which is strictly greater than 1. Finally, the maximum step size over all the components must satisfy

$$(\beta^k)_{\max} \leq \min\{(\beta_{ij}^k)_{\max}\}. \tag{3.37}$$

This choice ensures the non-negativity of all the components of $\mathbf{H}^k$ from iteration to iteration. Then, convergence of the algorithm is guaranteed by computing an appropriate step size, at each iteration, over the range $[0, (\beta^k)_{\max}]$ by means of a simplified line search such as the Armijo rule for example. Finally, it is important to notice that the use of a step size equal to 1 leads to the very simple and well-known multiplicative form:

$$[\mathbf{H}^{k+1}]_{ij} = [\mathbf{H}^k]_{ij} \frac{[\mathbf{R}^k]_{ij}}{[\mathbf{S}^k]_{ij}}. \tag{3.38}$$

This form is used because it is very easy to implement and it guarantees the non-negativity of successive iterates for an initial non-negative value $[\mathbf{H}^0]_{ij} \geq 0$. The main and important drawback is that the convergence of the algorithm is not assured.

3.4 Explicit expressions of the gradients

Before ending this section, let us compute $\nabla\mathcal{D}$ with respect to $\mathbf{H}$ and $\mathbf{W}$, using Equations (2.1) and (2.4). It can be expressed in matrix form as follows:

$$\nabla_H \mathcal{D} = \mathbf{W}^T \mathbf{A} \qquad \nabla_W \mathcal{D} = \mathbf{A}\mathbf{H}^T \tag{3.39}$$

where $\mathbf{A}$ is a matrix whose (i, j)-th entry is given by:

$$[\mathbf{A}]_{ij} = \frac{\partial d_{ij}}{\partial [\mathbf{W}\mathbf{H}]_{ij}}. \tag{3.40}$$

Equations (3.20), (3.38) associated to (3.39), (3.40), lead to the multiplicative algorithms described in (Cichoki *et al.* 2006; Févotte *et al.* 2009; Lee & Seung 2001). These are particular cases of the relaxed algorithms (3.15) (3.33), when a unit step size is used.

4 Minimization under non-negativity constraints and flux conservation

Let us now consider problem (2.5), which differs from (2.3) by additional flux constraints.

4.1 Flux conservation constraints

We make the following variable changes:

$$[\mathbf{W}]_{ij} = \frac{[\mathbf{Z}]_{ij}}{\sum_m [\mathbf{Z}]_{mj}}; \tag{4.1}$$

$$[\mathbf{H}]_{ij} = \left(\sum_m [\mathbf{V}]_{mj}\right) \frac{[\mathbf{T}]_{ij}}{\sum_m [\mathbf{T}]_{mj}}. \tag{4.2}$$

The term $(\sum_m [\mathbf{V}]_{mj})$ comes from the fact that $[\mathbf{H}]_{ij}$ is normalized to the column j of $\mathbf{V}$. In so doing, the problem becomes unconstrained with respect to the flux but we must search the solution in a domain where the denominator is a constant to ensure that the problem remains convex w.r.t. the new variables. It is an important point performed by our method. The flux conservation being provided by the change of variables, we can proceed the SGM on the new variables to ensure both the non-negativity and the flux conservation.

To deal with the non-negativity constraints, let us consider again the SGM algorithm and compute the gradient with respect to new variables.

4.2 Explicit expressions of the gradients

Let us compute expression of the gradients w.r.t. the new variables:

$$\frac{\partial \mathcal{D}}{\partial [\mathbf{Z}]_{lj}} = \sum_i \frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{ij}} \times \frac{\partial [\mathbf{W}]_{ij}}{\partial [\mathbf{Z}]_{lj}}, \tag{4.3}$$

$$\frac{\partial \mathcal{D}}{\partial [\mathbf{T}]_{lj}} = \sum_i \frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{ij}} \times \frac{\partial [\mathbf{H}]_{ij}}{\partial [\mathbf{T}]_{lj}} \tag{4.4}$$

where, in a compact form,

$$\frac{\partial [\mathbf{W}]_{ij}}{\partial [\mathbf{Z}]_{lj}} = \frac{1}{\sum_m [\mathbf{Z}]_{mj}} \times (\delta_{li} - [\mathbf{W}]_{ij}), \tag{4.5}$$

$$\frac{\partial [\mathbf{H}]_{ij}}{\partial [\mathbf{T}]_{lj}} = \frac{\sum_m [\mathbf{V}]_{mj}}{\sum_m [\mathbf{T}]_{mj}} \times \left(\delta_{li} - \frac{[\mathbf{H}]_{ij}}{\sum_m [\mathbf{V}]_{mj}}\right) \tag{4.6}$$

with δ_{li} the Kronecker symbol. As a consequence, the components of the opposite of the gradient of $\mathcal{D}$ with respect to the new variables can now be written as

$$-\frac{\partial \mathcal{D}}{\partial [\mathbf{Z}]_{lj}} = \frac{1}{\sum_m [\mathbf{Z}]_{mj}} \left(\left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{lj}}\right) - \sum_i [\mathbf{W}]_{ij} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{ij}}\right)\right) \tag{4.7}$$

and

$$-\frac{\partial \mathcal{D}}{\partial [\mathbf{T}]_{lj}} = \frac{\sum_m [\mathbf{V}]_{mj}}{\sum_m [\mathbf{T}]_{mj}} \left(\left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{lj}} \right) - \frac{\sum_i [\mathbf{H}]_{ij}}{\sum_m [\mathbf{V}]_{mj}} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{ij}} \right) \right) \tag{4.8}$$

4.3 SGM with the normalized variables

We solve both the split of the gradient between two positive functions and the conservation of the convexity w.r.t. to the new variables by making the shift of the form:

$$\begin{aligned} (-\partial \mathcal{D} / \partial [\mathbf{W}]_{ij})_s &\longleftarrow (-\partial \mathcal{D} / \partial [\mathbf{W}]_{ij}) + \eta, \quad \forall (i,j), \\ (-\partial \mathcal{D} / \partial [\mathbf{H}]_{ij})_s &\longleftarrow (-\partial \mathcal{D} / \partial [\mathbf{H}]_{ij}) + \mu, \quad \forall (i,j). \end{aligned}$$

Let us notice that this shift leaves Equations (4.9) and (4.14) unchanged. Consequently, using

$$\eta = -\min_{ij} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{ij}} \right) + \epsilon, \quad \mu = -\min_{ij} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{ij}} \right) + \epsilon$$

does not modify the gradient of $\mathcal{D}$ with respect to the new variables $\mathbf{Z}$ and $\mathbf{T}$, but ensures the non-negativity of $(-\partial \mathcal{D} / \partial [\mathbf{W}]_{ij})_s$ and $(-\partial \mathcal{D} / \partial [\mathbf{H}]_{ij})_s$. A constant ϵ is added to avoid numerical instability, however, it must be chosen small enough not to slow down the minimization. Let us note that this particular decomposition allows to ensure that the denominator in (4.1) and (4.2) remains constant and then we are always in the convexity domain. We shall now apply the SGM method.

4.4 Minimization with respect to $\mathbf{W}$

Consider the following gradient (4.9) decomposition:

$$[-\nabla_Z \mathcal{D}]_{ij} = [\mathbf{P}]_{ij} - [\mathbf{Q}]_{ij} \tag{4.9}$$

that involves the non-negative entries defined as follows

$$[\mathbf{P}]_{ij} = \left(-\frac{\partial D}{\partial [\mathbf{W}]_{ij}} \right)_s, \tag{4.10}$$

$$[\mathbf{Q}]_{ij} = [\mathbf{Q}]_{.j} = \sum_i [\mathbf{W}]_{ij} \left(-\frac{\partial D}{\partial [\mathbf{W}]_{ij}} \right)_s. \tag{4.11}$$

The relaxed form of the minimization algorithm can be expressed as

$$[\mathbf{Z}^{k+1}]_{lj} = [\mathbf{Z}^k]_{lj} + \alpha^k [\mathbf{Z}^k]_{lj} \left(\frac{(-\partial \mathcal{D} / \partial [\mathbf{W}^k]_{lj})_s}{\sum_i [\mathbf{W}^k]_{ij} (-\partial \mathcal{D} / \partial [\mathbf{W}^k]_{ij})_s} - 1 \right).$$

We clearly have $\sum_l [\mathbf{Z}^{k+1}]_{lj} = \sum_l [\mathbf{Z}^k]_{lj}$, for all α^k. This allows us to express the algorithm with respect to the initial variable $\mathbf{W}$, that is,

$$[\mathbf{W}^{k+1}]_{lj} = [\mathbf{W}]_{lj}^k + \alpha^k [\mathbf{W}]_{lj}^k \left(\frac{(-\partial \mathcal{D} / \partial [\mathbf{W}]_{lj}^k)_s}{\sum_i [\mathbf{W}]_{ij}^k (-\partial \mathcal{D} / \partial [\mathbf{W}]_{ij}^k)_s} - 1 \right). \tag{4.12}$$

Again, with a constant step size equal to 1, the algorithm takes a simple multiplicative form:

$$[\mathbf{W}^{k+1}]_{lj} = [\mathbf{W}^k]_{lj} \frac{(-\partial \mathcal{D} / \partial [\mathbf{W}^k]_{lj})_s}{\sum_i [\mathbf{W}^k]_{ij} (-\partial \mathcal{D} / \partial [\mathbf{W}^k]_{ij})_s}. \tag{4.13}$$

4.5 Minimization with respect to $\mathbf{H}$

In an analogous way, consider the following gradient (4.14) decomposition:

$$[-\nabla_T \mathcal{D}]_{ij} = [\mathbf{R}]_{ij} - [\mathbf{S}]_{ij} \tag{4.14}$$

that involves the non-negative entries given by

$$[\mathbf{R}]_{ij} = \frac{\sum_m [\mathbf{V}]_{mj}}{\sum_m [\mathbf{T}]_{mj}} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{ij}} \right)_s, \tag{4.15}$$

$$[\mathbf{S}]_{ij} = S_{.j} = \frac{\sum_m [\mathbf{V}]_{m,j}}{\sum_m [\mathbf{T}]_{mj}} \sum_i \frac{[\mathbf{H}]_{ij}}{\sum_m [\mathbf{V}]_{mj}} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{ij}} \right)_s. \tag{4.16}$$

This leads to the relaxed form of optimization algorithm with respect to variable $\mathbf{T}$, that is,

$$[\mathbf{T}^{k+1}]_{lj} = [\mathbf{T}^k]_{lj} + \alpha^k [\mathbf{T}^k]_{lj} \left(\frac{(-\partial \mathcal{D} / \partial [\mathbf{H}^k]_{lj})_s}{\sum_i \frac{[\mathbf{H}^k]_{ij}}{\sum_m [\mathbf{V}]_{mj}} (-\partial \mathcal{D} / \partial [\mathbf{H}^k]_{ij})_s} - 1 \right).$$

It can be seen that $\sum_l [\mathbf{T}^{k+1}]_{lj} = \sum_l [\mathbf{T}^k]_{lj}$, for all α^k, which implies that

$$[\mathbf{H}^{k+1}]_{lj} = [\mathbf{H}^k]_{lj} + \alpha^k [\mathbf{H}^k]_{lj} \left(\frac{(-\partial \mathcal{D} / \partial [\mathbf{H}^k]_{lj})_s}{\sum_i \frac{[\mathbf{H}^k]_{ij}}{\sum_m [\mathbf{V}]_{mj}} (-\partial \mathcal{D} / \partial [\mathbf{H}^k]_{ij})_s} - 1 \right). \tag{4.17}$$

The multiplicative form is obtained with a constant step size equal to 1, namely,

$$[\mathbf{H}^{k+1}]_{lj} = [\mathbf{H}^k]_{lj} \frac{(-\partial \mathcal{D} / \partial [\mathbf{H}^k]_{lj})_s}{\sum_i [\mathbf{H}^k]_{ij} (-\partial \mathcal{D} / \partial [\mathbf{H}^k]_{ij})_s} \sum_m [\mathbf{V}]_{mj}. \tag{4.18}$$

In the next section, we propose to illustrate this algorithm within the field of hyperspectral imaging.

5 Choice of the descent step size and convergence speed

On one hand, if the descent step size is fixed to one, there is no way to modify the convergence speed and the iterations number can be high, moreover, the convergence is not ensured but the algorithm takes a simple form. On the other hand, if the descent step size is searched by a simple rule, Armijo for example, the iterations number decreases but the duration of one iteration increases, from our experience, when the step size is computed, the overall gain is about ten or twenty percents and in this case the convergence is ensured.

6 Physical context: Hyperspectral imagery

Hyperspectral imaging has received considerable attention in the last few years. See for instance (Chang 2003), (Landgrebe 2003) and references therein. It consists of data acquisition with high sensitivity and resolution in hundreds contiguous spectral bands, geo-referenced within the same coordinate system. With its ability to provide extremely detailed data regarding the spatial and spectral characteristics of a scene, this technology offers immense new possibilities in collecting and managing information for civilian and military application areas.

Each vector pixel of an hyperspectral image characterizes a local spectral signature. Usually, one consider that each vector pixel can be modeled accurately as a linear mixture of different pure spectral components, called endmembers. Referring to our notations, each column of $\mathbf{V}$ can thus be interpreted as a spectral signature obtained by linear mixing of the spectra of endmembers, *i.e.*, the columns of $\mathbf{W}$. The problem is then to estimate the endmember spectra $\mathbf{W}$ and the abundance coefficients $\mathbf{H}$ from the spectral signatures $\mathbf{V}$.

In all the simulations presented in this paper, the end members are extracted from the ENVI library (ENVI 2003).

7 Simulation results

Many simulations have been performed to validate the proposed algorithm, Equations (4.13) and (4.18). The experiment presented in this paper corresponds to 10 linear mixtures of 3 endmembers, the length of each spectrum being 826. The three endmembers used in this example correspond to the spectra of the construction concrete, green grass, and micaceous loam. The chosen cost function for $\mathcal{D}$ is the Frobenius norm:

$$\mathcal{D}(\mathbf{V}, \mathbf{WH}) = \sum_{ij} ([\mathbf{WH}]_{ij} - [\mathbf{V}]_{ij})^2 = (\mathbf{WH} - \mathbf{V})^T (\mathbf{WH} - \mathbf{V}). \tag{7.1}$$

The used procedure is the following:

1. Take the spectra from a library (ENVI here).
2. Generate randomly the KN abundance coefficients H_{ij} in a given interval.
3. Compute $\mathbf{V}$.
4. Generate randomly H^0 and W^0 in the space constraints.
5. Compute the chosen cost function, here the Frobenius norm:
6. Compute the decomposition of the gradient w.r.t. Z, *i.e.* (4.10) and (4.11).
7. Compute $\mathbf{W}^{k+1}$, (4.12).
8. Compute the decomposition of the gradient w.r.t. T, *i.e.* (4.15) and (4.16).

9. Compute $\mathbf{H}^{k+1}$, (4.17).
10. Until the stopping criterion:

$$\frac{\mathcal{D}(\mathbf{V}, \mathbf{W}^{k+1}\mathbf{H}^{k+1}) - \mathcal{D}(\mathbf{V}, \mathbf{W}^{k}\mathbf{H}^{k})}{\mathcal{D}(\mathbf{V}, \mathbf{W}^{k}\mathbf{H}^{k})} \leq 10^{-10}. \tag{7.2}$$

Figure 1 shows the behaviour of the criterion $\mathcal{D}$ as a function of the number of iterations, and the 10 reconstructed spectra in comparison with the true ones. Figure 2 shows the estimated endmembers (columns of $\mathbf{W}$), and their abundance coefficients (rows of $\mathbf{H}$) after 12 000 iterations, and compared with the true values. Note that the initial values for $\mathbf{W}$ and $\mathbf{H}$ were chosen to satisfy the constraints, *i.e.*, positivity, sum to one of the columns of $\mathbf{W}$. We clearly see that the curves coincide almost perfectly. The normalization of the columns of matrix $\mathbf{W}$, as well as the flux conservation between $\mathbf{V}$ and $\mathbf{H}$, are satisfied at each iteration. Let us note that $\mathbf{H}$ and $\mathbf{W}$ could be estimated up to a permutation of the columns of $\mathbf{W}$, and to an analogous permutation of the rows of $\mathbf{H}$.

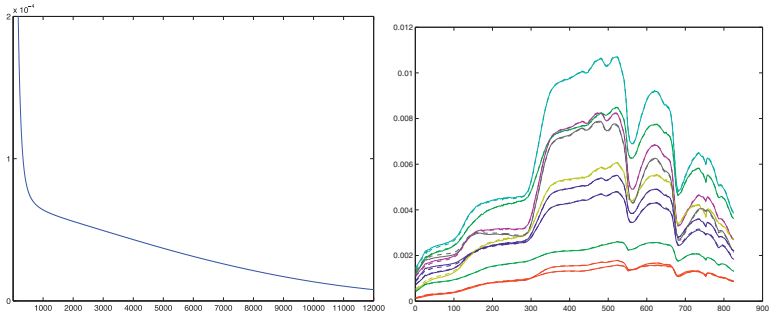


Fig. 1. Frobenius $\mathcal{D}(\mathbf{V}, \mathbf{WH})$ as a function of the number of iterations. Columns of $\mathbf{V}$ at the end of the iterations, solid line for true values, dashed line for estimated values.

8 Regularization

In full generality, we can add several regularization terms depending on one or two variables, the only constraint being that each regularization function must be convex w.r.t. the relevant variable. If the regularization term depends on the two variables, it must be convex w.r.t. one variable, the other being fixed. Here, we consider the case where the regularization penalty terms are incorporated separately on the columns of $\mathbf{W}$ and $\mathbf{H}$, and are added to the data consistency term $\mathcal{D}(\mathbf{V}, \mathbf{WH})$. Then the penalized objective function expresses as

$$\mathcal{D}_{\text{reg}}(\mathbf{V}, \mathbf{WH}) = \mathcal{D}(\mathbf{V}, \mathbf{WH}) + \gamma_1 \mathcal{F}_1(\mathbf{W}) + \gamma_2 \mathcal{F}_2(\mathbf{H}) \tag{8.1}$$

where $\mathcal{F}_1(\mathbf{W})$ and $\mathcal{F}_2(\mathbf{H})$ are penalty functions, and γ_1, γ_2 the respective regularization factors. The general rules given for SGM remain true for the regularized versions of the algorithms.

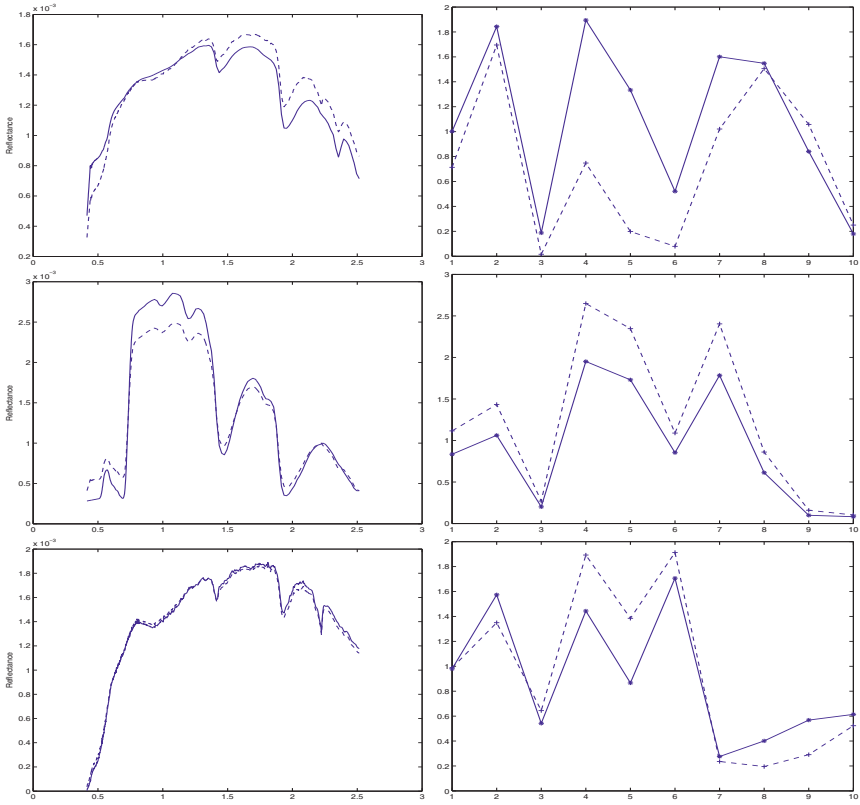


Fig. 2. On *the left*, columns of **W**. On *the right*, rows of **H**. On each plot: solid line for true values, dashed line for estimated values.

The minimization of $\mathcal{D}_{\text{reg}}$ w.r.t. the variable $\mathbf{Z}$ must take into account the regularization function $\mathcal{F}_1(\mathbf{W})$, Equation (4.1):

$$-\nabla_Z \mathcal{D}_{\text{reg}} = -\nabla_Z \mathcal{D} - \gamma_1 \nabla_Z \mathcal{F}_1, \tag{8.2}$$

and the minimization of $\mathcal{D}_{\text{reg}}$ w.r.t. the variable $\mathbf{T}$ must take into account the regularization function $\mathcal{F}_2(\mathbf{H})$, Equation (4.2).

$$-\nabla_T \mathcal{D}_{\text{reg}} = -\nabla_T \mathcal{D} - \gamma_2 \nabla_T \mathcal{F}_2. \tag{8.3}$$

In the following, we consider one regularization term at a time, that is, first on the spectra and then on the abundance coefficients.

8.1 Regularized SGM on the spectra $\mathbf{W}$

We develop in this section expressions of SGM for a regularization $\mathcal{F}_1$ on the normalized endmembers spectra $\mathbf{W}$, we have:

$$-\nabla_Z \mathcal{D}_{\text{reg}} = -\nabla_Z \mathcal{D} - \gamma_1 \nabla_Z \mathcal{F}_1, \tag{8.4}$$

$$-\nabla_T \mathcal{D}_{\text{reg}} = -\nabla_T \mathcal{D}. \tag{8.5}$$

The component of the opposite of the gradient of $\mathcal{D}_{\text{reg}}$ with respect to $\mathbf{Z}$ is:

$$\begin{aligned} -\frac{\partial \mathcal{D}_{\text{reg}}}{\partial [\mathbf{Z}]_{lj}} = \frac{1}{\sum_m [\mathbf{Z}]_{mj}} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{lj}} - \gamma_1 \frac{\partial \mathcal{F}_1}{\partial [\mathbf{W}]_{lj}} \right)_s - \\ \frac{1}{\sum_m [\mathbf{Z}]_{mj}} \sum_i [\mathbf{W}]_{ij} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{ij}} - \gamma_1 \frac{\partial \mathcal{F}_1}{\partial [\mathbf{W}]_{ij}} \right)_s \end{aligned} \tag{8.6}$$

In the same way that for the non regularized SGM, we solve both the split of the gradient between two positive functions and the conservation of the convexity w.r.t. the new variables by making the shift of the form:

$$\left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{ij}} - \gamma_1 \frac{\partial \mathcal{F}_1}{\partial [\mathbf{W}]_{ij}} \right)_s \longleftarrow \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{ij}} - \gamma_1 \frac{\partial \mathcal{F}_1}{\partial [\mathbf{W}]_{ij}} \right) + \eta + \epsilon \quad \forall (i,j)$$

with

$$\eta = -\min_{ij} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{ij}} - \gamma_1 \frac{\partial \mathcal{F}_1}{\partial [\mathbf{W}]_{ij}} \right).$$

The decomposition of the gradient of the regularized cost function w.r.t. $\mathbf{Z}$ is:

$$[-\nabla_Z \mathcal{D}_{\text{reg}}]_{ij} = [\mathcal{P}]_{ij} - [\mathcal{Q}]_{ij} \tag{8.7}$$

with

$$\begin{aligned} [\mathcal{P}]_{ij} =& \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{ij}} - \gamma_1 \frac{\partial \mathcal{F}_1}{\partial [\mathbf{W}]_{ij}} \right)_s, \\ [\mathcal{Q}]_{lj} = [\mathcal{Q}]_{.j} =& \sum_i [\mathbf{W}]_{.j} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{ij}} - \gamma_1 \frac{\partial \mathcal{F}_1}{\partial [\mathbf{W}]_{ij}} \right)_s \end{aligned} \tag{8.8}$$

and the iterate on $\mathbf{W}$ is:

$$[\mathbf{W}^{k+1}]_{lj} = [\mathbf{W}]^k_{lj} + \alpha^k [\mathbf{W}]^k_{lj} \left(\frac{\left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{lj}} - \gamma_1 \frac{\partial \mathcal{F}_1}{\partial [\mathbf{W}]_{lj}} \right)_s}{\sum_i [\mathbf{W}]_{ij} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{ij}} - \gamma_1 \frac{\partial \mathcal{F}_1}{\partial [\mathbf{W}]_{ij}} \right)_s} - 1 \right). \tag{8.9}$$

In the same way that for the non regularized SGM, with a constant step size equal to one, we get:

$$[\mathbf{W}^{k+1}]_{lj} = [\mathbf{W}]^k_{lj} \frac{\left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{lj}} - \gamma_1 \frac{\partial \mathcal{F}_1}{\partial [\mathbf{W}]_{lj}} \right)_s}{\sum_i [\mathbf{W}]_{ij} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{W}]_{ij}} - \gamma_1 \frac{\partial \mathcal{F}_1}{\partial [\mathbf{W}]_{ij}} \right)_s}. \tag{8.10}$$

The iterate on $\mathbf{H}$ is still given by Equation (4.17) or Equation (4.18) for a unit step size.

8.1.1 Tikhonov smoothness regularization

The well known Tikhonov regularization expresses some smoothness of the solution and is applied, here, on endmember spectra, *i.e.* on the columns of $\mathbf{W}$. This is justified by physical considerations, spectra varying slowly as a function of the wavelength. Consequently, the regularization function is:

$$\mathcal{F}_1(\mathbf{W}) = \frac{1}{2}\sum_{ij}([\mathbf{W}]_{ij} - c)^2 \tag{8.11}$$

with c a constant positive or zero, or more generally

$$\mathcal{F}_1(\mathbf{W}) = \frac{1}{2}\sum_{ij}[\partial_{1,2}\mathbf{W}]_{ij}^2 \tag{8.12}$$

where $\partial_{1,2}$ is the first or second-order derivative operator. For simplicity, we approximate $\partial_{1,2}\mathbf{W}$ in a closed numerical form as

$$[\partial_{1,2}\mathbf{W}]_{ij} = [\mathbf{W}]_{ij} - [\mathbf{A}\mathbf{W}]_{ij} \tag{8.13}$$

where $\mathbf{AW}$ stands for the convolution of each column of matrix $\mathbf{W}$ by a mask, *e.g.* $[1\ 0\ 0]$ and $[\frac{1}{2}\ 0\ \frac{1}{2}]$ for the first and second-order derivative operators, respectively. In this case, the opposite of the gradient can be expressed in matrix form as follows:

$$-[\nabla_{\mathbf{W}}\mathcal{F}_1]_{ij} = [(\mathbf{A} + \mathbf{A}^\top)\mathbf{W}]_{ij} - [(\mathbf{A}^\top\mathbf{A} + \mathbf{I})\mathbf{W}]_{ij}. \tag{8.14}$$

Note that Tikhonov regularization with the basic SGM algorithm was initially associated to the basic SGM algorithm in (Lantéri *et al.* 2011), *i.e.*, without flux constraint. The interested reader is invited to consult this reference for an overview of the results that have been obtained.

8.1.2 Simulations results

As for the non regularized SGM, many simulations have been performed to validate the proposed algorithm, Equations (8.10) and (4.18). Note that the different forms of the regularization term give approximatively the same practical results. The experiment corresponds to 10 linear mixtures of 3 endmembers, the length of each spectrum being 826. A noise vector distributed according to a Gaussian distribution with zero-mean and covariance matrix $\sigma^2\mathbf{I}_N$, where $\mathbf{I}_N$ is the $N \times N$ identity matrix has been added to each column of $\mathbf{V}$. Note that this statistical model assumes that the noise variances are the same in all bands. Results are given for a snr equal to $20dB$. Figure 3 shows the estimated endmembers (columns of $\mathbf{W}$) after 12 000 iterations, and compared with the true values with and without regularization. Figures 4 and 5 show the 10 reconstructed spectra in comparison with the true ones, respectively without and with regularization. We clearly see the interest of the regularization on the estimation.

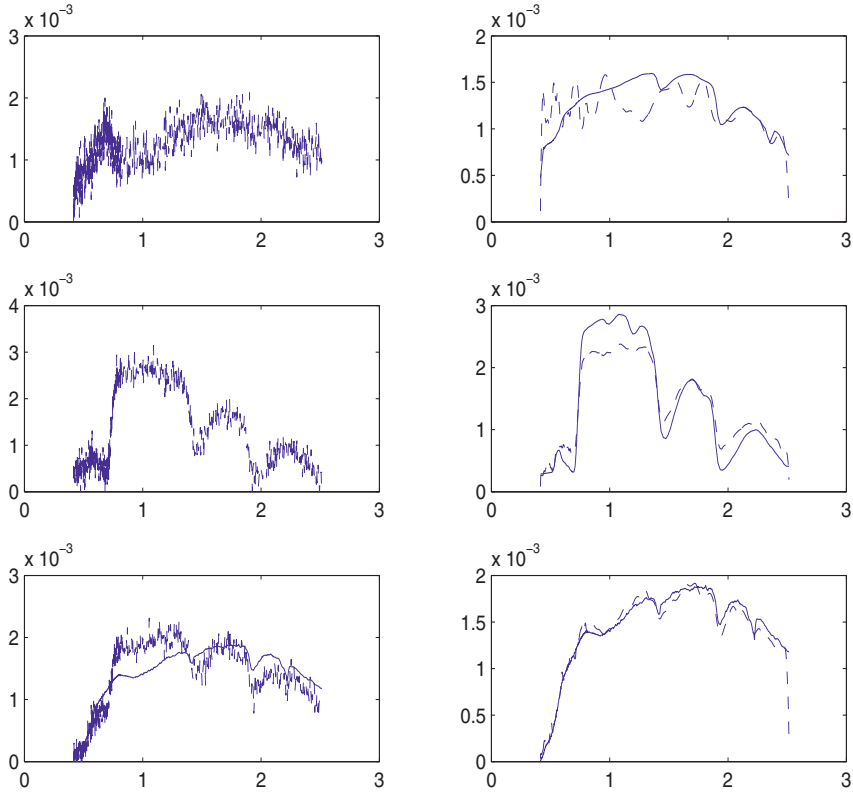


Fig. 3. Columns of $\mathbf{W}$. On each plot: solid line for true values, dashed line for estimated values. *Left column*: without regularization $\gamma = 0$. *Right column* with $\gamma = 0.1$.

8.2 Regularized SGM on the abundance coefficients $\mathbf{H}$

We develop in this section expressions of SGM for a regularization $\mathcal{F}_2$ on the normalized abundance coefficients $\mathbf{H}$, we have:

$$-\nabla_Z \mathcal{D}_{\text{reg}} = -\nabla_Z \mathcal{D} \tag{8.15}$$

$$-\nabla_T \mathcal{D}_{\text{reg}} = -\nabla_T \mathcal{D} - \gamma_2 \nabla_T \mathcal{F}_2. \tag{8.16}$$

In this case, the component of the opposite of the gradient of $\mathcal{D}_{\text{reg}}$ with respect to $\mathbf{T}$ is:

$$-\frac{\partial \mathcal{D}_{\text{reg}}}{\partial [\mathbf{T}]_{lj}} = \frac{1}{\sum_m [\mathbf{T}]_{mj}} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{lj}} - \gamma_2 \frac{\partial \mathcal{F}_2}{\partial [\mathbf{H}]_{lj}} \right)_s - \frac{1}{\sum_m [\mathbf{T}]_{mj}} \sum_i [\mathbf{H}]_{ij} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{ij}} - \gamma_2 \frac{\partial \mathcal{F}_2}{\partial [\mathbf{H}]_{ij}} \right)_s . \tag{8.17}$$

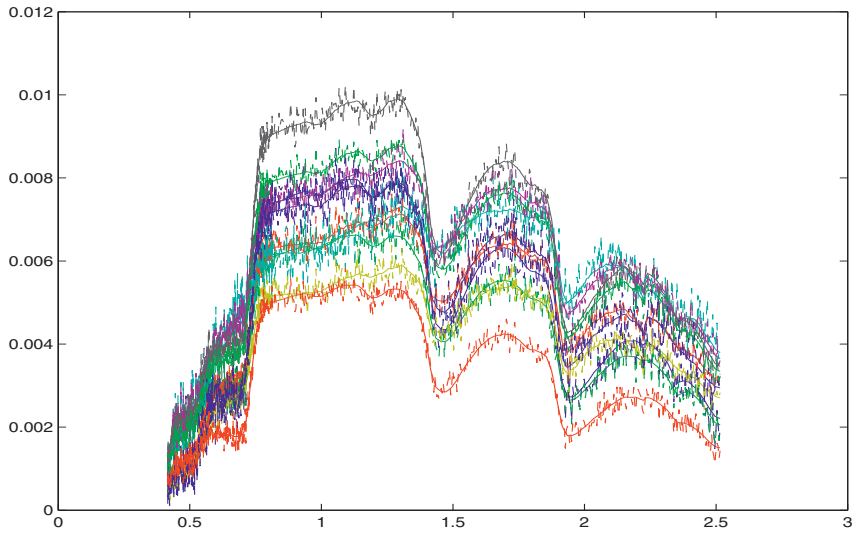


Fig. 4. Columns of $\mathbf{V}$, solid line for true values, dashed line for estimated values without regularization, $\gamma = 0$.

In the same way that for the non regularized SGM, we solve both the split of the gradient between two positive functions and the conservation of the convexity w.r.t. the new variables by making the shift of the form:

$$\left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{ij}} - \gamma_2 \frac{\partial \mathcal{F}_2}{\partial [\mathbf{H}]_{ij}}\right)_s \longleftarrow \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{ij}} - \gamma_2 \frac{\partial \mathcal{F}_2}{\partial [\mathbf{H}]_{ij}}\right) + \eta + \epsilon \quad \forall (i,j)$$

with

$$\eta = -\min_{ij}\left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{lj}} - \gamma_2 \frac{\partial \mathcal{F}_2}{\partial [\mathbf{H}]_{lj}}\right).$$

The decomposition of the gradient of the regularized cost function w.r.t. $\mathbf{T}$ is:

$$[-\nabla_T \mathcal{D}_{\text{reg}}]_{lj} = [\mathcal{R}]_{ij} - [\mathcal{S}]_{ij} \tag{8.18}$$

with

$$[\mathcal{R}]_{ij} = \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{ij}} - \gamma_2 \frac{\partial \mathcal{F}_2}{\partial [\mathbf{H}]_{lj}}\right)_s, \qquad [\mathcal{S}]_{ij} = \sum_i [\mathbf{H}]_{ij} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{ij}} - \gamma_2 \frac{\partial \mathcal{F}_2}{\partial [\mathbf{H}]_{ij}}\right)_s \tag{8.19}$$

and the iterate on $\mathbf{H}$ is:

$$[\mathbf{H}^{k+1}]_{lj} = [\mathbf{H}]_{lj}^k + \alpha^k [\mathbf{H}]_{lj}^k \left(\frac{\left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{lj}} - \gamma_2 \frac{\partial \mathcal{F}_2}{\partial [\mathbf{H}]_{lj}}\right)_s}{\sum_i [\mathbf{H}]_{ij} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{ij}} - \gamma_2 \frac{\partial \mathcal{F}_2}{\partial [\mathbf{H}]_{ij}}\right)_s} - 1\right). \tag{8.20}$$

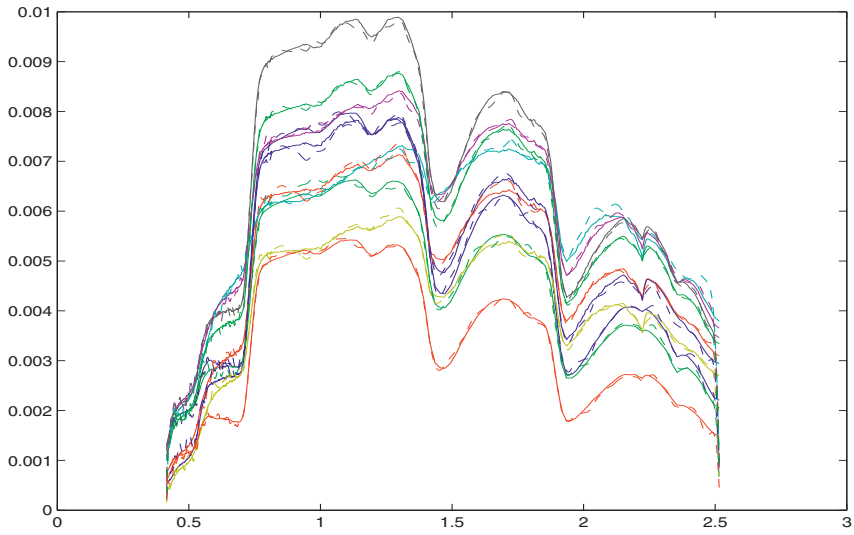


Fig. 5. Columns of $\mathbf{V}$, solid line for true values, dashed line for estimated values with $\gamma = 0.1$.

In the same way that for the non regularized SGM, with a constant step size equal to one, we get:

$$[\mathbf{H}^{k+1}]_{lj} = [\mathbf{H}]_{lj}^{k} \frac{\left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{lj}} - \gamma_2 \frac{\partial \mathcal{F}_2}{\partial [\mathbf{H}]_{lj}}\right)_s}{\sum_i [\mathbf{H}]_{ij} \left(-\frac{\partial \mathcal{D}}{\partial [\mathbf{H}]_{ij}} - \gamma_2 \frac{\partial \mathcal{F}_2}{\partial [\mathbf{H}]_{ij}}\right)_s}. \tag{8.21}$$

The iterate on $\mathbf{W}$ is still given by Equation (4.12) or Equation (4.13) for a unit step size.

8.2.1 Sparsity-enforcing regularization

Such a penalty, which expresses that most of information may be concentrated in a few coefficients, mainly applies to the abundance coefficients, that is, to the columns of $\mathbf{H}$. Keeping in mind that the algorithm satisfies flux conservation constraint, see (4.2), we are ready to consider the following sparsity measure σ introduced in (Hoyer (2004))

$$\sigma = \frac{\sqrt{K} - \frac{\|[\mathbf{H}]_{\bullet j}\|_1}{\|[\mathbf{H}]_{\bullet j}\|_2}}{\sqrt{K} - 1}, \qquad 0 \leq \sigma \leq 1 \tag{8.22}$$

with K the number of rows of $\mathbf{H}$, and $[\mathbf{H}]_{\bullet j}$ its j-th row. This clearly defines a relation between the ℓ_2-norm and the ℓ_1-norm of $[\mathbf{H}]_{\bullet j}$, the sum constraint on $\mathbf{H}$ associated with non negativity inducing a constant ℓ_1-norm.

$$\|[\mathbf{H}]_{\bullet j}\|_2^2 = \alpha^2 \|[\mathbf{H}]_{\bullet j}\|_1^2 \tag{8.23}$$

with

$$\alpha = \frac{1}{\sqrt{K} - \sigma(\sqrt{K}-1)}, \qquad \frac{1}{\sqrt{K}} \leq \alpha \leq 1. \tag{8.24}$$

Note that only two values for σ lead to unambiguous situations; if α is equal to one, only one entry of $[\mathbf{H}]_{\bullet j}$ is nonzero; if $\alpha = 1/\sqrt{K}$, all the entries of $[\mathbf{H}]_{\bullet j}$ are equal. Any other value for α can correspond to different sets of entries. As a consequence, we suggest to consider the following penalty function[2]:

$$\mathcal{F}_2(\mathbf{H}) = \frac{1}{2}\sum_j \left(\|[\mathbf{H}]_{\bullet j}\|_2^2 - \alpha^2\|[\mathbf{H}]_{\bullet j}\|_1^2\right)^2 \tag{8.25}$$

with α equal to one, and use of the regularization factor γ_2 in (8.1) to push $[\mathbf{H}]_{\bullet j}$ toward a sparse solution. For convenience, let us provide the opposite of the gradient of $\mathcal{F}_2(\mathbf{H})$

$$\begin{aligned} -[\nabla_{\mathbf{H}}\mathcal{F}_2]_{ij} = (\alpha^2\|[\mathbf{H}]_{\bullet j}\|_1^2 - \|[\mathbf{H}]_{\bullet j}\|_2^2) \\ ([\mathbf{H}]_{ij} - \alpha^2\|[\mathbf{H}]_{\bullet j}\|_1) \end{aligned} \tag{8.26}$$

to be used in (8.21). In the next section, we shall test this algorithm for hyperspectral data unmixing.

8.2.2 Simulations results

To test interest of sparsity regularization on the abundance coefficients, we take 20 linear mixtures of 6 endmembers, the length of each spectrum being 826. The six endmembers correspond to the construction concrete, green grass, micaceous loam, olive green paint, bare red brick and galvanized steel metal.

In order to characterize the performance of our approach, and show that it tends to provide sparse solutions, we considered a matrix $\mathbf{H}$ with only one nonzero entry per column. This entry was selected randomly and set to one. See Figure 6. Each observed spectrum was corrupted by an additive white Gaussian noise at a signal-to-noise ratio equal to 20 dB. The matrices $\mathbf{H}$ obtained for $\gamma_2 = 0$ and $\gamma_2 = 10^{-3}$, respectively, are presented in Figures 7 and 8.

We clearly observe that the sparsity-enforcing function allowed us to recover, in most cases, the endmembers involved in each observed spectrum. On the contrary, when no sparsity penalty term was used, all the entries of the estimated matrix $\mathbf{H}$ were nonzero. Finally, we checked that normalization of the columns of the matrix $\mathbf{W}$, as well as the flux conservation between $\mathbf{V}$ and $\mathbf{H}$, were satisfied at each iteration in both cases. On Figure 9, the behaviour of s_j is plotted as a function of the number of iterations, one see clearly that s_j tends to 1, whatever j after a small number of iterations.

[2]Using (4.2), note that $[\mathbf{H}]_{\bullet j}\|_1^2$ remains constant along iterations.

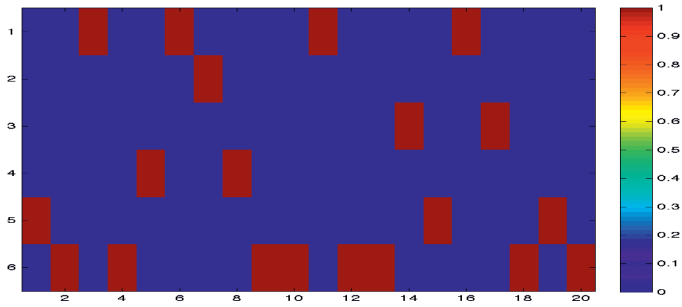


Fig. 6. True **H** with a sparsity $s_j = 1$, $\alpha = 1$.

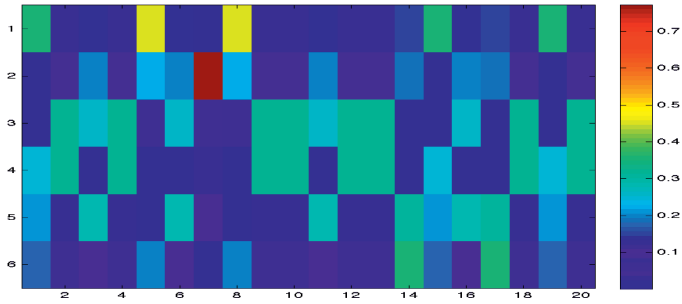


Fig. 7. Estimated **H** without sparsity constraint, $\mu = 0$.

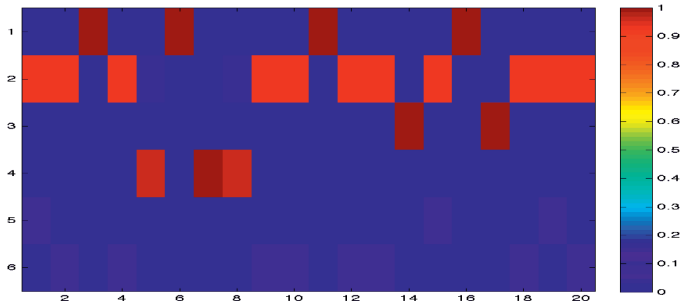


Fig. 8. Estimated **H** with a sparsity constraint, $\mu = 0.001$.

9 Conclusion

In this paper, we proposed a (split) gradient-descent method to solve the nonnegative matrix factorization problem subject to flux conservation constraints on each column of the estimated matrices. Tikhonov regularization and sparsity-enforcing

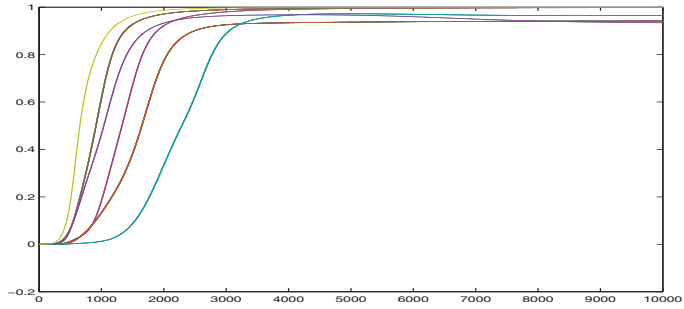


Fig. 9. s_j as a function of the number of iterations, $\mu = 0.001$.

regularization have been also considered. Application in the context of hyperspectral data unmixing shows the effectiveness and the interest of the proposed algorithms.

References

Chang, C.I., 2003, Hyperspectral Imaging: Techniques for Spectral Detection and Classification (New York: Plenum Publishing Co.)

Cichocki, A., Zdunek, R., & Amari, S., 2006, Csiszár's Divergences for Non-negative Matrix Factorization: Family of New Algorithms, Ser. Lectures Notes in Computer Science (Springer Berlin/ Heidelberg), Vol. 3889

Daube-Witherspoon, M.E., & Muehllehner, G., 1986, IEEE Trans. Medical Imaging, 5, 61

Dempster, A.D., Laird, N.M., & Rubin, D.B., 1977, J. R. Stat. Soc., B 39, 1

Desidera, G., Anconelli, B., Bertero, M., Boccacci, P., & Carbillet, M., 2006, "Application of iterative blind deconvolution to the reconstruction of lbt linc-nirvana images", A&A

Févotte, C., Bertin, N., & Durrieu, J.-L., 2009, "Nonnegative matrix factorization with the itakura-saito divergence, with application to music analysis", Neural Computation

Hoyer, P.O., 2004, J. Machine Learning, 5, 1457

Landgrebe, D.A., 2003, Signal Theory Methods in Multispectral Remote Sensing (New York: Wiley)

Lantéri, H., Roche, M., Cuevas, O., & Aime, C., 2001, Signal Processing, 54, 945

Lantéri, H., Roche, M., & Aime, C., 2002, Inverse Probl., 18, 1397

Lantéri, H., Theys, C., Benvenuto, F., & Mary, D., 2009, "Méthode algorithmique de minimisation de fonctions d'écart entre champs de données, application à la reconstruction d'images astrophysiques", in GRETSI

Lantéri, H., Aime, C., Beaumont, H., & Gaucherel, P., 1994, "Blind deconvolution using the richardson-lucy algorithm", in European Symposium on Satellite and Remote Sensing

Lantéri, H., Theys, C., & Richard, C., 2011, "Regularized split gradient method for non negative matrix factorization", in ICASSP, Prague

Lantéri, H., Theys, C., & Richard, C., 2011, "Nonnegative matrix factorization with regularization and sparsity-enforcing terms", in CAMSAP, Porto Rico

Lee, D.D., & Seung, H.S., 2001, Adv. NIPS, 13, 556

Lucy, L.B., 1974, AJ, 79, 745

Richardson, W.H., 1972, J. Opt. Soc. Am., 1, 55

RSI (Research Systems Inc.), 2003, ENVI User's guide Version 4.0, Boulder, CO 80301 USA, Sep.

Theys, C., Dobigeon, N., Tourneret, J.-Y., & Lantéri, H., 2009, "Linear unmixing of hyperspectral images using a scaled gradient method", in SSP, Cardiff

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 381–401*

MCMC ALGORITHMS FOR SUPERVISED AND UNSUPERVISED LINEAR UNMIXING OF HYPERSPECTRAL IMAGES

N. Dobigeon[1], S. Moussaoui[2], M. Coulon[1], J.-Y. Tourneret[1] and A.O. Hero[3]

Abstract. In this paper, we describe two fully Bayesian algorithms that have been previously proposed to unmix hyperspectral images. These algorithms relies on the widely admitted linear mixing model, *i.e.* each pixel of the hyperspectral image is decomposed as a linear combination of pure endmember spectra. First, the unmixing problem is addressed in a supervised framework, *i.e.*, when the endmembers are perfectly known, or previously identified by an endmember extraction algorithm. In such scenario, the unmixing problem consists of estimating the mixing coefficients under positivity and additivity constraints. Then the previous algorithm is extended to handle the unsupervised unmixing problem, *i.e.*, to estimate the endmembers and the mixing coefficients jointly. This blind source separation problem is solved in a lower-dimensional space, which effectively reduces the number of degrees of freedom of the unknown parameters. For both scenarios, appropriate distributions are assigned to the unknown parameters, that are estimated from their posterior distribution. Markov chain Monte Carlo (MCMC) algorithms are then developed to approximate the Bayesian estimators.

1 Abstract

For several decades, hyperspectral imagery has been demonstrating its high interest in numerous research works devoted to Earth monitoring. This interest can be easily explained by the high spectral resolution of the images provided by the

[1] University of Toulouse, IRIT/INP-ENSEEIHT, 2 rue Camichel, BP. 7122, 31071 Toulouse Cedex 7, France
[2] IRCCyN - CNRS UMR 6597, ECN, 1 rue de la Noë, BP. 92101, 44321 Nantes Cedex 3, France
[3] University of Michigan, Department of EECS, 1301 Beal Avenue, Ann Arbor, 48109-2122, USA

DOI: 10.1051/eas/1359017

hyperspectral sensors. For instance, hyperspectral images can provide automatic classification maps for mineralogic surveys, avoiding long and tedious sampling campaigns (Jackson & Landgrebe 2002; Rellier *et al.* 2004). When environmental issues are on the front of the stage, hyperspectral imaging enables to provide crucial information related to macroscopic parameters, *e.g.*, the status of ecosystems or plants. Obviously, the price to pay for extracting the information contained in these images is to develop new methods exploiting the data provided by hyperspectral sensors efficiently.

Since the first hyperspectral images were acquired, spectral unmixing has been of considerable interest, not only in the remote sensing community, but also in the signal and image processing community. Solving this problem can indeed provide answers to various important issues such as classification (Chang 2003), material quantification (Plaza *et al.* 2005) and sub-pixel detection (Manolakis *et al.* 2001). Spectral unmixing consists of decomposing each pixel of the observed scene into a collection of reference spectra, usually referred to as endmembers, and estimating their proportions, or abundances, in each pixel (Bioucas-Dias *et al.* 2012). To formally describe the mixture, the most frequently encountered model is the macroscopic model which gives a good approximation of the nonlinear model introduced by Hapke (Hapke 1981) in the reflective spectral domain from visible to near-infrared (0.4 μm to 2.5 μm) (Johnson *et al.* 1983). This linear model assumes that the observed pixel spectrum is a weighted linear combination of the endmember spectra.

As noticed in (Keshava & Mustard 2002), linear spectral unmixing has often been handled as a two-step procedure: the endmember extraction step and the inversion step, respectively. In the first step of the analysis, the macroscopic materials that are present in the observed scene are identified by using an endmember extraction algorithm (EEA). The most popular EEAs include pixel purity index (PPI), N-FINDR (Winter 1999), and more recently the VCA algorithm (Nascimento & Bioucas-Dias 2005a) which proposes to recover the vertices of the biggest simplex in the observed data. A common assumption in these EEAs is that they require the presence of pure pixels in the observed image. Conversely, (Craig 1994) and (Bowles *et al.* 1995) proposed minimum volume transforms to recover the smallest simplex that contains all the dataset.

The second step of spectral unmixing is devoted to the abundance estimation. These abundances have to ensure constraints inherent to hyperspectral imagery: as they represent proportions, the abundances have to satisfy positivity and additivity constraints. Several algorithms proposed in the literature to solve this inversion step rely on constrained optimization techniques (Heinz & Chang 2001; Theys *et al.* 2009; Tu *et al.* 1998).

This paper studies alternatives based on Bayesian inference for supervised and unsupervised unmixing problems. In the first part of this work, the endmembers are assumed to be previously identified, *e.g.*, using *a priori* knowledge regarding the observed scene or using results provided by an EEA. In this case, the unmixing algorithm performs the inversion step, *i.e.*, it estimates the abundance coefficients under positivity and additivity constraints. In a second part of this paper, we

introduce a spectral unmixing algorithm in a fully unsupervised framework to estimate the pure component spectra and their proportions jointly.

In both frameworks, Bayesian formulation allows the constraints within the model to be satisfied. Indeed, appropriate prior distributions are chosen to take into account the positivity and additivity of the abundances, as well as the positivity of the endmember spectra. To overcome the complexity of the posterior distribution, Markov chain Monte Carlo algorithms are proposed to approximate the standard minimum mean squared error estimator. Moreover, as the full posterior distribution of all the unknown parameters is available, confidence intervals can be easily computed. These measures allow the accuracy of the different estimates to be quantified.

2 Linear mixing model and problem statement

Let consider P pixels of an hyperspectral image acquired in L spectral bands. According to the linear mixing model, described for instance in (Bioucas-Dias *et al.* 2012), the observed spectrum $\mathbf{y}_p = [y_{p,1}, \ldots, y_{p,L}]^T$ of the pth pixel ($p = 1, \ldots, P$) is written as an the linear combination of R spectral signatures $\mathbf{m}_r$, corrupted by an additive noise $\mathbf{n}_p$:

$$\mathbf{y}_p = \sum_{r=1}^{R} \mathbf{m}_r a_{p,r} + \mathbf{n}_p, \tag{2.1}$$

where $\mathbf{m}_r = [m_{r,1}, \ldots, m_{r,L}]^T$ is the pure spectrum that is characteristic of the rth material and $a_{p,r}$ is the abundance of the rth material in the pth pixel. Moreover, in (2.1), $\mathbf{n}_p = [n_{p,1}, \ldots, n_{p,L}]^T$ is an noise sequence whose components are assumed to be independent and identically distributed (i.i.d.) according to a centered Gaussian distribution with covariance matrix[4] $\mathbf{\Sigma}_{\mathrm{n}} = \sigma^2 \mathbf{I}_L$, where $\mathbf{I}_L$ is the identity matrix of size $L \times L$

$$\mathbf{n}_p | \sigma^2 \sim \mathcal{N}(\mathbf{0}_L, \mathbf{\Sigma}_{\mathrm{n}}). \tag{2.2}$$

Due to physical considerations (Keshava & Mustard 2002), the abundance vectors $\mathbf{a}_p = [a_{p,1}, \ldots, a_{p,R}]^T$ in (2.1) satisfy the following positivity and additivity constraints

$$\begin{cases} a_{p,r} \geq 0, \ \forall r = 1, \ldots, R, \\ \sum_{r=1}^{R} a_{p,r} = 1. \end{cases} \tag{2.3}$$

In other words, the P abundance vectors belong to the space

$$\mathcal{A} = \{\mathbf{a}_p : \ \|\mathbf{a}\|_1 = 1 \text{ and } \mathbf{a}_p \succeq \mathbf{0}\}, \tag{2.4}$$

where $\|\cdot\|_1$ is the ℓ_1 norm such that $\|\mathbf{x}\|_1 = \sum_i |x_i|$, and $\mathbf{a}_p \succeq \mathbf{0}$ stands for the set of inequalities $\{a_{p,r} \geq 0\}_{r=1,\ldots,R}$. In addition, the spectral signatures $\mathbf{m}_r$

[4]The proposed model can be easily extended to more complex noise models, following for instance (Dobigeon *et al.* 2008a).

correspond to reflectance measures and, as a consequence, need to ensure the positivity constraints

$$m_{r,l} \geq 0, \ \forall r = 1, \ldots, R, \ \forall l = 1, \ldots, L. \tag{2.5}$$

If we consider all the pixels in the hyperspectral image, the set of Equations (2.1) can be rewritten using the following matrix notations

$$\mathbf{Y} = \mathbf{MA} + \mathbf{N} \tag{2.6}$$

where $\mathbf{Y}$ is a $L \times P$ matrix that contains all the observations associated with the image pixels, $\mathbf{M}$ is the $L \times R$ matrix of the spectral signatures, $\mathbf{A}$ is the $R \times P$ matrix of the abundances and $\mathbf{N}$ is a $L \times P$ matrix of the noise vectors

$$\begin{array}{lll} \mathbf{Y} & = [\mathbf{y}_1, \ldots, \mathbf{y}_P], & \mathbf{M} & = [\mathbf{m}_1, \ldots, \mathbf{m}_R], \\ \mathbf{A} & = [\mathbf{a}_1, \ldots, \mathbf{a}_P], & \mathbf{N} & = [\mathbf{n}_1, \ldots, \mathbf{n}_P]. \end{array} \tag{2.7}$$

This paper proposes a Bayesian approach to first estimate the abundance coefficient under the constraints (2.3) when the spectral signatures are known. Then, the spectra of the pure components will be assumed unknown and will be included within the estimation procedure.

3 Supervised unmixing: The spectral components are known

When the pure spectral components (also known as *endmembers*) are perfectly known, the problem of linear unmixing reduces to the inversion step, *i.e.*, the constrained estimation of the abundances. This problem can be formulated as a linear regression under constraints whose resolution can be conducted within a Bayesian framework. Indeed, Bayesian models are very convenient in such situation since the constraints are conveniently handled when defining a priori distributions for the unknown parameters. Several constraints have been studied in the literature, including monotony (Chen & Deely 1996), positivity (Moussaoui *et al.* 2006) or sparsity (Blumensath & Davies 2007; Févotte & Godsill 2006). Constraints inherent to hyperspectral imagery are positivity and additivity, as explained in Section 2. In what follows, the Bayesian model to solve the supervised spectral unmixing model is described. Note that in this supervised scenario, spectral unmixing is conducted pixel-by-pixel. As consequence, to lighten the notations, the dependence of the quantity $\mathbf{y}_p$, $\mathbf{a}_p$, $\mathbf{c}_p$ on the pixel p will be omitted.

3.1 Bayesian model

3.1.1 Likelihood

The linear mixing model defined by (2.1) and the statistical properties (2.2) of the noise vector $\mathbf{n}$ lead to a Gaussian distribution for the observed spectrum for the pth pixel:

$$\mathbf{y}|\mathbf{a}, \sigma^2 \sim \mathcal{N}\left(\mathbf{Ma}, \sigma^2 \mathbf{I}_L\right). \tag{3.1}$$

As a consequence, the likelihood function of the vector $\mathbf{y}$ can be written

$$f\left(\mathbf{y}|\mathbf{a},\sigma^2\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{L}{2}} \exp\left[-\frac{\|\mathbf{y}-\mathbf{M}\mathbf{a}\|^2}{2\sigma^2}\right], \tag{3.2}$$

where $\|\mathbf{x}\| = \left(\mathbf{x}^T\mathbf{x}\right)^{\frac{1}{2}}$ is the ℓ_2-norm of the vector $\mathbf{x}$.

3.1.2 Parameter prior distributions

When the spectra of the pure components $\mathbf{m}_1,\ldots,\mathbf{m}_R$ are known, the vector of unknown vector denoted as $\boldsymbol{\theta}$ is composed of the abundance vector and the noise variance $\boldsymbol{\theta} = \left\{\mathbf{a},\sigma^2\right\}$.

Abundance coefficients. For each pixel p, thanks to the additivity constraints enounced in (2.3), the abundance vector $\mathbf{a}$ can be rewritten[5]

$$\mathbf{a} = \begin{bmatrix} \mathbf{c} \\ a_R \end{bmatrix} \qquad \text{with} \qquad \mathbf{c} = \begin{bmatrix} a_1 \\ \vdots \\ a_{R-1} \end{bmatrix}, \tag{3.3}$$

and $a_R = 1 - \sum_{r=1}^{R-1} a_r$. According to the model proposed in (Dobigeon *et al.* 2008b), the prior distribution chosen for $\mathbf{c}$ is a uniform distribution defined on the simplex $\mathcal{S}$

$$\mathcal{S} = \left\{\mathbf{c};\ \|\mathbf{c}\|_1 \leq 1 \text{ and } \mathbf{c} \succeq \mathbf{0}\right\}. \tag{3.4}$$

Choosing this prior distribution for $\mathbf{c}$ is fully equivalent of choosing a Dirichlet prior $\mathcal{D}\left(1,\ldots,1\right)$ for $\mathbf{a}$, *i.e.*, a uniform distribution on the the set $\mathcal{A}$ of admissible values for $\mathbf{a}$ (defined by (2.4)) (Robert 2007, Appendix A).

Noise variance. A conjugate inverse-gamma distribution is chosen as a prior distribution for the noise variance σ^2

$$\sigma^2\,|\nu,\gamma \sim \mathcal{IG}\left(\frac{\nu}{2},\frac{\gamma}{2}\right), \tag{3.5}$$

where $\mathcal{IG}\left(\frac{\nu}{2},\frac{\gamma}{2}\right)$ is an inverse-gamma distribution of parameters $\frac{\nu}{2}$ and $\frac{\gamma}{2}$. This distribution has been successfully used in several works of the literature, *e.g.*, (Punskaya *et al.* 2002) and (Dobigeon *et al.* 2007). As in the references above, the hyperparameter ν will be fixed to $\nu = 2$.

Moreover, γ is an hyperparameter assumed to be unknown, and a non-informative Jeffreys' distribution is chosen as prior distribution (Jeffreys 1961)

$$f\left(\gamma\right) \propto \frac{1}{\gamma}\mathbf{1}_{\mathbb{R}^+}\left(\gamma\right), \tag{3.6}$$

where $\propto$ stands for "proportional to".

[5]For writing conciseness, the last component of $\mathbf{a}$ will be always expressed as a function of the others. Note however that in the algorithm described in the following section, the discarded component can be randomly chosen at each iteration of the Gibbs sampler.

3.1.3 Posterior distribution

The posterior distribution of the unknown parameter vector $\boldsymbol{\theta} = \left\{\mathbf{c}, \sigma^2\right\}$ is computed from the following hierarchical model

$$f(\boldsymbol{\theta}|\mathbf{y}) \propto \int f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\gamma) f(\gamma) d\gamma, \tag{3.7}$$

where $f\left(\mathbf{y}|\boldsymbol{\theta}\right)$ and $f\left(\gamma\right)$ are defined in (3.2) and (3.6), respectively. By assuming prior independence between σ^2 and $\mathbf{c}$, *i.e.* $f\left(\boldsymbol{\theta}|\gamma\right) = f\left(\mathbf{c}\right) f(\sigma^2|\gamma)$, the hyperparameter γ can be integrated out from the joint distribution $f\left(\boldsymbol{\theta}, \gamma|\mathbf{y}\right)$ in (3.7), which leads to

$$f\left(\mathbf{c}, \sigma^2|\mathbf{y}\right) \propto \frac{1}{\sigma^{L+2}} \exp\left[-\frac{\|\mathbf{y} - \mathbf{M}\mathbf{a}\|^2}{2\sigma^2}\right] \mathbf{1}_{\mathcal{S}}\left(\mathbf{c}\right). \tag{3.8}$$

Note that this posterior distribution is defined on the simplex $\mathcal{S} \times \mathbb{R}^+$, *i.e.*, $\mathbf{c}$ satisfies the constraints resulting from the positivity and additivity constraints of $\mathbf{a}$. We introduce in the following section a Gibbs sampler that allows samples to be generated according to the joint distribution $f\left(\mathbf{c}, \sigma^2|\mathbf{y}\right)$.

3.2 Gibbs sampler

Samples (denoted as $\cdot^{(t)}$ where t is the iteration index) can be generated according to $f\left(\mathbf{c}, \sigma^2|\mathbf{y}\right)$ thanks to a Gibbs sampler described below. It successively generates samples according to the conditional distributions $f\left(\mathbf{c}|\sigma^2, \mathbf{y}\right)$ and $f(\sigma^2|\mathbf{c}, \mathbf{y})$. The main steps of this algorithm are detailed below and are summarized by Algo. 1. The interested reader can refer to (Robert & Casella 1999) for more details on MCMC methods.

Algorithm 1 Gibbs sampler for supervised unmixing

1: *% Initialization*
2: Sampling the parameters $\widetilde{\sigma}^{2(0)}$ and $\widetilde{\mathbf{c}}^{(0)}$ from the prior distributions defined in Section 3.1.2,
3: *% Iterations*
4: **for** $t = 1, 2, \ldots,$ **do**
5: Sampling $\widetilde{\mathbf{c}}^{(t)}$ according to the distribution (3.11),
6: Sampling $\widetilde{\sigma}^{2(t)}$ according to the distribution (3.12),
7: **end for**

3.2.1 Sampling according to $f\left(\mathbf{c}|\sigma^2, \mathbf{y}\right)$

The conditional posterior distribution of the partial abundance vector is

$$f\left(\mathbf{c}\left|\sigma^2, \mathbf{y}\right.\right) \propto \exp\left[-\frac{\left(\mathbf{c} - \boldsymbol{\upsilon}\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{c} - \boldsymbol{\upsilon}\right)}{2}\right] \mathbf{1}_{\mathcal{S}}\left(\mathbf{c}\right), \tag{3.9}$$

where

$$\begin{cases} \boldsymbol{\Sigma} = \left[\left(\mathbf{M}_{\text{-}R} - \mathbf{m}_R \mathbf{1}_{R-1}^T \right)^T \boldsymbol{\Sigma}_{\text{n}}^{-1} \left(\mathbf{M}_{\text{-}R} - \mathbf{m}_R \mathbf{1}_{R-1}^T \right) \right]^{-1}, \\ \boldsymbol{\upsilon} = \boldsymbol{\Sigma} \left[\left(\mathbf{M}_{\text{-}R} - \mathbf{m}_R \mathbf{1}_{R-1}^T \right)^T \boldsymbol{\Sigma}_{\text{n}}^{-1} \left(\mathbf{y} - \mathbf{m}_R \right) \right], \end{cases} \tag{3.10}$$

with $\boldsymbol{\Sigma}_{\text{n}}^{-1} = \frac{1}{\sigma^2}\mathbf{I}_L$, $\mathbf{1}_{R-1} = [1, \ldots, 1]^T \in \mathbb{R}^{R-1}$ and where $\mathbf{M}_{\text{-}R}$ is the matrix $\mathbf{M}$ whose Rth column has been removed. As a consequence, the vector $\mathbf{c}|\sigma^2, \mathbf{y}$ is distributed according to a multivariate Gaussian distribution truncated on the simplex $\mathcal{S}$ defined by (3.4)

$$\mathbf{c}\,|\sigma^2, \mathbf{y} \sim \mathcal{N}_{\mathcal{S}}\left(\boldsymbol{\upsilon}, \boldsymbol{\Sigma}\right). \tag{3.11}$$

Sampling according to this truncated Gaussian distribution can be conducted following the strategy described in (Dobigeon & Tourneret 2007).

3.2.2 Sampling according to $f(\sigma^2|\mathbf{c}, \mathbf{y})$

By looking at the joint distribution $f\left(\sigma^2, \mathbf{c}|\mathbf{y}\right)$, it can be stated that the conditional distribution of $\sigma^2|\mathbf{c}, \mathbf{y}$ is the following inverse-gamma distribution

$$\sigma^2|\mathbf{c}, \mathbf{y} \sim \mathcal{IG}\left(\frac{L}{2}, \frac{\|\mathbf{y} - \mathbf{M}\mathbf{a}\|^2}{2}\right). \tag{3.12}$$

3.3 Simulation results on synthetic data

To illustrate the algorithm performance, a synthetic mixture of $R = 3$ pure components is generated. These spectral signatures are extracted from the library provided with the ENVI software (RSI (Research Systems Inc.) 2003, p. 1035) and are characteristics of a urban or sub-urban scene: construction concrete, green grass and micaceous loam. The mixing coefficients are defined as $a_1 = 0.3$, $a_2 = 0.6$ and $a_2 = 0.1$. The observed spectrum has been corrupted by an additive Gaussian noise with variance $\sigma^2 = 0.025$, which corresponds to a signal-to-noise ratio RSB $\approx$ 15dB where RSB $= L^{-1}\sigma^{-2}\left\|\sum_{r=1}^{R} \mathbf{m}_r a_r\right\|^2$. The endmembers and the resulting observed spectrum are represented in Figure 1.

Figure 2 shows the posterior distributions of the abundance coefficients a_r $(r = 1, 2, 3)$ estimated by the proposed Gibbs sampler for $N_{\text{MC}} = 20\,000$ iterations (with $N_{\text{bi}} = 100$ burn-in iterations). These distributions are in good agreement with the actual values of the coefficients $\mathbf{a} = [0.3, 0.6, 0.1]^T$. As a comparison, the results obtained with the FCLS algorithm (Chang & Ji 2001; Heinz & Chang 2001) have been also depicted in this figure for N_{MC} Monte Carlo simulations (*i.e.*, for N_{MC} realizations of the noise sequence).

3.4 Results on real data

This paragraph presents the analysis of an hyperspectral image that has been widely studied in the literature (Akgun *et al.* 2005; Chen 2005; Christophe *et al.* 2005; Tang & Pearlman 2004). This image, depicted in Figure 3, is initially

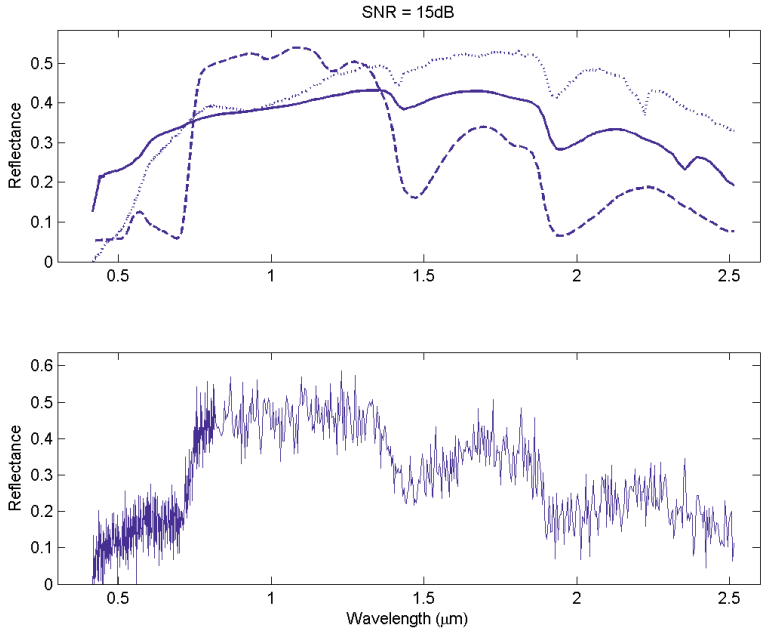


Fig. 1. *Top*: endmember spectra: construction concrete (line), green grass (dashed line), loam (dotted line). *Bottom*: spectrum of the observed pixel.

composed of 189 spectral bands (when the water absorption bands have been removed). It has been acquired by the spectro-imager AVIRIS (Jet Propulsion Lab. (JPL) 2006) in 1997 over Moffett Field, CA. It is composed of a lake and a coastal area. The spectral unmixing algorithm has been applied on a 50×50 scene. The analyzed image area is depicted in Figure 3.

3.4.1 Endmember identification

First, the pure materials that are present in the image have been identified. Since no prior knowledge is available for the analyzed scene, an endmember extraction algorithm has been used to recover to identify the endmember spectra. More precisely, N-FINDR (Winter 1999) has been used to identify $R = 3$ endmembers that are represented in Figure 4: vegetation, water and soil. Note that the number of endmembers has been determined by a principal component analysis, as explained in (Keshava & Mustard 2002).

3.4.2 Abundance estimation

The supervised unmixing algorithm introduced in Sections 3.1 and 3.2 has been applied on each pixel of the AVIRIS hyperspectral image using the endmembers

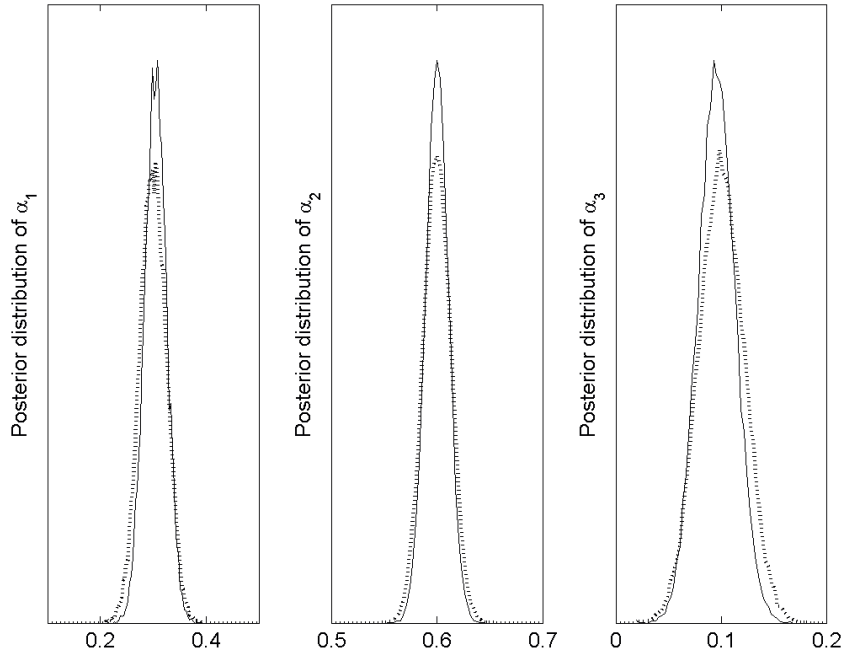


Fig. 2. Posterior distributions of the abundance coefficients $[a_1, a_2, a_3]^T$ estimated by the proposed algorithm (continuous lines) and histograms of the estimated values by FCLS (dashed lines).

previously identified. The abundance maps estimated by the proposed algorithm for the $R = 3$ materials are depicted in Figure 5 (top). Note that a white (resp. black) pixel corresponds to a high (resp. low) proportion of the corresponding material. The lake area (that appears as white pixels on the water abundance map) has been clearly recovered. The results obtained with the unmixing algorithm provided with the ENVI software (RSI (Research Systems Inc.) 2003, p. 739) are also depicted in Figure 5 (bottom). These results obtained with constrained least square algorithm are in good agreement with those of Figure 5 (top). Note however that the proposed algorithm also allows posterior distributions to be estimated. These posterior distributions can be useful to derive confidence intervals.

4 Unsupervised unmixing

As explained in (Bioucas-Dias *et al.* 2012; Keshava & Mustard 2002), linear spectral unmixing has been often addressed in a two-step procedure: i) endmember identification by an EEA and ii) abundance estimation. However, solving the unmixing problem in two distinct and successive steps may lead to poor estimation performance. In particular, when no pure pixels are present in the image,

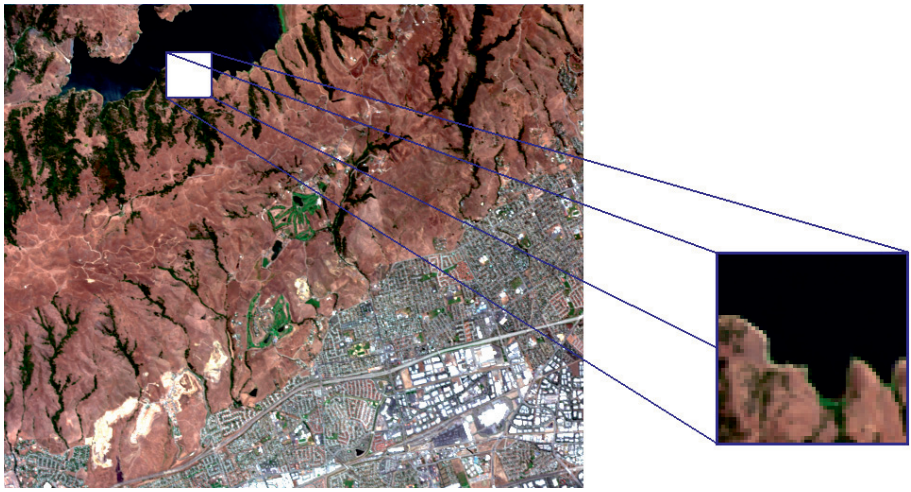


Fig. 3. Real hyperspectral image acquired by AVIRIS over Moffett Field in 1997 (*left*) and the region of interest (*right*).

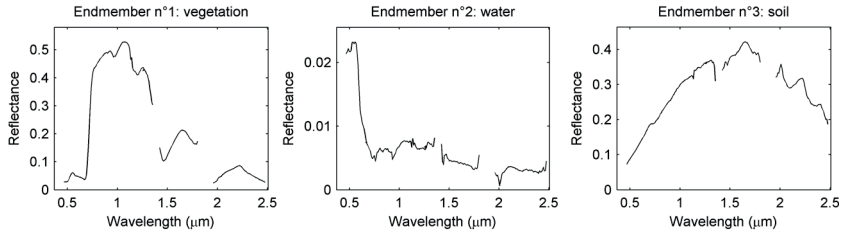


Fig. 4. The $R = 3$ endmember spectra obtained by N-FINDR.

the geometric EEA such as VCA, N-FINDR or PPI provide inadequate endmember estimates. To overcome this issue, we propose to solve the linear unmixing problem in a fully unsupervised Bayesian framework, by estimating the endmember spectra and the corresponding abundances jointly. This approach casts the unmixing problem as an blind source separation (BSS), that as received a huge interest in the signal processing literature. In particular, it is well known that independent component analysis (ICA) (Hyvärinen *et al.* 2001) is a powerful solution of BSS problems. However, as noticed in (Nascimento & Bioucas-Dias 2005b) and (Dobigeon & Achard 2005), ICA-based algorithms fails to solve the unmixing problem, mainly due to the high correlation between the source signals. Other strategies, based on non-negative matrix factorization techniques (Paatero & Tapper 1994), can be used to jointly estimate the endmember spectra and the abundance coefficients. However, these algorithms do not take explicitly into account the sum-to-one constraint on the abundance coefficients. Conversely, the Bayesian framework is a convenient way to ensure all the constraints (positivity

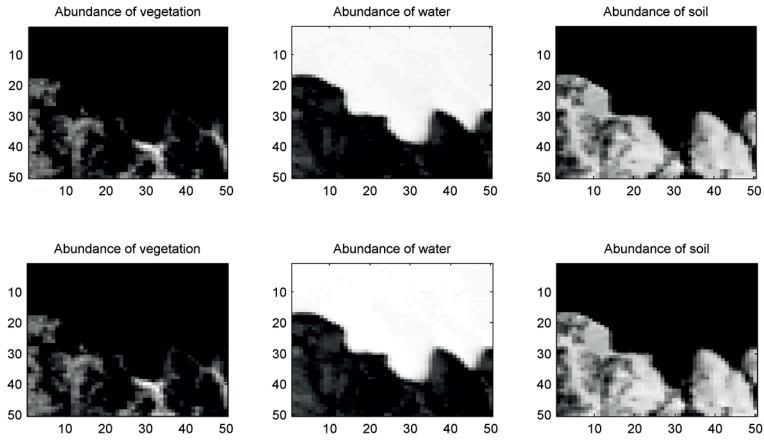


Fig. 5. *Top*: abundance maps estimated by the proposed algorithm. *Bottom*: abundance maps estimated by the unmixing routine provided with the ENVI software.

on the abundance coefficients and endmember spectra, additivity on the abundance coefficients) by defining appropriate prior distributions for the unknown parameters.

Moreover, a geometrical interpretation of the linear unmixing problem allows one to show that the spectral signatures can be estimated in an appropriate lower-dimensional subspace. This estimation in a subspace allows the number of degree of freedom to be significantly reduced for the parameters, while ensuring the physical constraints.

4.1 Bayesian model

Unsupervised spectral unmixing can be formulated as a blind source separation problem. Thus, the joint estimation of the endmembers and the abundances requires to consider all the image pixels $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_P]$ during the analysis. From a pixel-wise analysis in Section 3.1, spectral unmixing is now conducted on a whole hyperspectral image. More specifically, the previous Bayesian model introduced in 3.1 is extended by defining a prior model for the endmember spectra. The new posterior distribution associated with the new set of unknown parameters is finally derived.

4.1.1 Likelihood function

By assuming the independence of the noise vector, the new likelihood function associated with the observed pixel matrix $\mathbf{Y}$ is the product of the marginal likelihood functions (3.1.1)

$$f\left(\mathbf{Y}|\mathbf{M},\mathbf{C}\right) = \prod_{p=1}^{P} f\left(\mathbf{y}_p|\mathbf{M},\mathbf{c}_p\right)$$

where $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_P]^T$ is a matrix coming from the reparametrization (3.3) of the abundance vectors and $f(\mathbf{y}_p|\mathbf{M}, \mathbf{c}_p)$ has been defined in (3.1.1).

4.1.2 Prior of the abundance coefficients

The coefficient vectors $\mathbf{c}_1, \ldots, \mathbf{c}_P$ are assumed to be *a priori* independent. Thus, the prior distribution for the coefficient matrix $\mathbf{C}$ can be written as the product of the prior chosen in paragraph 3.1.2

$$f(\mathbf{C}) = \prod_{p=1}^{P} f(\mathbf{c}_p)$$

with

$$f(\mathbf{c}_p) \propto \mathbf{1}_{\mathcal{S}}(\mathbf{c}_p)$$

where $\mathcal{S}$ has been defined in (3.4). This prior allows the constraints inherent to the linear mixing model to be ensured. Moreover, this prior has the great advantage of imposing a constraint on the size of the simplex spanned by the endmembers in the hyperspectral space. Indeed, as demonstrated in (Dobigeon *et al.* 2009), among two admissible solutions for the unmixing problem, this prior will favor the solution that corresponds to the simplex of minimum volume. Note that this property has been exploited also in (Arngren *et al.* 2011; Bowles *et al.* 1995; Craig 1994).

4.1.3 Prior model for the endmembers

Dimensionality reduction. First, notice that the set

$$\mathcal{S}_{\mathbf{M}} = \left\{ \mathbf{x} \in \mathbb{R}^L;\ \mathbf{x} = \sum_{r=1}^{R} \lambda_r \mathbf{m}_r,\ \sum_{r=1}^{R} \lambda_r = 1,\ \lambda_r \geq 0 \right\}$$

is a convex polytope of $\mathbb{R}^L$ whose vertices $\mathbf{m}_1, \ldots, \mathbf{m}_R$ are the $R \ll L$ spectral signatures to be estimated. As a consequence, the unobserved data $\mathbf{X} = \mathbf{MA} = \mathbf{Y} - \mathbf{N}$ can be represented in a lower-dimensional subspace $\mathcal{V}_K$ of $\mathbb{R}^K$ with $R - 1 \leq K \ll L$ without any loss of information. In this subspace, the noise-free data $\mathbf{X}$ span a $(R-1)$-simplex whose vertices are the projections of the endmembers. As stated in (Keshava & Mustard 2002), dimensional reduction is a classical step while performing spectral unmixing, required by numerous EEAs, such as N-FINDR (Winter 1999) and PPI (Boardman 1993). In this paper, we propose to estimate the projections $\mathbf{t}_r$ $(r = 1, \ldots, R)$ of the spectral signatures $\mathbf{m}_r$ onto the subspace $\mathcal{V}_K$. This approach allows the number of degrees of freedom to be significantly reduced. We assume that this subspace has been previously estimated by a dimensional reduction technique (*e.g.*, principal component analysis, PCA).

PCA-based dimensional reduction. The empirical covariance matrix of $\boldsymbol{\Upsilon}$ of the observed data $\mathbf{Y}$ is

$$\boldsymbol{\Upsilon} = \frac{1}{P}\sum_{p=1}^{P}\left(\mathbf{y}_p - \bar{\mathbf{y}}\right)\left(\mathbf{y}_p - \bar{\mathbf{y}}\right)^T \tag{4.1}$$

where $\bar{\mathbf{y}}$ is the empirical mean

$$\bar{\mathbf{y}} = \frac{1}{P}\sum_{p=1}^{P}\mathbf{y}_p. \tag{4.2}$$

Let

$$\begin{cases} \mathbf{D} = \operatorname{diag}\left(\lambda_1, \ldots, \lambda_K\right), \\ \mathbf{V} = \left[\mathbf{v}_1, \ldots, \mathbf{v}_K\right]^T \end{cases} \tag{4.3}$$

denote the diagonal matrix of the K highest eigenvalues and the corresponding matrix of eigenvectors of $\boldsymbol{\Upsilon}$, respectively. The projected vector $\mathbf{t}_r \in \mathbb{R}^K$ of the endmember spectrum $\mathbf{m}_r \in \mathbb{R}^L$ is then obtained by the affine transformation

$$\mathbf{t}_r = \mathbf{P}\left(\mathbf{m}_r - \bar{\mathbf{y}}\right) \tag{4.4}$$

where $\mathbf{P} = \mathbf{D}^{-\frac{1}{2}}\mathbf{V}$. Equivalently,

$$\mathbf{m}_r = \mathbf{U}\mathbf{t}_r + \bar{\mathbf{y}} \tag{4.5}$$

where $\mathbf{U} = \mathbf{V}^T\mathbf{D}^{\frac{1}{2}}$. Note that in the subspace $\mathcal{V}_{R-1}$ obtained for $K = R-1$, the vectors $\{\mathbf{t}_r\}_{r=1,\ldots,R}$ span a simplex that the classical EEAs (*e.g.*, N-FINDR Winter 1999, MVT Craig 1994 and ICE Berman *et al.* 2004) try to estimate. We propose to estimate the projected vertices $\mathbf{t}_r$ $(r = 1, \ldots, R)$ of this simplex in a Bayesian setting. The prior distributions assigned to the projections $\mathbf{t}_r$ $(r = 1, \ldots, R)$ are detailed in the following paragraph.

Prior distributions of the projected endmembers. The spectral signature $\mathbf{m}_r \in \mathbb{R}^L$ and its projection $\mathbf{t}_r \in \mathbb{R}^K$ onto $\mathcal{V}_K$ are related by $\mathbf{t}_r = \mathbf{P}\left(\mathbf{m}_r - \bar{\mathbf{y}}\right)$ and $\mathbf{m}_r = \mathbf{U}\mathbf{t}_r + \bar{\mathbf{y}}$, where $\mathbf{P}$ is a projection matrix, $\mathbf{U}$ is the pseudo-inverse of $\mathbf{P}$ and $\bar{\mathbf{y}}$ is the empirical mean of the observations. The prior distributions chosen for the endmember spectra should be chosen such that the endmember spectra satisfy the positivity constraints (2.5). Straightforward computations allows the space $\mathcal{T}_r \subset \mathcal{V}_K$ to be identified such that

$$\left\{m_{l,r} \geq 0,\ \forall l = 1, \ldots, L\right\} \quad \Leftrightarrow \quad \left\{\mathbf{t}_r \in \mathcal{T}_r\right\} \tag{4.6}$$

thanks to the L following inequalities

$$\mathcal{T}_r = \left\{\mathbf{t}_r; \bar{y}_l + \sum_{k=1}^{K} u_{l,k}t_{k,r} \geq 0, l = 1, \ldots, L\right\}, \tag{4.7}$$

One of the originality of the proposed blind source separation method consists of defining prior distributions for the endmember projections $\mathbf{t}_r$ onto the subspace $\mathcal{V}_K$ instead of the endmembers $\mathbf{m}_r$ themselves. More precisely, a multivariate Gaussian distribution

$$\mathbf{t}_r \sim \mathcal{N}_{\mathcal{T}_r}\left(\mathbf{e}_r, s_r^2 \mathbf{I}_K\right) \tag{4.8}$$

truncated to the set $\mathcal{T}_r$ is chosen as prior distribution for each vector $\mathbf{t}_r$ ($r = 1, \ldots, R$). The mean vectors $\mathbf{e}_r$ of these prior distributions are fixed to some values corresponding to the solutions provided by fast EEA, such as N-FINDR and VCA. In absence of additional prior information, the variances s_r^2 ($r = 1, \ldots, R$) are fixed to large values $s_1^2 = \ldots = s_R^2 = 50$, which allows some deviations to be modeled between the actual endmember projections $\mathbf{t}_r$ and the crude estimations $\mathbf{e}_r$ provided by N-FINDR or VCA.

4.1.4 Posterior distribution

Following the the Bayes rule, the prior distributions of unknown parameters defined in the paragraphs 4.1.3 and 4.1.2, associated with the likelihood function defined in paragraph 4.1.1, lead to the following joint posterior distribution

$$\begin{aligned} f\left(\mathbf{C}, \mathbf{T}, \sigma^2 \middle| \mathbf{Y}\right) \propto & \prod_{r=1}^{R} \exp\left[-\frac{\|\mathbf{t}_r - \mathbf{e}_r\|^2}{2s_r^2}\right] \mathbf{1}_{\mathcal{T}_r}(\mathbf{t}_r) \\ & \times \prod_{p=1}^{P} \left[\left(\frac{1}{\sigma^2}\right)^{\frac{L}{2}+1} \exp\left(-\frac{\|\mathbf{y}_p - (\mathbf{U}\mathbf{T} + \bar{\mathbf{y}}\mathbf{1}_{R-1})\,\mathbf{a}_p\|^2}{2\sigma^2}\right)\right] \\ & \times \prod_{p=1}^{P} \mathbf{1}_{\mathcal{S}}(\mathbf{c}_p)\,. \end{aligned} \tag{4.9}$$

Since the standard Bayesian estimators (*e.g.*, minimum mean square error (MMSE) or maximum *a posteriori* (MAP) estimators) are difficult to derive from (4.9), a Gibbs algorithm, detailed in the following paragraph, allows samples $\left\{\mathbf{C}^{(t)}, \mathbf{T}^{(t)}, \sigma^{2(t)}\right\}$ to be generated according to this distribution. These samples are then used to approximate the Bayesian estimators.

4.2 Gibbs sampler

The Gibbs sampler that allows samples to be asymptotically distributed according to the posterior (4.9) is detailed below. This algorithm is similar to the Gibbs sampler introduced in paragraph 3.2 (Algo. 1) with an additional step that consists of sampling according to the conditional distribution $f\left(\mathbf{T}|\mathbf{C}, \sigma^2, \mathbf{Y}\right)$ (see also Algo. 2).

Algorithm 2 Gibbs sampler for **unsupervised** unmixing

% Pre-processing
Computing the empirical mean $\bar{\mathbf{y}}$ following (4.2),
Computing the matrices $\mathbf{D}$ and $\mathbf{V}$ following (4.3) thanks to a PCA,
Set $\mathbf{U} = \left(\mathbf{V}^T\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{D}^{\frac{1}{2}}$,
Choose the crude estimations $\mathbf{e}_r \in \mathcal{V}_K$ required in (4.8),
% Initialization
for $r = 1, \ldots, R$ **do**
  Sampling $\mathbf{t}_r^{(0)}$ according to (4.8),
  Set $\mathbf{m}_r^{(0)} = \mathbf{U}\mathbf{t}_r^{(0)} + \bar{\mathbf{y}}$,
end for
Sampling $\sigma^{2(0)}$ according to (3.5),
% Iterations
for $t = 1, 2, \ldots,$ **do**
  for $p = 1, \ldots, P$ **do**
    Sampling $\mathbf{c}_p^{(t)}$ according to (4.11),
  end for
  for $r = 1, \ldots, R$ **do**
    for $k = 1, \ldots, K$ **do**
      Sampling $t_{k,r}^{(t)}$ according to (4.15),
    end for
    Set $\mathbf{m}_r^{(t)} = \mathbf{U}\mathbf{t}_r^{(0)} + \bar{\mathbf{y}}$,
  end for
  Sampling $\sigma^{2(t)}$ according to (4.17).
end for

4.2.1 Sampling according to $f\left(\mathbf{C}|\mathbf{T}, \sigma^2, \mathbf{Y}\right)$

For each pixel p, as in paragraph 3.2.1, the conditional distribution of the coefficient vector $\mathbf{c}_p$ is

$$f\left(\mathbf{c}_p \left| \mathbf{T}, \sigma^2, \mathbf{y}_p\right.\right) \exp\left[-\frac{\left(\mathbf{c}_p - \boldsymbol{\upsilon}_p\right)^T \boldsymbol{\Sigma}_p^{-1}\left(\mathbf{c}_p - \boldsymbol{\upsilon}_p\right)}{2}\right] \mathbf{1}_{\mathcal{S}}\left(\mathbf{c}_p\right), \tag{4.10}$$

where $\boldsymbol{\Sigma}_p$ and $\boldsymbol{\upsilon}_p$ have been defined in 3.2.1. Consequently, the vector $\mathbf{c}_p \left| \mathbf{T}, \sigma^2, \mathbf{y}_p\right.$ is distributed according to a multivariate Gaussian distribution truncated onto the simplex $\mathcal{S}$ defined by (3.4)

$$\mathbf{c}_p \left| \mathbf{T}, \sigma^2, \mathbf{y}_p\right. \sim \mathcal{N}_{\mathcal{S}}\left(\boldsymbol{\upsilon}_p, \boldsymbol{\Sigma}_p\right). \tag{4.11}$$

4.2.2 Sampling according to $f\left(\mathbf{T}|\mathbf{C},\sigma^2,\mathbf{Y}\right)$

Let $\mathbf{T}_{\text{-}r}$ denote the $\mathbf{T}$ whose rth column has been removed. The conditional posterior distribution of $\mathbf{t}_r$ $(r=1,\ldots,R)$ is

$$f\left(\mathbf{t}_r|\mathbf{T}_{\text{-}r},\mathbf{c}_r,\sigma^2,\mathbf{Y}\right)\propto\exp\left[-\frac{1}{2}\left(\mathbf{t}_r-\boldsymbol{\tau}_r\right)^T\boldsymbol{\Lambda}_r^{-1}\left(\mathbf{t}_r-\boldsymbol{\tau}_r\right)\right]\mathbf{1}_{\mathcal{T}_r}\left(\mathbf{t}_r\right), \tag{4.12}$$

with

$$\begin{cases}\boldsymbol{\Lambda}_r=\left[\displaystyle\sum_{p=1}^{P}a_{p,r}^2\mathbf{U}^T\boldsymbol{\Sigma}_{\text{n}}^{-1}\mathbf{U}+\frac{1}{s_r^2}\mathbf{I}_K\right]^{-1},\\ \boldsymbol{\tau}_r=\boldsymbol{\Lambda}_r\left[\displaystyle\sum_{p=1}^{P}a_{p,r}\mathbf{U}^T\boldsymbol{\Sigma}_{\text{n}}^{-1}\boldsymbol{\epsilon}_{p,r}+\frac{1}{s_r^2}\mathbf{e}_r\right],\end{cases} \tag{4.13}$$

and

$$\boldsymbol{\epsilon}_{p,r}=\mathbf{y}_p-a_{p,r}\bar{\mathbf{y}}-\sum_{j\neq r}a_{p,j}\mathbf{m}_j. \tag{4.14}$$

Generating vectors distributed according to this distribution is not straightforward, mainly due to the truncature on the set $\mathcal{T}_r$. One alternative strategy consists of generating each component $t_{k,r}$ of $\mathbf{t}_r$ conditionally upon the others $\mathbf{t}_{\text{-}k,r}=\{t_{j,r}\}_{j\neq k}$. By denoting $\mathcal{U}_k^+=\{l;u_{l,k}>0\}$, $\mathcal{U}_k^-=\{l;u_{l,k}<0\}$ and $\varepsilon_{l,k,r}=\bar{y}_l+\sum_{j\neq k}u_{l,j}t_{j,r}$, it follows

$$t_{k,r}|\mathbf{t}_{\text{-}k,r},\mathbf{T}_{\text{-}r},\mathbf{c}_r,\sigma^2,\mathbf{Y}\sim\mathcal{N}_{\left[t_{k,r}^-,t_{k,r}^+\right]}\left(w_{k,r},z_{k,r}^2\right), \tag{4.15}$$

with

$$\begin{cases}t_{k,r}^-=\max\limits_{l\in\mathcal{U}_k^+}-\dfrac{\varepsilon_{l,k,r}}{u_{l,k}},\\ t_{k,r}^+=\min\limits_{l\in\mathcal{U}_k^-}-\dfrac{\varepsilon_{l,k,r}}{u_{l,k}},\end{cases} \tag{4.16}$$

where $w_{k,r}$ and $z_{k,r}^2$ are the conditional mean and variance computed following (Kay 1988, p. 324) (see also similar computations in Dobigeon & Tourneret 2007). Generating samples according to the truncated Gaussian distribution (4.15) can be performed using various strategies, such as (Robert 1995).

4.2.3 Sampling according to $f\left(\sigma^2|\mathbf{C},\mathbf{T},\mathbf{Y}\right)$

The conditional distribution of $\sigma^2|\mathbf{C},\mathbf{T},\mathbf{Y}$ is the inverse-gamma distribution

$$\sigma^2|\mathbf{C},\mathbf{T},\mathbf{Y}\sim\mathcal{IG}\left(\frac{PL}{2},\frac{1}{2}\sum_{p=1}^{P}\left\|\mathbf{y}_p-\mathbf{M}\mathbf{a}_p\right\|^2\right). \tag{4.17}$$

4.3 Simulation results on synthetic data

To illustrate the performance of the proposed method, the algorithm has been applied on a 100×100-pixel image, where $R = 3$ spectral signatures have been linearly mixed: construction concrete, green grass, red bare brick. These signatures have been measured in $L = 413$ spectral bands and are depicted in Figure 6 (top, in black). These materials have been linearly mixed with random proportions (ensuring the sum-to-one and positivity constraints), with an i.i.d. noise corresponding to signal-to-noise ratio SNR = 15dB.

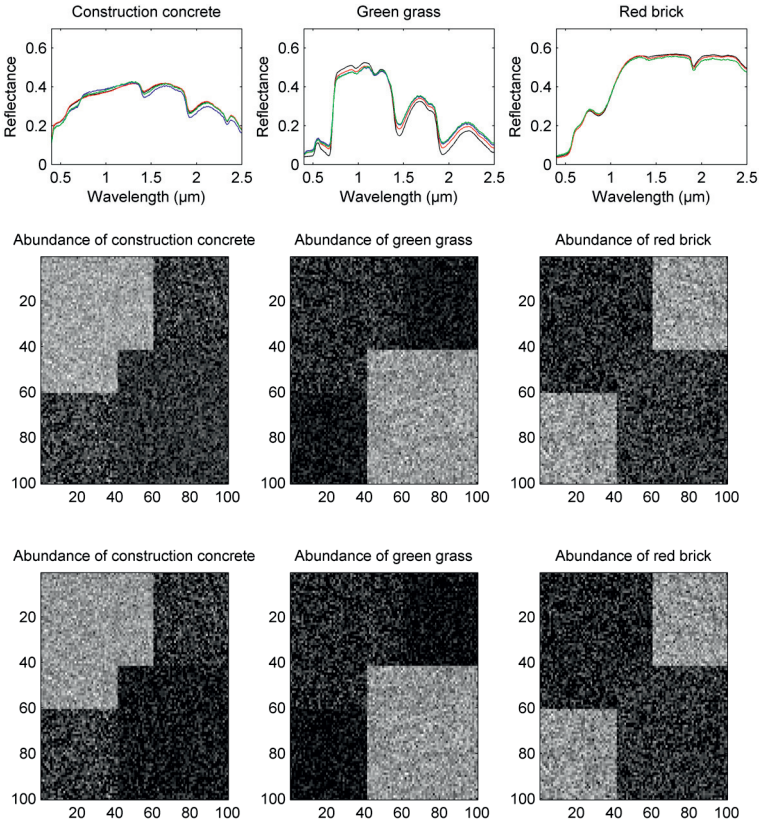


Fig. 6. *Top*: actual spectra (black), spectra estimated by N-FINDR (blue), estimated by VCA (green) and estimated by the proposed approach (red). *Middle* and *bottom*: actual and estimated abundance maps.

The estimation results for the spectral signatures obtained by the proposed algorithm, depicted in Figure 6 (top, in red) have been compared with the results provided by two geometrical EEAs: VCA and N-FINDR. Table 1 (top) reports the mean square errors defined by

$$\mathrm{MSE}_r^2 = \|\hat{\mathbf{m}}_r - \mathbf{m}_r\|^2, \quad r = 1, \ldots, R. \tag{4.18}$$

The results regarding the estimation of the 10^4 abundance vectors (see Fig. 6, bottom), are reported in the Table 1 (bottom) in terms of mean square errors for each component:

$$\mathrm{MSE}_r^2 = \sum_{p=1}^{P} \left(\hat{a}_{p,r} - a_{p,r}\right)^2, \quad r = 1, \ldots, R, \tag{4.19}$$

where $\hat{a}_{p,r}$ is the estimated abundance coefficient for the $\#r$ material in the $\#p$ pixel. These results demonstrate that the proposed algorithm provides better estimation performance than the two other algorithms.

Table 1. Estimation performance comparison between the algorithms VCA, N-FINDR and the proposed Bayesian approach: mean square errors between the $R = 3$ estimated and actual spectra (*top*), global mean square errors between the estimated and actual abundances (*bottom*).

Spectra	Bayesian algo.	VCA	N-FINDR
Endmember #1	0.10	1.29	0.54
Endmember #2	2.68	15.59	5.19
Endmember #3	0.16	4.35	0.57
Abundances	Bayesian algo.	VCA	N-FINDR
Endmember #1	25.68	57.43	30.66
Endmember #2	29.97	74.48	46.45
Endmember #3	3.19	83.02	11.22

4.4 Results on real data

The proposed algorithm is finally applied to the Moffett Field image introduced in Section 3.4. The $R = 3$ endmembers identified by the Bayesian algorithm are depicted in Figure 7 (top). The corresponding estimated abundance maps are represented in Figure 7 (bottom). Both results are in good agreement with those of Figures 4 and 5 obtained using a supervised unmixing approach.

5 Conclusion

This article presented two Bayesian algorithms to solve the problem of linear unmixing of hyperspectral images in supervised and unsupervised frameworks. For each scenario, suitable prior distributions were assigned to the unknown parameters. In particular, these distributions were chosen to ensure constraints inherent to the mixing model: positivity and additivity for the abundance coefficients and positivity for the endmember spectra. MCMC algorithms were designed to generate samples distributed according to the posterior distribution of the unknown parameters. Simulation results, obtained on synthetic and real hyperspectral images, demonstrated the interest of the proposed methods. Both of the strategies

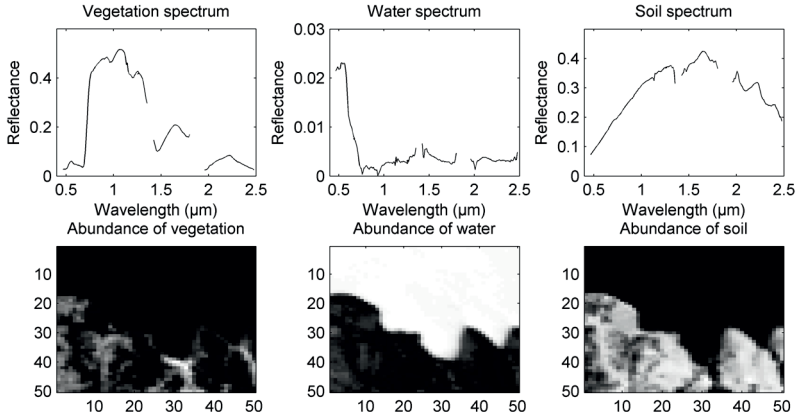


Fig. 7. *Top*: the $R = 3$ endmembers estimated by the unsupervised algorithm in the Moffett Field image. *Bottom*: the corresponding estimated abundance maps.

detailed in this article ignore any spatial correlations between the observed pixels. To improve the unmixing performance, intrinsic dependencies between the parameters of interest, *e.g.*, the abundance vectors, could be exploited. Extending the previous approaches, a hidden Markov model has been introduced in (Eches *et al.* 2011). Conversely, Mittelman *et al.* have proposed a nonparametric Bayesian algorithm to jointly unmix and classify hyperspectral images (Mittelman *et al.* 2012). Future works also include the design of efficient unmixing algorithms to analyze hyperspectral data resulting from non-linear mixtures. Encouraging results have been obtained in (Altmann *et al.* 2012; Halimi *et al.* 2011).

Part of this work was conducted in collaboration with Prof. C.-I. Chang, University of Maryland. Some results were obtained during a "Young researcher" Project founded by GdR-ISIS. The authors would also like to thank Jérôme Idier and Eric le Carpentier for fruitful discussion regarding this work.

References

Akgun, T., Altunbasak, Y., & Mersereau, R.M., 2005, IEEE Trans. Image Process., 14, 1860

Altmann, Y., Halimi, A., Dobigeon, N., & Tourneret, J.-Y., 2012, IEEE Trans. Image Process., 21, 3017

Arngren, M., Schmidt, M.N., & Larsen, J., 2011, J. Signal Proc. Syst., 65, 479

Berman, M., Kiiveri, H., Lagerstrom, R., Ernst, A., Dunne, R., & Huntington, J.F., 2004, IEEE Trans. Geosci. Remote Sens., 42, 2085

Bioucas-Dias, J.M., Plaza, A., Dobigeon, N., *et al.*, 2012, IEEE J. Sel. Topics Appl. Earth Observations Remote Sens., 5, 354

Blumensath, T., & Davies, M.E., 2007, IEEE Trans. Signal Process., 55, 4474

Boardman, J., 1993, in Summaries 4th Annu. JPL Airborne Geoscience Workshop, Vol. 1 (JPL Pub., Washington, D.C.), 11

Bowles, J.H., Palmadesso, P.J., Antoniades, J.A., Baumback, M.M., & Rickard, L.J., 1995, ed. M. Strojnik & B.F. Andresen, Infrared Spaceborne Remote Sensing III, SPIE, 2553, 148

Chang, C.-I., 2003, Hyperspectral Imaging: Techniques for Spectral detection and classification (Kluwer, New York)

Chang, C.-I., & Ji, B., 2001, IEEE Trans. Geosci. Remote Sensing, 44, 378

Chen, F.W., 2005, IEEE Geosci. Remote Sensing Lett., 2, 64

Chen, M.-H., & Deely, J.J., 1996, J. Agricultural, Biological Environmental Stat., 1, 467

Christophe, E., Léger, D., & Mailhes, C., 2005, IEEE Trans. Geosci. Remote Sensing, 43, 2103

Craig, M., 1994, IEEE Trans. Geosci. Remote Sens., 542

Dobigeon, N., & Achard, V., 2005, ed. L. Bruzzone, Image and Signal Processing for Remote Sensing XI, SPIE, 5982, 335

Dobigeon, N., Moussaoui, S., Coulon, M., Tourneret, J.-Y., & Hero, A.O., 2009, IEEE Trans. Signal Process., 57, 4355

Dobigeon, N., & Tourneret, J.-Y., 2007, Efficient sampling according to a multivariate Gaussian distribution truncated on a simplex, Technical report, IRIT/ENSEEIHT/TéSA, France

Dobigeon, N., Tourneret, J.-Y., & Hero III, A.O., 2008a, in Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP, Las Vegas, USA), 3433

Dobigeon, N., Tourneret, J.-Y., & Chang, C.-I., 2008b, IEEE Trans. Signal Process., 56, 2684

Dobigeon, N., Tourneret, J.-Y., & Davy, M., 2007, IEEE Trans. Signal Process., 55, 1251

Eches, O., Dobigeon, N., & Tourneret, J.Y., 2011, IEEE Trans. Geosci. Remote Sensing, 49, 4239

Févotte, C., & Godsill, S.J., 2006, IEEE Trans. Audio, Speech, Language Process., 14, 2174

Halimi, A., Altmann, Y., Dobigeon, N., & Tourneret, J.-Y., 2011, IEEE Trans. Geosci. Remote Sensing, 49, 4153

Hapke, B.W., 1981, J. Geophys. Res., 86, 3039

Heinz, D.C., & Chang, C.-I., 2001, IEEE Trans. Geosci. Remote Sens., 29, 529

Hyvärinen, A., Karhunen, J., & Oja, E., 2001, Independent Component Analysis (John Wiley, New York)

Jackson, Q., & Landgrebe, D.A., 2002, IEEE Trans. Geosci. Remote Sens., 40, 1082

Jeffreys, H., 1961, Theory of Probability, 3 edition (Oxford University Press, London)

Jet Propulsion Lab. (JPL), 2006, AVIRIS Free Data

Johnson, P.E., Smith, M.O., Taylor-George, S., & Adams, J.B., 1983, J. Geophys. Res., 88, 3557

Kay, S.M., 1988, Modern spectral estimation (Prentice Hall)

Keshava, N., & Mustard, J.F., 2002, IEEE Signal Process. Mag., 19, 44

Manolakis, D., Siracusa, C., & Shaw, G., 2001, IEEE Trans. Geosci. Remote Sens., 39, 1392

Mittelman, R., Dobigeon, N., & Hero III, A.O., 2012, IEEE Trans. Signal Process., 60, 1656

Moussaoui, S., Brie, D., Mohammad-Djafari, A., & Carteret, C., 2006, IEEE Trans. Signal Process., 54, 4133

Nascimento, J.M., & Bioucas-Dias, J.M., 2005a, IEEE Trans. Geosci. Remote Sens., 43, 898

Nascimento, J.M.P., & Bioucas-Dias, J.M., 2005b, IEEE Trans. Geosci. Remote Sens., 43, 175

Paatero, P., & Tapper, U., 1994, Environmetrics, 5, 111

Plaza, J., Pérez, R., Plaza, A., Martínez, P., & Valencia, D., 2005, ed. J.O. Jensen & J.-M. Thériault, Chemical and Biological Standoff Detection III, SPIE, 5995, 79

Punskaya, E., Andrieu, C., Doucet, A., & Fitzgerald, W., 2002, IEEE Trans. Signal Process., 50, 747

Rellier, G., Descombes, X., Falzon, F., & Zerubia, J., 2004, IEEE Trans. Geosci. Remote Sens., 42, 1543

Robert, C.P., 1995, Stat. Comput., 5, 121

Robert, C.P., 2007, The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation, 2 edition (Springer Texts in Statistics, Springer-Verlag, New York)

Robert, C.P., & Casella, G., 1999, Monte Carlo Statistical Methods (Springer-Verlag, New York)

RSI (Research Systems Inc.), 2003, ENVI User's guide Version 4.0, Boulder, CO 80301 USA

Tang, X., & Pearlman, W.A., 2004, Proc. IEEE Int. Conf. Image Proc. (ICIP), 5, 3283

Theys, C., Dobigeon, N., Tourneret, J.-Y., & Lantéri, H., 2009, in Proc. IEEE-SP Workshop Stat. and Signal Processing (SSP) (Cardiff, UK), 729

Tu, T.M., Chen, C.H., & Chang, C.-I., 1998, IEEE Trans. Geosci. Remote Sens., 36, 171

Winter, M., 1999, in Proc. 13th Int. Conf. Appl. Geologic Remote Sensing, 2, 337, Vancouver

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 403–416*

RESTORATION OF HYPERSPECTRAL ASTRONOMICAL DATA WITH SPECTRALLY VARYING BLUR

F. Soulez[1], E. Thiébaut[1] and L. Denis[2]

Abstract. In this paper we present a method for hyper-spectral image restoration for integral field spectrographs (IFS) data. We specifically address two topics: (i) the design of a fast approximation of spectrally varying operators and (ii) the comparison between two kind of regularization functions: quadratic and spatial sparsity functions. We illustrate this method with simulations coming from the Multi Unit Spectroscopic Explorer (MUSE) instrument. It shows the clear increase of the spatial resolution provided by our method as well as its denoising capability.

1 Introduction

In the last decade, integral field spectrographs (IFS) have become popular tools for astronomical observations. Such instruments are now installed on all the main optical telescope facilities around the world. They provide spatially resolved spectra of a whole region of the sky, yielding $(\boldsymbol{\theta}, \lambda)$ data cubes – with $\boldsymbol{\theta}$ the 2D angular position and λ the wavelength – with several hundreds of wavelength bins. With IFS, astronomical data enters the hyper-spectral era. New dedicated image reconstruction techniques are needed to take full advantage of the data gathered by these instruments. Because the light is split on multiple channels instead of being integrated on a single image, the information contents is increased at the cost of a lower signal to noise for the same exposure time. Furthermore, atmospheric turbulence and instrumental response spatially blur the observations, degrading the spatial resolution.

First attempts to restore multi-channel images consisted in applying classical 2D restoration techniques like Wiener filter or Richardson-Lucy algorithm on

[1] Université de Lyon, 69003 Lyon, France; Université Lyon 1, Observatoire de Lyon, 9 avenue Charles André, 69230 Saint-Genis Laval, France; CNRS, UMR 5574, Centre de Recherche Astrophysique de Lyon; École Normale Supérieure de Lyon, 69007 Lyon, France

[2] Université de Lyon, 42023 Saint-Etienne, France,
CNRS, UMR 5516, Laboratoire Hubert Curien, 42000 Saint-Etienne, France,
Université de Saint-Etienne, Jean Monnet, 42000 Saint-Etienne, France

DOI: 10.1051/eas/1359018

each individual channel. The caveat of these approaches is to ignore the natural spectral correlations present in the data. The first restoration technique specifically dedicated to multichannel data (Hunt & Kubler 1984) was a Minimum Mean Square Error (MMSE) restoration filter based on the assumption that signal autocorrelation is spatially and spectrally separable. This assumption was later relaxed (Galatsanos & Chin 1989) and many other multichannel linear restoration filters have been proposed since (Galatsanos *et al.* 1991; Gaucel *et al.* 2006; Katsaggelos *et al.* 1993; Tekalp & Pavlovic 1990). More recently, Fourier/wavelet restoration techniques (Neelamani *et al.* 2004) have been adapted to multispectral data (Benazza-Benyahia & Pesquet 2006; Duijster *et al.* 2009). In remote sensing, some authors (Akgun *et al.* 2005; Bobin *et al.* 2009) combine demixing and restoration to achieve enhanced spatial resolution given the strong assumption that the observed scene is composed of only a few materials with unknown spectrum.

Most of these developments on restoration of multi-spectral images are dedicated to remote sensing and color (RGB) images. Those methods can't easily be directly applied to astronomical data with its specific features like large dynamic range and strong sharp features (*e.g.* narrow emission lines or peaked sources). Few restoration techniques for multi-spectral astronomical images have been proposed for (x, λ) data (slit spectrography) (Courbin *et al.* 2000; Lucy & Walsh 2003) or (x, y, λ) data composed of slit spectrography scans (Rodet *et al.* 2008). However astronomical hyperspectral processing is gaining more and more attention as it is becoming mandatory to fully exploit the capabilities of new integral field spectrographs (*e.g.* second generation VLT instruments MUSE and KMOS) and restoration algorithms dedicated to IFS begin to appear (Bongard *et al.* 2009; Bongard *et al.* 2011; Bourguignon *et al.* 2011a,b; Soulez *et al.* 2008).

Following the work we have done in Soulez *et al.* (2011) and Bongard *et al.* (2011), we present in this paper a deconvolution method based on a so called *inverse problem* approach. It is very generic and exploits intrinsic regularities of hyper-spectral data. We suppose that a good estimation of the point spread function (PSF) is provided by other means (*e.g.* by calibration on the telescope guiding stars or on information from the adaptive optics system) and defer the blind restoration problem to a later time.

Our approach will be illustrated on data provided by the MUSE IFS simulator. Still in integration, the MUSE IFS (Henault *et al.* 2003) will be installed on the ESO Very Large Telescope (VLT) in 2013. It is a "slicer" based IFS that covers in its wide field mode a $60'' \times 60''$ spectroscopic field-of-view subdivided into a grid of about 300×300 spatial elements (spaxels). To each spaxel corresponds a spectrum, obtained by dispersing the light on 3463 equally spaced spectral bins from 480 nm to 930 nm.

2 Problem formulation

2.1 Model description

The direct model describes how the observed data $\boldsymbol{y}$ is related to the 3D intensity distribution of the object of interest $I(\boldsymbol{\theta}, \lambda)$ with $\boldsymbol{\theta} = (\theta_1, \theta_2)$ the 2-D position

angle. This data cube $\boldsymbol{y}$ is composed of N_λ monochromatic images of N_Ω pixels. It writes:

$$y_{\boldsymbol{k},\ell} = g(\boldsymbol{\theta}_{\boldsymbol{k}}, \lambda_\ell) + e_{\boldsymbol{k},\ell} \tag{2.1}$$

where $g(\boldsymbol{\theta}, \lambda)$ is the distribution sampled by the detector, $(\boldsymbol{\theta}_{\boldsymbol{k}}, \lambda_\ell)$ are the spatio-spectral coordinates of the pixel at the 2D spatial index $\boldsymbol{k}$ and spectral index ℓ, and $e_{\boldsymbol{k},\ell}$ accounts for the errors (noise and model approximations). The sampled distribution $g(\boldsymbol{\theta}, \lambda)$ writes:

$$g(\boldsymbol{\theta}, \lambda) = \iiint h(\boldsymbol{\theta} - \boldsymbol{\theta}', \lambda - \lambda'; \boldsymbol{\theta}', \lambda')\, I(\boldsymbol{\theta}', \lambda')\, \mathrm{d}^2\boldsymbol{\theta}'\, \mathrm{d}\lambda' \tag{2.2}$$

where $h(\Delta\boldsymbol{\theta}, \Delta\lambda; \boldsymbol{\theta}, \lambda)$ is the recentred PSF at position $\boldsymbol{\theta}$ and at wavelength λ. In words, the PSF is the linear response of the total observing system (atmosphere + optics + detector) for a monochromatic point-like source at $(\boldsymbol{\theta}, \lambda)$.

At best, we can only recover an approximation of the true object brightness distribution, we choose to represent the sought distribution by:

$$I(\boldsymbol{\theta}, \lambda) = \sum_{\boldsymbol{k},\ell} x_{\boldsymbol{k},\ell}\, b_{\boldsymbol{k},\ell}(\boldsymbol{\theta}, \lambda) \tag{2.3}$$

where $\boldsymbol{x}$ are the unknown parameters and $b_{\boldsymbol{k},\ell}(\boldsymbol{\theta}, \lambda)$ are basis functions. Using interpolation functions for the basis functions and the same spatio-spectral sampling for the model and the data yields:

$$x_{\boldsymbol{k},\ell} \approx I(\boldsymbol{\theta}_{\boldsymbol{k}}, \lambda_\ell). \tag{2.4}$$

The direct model then writes:

$$y_{\boldsymbol{k},\ell} = \sum_{\boldsymbol{k}',\ell'} H_{\boldsymbol{k},\ell,\boldsymbol{k}',\ell'}\, x_{\boldsymbol{k}',\ell'} + e_{\boldsymbol{k},\ell} \tag{2.5}$$

with $\mathbf{H}$ the linear operator corresponding to the system response. Using compact matrix notation:

$$\boldsymbol{y} = \mathbf{H} \cdot \boldsymbol{x} + \boldsymbol{e}. \tag{2.6}$$

Under the same assumption as those leading to Equation (2.4):

$$H_{\boldsymbol{k},\ell,\boldsymbol{k}',\ell'} \approx h(\boldsymbol{\theta} - \boldsymbol{\theta}', \lambda - \lambda'; \boldsymbol{\theta}', \lambda')\Pi\boldsymbol{\theta}^2\Pi\lambda\,, \tag{2.7}$$

with $\Pi\boldsymbol{\theta}^2$ and $\Pi\lambda$ the pixel size and the effective spectral bandwidth respectively.

The linear operator $\mathbf{H}$ models the linear response of the observation system. It can be described by a PSF which varies both spatially and spectrally. As the telescope and the atmosphere don't have any effect along the spectral dimension, blur along spectral dimension is only due to the IFS. Conversely, without adaptive optics system, the atmosphere is responsible for most of the blur along spatial dimensions. As the field of view (FOV) is limited, we can assume that this PSF is spatially shift invariant:

$$h_\lambda((\Delta\boldsymbol{\theta}, \Delta\lambda) = h(\Delta\boldsymbol{\theta}, \Delta\lambda; \lambda). \tag{2.8}$$

However wavelength-wise PSF's h_λ may be centered at a location $\boldsymbol{\theta}_\lambda$ which depends on the wavelength so as to account for imperfect instrumental alignment and atmospheric differential refractive index (ADR). Furthermore PSF is not necessarily normalized in order to account for the variable throughput (atmospheric and instrumental transmission). Finally, the operator $\mathbf{H}$ can be described as a *spectrally varying convolution.*

2.2 Spectrally varying PSF approximation

If the observing system is spatially and spectrally shift-invariant, $\mathbf{H}$ is a block Toeplitz matrix with Toeplitz block that can be diagonalized by means of discrete Fourier transforms (under a circulant approximation or providing a proper processing of the egdes as explained later). Such transforms being efficiently computed thanks to the FFT (Fast Fourier Transform) algorithm. In the considered case, the PSF is spatially shift-invariant but depends on the wavelength of the source. In order to implement a fast version of such an operator $\mathbf{H}$ storing the full $\mathbf{H}$ ($\geq 10^{12}$ elements) is not possible and, even so, applying it in this form would take to much CPU time, we propose to follow the prescription of Denis *et al.* (2011) and write:

$$h_\lambda(\Delta\boldsymbol{\theta}, \Delta\lambda) \approx \sum_p \phi_p(\lambda)\, h_p(\Delta\boldsymbol{\theta}, \Delta\lambda) \tag{2.9}$$

where:

$$h_p(\Delta\boldsymbol{\theta}, \Delta\lambda) \stackrel{\text{def}}{=} h(\Delta\boldsymbol{\theta}, \Delta\lambda; \lambda_p) \tag{2.10}$$

are samples at differents wavelengts $\{\lambda_p\}_{p=1}^P$ of the recentered spectrally-varying PSF and $\{\phi_p(\lambda) : \mathbb{R} \mapsto \mathbb{R}\}_{p=1}^P$ are spectral interpolation functions. With this modeling of the PSF, the operator $\mathbf{H}$ becomes:

$$\mathbf{H} = \sum_{p=1}^{P} \mathbf{H}_p \cdot \mathbf{K}_p \tag{2.11}$$

with $\mathbf{H}_p$ the discrete 3D convolution by $h_p(\Delta\boldsymbol{\theta}, \Delta\lambda)$ and $\mathbf{K}_p$ an operator which extracts a subset of the spectral range (around λ_p) and weights the selected spaxels by the interpolation function $\phi_p(\lambda)$. Operators $\mathbf{H}_p$ are implemented using 3D FFT's while $\mathbf{K}_p$'s are very sparse as their only non-zero coefficients are along their diagonal. Thus, as long as the spectral support of $h_p(\Delta\boldsymbol{\theta}, \Delta\lambda)$ is sufficiently small compared to the patch selected by $\mathbf{K}_p$, applying $\mathbf{H}$ (or its adjoint $\mathbf{H}^* = \sum_p \mathbf{K}_p^{\mathrm{T}} \cdot \mathbf{H}_p^*$) is very fast. The computations are dominated by the calculus of the FFT.

First order (linear) interpolation with a subset of PSF built by sampling on a uniform grid $\{\lambda_1, ..., \lambda_P\}$ leads to interpolation weights supported on a patch twice the grid step along. Each patch extracted by $\mathbf{K}_p$ are convolved only with the coresponding PSF $h_p(\Delta\boldsymbol{\theta}, \Delta\lambda)$. As a consequence, the computational cost for applying our spectrally varying operator is only roughly twice the cost the applying a non-varying operator: (4 *versus* 2 FFTs). As stated in Denis *et al.* (2011), such

approximation of the shift varying PSF preserves some good properties of the PSF, namely normalization, positivity and symmetry.

3 Maximum *a posteriori* approach

Restoration is a typical ill-posed problem (Bertero & Boccacci 1998). We choose to solve it by adding priors in a classical Maximum A Posteriori (MAP) framework. This is achieved by estimating the object $\boldsymbol{x}^+$ that minimizes a cost function $f(\boldsymbol{x})$:

$$\boldsymbol{x}^+ = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x}) \,, \tag{3.1}$$

$$f(\boldsymbol{x}) = f_{\mathsf{data}}(\boldsymbol{x}) + f_{\mathsf{prior}}(\boldsymbol{x}) \,. \tag{3.2}$$

This cost function $f(\boldsymbol{x})$ is the sum of a *likelihood penalty* $f_{\mathsf{data}}(\boldsymbol{x})$ ensuring the agreement between the model and the data $\boldsymbol{y}$, and a *regularization penalty* $f_{\mathsf{prior}}(\boldsymbol{x})$ introducing subjective *a priori* knowledge about the object.

3.1 Likelihood and noise statistics

Assuming Gaussian noise, the likelihood penalty reads:

$$f_{\mathsf{data}}(\boldsymbol{x}) = [\boldsymbol{y} - \mathbf{H} \cdot \boldsymbol{x}]^{\mathrm{T}} \cdot \mathbf{W}_{\mathsf{err}} \cdot [\boldsymbol{y} - \mathbf{H} \cdot \boldsymbol{x}] \,, \tag{3.3}$$

where the weighting matrix $\mathbf{W}_{\mathsf{err}} = \mathbf{C}_{\mathsf{err}}^{-1}$ is the inverse of the angular-spectral covariance of the errors (noise + approximations). Assuming uncorrelated noise, $\mathbf{W}_{\mathsf{err}}$ is diagonal and Equation (3.3) simplifies to:

$$f_{\mathsf{data}}(\boldsymbol{x}) = \sum_{\boldsymbol{k},\ell} w_{\boldsymbol{k},\ell} \, [\boldsymbol{y} - \mathbf{H} \cdot \boldsymbol{x}]^2_{\boldsymbol{k},\ell}$$

where $1/w_{\boldsymbol{k},\ell}$ is the noise variance of the measurements at pixel $\boldsymbol{k}$ and channel ℓ. This model can cope with non-stationary noise and can be used to express confidence on each measurements. Since unmeasured data can be considered as having infinite variance, we readily deal with missing or bad pixels as follows:

$$w_{\boldsymbol{k},\ell} \stackrel{\text{def}}{=} \begin{cases} \mathrm{Var}(y_{\boldsymbol{k},\ell})^{-1} & \text{if } y_{\boldsymbol{k},\ell} \text{ is measured,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

This treatment of missing data is rigorous because (i) it consistently accounts for unmeasured data and bad pixels, and (ii) it allows to properly expand the synthesized FOV to avoid field aliasing and border artifacts caused by convolution using Fourier transform. This formulation provides a rigorous scheme to take into account photons emitted by sources outside the FOV that are measured because of the blurring. As a consequence, restored object has to be estimated even outside of the field of view, by extending the size of the FOV by at least the PSF size. As we showed in Soulez *et al.* (2008) and Bongard *et al.* (2011), this may lead to a small extension of the FOV of the instrument that can be relatively significant

when this FOV is small like in the SNIFS instrument considered by Bongard *et al.* (2011).

Except for very low detector noise ($<$ few e^- per pixel), the total noise (Gaussian detector noise plus Poisson noise) is approximated by a non stationary uncorrelated Gaussian noise (Mugnier *et al.* 2004):

$$w_{\boldsymbol{k},\ell} \stackrel{\text{def}}{=} \begin{cases} \left(\gamma \max(y_{\boldsymbol{k},\ell}, 0) + \sigma^2_{\boldsymbol{k},\ell}\right)^{-1} & \text{if } y_{\boldsymbol{k},\ell} \text{ is measured,} \\ 0 & \text{otherwise,} \end{cases} \tag{3.5}$$

where γ accounts for the quantization factor of the detector (*i.e.* number of photon per quantization level) and $\sigma^2_{\boldsymbol{k},\ell}$ is the variance on the pixel $(\boldsymbol{k}, \ell)$ of other sources of noise than the signal, like for example read-out noise for instance.

3.2 Priors

In our MAP framework, priors on the object are enforced by the *regularization penalty* $f_{\mathsf{prior}}(\boldsymbol{x})$ term of the total cost function $f(\boldsymbol{x})$. It introduces in the solution generic knowledge about the observed objects. In addition, we enforce strict priors to ensure the non negativity of the parameters $\boldsymbol{x}$.

As in hyper-spectral imaging the spatial and the spectral dimension have different physical meaning we split the regularization function as the sum of a spatial regularization $f_{\mathsf{spatial}}(\boldsymbol{x})$ and a spectral regularization $f_{\mathsf{spectral}}(\boldsymbol{x})$:

$$f_{\mathsf{prior}}(\boldsymbol{x}) = \alpha f_{\mathsf{spatial}}(\boldsymbol{x}) + \beta f_{\mathsf{spectral}}(\boldsymbol{x}). \tag{3.6}$$

where α and β are hyper-parameters that have to be tuned to set the importance of the priors.

In this work we propose two kind of regularization functions: (i) a quadratic regularisation and (ii) a spatial sparsity regularisation.

3.2.1 Quadratic regularization

Quadratic regularization (so called Tikhonov) is the most simple prior that can be introduced in our MAP scheme. In that case and with the least square likelihood function defined in 3.3, the minimization of Equation (3.1) shows good convergence property since the total cost function $f(\boldsymbol{x})$ is strictly convex and quadratic.

As stated in (Bongard *et al.* 2011), it is customary to minimize the quadratic norm of finite differences to account for continuities along the three dimensions of the brightness distribution. The regularization functions are thus:

$$f_{\mathsf{spatial}}(\boldsymbol{x}) = \sum_{\boldsymbol{k},\ell} (\nabla_{k_1} \boldsymbol{x})^2 + (\nabla_{k_2} \boldsymbol{x})^2 , \tag{3.7}$$

$$f_{\mathsf{spectral}}(\boldsymbol{x}) = \sum_{\boldsymbol{k},\ell} (\nabla_{\ell} \boldsymbol{x})^2 , \tag{3.8}$$

where $\nabla_i \boldsymbol{x}$ is the finite differential operator along the dimension indexed by the letter i.

3.2.2 Spatial only sparsity regularization

In wide field observations, astronomical data is mainly composed of bright objects (stars, galaxy) over a flat background. Most of the quite large MUSE field of view will thus contain only background. As a consequence, the observed scene is intrinsically spatially sparse. This spatial sparsity prior can be enforced by means of structured norms (Fornasier & Rauhut 2008; Kowalski & Torrésani 2009):

$$f_{\mathsf{sparsity}}(\boldsymbol{x}) = \sum_{\boldsymbol{k}} \left[\sqrt{\sum_{\ell} x_{\boldsymbol{k},\ell}^2 + \epsilon^2} - \epsilon \right] \tag{3.9}$$

where ϵ is a small real number ($\epsilon \approx 10^{-9}$) that ensures the derivability in 0 (hyperbolic approximation of the ℓ_1 norm). This regularization enforces spatial sparsity and spectral correlation since it favors solutions where bright features in each spectral channel are at the same spatial location.

The regularization defined in Equation (3.9) does not ensure the spectral continuity of the solution whereas in practice the spectral energy distribution (SED) of a pixel should be relatively smooth excepted near emission and absorption lines. For that reason we introduce an additional regularization function:

$$f_{\mathsf{spectral}}(\boldsymbol{x}) = \sum_{\boldsymbol{k},\ell} \left[\sqrt{\left(\nabla_\ell \boldsymbol{x}_{\boldsymbol{k}}\right)^2 + \zeta^2} - \zeta \right]. \tag{3.10}$$

This regularization tends to smooth the spectra $x_{\boldsymbol{k}}$ but preserve discontinuities where $|\nabla_\ell \boldsymbol{x}_{\boldsymbol{k}}| \gg \zeta$. This situation is for example encountered at absorption or emission lines, which shall not be smoothed.

3.2.3 Renormalization

Owing to the large variations of the dynamical range between spectral channel of astronomical images, these regularizations lead to over-regularize bright features or under-regularize faint ones. For that reason, as Bongard *et al.* (2011) we rather suggest to apply these regularizations to *spectrally whitened* object $\boldsymbol{x}'$:

$$x'_{\boldsymbol{k},\ell} = x_{\boldsymbol{k},\ell} / s_\ell \tag{3.11}$$

with $s_\ell = \langle x_{\boldsymbol{k},\ell} \rangle_{\boldsymbol{k}}$ the spatially averaged object spectrum – $\langle\ \rangle_{\boldsymbol{k}}$ denotes averaging over pixel index $\boldsymbol{k}$. To avoid introducing more non-linearity in regularizations, we estimate the mean object spectrum directly from the data:

$$s_\ell = \langle y_{\boldsymbol{k},\ell} \rangle_{\boldsymbol{k}} / \eta_\ell \tag{3.12}$$

with $\eta_\ell = \eta(\lambda_\ell)$ the effective throughput in ℓ-th spectral channel:

$$\eta_\ell = \iiint h_\ell(\Delta\boldsymbol{\theta}, \Delta\lambda)\, \mathrm{d}^2\Delta\boldsymbol{\theta}\, \mathrm{d}\Delta\lambda. \tag{3.13}$$

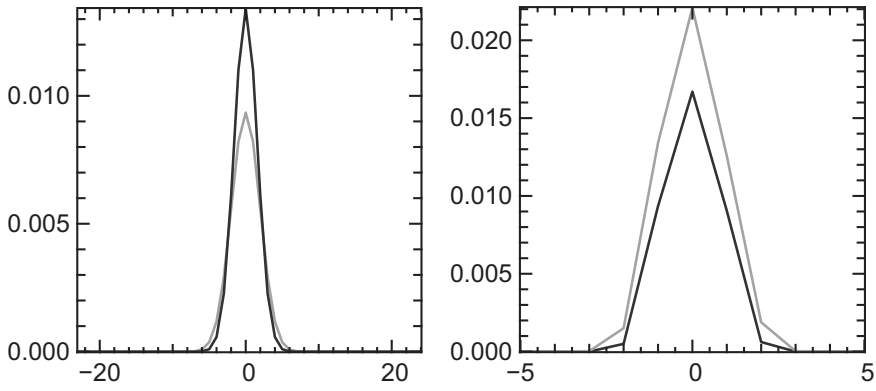


Fig. 1. *Left*: profile of the PSF along a spatial dimension. *Right*: profile of the PSF along the spectral dimension. Grey and black profile correspond to the blue and the red ends of the IFS respectively.

3.3 Algorithm description

As discussed in Section 3.1, due to the convolution process, flux from the object just outside of the field of view may have an impact on the data. To take this fact correctly into account, the estimated object has to be spatially larger than the observed field of view. At least half of the PSF support must be added on each side of the observed field of view to form the restored field of view.

The level of priors introduced in the restoration is balanced by hyper-parameters α and β that are estimated by trial and error. The restored data cube $\boldsymbol{x}^+$ is the solution of Equation (3.1). It requires the minimization of the cost function $f(\boldsymbol{x})$ that involves a large number of parameters ($> 10^6$) with positivity constraints. To that end, we use the VMLM-B algorithm (Thiébaut 2002) which is a limited memory variant of the variable metric method with BFGS updates (Nocedal & Wright 1999). This algorithm has proved its effectiveness for image reconstruction and only requires the computation of the penalty function being minimized together with its gradient. The memory requirement is a few times the size of the problem.

4 Results

The quality of the presented algorithm was assessed on data from the MUSE IFS simulator. This data is a part of 51×36 spaxels (pixels size: $0\farcs2 \times 0\farcs2$) of the whole MUSE data cube. It contains 3463 spectral channels comprised between 480 nmand 930 nm. The PSF was computed for a seeing of $1.1''$. This PSF shown on Figure 1 is supposed to be separable and composed of a spatial field spread function (FSF) and a spectral line spread function (LSF). As shown in Figure 1, both of them vary spectrally. FSF is Gaussian with a full width at half maximum

that varies from 0.″75 (3.75 pixels) at the red end to 0.″92 (4.6 pixels) at the blue end. In addition, the MUSE IFS simulator provided a cube of the variance for each pixel as it will be estimated by the data reduction software of the instrument.

To perform the restoration, we first have to build the fast approximation of the operator $\mathbf{H}$ as defined Equation (2.11). For these experiments, the PSF was sampled on a grid of $P = 350$ evenly spaced wavelengths to give the h_p and linear interpolation along the spectral dimension that was used for the weights ϕ_p. As linear interpolation is used, the PSF h_ℓ centered on spectral channel ℓ with $\lambda_p \leq \lambda_\ell \leq \lambda_{p+1}$ is interpoled only using h_p and h_{p+1}. The Euclidean norm of the differences between the true PSF h_λ and our approximation is less than 8×10^{-5} (0.08% relatively to the euclidean norm of the PSF). That gives a quantitative estimate of the good quality of our approximation.

As stated in Section 3.1, the restored field of view must be larger than the data FOV. In the presented experiments, the size of the restored FOV is extended to 64×48 spaxels and 3481 wavelengths.

The data were processed with both quadratic and spatial sparsity regularizations. The effectiveness is qualitatively evaluated by visual inspection and quantitatively by the root mean square error (RMSE):

$$\mathrm{RMSE}(\boldsymbol{x}) = \sqrt{\frac{1}{N_\Omega\, N_\lambda} \sum_{\boldsymbol{k},\ell} [\boldsymbol{x} - \boldsymbol{o}]^2_{\boldsymbol{k},\ell}}\,,$$

with $\boldsymbol{o}$ the truth. In both cases, the hyper-parameters α and β were set to minimize the RMSE. For the quadratic case, with the hyper-parameters $\alpha = 1$ and $\beta = 1$, the algorithm converged in about 5 hours to the solution $\boldsymbol{x}^+_{\mathsf{quad}}$ with $\mathrm{RMSE}(\boldsymbol{x}^+_{\mathsf{quad}}) = 0.418$. For the spatial sparsity case, the algorithm converged in about 8 hours to the solution $\boldsymbol{x}^+_{\mathsf{Spar}}$ with $\mathrm{RMSE}(\boldsymbol{x}^+_{\mathsf{Spar}}) = 0.344$ with the hyper-parameters $\alpha = 15\,000$, $\beta = 0.05$ and $\zeta = 1$.

The results are shown on Figures 2 and 3. Figure 2 shows the data, the results and the true object integrated over the whole spectral range of the instrument. It clearly illustrates the gain in term of spatial resolution provided by our method. Both the shapes of the central galaxy and of the one near the upper left corner are recovered. Compared to the solution with spatial sparsity regularization $\boldsymbol{x}^+_{\mathsf{Spar}}$, the solution with quadratic regularization $\boldsymbol{x}^+_{\mathsf{quad}}$ shows more artifacts (*e.g.* on the bottom left part of the central galaxy) and bright spots are a bit over-smoothed.

We display in Figure 3 spectral cuts through the heart of the central galaxy materialized by the dashed line in 2. These figures show (θ, λ) images zoomed between 567 nmand 574 nmfor the data, both restorations and the true object. These plots show the resolution gain provided by our algorithm: the two brightest objects are well separated, with the spectrum at 43rd column visible in the restoration that was not visible in the data. Once again, the solution with quadratic regularization $\boldsymbol{x}^+_{\mathsf{quad}}$ shows much more artifacts (*e.g.* on the bright emission line at $\lambda = 573$ nmand $\theta = 22$). Furthermore, the noise has been drastically reduced by our method as this can clearly be seen by looking at the background.

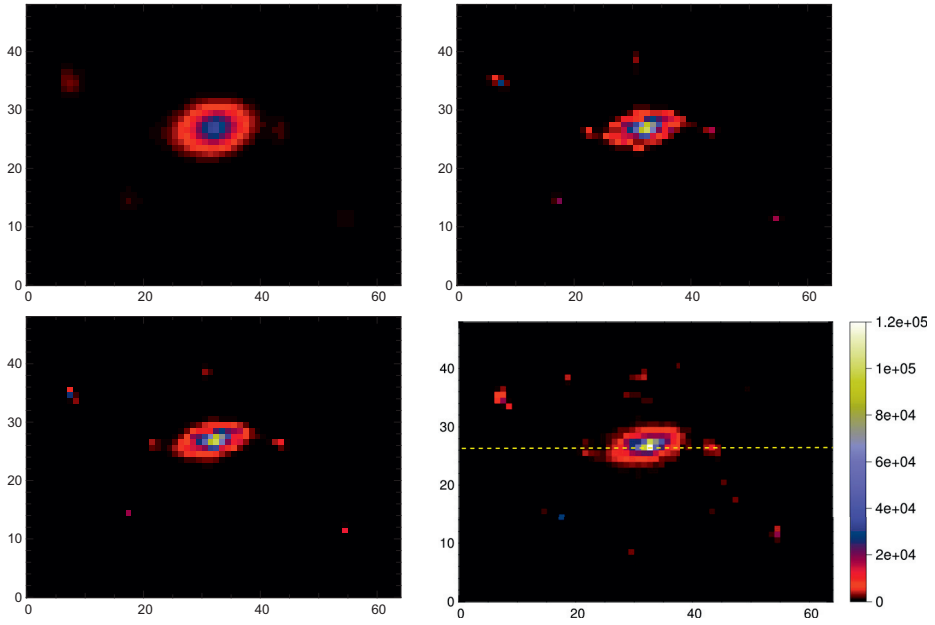


Fig. 2. Images spectrally integrated. *Top left*: raw data. *Top right*: restored object $\boldsymbol{x}^+_{\mathsf{quad}}$ with quadratic regularization. *Bottom left*: restored object $\boldsymbol{x}^+_{\mathsf{Spar}}$ with spatial sparsity regularization. *Bottom right*: true object $\boldsymbol{o}$.

Figure 4 displays the spectra of the brighter spaxel of the galaxy ($\boldsymbol{\theta} = (33, 27)$) of the data, the quadratic restoration $\boldsymbol{x}^+_{\mathsf{quad}}$ and the spatial sparsity restoration $\boldsymbol{x}^+_{\mathsf{Spar}}$ and the ground truth. Even though regularizations introduce some expected bias, the restored spectra are closer to the ground truth and far less noisy than the measured spectrum. In the spatial sparsity restoration $\boldsymbol{x}^+_{\mathsf{Spar}}$ (red), most of the spectral features are preserved. These features are over-smoothed in the quadratic restoration $\boldsymbol{x}^+_{\mathsf{quad}}$ (green). The bias between restoration and is quite strong as it is the spectrum of the spaxel with the higher dynamical range and it tends to be flattened by the regularization. The hyper-parameters were tuned to provide the minimal RMSE for the whole field of view. As a consequence, the hyper-parameters setting for a sufficient regularization of the faint sources is strong for bright sources and tends to smooth them. However, this bias disappears if we integrate spatially on few spaxels as we show in Figure 5 on the spectra of the central 3×3 region of the central galaxy. This means that the bias is mainly imputable to the remaining blur.

5 Conclusion

In this paper, we present a method for restoring hyperspectral data. We especially focused on two points: (i) the design of an efficient operator modelling spectrally

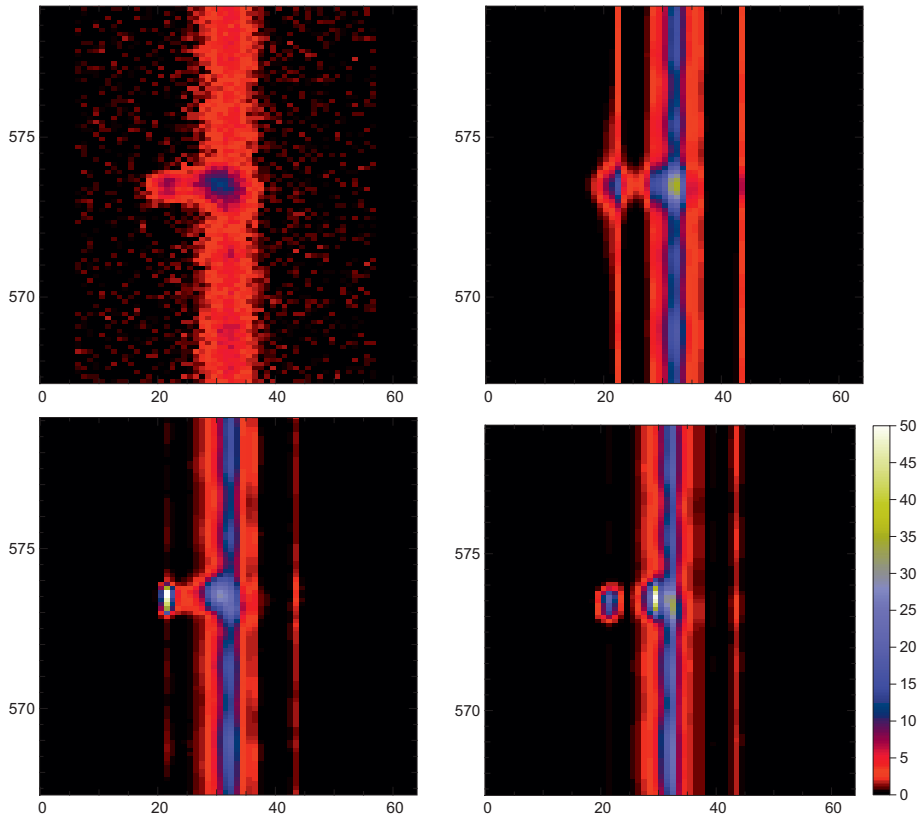


Fig. 3. (θ, λ) images of the cut materialized by the yellow line in Figure 2 magnified between 567 nm and 574 nm. *Top left*: raw data. *Top right*: restored object $\boldsymbol{x}^+_{\text{quad}}$ with quadratic regularization. *Bottom left*: restored object $\boldsymbol{x}^+_{\text{Spar}}$ with spatial sparsity regularization. *Top right*: true object $\boldsymbol{o}$.

varying blur and (ii) a comparison between quadratic and spatial sparsity regularization functions.

We have shown that using PSF interpolation it is possible to design an effective operator approximating spectrally varying blur. Our formulation preserves the positivity, the normalization and the symmetry of the PSF. The computational cost of such approximation, that is twice as much as spectrally invariant convolution, remains tractable and it is possible to consider the processing of whole MUSE data cubes (size: $300 \times 300 \times 3463$) with nowadays CPU power. Furthermore, this type of operator can be easily extended to blurs that vary both spatially and spectrally as in wide field observations with adaptive optics.

By exploiting jointly spatial and spectral correlations present in the data, our method provides a strong spatial resolution enhancement and an effective denoising

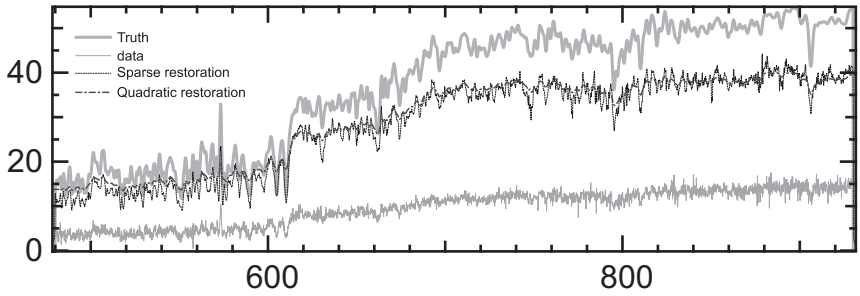


Fig. 4. Spectra of the brighter spaxel of the central galaxy of the data (thin grey line), the quadratic restoration $\boldsymbol{x}^+_{\mathsf{quad}}$ (dark grey dash dotted line) and the spatial sparsity restoration $\boldsymbol{x}^+_{\mathsf{Spar}}$ (thin dashed black line) compared to the true spectrum (thick grey line).

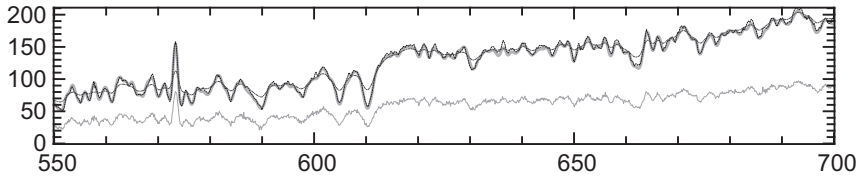


Fig. 5. Details (550 to 700 nm) of spectra integrated on a 3×3 region centered on the brighter spaxel of the central galaxy of the data (thin grey line), the quadratic restoration $\boldsymbol{x}^+_{\mathsf{quad}}$ (dark grey dash dotted line) and the spatial sparsity restoration $\boldsymbol{x}^+_{\mathsf{Spar}}$ (thin dashed black line) compared to the true spectrum (thick grey line).

along the spectral dimension. Its deblurring performance is assessed on simulations showing the clear improvement in terms of both resolution and denoising. The comparison of a quadratic and a spatial sparsity regularization, shows that spatial sparsity regularization are less prone to artifacts and preserves most of the spatial and spectral features. However, the non linearity introduced by such regularization slows down the convergence of the optimization algorithm. In that case, optimization algorithms as Alternating Direction Method of Multiplier (ADMM) seem to provides faster convergence than our VMLMB algorithm as we shown in Thiébaut & Soulez (2012).

This study as well as the one of Bourguignon *et al.* (2011a) show clearly the improvement given by a rigorous processing of hyperspectral astronomical data cube. However, two main problems remains in this field (i) the settings of the hyper-parameters, (ii) the estimation of the PSF. Our experience on SNIFS real data cube (Bongard *et al.* 2011) indicates that the hyper-parameters remains approximately identical for similar observations conditions. For the problems of the PSF estimation, we are currently studying blind deconvolution method where PSFs is estimated conjointly with the restoration only using the observations.

The authors would like to thank Roland Bacon P.I. of the MUSE instrument for providing the simulated data. This work is supported by the French ANR (*Agence Nationale de la Recherche*), Éric Thiébaut and Loïc Denis work for the MiTiV project (*Méthodes Inverses de Traitement en Imagerie du Vivant*, ANR-09-EMER-008) and Ferréol Soulez is funded by the POLCA project (*Percées astrophysiques grâce au traitement de données interférométriques polychromatiques*, ANR-10-BLAN-0511).

Our algorithm has been implemented and tested with YORICK(`http://yorick.sourceforge.net/`) which is freely available.

References

Akgun, T., Altunbasak, Y., & Mersereau, R., 2005, Image Proc., IEEE Trans., 14, 1860

Benazza-Benyahia, A., & Pesquet, J.C., 2006, in European Signal and Image Processing Conference, EUSIPCO'06, 5, 4 (Firenze, Italy)

Bertero, M., & Boccacci, P., 1998, Introduction to Inverse Problems in Imaging (Taylor & Francis)

Bobin, J., Moudden, Y., Starck, J.L., & Fadili, J., 2009, in Soc. Photo-Opt. Instrum. Eng. (SPIE) Conf. Ser., 7446, 42

Bongard, S., Thiébaut, E., Soulez, F., & Pecontal, E., 2009, in Proceedings of the First IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing (WHISPERS'09), cdrom (Grenoble, France)

Bongard, S., Soulez, F., Thiébaut, É., & Pecontal, E., 2011, MNRAS, 418, 258

Bourguignon, S., Mary, D., & Slezak, É., 2011a, Statistical Methodology

Bourguignon, S., Mary, D., & Slezak, É., 2011b, Selected Topics Signal Proc., IEEE J., 5, 1002

Courbin, F., Magain, P., Kirkove, M., & Sohy, S., 2000, ApJ, 529, 1136

Denis, L., Thiébaut, E., & Soulez, F., 2011, in 18th IEEE International Conference on Image Processing (Bruxelles, France), 2873

Duijster, A., Scheunders, P., & Backer, S.D., 2009, IEEE Trans. Geosc. Remote Sens., 47, 3892

Fornasier, M., & Rauhut, H., 2008, SIAM J. Numer. Anal., 46, 577

Galatsanos, N., & Chin, R., 1989, IEEE Trans. Acoustics, Speech, Signal Proc., 37, 415

Galatsanos, N., Katsaggelos, A., Chin, R., & Hillery, A., 1991, IEEE Trans. Signal Proc., 39, 2222

Gaucel, J.M., Guillaume, M., & Bourennane, S., 2006, European Signal Processing Conference

Henault, F., Bacon, R., Bonneville, C., *et al.*, 2003, Proc. SPIE, 4841, 1096

Hunt, B.R., & Kubler, O., 1984, IEEE Trans. Acoustics, Speech, Signal Proc., 32, 592

Katsaggelos, A., Lay, K., & Galatsanos, N., 1993, Image Proc., IEEE Trans., 2, 417

Kowalski, M., & Torrésani, B., 2009, Signal, Image Video Proc., 3, 251

Lucy, L., & Walsh, J., 2003, AJ, 125, 2266

Mugnier, L., Fusco, T., & Conan, J.-M., 2004, J. Opt. Soc. Am. A, 21, 1841

Neelamani, R., Choi, H., & Baraniuk, R., 2004, IEEE Trans. Signal Process., 52, 418

Nocedal, J., & Wright, S., 1999, Numerical optimization (Springer)

Rodet, T., Orieux, F., Giovannelli, J., & Abergel, A., 2008, IEEE J. Selected Topics Signal Proc., 2, 802

Soulez, F., Bongard, S., Thiébaut, E., & Bacon, R., 2011, in Proceedings of the Third IEEE-GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, cdrom (Lisbonne, Portugal)

Soulez, F., Thiébaut, E., Gressard, A., Dauphin, R., & Bongard, S., 2008, Proceeding of the 16th European Signal Processing Conference EUSIPCO (Lausanne, Suisse)

Tekalp, M., & Pavlovic, G., 1990, Signal Proc., 19, 221

Thiébaut, E., 2002, ed. J.-L. Starck, Astron. Data Anal. II., 4847, 174

Thiébaut, É., & Soulez, F., 2012, in SPIE Astronomical Telescopes+ Instrumentation, 84451C, International Society for Optics and Photonics

New Concepts in Imaging: Optical and Statistical Models
D. Mary, C. Theys and C. Aime (eds)
EAS Publications Series, **59** *(2013) 417–437*

SUPERVISED NONLINEAR UNMIXING OF HYPERSPECTRAL IMAGES USING A PRE-IMAGE METHODS

N.H. Nguyen[1], J. Chen[1,2], C. Richard[1], P. Honeine[2] and C. Theys[1]

Abstract. Spectral unmixing is an important issue to analyze remotely sensed hyperspectral data. This involves the decomposition of each mixed pixel into its pure endmember spectra, and the estimation of the abundance value for each endmember. Although linear mixture models are often considered because of their simplicity, there are many situations in which they can be advantageously replaced by nonlinear mixture models. In this chapter, we derive a supervised kernel-based unmixing method that relies on a pre-image problem-solving technique. The kernel selection problem is also briefly considered. We show that partially-linear kernels can serve as an appropriate solution, and the nonlinear part of the kernel can be advantageously designed with manifold-learning-based techniques. Finally, we incorporate spatial information into our method in order to improve unmixing performance.

1 Introduction

Pixel-vectors in hyperspectral images are usually mixtures of spectral components associated with a number of pure materials present in the scene (Keshava & Mustard 2002). In order to reveal embedded information, one needs to identify the endmembers present in each pixel and derive the relative proportions of different materials. Under the assumption that the endmembers have been determined *a priori* using some appropriate extraction approaches, see *e.g.*, (Boardman 1993; Nascimento & Bioucas-Dias 2005; Winter 1999), unmixing of hyperspectral images then consists of estimating the fractional abundances.

The abundance estimation problem has most often been solved based on the linear mixing model. Some examples are described in (Dobigeon *et al.* 2009; Heinz & Chang 2001; Honeine & Richard 2012; Theys *et al.* 2009). For instance, the

[1] Université de Nice Sophia-Antipolis, CNRS, Observatoire de la Côte d'Azur, France
[2] Université de Technologie de Troyes, CNRS, France

DOI: 10.1051/eas/1359019

FCLS method presented in (Heinz & Chang 2001) estimates the abundances by minimizing a mean-square-error criterion subject to linear equality and inequality constraints. The geometric strategy described in (Honeine & Richard 2012) reduces to calculate ratios of polyhedra volumes in the space spanned by the hyperspectral pixel-vectors. The main advantage of the former is the convexity of the optimization problem. A very low computational cost characterizes the latter.

In real-world scenes, the interaction between materials can generate nonlinear effects that influence the precision in abundance calculation, and can cause the abundance vectors to violate the non-negativity and the sum-to-one constraints. Nonlinear models can then be introduced to account for these effects, *e.g.*, the generalized bilinear model (Halimi *et al.* 2011), the post non-linear mixing model Jutten & Karhunen (2003), and the intimate model (Hapke 1981). Nonlinear unmixing methods attempt to invert these models and estimate the abundances. In (Halimi *et al.* 2011), a nonlinear unmixing algorithm for general bilinear mixture model was proposed. Based on Bayesian inference, this method however has a high computational complexity and is dedicated to the bilinear model. In (Nascimento & Bioucas-Dias 2009; Raksuntorn & Du 2010), the authors extended the collection of endmembers by adding artificial cross-terms of pure signatures to model light scattering effects on different materials. However, it is not easy to identify which cross-terms should be selected and added to the endmember dictionary. If all the possible cross-terms were considered, the set of endmembers would expand dramatically. Another possible strategy is to use manifold learning approaches such as Isomap (Tenenbaum *et al.* 2000), and LLE (Roweis & Saul 2000), which allow the use of linear methods in a linear space of non-linearly mapped data. Finally, in (Chen *et al.* 2013b), the authors formulated a new kernel-based paradigm that relies on the assumption that the mixing mechanism can be described by a linear mixture of endmember spectra, with additive nonlinear fluctuations defined in a reproducing kernel Hilbert space. This family of models has a clear physical interpretation, and allows to take complex interactions of endmembers into account.

The abundance estimation stage can be accomplished within the context where the abundances of the endmembers are known for some pixels, called training data. A learning process is then applied to estimate the abundances for the remaining pixels. See, *e.g.*, (Altmann *et al.* 2011b; Themelis *et al.* 2010; Tourneret *et al.* 2008). In (Altmann *et al.* 2011b), the map that approximates the abundances for any pixel-vector is a linear combination of radial basis functions. Its weights are estimated based on training samples. An orthogonal least-squares algorithm is then applied to reduce the number of radial basis functions in the model. In this chapter, we show that the learning process for abundance estimation based on training data can be viewed as a pre-image problem (Honeine & Richard 2011). While the mapping from input space to feature space is of primary importance in kernel methods, the reverse mapping from feature space back to input space can be also useful. Solving the pre-image problem within the context of our application consists of approximating the reverse mapping from the high-dimensional space of hyperspectral pixel-vectors to the low-dimensional space of abundance vectors.

We also consider the problem of kernel selection. As in (Chen *et al.* 2013b), we show that partially-linear kernels can serve as an appropriate solution. In this case, the nonlinear part of the kernel can be advantageously designed with manifold-learning-based techniques. We also investigate how to incorporate spatial correlation into the abundance estimation process. Total-variation regularization was introduced with success in (Iordache *et al.* 2011) to perform this task within the context of linear unmixing, and used in (Chen *et al.* 2013a) to extend the kernel-based framework presented in (Chen *et al.* 2013b). In the spirit of these recent results, a pre-image method for nonlinear spectral unmixing coupled with a ℓ_1-type spatial regularization is derived in this chapter.

This chapter is organized as follows. Section 2 describes the problem of nonlinear unmixing of hyperspectral data. It also introduces the pre-image problem within the context of kernel-based data processing. Section 3 solves the pre-image problem with kernel matrix regression in order to perform nonlinear unmixing of hyperspectral data. Section 4 addresses the question of kernel selection. Section 5 aims at solving the same problem with spatial regularization. Section 6 shows experimental results. Finally, Section 7 concludes the chapter.

2 Hyperspectral data unmixing formulated as a pre-image problem

2.1 Hyperspectral image mixing model

Let $\boldsymbol{r} = [r_1, r_2, \ldots, r_L]^\top$ be an observed hyperspectral pixel-vector, with L the number of spectral bands. We shall assume that $\boldsymbol{r}$ is a mixture of R endmember spectra $\boldsymbol{m}_i$. Let us denote by $\boldsymbol{M} = [\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_R]$ the L-by-R endmember matrix, and by $\boldsymbol{\alpha}$ the R-dimensional abundance vector associated with $\boldsymbol{r}$.

We first consider the linear mixing model where any observed pixel is a linear combination of the endmembers, weighted by the fractional abundances, that is,

$$\boldsymbol{r} = \boldsymbol{M}\boldsymbol{\alpha} + \boldsymbol{v} \tag{2.1}$$

where $\boldsymbol{v}$ is a noise vector. The abundance vector $\boldsymbol{\alpha}$ is usually determined by minimizing a cost function, *e.g.*, the mean-square reconstruction error, under the non-negativity and sum-to-one constraints

$$\begin{aligned} &\alpha_i \geq 0, \qquad \forall i \in 1, \ldots, R \\ &\sum_{i=1}^{R} \alpha_i = 1. \end{aligned} \tag{2.2}$$

The above model assumes that abundance vector $\boldsymbol{\alpha}$ lies on a simplex of R vertices. A direct consequence is that pixel-vectors $\boldsymbol{r}$ also lie in a simplex with vertices the R endmember spectra. There are many situations, involving multiple scattering effects, in which model (2.1) may be inappropriate and could be advantageously replaced by a nonlinear one. Consider the general mixing mechanism

$$\boldsymbol{r} = \boldsymbol{\Psi}(\boldsymbol{\alpha}, \boldsymbol{M}) + \boldsymbol{v} \tag{2.3}$$

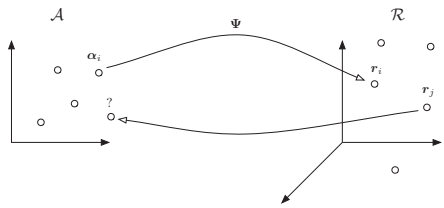


Fig. 1. The basic pre-image problem.

with $\boldsymbol{\Psi}$ an unknown function that defines the interactions between the endmembers in matrix $\boldsymbol{M}$ subject to conditions (2.2).

As illustrated in Figure 1, models (2.1) and (2.3) both rely on a mapping from the low-dimensional input space $\mathcal{A}$ of abundance vectors $\boldsymbol{\alpha}$ into the high-dimensional output space $\mathcal{R}$ of hyperspectral data $\boldsymbol{r}$. In this paper, we consider the problem of estimating abundances as a pre-image problem (Honeine & Richard 2011). Solving the pre-image problem, within a supervised learning context, consists of approximating the reverse mapping that allows to recover the abundance vector $\boldsymbol{\alpha}$ given any pixel-vector $\boldsymbol{r}$, based on training data.

2.2 Estimating a pre-image

This section introduces an original framework, based on the pre-image problem, for supervised unmixing of hyperspectral data. See Figure 2. In order to allow the model to better capture some complex mixing phenomena, we use a reproducing kernel Hilbert space (RKHS) framework in place of $\mathcal{R}$. We shall now review the main definitions and properties related to reproducing kernel Hilbert spaces (Aronszajn 1950).

Let $\mathcal{H}$ denote a Hilbert space of real-valued functions ψ on $\mathcal{R}$, and let $\langle \cdot\,, \cdot \rangle_{\mathcal{H}}$ be the inner product in $\mathcal{H}$. Suppose that the evaluation functional δ_r defined by $\delta_r[\psi] = \psi(\boldsymbol{r})$ is linear with respect to ψ and bounded, for all $\boldsymbol{r}$ in $\mathcal{R}$. By virtue of the Riesz representation theorem, there exists a unique positive definite function $\boldsymbol{r} \mapsto \kappa(\boldsymbol{r}, \boldsymbol{r}')$ in $\mathcal{H}$, denoted by $\kappa(\cdot, \boldsymbol{r}')$ and called *representer of evaluation* at $\boldsymbol{r}'$, which satisfies (Aronszajn 1950)

$$\psi(\boldsymbol{r}') = \langle \psi, \kappa(\cdot, \boldsymbol{r}') \rangle_{\mathcal{H}}, \quad \forall \psi \in \mathcal{H} \tag{2.4}$$

for every fixed $\boldsymbol{r}' \in \mathcal{R}$. A proof of this may be found in (Aronszajn 1950). Replacing ψ by $\kappa(\cdot, \boldsymbol{r})$ in (2.4) yields

$$\kappa(\boldsymbol{r}, \boldsymbol{r}') = \langle \kappa(\cdot, \boldsymbol{r}), \kappa(\cdot, \boldsymbol{r}') \rangle_{\mathcal{H}} \tag{2.5}$$

for all $\boldsymbol{r}, \boldsymbol{r}' \in \mathcal{R}$. Equation (2.5) is the origin of the generic term *reproducing kernel* to refer to κ. Denoting by $\boldsymbol{\Phi}$ the map that assigns the kernel function $\kappa(\cdot, \boldsymbol{r})$ to each input data $\boldsymbol{r}$, Equation (2.5) implies that

$$\kappa(\boldsymbol{r}, \boldsymbol{r}') = \langle \boldsymbol{\Phi}(\boldsymbol{r}), \boldsymbol{\Phi}(\boldsymbol{r}') \rangle_{\mathcal{H}}. \tag{2.6}$$

The kernel thus evaluates the inner product of any pair of elements of $\mathcal{R}$ mapped to the space $\mathcal{H}$ without any explicit knowledge of $\boldsymbol{\Phi}$ and $\mathcal{H}$. Within the machine learning area, this key idea is known as the *kernel trick*.

As shown in Figure 2, mapping back to the space $\mathcal{A}$ in order to recover $\boldsymbol{\alpha}$, given any $\kappa(\cdot, \boldsymbol{r})$ in $\mathcal{H}$, is a critical task. Generally, most of the features in $\mathcal{H}$ have no exact pre-image in $\mathcal{A}$. The pre-image problem in kernel-based machine learning has attracted a considerable interest in the last fifteen years. See (Honeine & Richard 2011) for an overview. In (Mika *et al.* 1999), Mika *et al.* introduced the problem and its ill-posedness. They also derived a fixed-point iteration strategy, potentially unstable, to find a solution without any guarantee of optimality. In (Kwok & Tsang 2003), Kwok *et al.* suggested a relationship between the distances in the feature space $\mathcal{H}$ and in the input space $\mathcal{A}$. Applying a multidimensional scaling technique yields an inverse map estimate, and thus a pre-image. This approach has opened the way to a range of other techniques that use training data in both spaces as prior information, such as manifold learning (Roweis & Saul 2000; Tenenbaum *et al.* 2000) and out-of-sample methods (Arias *et al.* 2007; Bengio *et al.* 2003).

In this chapter, we shall use an efficient method for solving the pre-image problem that was recently proposed in (Honeine & Richard 2011). It consists of deriving a transformation that preserves the inner products between training data, in the input space $\mathcal{A}$ and, with some abuse of notation, in the feature space $\mathcal{H}$. Given any $\boldsymbol{r}$, it thus allows to estimate $\boldsymbol{\alpha}$ from $\kappa(\cdot, \boldsymbol{r})$. The next section is dedicated to this approach, and its application to supervised unmixing.

3 Supervised unmixing

Given a set of training data $\{(\boldsymbol{\alpha}_1, \boldsymbol{r}_1), \ldots, (\boldsymbol{\alpha}_n, \boldsymbol{r}_n)\}$, we seek the pre-image $\boldsymbol{\alpha}$ in $\mathcal{A}$ of some arbitrary $\kappa(\cdot, \boldsymbol{r})$ of $\mathcal{H}$. The proposed approach consists of two stages: First, learning the reverse map; Then, estimating the pre-image.

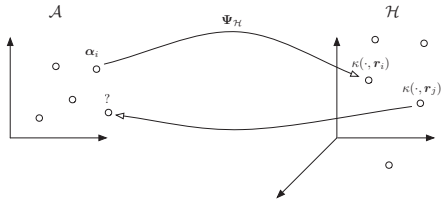


Fig. 2. The pre-image problem.

3.1 Stage 1: Learning the reverse map

By virtue of the Representer Theorem (Schölkopf *et al.* 2000), we know that we can limit our investigation to the space spanned by the n kernel functions

$\{\kappa(\cdot, \boldsymbol{r}_1), \ldots, \kappa(\cdot, \boldsymbol{r}_n)\}$. Let us focus on only a subspace spanned by ℓ functions to be determined, denoted by $\{\psi_1, \ldots, \psi_\ell\}$ with $\ell \leq n$, of the form

$$\psi_k = \sum_{i=1}^{n} \lambda_{ki}\, \kappa(\cdot, \boldsymbol{r}_i), \quad k = 1, \ldots, \ell. \tag{3.1}$$

We consider the analysis operator C: $\mathcal{H} \to \mathbb{R}^\ell$ defined as

$$C\varphi = [\langle \varphi, \psi_1 \rangle_{\mathcal{H}} \ldots \langle \varphi, \psi_\ell \rangle_{\mathcal{H}}]^\top. \tag{3.2}$$

Note that the k-th entry of the representation of any kernel function $\kappa(\cdot, \boldsymbol{r})$ is given by

$$\langle \kappa(\cdot, \boldsymbol{r}), \psi_k \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \lambda_{ki}\, \kappa(\boldsymbol{r}, \boldsymbol{r}_i). \tag{3.3}$$

It is interesting to note that $\langle \kappa(\cdot, \boldsymbol{r}), \psi_k \rangle_{\mathcal{H}} = \psi_k(\boldsymbol{r})$ by the reproducing property of the space $\mathcal{H}$. The kernel function $\kappa(\cdot, \boldsymbol{r})$ is thus represented by the ℓ-length vector

$$\boldsymbol{\psi}_r = [\psi_1(\boldsymbol{r})\, \psi_2(\boldsymbol{r})\, \ldots\, \psi_\ell(\boldsymbol{r})]^\top \tag{3.4}$$

with $\psi_k(\boldsymbol{r})$ defined in (3.3). In order to fully define the analysis operator C, that is, to estimate the λ_{ki}, we suggest to consider the following relationship between any inner product in the input space $\mathcal{A}$ and, with some abuse of notation, with its counterpart in the feature space $\mathcal{H}$

$$\boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_j = \boldsymbol{\psi}_{r_i}^\top \boldsymbol{\psi}_{r_j} + \epsilon_{ij}, \quad \forall\, i, j = 1, \ldots, n \tag{3.5}$$

where ϵ_{ij} denotes the lack-of-fit of the above model. Note that there is no constraint on the analysis functions ψ_k, except their form (3.1) and the goodness-of-fit constraint (3.5), because reconstruction from expansion coefficients is not considered. Let us now estimate the λ_{ki} in (3.3) so that the empirical variance of ϵ_{ij} is minimal, that is,

$$\min_{\lambda_{11}, \ldots, \lambda_{\ell n}} \frac{1}{2} \sum_{i,j=1}^{n} (\boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_j - \boldsymbol{\psi}_{r_i}^\top \boldsymbol{\psi}_{r_j})^2 + \eta\, P(\psi_1, \ldots, \psi_\ell) \tag{3.6}$$

where P is a regularization function, and η a tunable parameter used to control the tradeoff between fitting the data and smoothness of the solution. We shall use ℓ_2-norm penalization in this paper, defined as

$$P(\psi_1, \ldots, \psi_\ell) = \sum_{k=1}^{\ell} \|\psi_k\|_{\mathcal{H}}^2. \tag{3.7}$$

The optimization problem can be expressed in matrix form as

$$\min_{\boldsymbol{L}} \frac{1}{2} \|\boldsymbol{A} - \boldsymbol{K}\boldsymbol{L}^\top \boldsymbol{L}\boldsymbol{K}\|_F^2 + \eta\, \mathrm{trace}(\boldsymbol{L}^\top \boldsymbol{L}\boldsymbol{K}) \tag{3.8}$$

where $\boldsymbol{A}$ and $\boldsymbol{K}$ are the Gram matrices with (i,j)-th entries defined as $\boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_j$ and $\kappa(\boldsymbol{r}_i, \boldsymbol{r}_j)$, respectively, and $\boldsymbol{L}$ is the matrix with (i,j)-th entry given by λ_{ij}.

Taking the derivative of this cost with respect to $\boldsymbol{L}^\top \boldsymbol{L}$, rather than $\boldsymbol{L}$, we get

$$\hat{\boldsymbol{L}}^\top \hat{\boldsymbol{L}} = \boldsymbol{K}^{-1}(\boldsymbol{A} - \eta \boldsymbol{K}^{-1})\boldsymbol{K}^{-1}. \tag{3.9}$$

In the following, we shall show that only $\hat{\boldsymbol{L}}^\top \hat{\boldsymbol{L}}$ is needed to calculate the pre-image.

3.2 Stage 2: Estimate the pre-image

Let us first consider the case of any function φ of $\mathcal{H}$, which can be written as follows

$$\varphi = \sum_{i=1}^{n} \phi_i \, \kappa(\cdot, \boldsymbol{r}_i) + \varphi^\perp \tag{3.10}$$

with $\varphi^\perp$ an element of the orthogonal complement to the subspace spanned by the kernel functions $\kappa(\cdot, \boldsymbol{r}_i)$. Note at this point that the parameters ϕ_i are supposed to be known. In any case, they can be evaluated by projecting φ onto the subspace spanned the n kernel functions $\kappa(\cdot, \boldsymbol{r}_i)$, that is, by solving

$$\min_{\boldsymbol{\phi}} \|\varphi - \sum_{i=1}^{n} \phi_i \, \kappa(\cdot, \boldsymbol{r}_i)\|_{\mathcal{H}}^2. \tag{3.11}$$

This yields the n-by-n linear system of equations $\boldsymbol{K}\boldsymbol{\phi} = \boldsymbol{\varphi}_0$, where $\boldsymbol{\varphi}_0$ is the vector with i-th entry $\varphi(\boldsymbol{r}_i)$, and $\boldsymbol{\phi}$ stands for the vector with i-th entry ϕ_i, for $i = 1, \ldots, n$. Referring back to Equation (3.10), the k-th entry of the representation of φ by the analysis operator C, denoted by $\boldsymbol{\varphi}$, is given by

$$\langle \varphi, \psi_k \rangle_{\mathcal{H}} = \sum_{i,j=1}^{n} \phi_i \, \hat{\lambda}_{kj} \, \kappa(\boldsymbol{r}_i, \boldsymbol{r}_j), \tag{3.12}$$

where $\hat{\lambda}_{kj}$ is the (k,j)-th entry of the matrix $\hat{\boldsymbol{L}}$ estimated during Stage 1. This directly implies that $\boldsymbol{\varphi} = \hat{\boldsymbol{L}}\boldsymbol{K}\boldsymbol{\phi}$. Minimizing now the lack-of-fit (3.5), with respect to the pre-image $\boldsymbol{\alpha}$ given $\boldsymbol{\varphi}$, between $\boldsymbol{\alpha}^\top \boldsymbol{\alpha}_i$ and $\boldsymbol{\varphi}^\top \boldsymbol{\psi}_{r_i}$ for $i = 1, \ldots, n$, leads to the optimization problem

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= \arg\min_{\boldsymbol{\alpha}} \frac{1}{2} \|\boldsymbol{\Lambda}^\top \boldsymbol{\alpha} - \boldsymbol{K}\hat{\boldsymbol{L}}^\top \hat{\boldsymbol{L}}\boldsymbol{K}\boldsymbol{\phi}\|^2 \\ &= \arg\min_{\boldsymbol{\alpha}} \frac{1}{2} \|\boldsymbol{\Lambda}^\top \boldsymbol{\alpha} - (\boldsymbol{A} - \eta \boldsymbol{K}^{-1})\boldsymbol{\phi}\|^2 \end{aligned} \tag{3.13}$$

subject to the non-negativity and sum-to-one constraints (2.2). Here $\boldsymbol{\Lambda}$ is the matrix with i-th column the vector $\boldsymbol{\alpha}_i$.

Let us now consider the particular case where one seeks the pre-image $\boldsymbol{\alpha}$ of some kernel function $\kappa(\cdot, \boldsymbol{r}_0)$. Substituting φ by $\kappa(\cdot, \boldsymbol{r}_0)$ in Equation (3.11) leads us to the system $\boldsymbol{K}\boldsymbol{\phi} = \boldsymbol{\kappa}_0$, where $\boldsymbol{\kappa}_0$ is the vector with i-th entry $\kappa(\boldsymbol{r}_i, \boldsymbol{r}_0)$.

Minimizing the appropriate lack-of-fit (3.5) with respect to the pre-image $\boldsymbol{\alpha}$ leads us to the optimization problem

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\boldsymbol{\Lambda}^\top \boldsymbol{\alpha} - (\boldsymbol{A} - \eta \boldsymbol{K}^{-1})\boldsymbol{K}^{-1}\boldsymbol{\kappa}_0\|^2 \tag{3.14}$$

subject to the non-negativity and sum-to-one constraints (2.2). This convex optimization problem can be solved using the FCLS strategy to deal with the equality constraints (Heinz & Chang 2001), associated with a nonnegative least-mean-square algorithm. See, *e.g.*, (Chen *et al.* 2011) for an overview.

4 Kernel selection

The kernel function $\kappa(\cdot, \boldsymbol{r})$ maps the measurements $\boldsymbol{r}$ into a very high, even infinite, dimensional space $\mathcal{H}$. It characterizes the solution space for the possible nonlinear relationships between input data $\boldsymbol{\alpha}$ and output data $\boldsymbol{r}$. Classic examples of kernels are the Gaussian kernel $\kappa(\boldsymbol{r}_i, \boldsymbol{r}_j) = \exp\bigl(-\|\boldsymbol{r}_i - \boldsymbol{r}_j\|^2/2\sigma^2\bigr)$, with σ the kernel bandwidth, and the q-th degree non-homogeneous polynomial kernel $\kappa(\boldsymbol{r}_i, \boldsymbol{r}_j) = (1 + \boldsymbol{r}_i^\top \boldsymbol{r}_j)^q$, with $q \in \mathbb{N}^*$. We shall now make some suggestions for selecting specific kernels, before testing it in the next section. On the one hand, we shall briefly propose to design the kernel directly from data by using manifold learning techniques. On the other hand, we shall present a partially-linear kernel that has proved its efficiency for nonlinear unmixing (Chen *et al.* 2013b).

4.1 Kernel selection based on manifold learning techniques

In (Ham *et al.* 2003), the manifold learning problem is treated within the context of kernel PCA. The process of revealing the underlying structure of data is viewed as a nonlinear dimensionality reduction method, based on local information with LLE (Roweis & Saul 2000), or geodesic distance with Isomap (Tenenbaum *et al.* 2000). These techniques can be used to design kernels that preserve some aspects of the manifold structure of the space $\mathcal{R}$ to which the vectors $\boldsymbol{r}_i$ belong, in the feature space $\mathcal{H}$ of the functions $\kappa(\cdot, \boldsymbol{r}_i)$. We used such techniques in (Nguyen *et al.* 2012) for unmixing of hyperspectral data.

As an example, we consider radial basis kernels of the form $\kappa(\boldsymbol{r}_i, \boldsymbol{r}_j) = f(\|\boldsymbol{r}_i - \boldsymbol{r}_j\|)$ with $f \in \mathcal{C}_\infty$. A sufficient condition for this class of kernels to be positive-definite, and thus valid, is the complete monotonicity of the function f, which can be expressed as follows,

$$(-1)^k f^{(k)}(r) \geq 0, \quad \forall r \geq 0 \tag{4.1}$$

where $f^{(k)}$ denotes the k-th order derivative of f (Cucker & Smale 2002). Instead of using the euclidean distance $d_{ij} = \|\boldsymbol{r}_i - \boldsymbol{r}_j\|$ with f, we can use pairwise distances $d_{\text{iso},ij} = \|\boldsymbol{r}_i - \boldsymbol{r}_j\|_{\text{iso}}$ provided by Isomap. This approach consists of constructing a symmetric adjacency graph using a nearest neighborhood based criterion, and applying Dijkstra algorithm to compute the shortest path along

edges of this graph, between each pair of data. Unfortunately, the Gram matrix $\boldsymbol{K}_{\text{iso}}$ constructed in such a way has no guarantee of being positive definite. This difficulty can be overcome by using multidimensional scaling, which maps the data into a low-dimensional euclidean subspace where edge lengths are best preserved. An alternative is to force matrix $\boldsymbol{K}_{\text{iso}}$ to be positive definite using one of the approaches describes in (Muñoz & Diego 2006).

4.2 Partially-linear Kernel

Model (2.1)-(2.2) assumes that the relationship between the abundance vectors $\boldsymbol{\alpha}_i$ and the hyperspectral pixel-vectors $\boldsymbol{r}_i$ is linear. There are however many situations, involving multiple scattering effects, in which this model may be inappropriate and could be advantageously replaced by a nonlinear one. In (Chen *et al.* 2013a,b), we studied mixing models defined by a linear trend parameterized by the abundance vector, combined with a nonlinear fluctuation term. Extensive experiments, both with synthetic and real scenes, illustrated the flexibility and and the effectiveness of this class of models. In the spirit of these derivations, we suggest to consider kernels of the form

$$\kappa(\boldsymbol{r}_i, \boldsymbol{r}_j) = (1-\gamma)\, \boldsymbol{r}_i^\top \boldsymbol{\Sigma}\, \boldsymbol{r}_j + \gamma\, \kappa'(\boldsymbol{r}_i, \boldsymbol{r}_j) \tag{4.2}$$

with $\kappa'(\boldsymbol{r}_i, \boldsymbol{r}_j)$ a reproducing kernel, $\boldsymbol{\Sigma}$ a non-negative matrix, and γ a parameter in $[0, 1]$ to adjust the balance between the linear and the nonlinear kernels.

In all the experiments, we shall use the above kernel with $\boldsymbol{\Sigma} = (\boldsymbol{M}\boldsymbol{M}^\top)^\dagger$

$$\kappa(\boldsymbol{r}_i, \boldsymbol{r}_j) = (1-\gamma)\, \boldsymbol{r}_i^\top (\boldsymbol{M}\boldsymbol{M}^\top)^\dagger\, \boldsymbol{r}_j + \gamma\, \kappa'(\boldsymbol{r}_i, \boldsymbol{r}_j) \tag{4.3}$$

where $(\cdot)^\dagger$ stands for the pseudo-inverse. Indeed, for $\gamma = 0$, it can be shown that this kernel leads to the least-mean-square estimate of the abundance vector in the case of a linear mixing scenario.

5 Spatial regularization applied to supervised unmixing

5.1 Formulation

In the previous section, we showed how to estimate the abundances by learning a reverse mapping. This approach consisted of considering pixel vectors as if they were independent from their neighboring pixels. However, a fundamental property of remotely sensed data is that they convey multivariate information into a 2D pictorial representation. Hyperspectral analysis techniques can thus benefit from the inherent spatial-spectral duality in hyperspectral scenes. Following this idea, researchers exploited spatial information for endmember estimation (Martin & Plaza 2011; Rogge *et al.* 2007; Zortea & Plaza 2009) and pixel vectors classification (Fauvel *et al.* 2012, to appear; Li *et al.* 2011). Recently, spatial processing methods were also derived for semi-supervised unmixing (Chen *et al.* 2013a). In this section, we aim at improving the pre-image method by incorporating such information.

Following (Iordache *et al.* 2011), an optimization method based on split variable iteration is proposed to deal with this problem that suffers the non-smoothness of the regularization term.

Let us denote by $\boldsymbol{\Delta}$ the matrix of the abundance vectors, that is, $\boldsymbol{\Delta} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]$. In order to take the spatial relationship among pixels into consideration, we suggest to consider a general cost function of the form

$$J(\boldsymbol{\Delta}) = J_{\text{err}}(\boldsymbol{\Delta}) + \nu J_{\text{sp}}(\boldsymbol{\Delta}) \tag{5.1}$$

subject to the non-negativity constraint imposed on each entry of $\boldsymbol{\Delta}$, and the sum-to-one constraint imposed on each column of the matrix $\boldsymbol{\Delta}$, namely, on each $\boldsymbol{\alpha}_i$. For ease of notation, these two physical constraints will be expressed by

$$\begin{aligned} &\boldsymbol{\Delta} \succeq \mathbf{0} \\ &\boldsymbol{\Delta}^\top \mathbf{1}_R = \mathbf{1}_N. \end{aligned} \tag{5.2}$$

The function $J_{\text{err}}(\boldsymbol{\Delta})$ represents the modeling error, and $J_{\text{sp}}(\boldsymbol{\Delta})$ is a regularization term to promote similarity of the fractional abundances within neighboring pixels. The non-negative parameter ν is used to control the trade-off between data fidelity and pixel similarity.

To take spatial relationships among pixels into consideration, let us consider the following regularization function

$$J_{\text{sp}}(\boldsymbol{\Delta}) = \sum_{i=1}^{n} \sum_{j \in \mathcal{N}(i)} \|\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j\|_1 \tag{5.3}$$

where $\| \|_1$ denotes the vector ℓ_1-norm, and $\mathcal{N}(i)$ is the set of neighbors of the pixel i. This regularization term promotes spatial homogeneity as neighboring pixels may be characterized by similar abundances for most materials. Without any loss of generality, in this paper, we restrict the neighborhood of the pixel i by taking the 4 nearest pixels $i-1$ and $i+1$ (row adjacency), $i-w$ and $i+w$ (column adjacency). In this case, let us define the $(n \times n)$ matrices $\boldsymbol{H}_{\leftarrow}$ and $\boldsymbol{H}_{\rightarrow}$ as the two linear operators that compute the difference between any abundance vector and its left-hand neighbor, and right-hand neighbor, respectively. Similarly, let $\boldsymbol{H}_{\uparrow}$ and $\boldsymbol{H}_{\downarrow}$ be the linear operators that compute that difference with the top neighbor and the down neighbor, respectively. With these notations, the regularization function (5.3) can be rewritten in matrix form as

$$J_{\text{sp}}(\boldsymbol{\Delta}) = \|\boldsymbol{\Delta}\boldsymbol{H}\|_{1,1} \tag{5.4}$$

with $\boldsymbol{H}$ the $(n \times 4n)$ matrix $\left(\boldsymbol{H}_{\leftarrow}\, \boldsymbol{H}_{\rightarrow}\, \boldsymbol{H}_{\uparrow}\, \boldsymbol{H}_{\downarrow}\right)$ and $\| \|_{1,1}$ the sum of the ℓ_1-norms of the columns of a matrix. Note that this regularization function is convex but non-smooth.

Considering both the modeling error and the regularization term, the optimization problem becomes

$$\min_{\boldsymbol{\Delta}} \sum_{i=1}^{n} \frac{1}{2} \|\boldsymbol{\Lambda}^\top \boldsymbol{\alpha}_i - (\boldsymbol{A} - \eta \boldsymbol{K}^{-1}) \boldsymbol{K}^{-1} \boldsymbol{\kappa}_i\|^2 + \nu \|\boldsymbol{\Delta} \boldsymbol{H}\|_{1,1}$$
$$\text{subject to} \quad \boldsymbol{\Delta} \succeq 0 \quad \text{and} \quad \boldsymbol{\Delta}^\top \mathbf{1}_R = \mathbf{1}_N \tag{5.5}$$

where ν controls the trade-off between model fitting in each pixel and similarity among neighboring pixels. For ease of notation, in the following, we shall write $\boldsymbol{\Delta} \in \mathcal{S}_{+1}$ to denote the non-negativity and sum-to-one constraints.

5.2 Solution

Even though the optimization problem (5.5) is convex, it cannot be solved easily because of the non-smooth regularization term. In order to overcome this drawback, we rewrite it in the following equivalent form

$$\min_{\boldsymbol{\Delta} \in \mathcal{S}_{+1}} \sum_{i=1}^{n} \frac{1}{2} \|\boldsymbol{\Lambda}^\top \boldsymbol{\alpha}_i - (\boldsymbol{A} - \eta \boldsymbol{K}^{-1}) \boldsymbol{K}^{-1} \boldsymbol{\kappa}_i\|^2 + \nu \|\boldsymbol{U}\|_{1,1}$$
$$\text{subject to} \quad \boldsymbol{V} = \boldsymbol{\Delta} \quad \text{and} \quad \boldsymbol{U} = \boldsymbol{V} \boldsymbol{H} \tag{5.6}$$

where we have introduced two new matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, and two additional constraints. The matrix $\boldsymbol{U}$ will allow us to decouple the non-smooth ℓ_1-norm regularization functional from the main quadratic problem. The matrix $\boldsymbol{V}$ will relax connections between pixels. This variable-splitting approach was initially introduced in (Goldstein & Osher 2009).

As studied in (Goldstein & Osher 2009), the split Bregman iteration algorithm is an efficient method to deal with a broad class of ℓ_1-regularized problems. By applying this framework to (5.5), the following formulation is obtained

$$\boldsymbol{\Delta}^{(k+1)}, \boldsymbol{V}^{(k+1)}, \boldsymbol{U}^{(k+1)} = \underset{\boldsymbol{\Delta} \in \mathcal{S}_{+1}, \boldsymbol{V}, \boldsymbol{U}}{\arg\min} \sum_{i=1}^{n} \frac{1}{2} \|\boldsymbol{\Lambda}^\top \boldsymbol{\alpha}_i - (\boldsymbol{A} - \eta \boldsymbol{K}^{-1}) \boldsymbol{K}^{-1} \boldsymbol{\kappa}_i\|^2$$
$$+ \nu \|\boldsymbol{U}\|_{1,1} + \frac{\zeta}{2} \|\boldsymbol{\Delta} - \boldsymbol{V} - \boldsymbol{D}_1^{(k)}\|_F^2 + \frac{\zeta}{2} \|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{H} - \boldsymbol{D}_2^{(k)}\|_F^2 \tag{5.7}$$

with

$$\boldsymbol{D}_1^{(k+1)} = \boldsymbol{D}_1^{(k)} + \left(\boldsymbol{V}^{(k+1)} - \boldsymbol{\Delta}^{(k+1)} \right)$$
$$\boldsymbol{D}_2^{(k+1)} = \boldsymbol{D}_2^{(k)} + \left(\boldsymbol{V}^{(k+1)} \boldsymbol{H} - \boldsymbol{U}^{(k+1)} \right) \tag{5.8}$$

where $\| \|_F^2$ denotes the matrix Frobenius norm, and ζ is a positive parameter. Because we have split the components of the cost function, we can now solve the above minimization problem efficiently by iteratively minimizing the cost function with respect to $\boldsymbol{\Delta}$, $\boldsymbol{V}$ and $\boldsymbol{U}$ separately. We shall now describe the three steps that have to be performed.

5.2.1 Step 1: Optimization with respect to $\boldsymbol{\Delta}$

The optimization problem (5.7) reduces to

$$\boldsymbol{\Delta}^{(k+1)} = \underset{\boldsymbol{\Delta}\in\mathcal{S}_{+1}}{\arg\min} \sum_{i=1}^{n} \frac{1}{2} \left(\|\boldsymbol{\Lambda}^\top \boldsymbol{\alpha}_i - (\boldsymbol{A} - \eta \boldsymbol{K}^{-1})\boldsymbol{K}^{-1}\boldsymbol{\kappa}_i\|^2 + \zeta \|\boldsymbol{\alpha}_i - \boldsymbol{\xi}_i^{(k)}\|^2 \right) \tag{5.9}$$

where $\boldsymbol{\xi}_i^{(k)} = \boldsymbol{V}_i^{(k)} + \boldsymbol{D}_{1,i}^{(k)}$. Here, $\boldsymbol{V}_i^{(k)}$ and $\boldsymbol{D}_{1,i}^{(k)}$ denote the i-th column of $\boldsymbol{V}^{(k)}$ and $\boldsymbol{D}_1^{(k)}$, respectively. It can be observed that this problem can be decomposed into subproblems, each one involving an abundance vector $\boldsymbol{\alpha}_i$. This results from the use of the matrix $\boldsymbol{V}$ in the split iteration algorithm (5.7).

Let us now solve the local optimization problem

$$\begin{aligned} \boldsymbol{\alpha}_i^{(k+1)} = \underset{\boldsymbol{\alpha}_i}{\arg\min}\ & \frac{1}{2}\|\boldsymbol{\Lambda}^\top \boldsymbol{\alpha}_i - (\boldsymbol{A} - \eta \boldsymbol{K}^{-1})\boldsymbol{K}^{-1}\boldsymbol{\kappa}_i\|^2 + \zeta \|\boldsymbol{\alpha}_i - \boldsymbol{\xi}_i^{(k)}\|^2 \\ \text{subject to}\quad & \boldsymbol{\alpha}_i \succeq 0 \\ & \boldsymbol{\alpha}_i^\top \mathbf{1}_R = 1. \end{aligned} \tag{5.10}$$

Estimating $\boldsymbol{\alpha}_i$ reduces to a quadratic optimization problem with linear equality and inequality constraints, which can be efficiently solved by off-the-shelf methods. This process has to be repeated for $i = 1, \ldots, n$ in order to get $\boldsymbol{\Delta}^{(k+1)}$.

5.2.2 Step 2: Optimization with respect to $\boldsymbol{V}$

The optimization problem (5.7) now reduces to

$$\boldsymbol{V}^{(k+1)} = \underset{\boldsymbol{V}}{\arg\min} \|\boldsymbol{\Delta}^{(k+1)} - \boldsymbol{V} - \boldsymbol{D}_1^{(k)}\|_F^2 + \|\boldsymbol{U}^{(k)} - \boldsymbol{V}\boldsymbol{H} - \boldsymbol{D}_2^{(k)}\|_F^2. \tag{5.11}$$

Equating to zero the derivative of (5.11) with respect to $\boldsymbol{V}$ leads to

$$\left(\boldsymbol{\Delta}^{(k+1)} - \boldsymbol{V} - \boldsymbol{D}_1^{(k)}\right) + \left(\boldsymbol{U}^{(k)} - \boldsymbol{V}\boldsymbol{H} - \boldsymbol{D}_2^{(k)}\right)\boldsymbol{H}^\top = 0 \tag{5.12}$$

whose solution is then given by

$$\boldsymbol{V}^{(k+1)} = \left(\boldsymbol{\Delta}^{(k+1)} - \boldsymbol{D}_1^{(k)} + (\boldsymbol{U}^{(k)} - \boldsymbol{D}_2^{(k)})\boldsymbol{H}^\top\right)(\boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^\top)^{-1}. \tag{5.13}$$

As a conclusion, this subproblem has an explicit solution that involves the inverse of the matrix $(\boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^\top)$. The latter can be evaluated once the neighborhood relationship is defined.

5.2.3 Step 3: Optimization with respect to $\boldsymbol{U}$

The last optimization problem we have to consider is as follows

$$\boldsymbol{U}^{(k+1)} = \underset{\boldsymbol{U}}{\arg\min}\, \nu\|\boldsymbol{U}\|_{1,1} + \frac{\zeta}{2}\|\boldsymbol{U} - \boldsymbol{V}^{(k+1)}\boldsymbol{H} - \boldsymbol{D}_2^{(k)}\|_F^2. \tag{5.14}$$

Its solution can be expressed via the well-known soft threshold function

$$\boldsymbol{U}^{(k+1)} = \text{Thresh}\left(\boldsymbol{V}^{(k+1)}\boldsymbol{H} + \boldsymbol{D}_2^{(k)}, \frac{\nu}{\zeta}\right) \tag{5.15}$$

where $\text{Thresh}(\cdot, \tau)$ denotes the component-wise application of the soft threshold function defined as

$$\text{Thresh}(x, \tau) = \text{sign}(x)\ \max(|x| - \tau, 0). \tag{5.16}$$

As in Step 2, the third subproblem has an explicit solution. The computational time is also almost negligible.

To conclude, the problem (5.6) is solved by iteratively applying (5.7) and (5.8), where the optimization of (5.7) can be performed by applying Steps 1 to 3. These iterations continue until some stopping criterion is satisfied. It can be shown that, if the problem (5.7) has a solution $\boldsymbol{\Delta}^*$ given any $\zeta > 0$, then the generated sequence $\boldsymbol{\Delta}^{(k)}$ converges to $\boldsymbol{\Delta}^*$ (Eckstein & Bertsekas 1992).

6 Simulation results

In this section, we shall experiment the pre-image method with and without spatial regularization in order to evaluate the benefit of using the latter. We shall compare it with state-of-the-art methods.

6.1 Experiments with the pre-image method

Spatial regularization is not addressed in this subsection. Two synthetic scenes were generated with real material spectra, on the one hand from abundance vectors uniformly distributed in the simplex defined by the non-negativity and the sum-to-one constraints, and on the other hand from abundance vectors lying on a manifold.

6.1.1 Experiments on synthetic images with uniformly-distributed abundances

We shall first report some experimental results on synthetic images, which were generated by linear and nonlinear mixing of several endmember signatures. The materials that were considered are alunite, calcite, epidote, kaolinite, and buddingtonite. There spectra were extracted from the ENVI software library, and consisted of 420 contiguous bands, covering wavelength ranging from 0.3951 to 2.56 micrometers. They were used to synthesize 50×50 images with different mixture models, each providing $n = 2500$ pixels for evaluating and comparing several unmixing algorithms. These three models were: the linear model, the bilinear mixture model with attenuation factors $\gamma_{ij} = 1$ (Halimi *et al.* 2011), and the post-nonlinear mixing model (PNMM) defined by (Jutten & Karhunen 2003)

$$\boldsymbol{r} = (\boldsymbol{M}\boldsymbol{\alpha})^{\xi} + \boldsymbol{v} \tag{6.1}$$

where $(\cdot)^{\xi}$ denotes the exponential value ξ applied to each entry of the input vector. This parameter was set to 0.7. The abundance vectors $\boldsymbol{\alpha}_i$, with $i = 1, \ldots, 2500$, were uniformly generated in the simplex defined by the non-negativity and the sum-to-one constraints. In the first scene, only three materials were selected to generate images: epidote, kaolinite, buddingtonite. In the second scene, five materials were used: alunite, calcite, epidote, kaolinite, buddingtonite. These scenes were corrupted with an additive white Gaussian noise $\boldsymbol{v}$ with two levels of SNR, 15 dB and 30 dB.

The following algorithms were considered in our experiments.

- **The Fully Constrained Least Square method (FCLS)** (Heinz & Chang 2001): This algorithm relies on a semi-supervised learning setting in the sense that unmixing is performed using endmember spectra as prior information. It is based on a linear mixture model, and provides the optimal solution in the least-mean-square sense subject to the non-negativity and the sum-to-one constraints.
- **The Kernel Fully Constrained Least Square method (KFCLS)** (Broadwater *et al.* 2007): This semi-supervised nonlinear algorithm is the kernel-based counterpart of FCLS, obtained by replacing all the inner products in FCLS by kernel functions. In the experiments, as for our pre-image algorithm, we used the Gaussian kernel with kernel bandwidth $\sigma = 4$.
- **The Bayesian algorithm derived for generalized bilinear model (BilBay)** (Halimi *et al.* 2011): This semi-supervised method is based on appropriate prior distributions for the unknown abundances, which must satisfy the non-negativity and sum-to-one constraints, and then derives joint posterior distribution of these parameters. A Metropolis-within-Gibbs algorithm is used to estimate the unknown model parameters.
- **The RBF-with-OLS method (RBF-OLS)** (Altmann *et al.* 2011a): As our pre-image method, this supervised algorithm aims at learning a nonlinear reverse mapping from $\mathcal{R}$ to $\mathcal{A}$. The estimator is a linear combination of radial basis functions with centers chosen from the training data through an OLS procedure.
- **The pre-image algorithm proposed in this paper**: The inhomogeneous polynomial kernel (P) of degree $d = 2$, the Gaussian kernel (G) with kernel bandwidth $\sigma = 4$, and the partially-linear kernel (PL) associating a linear kernel and a Gaussian kernel with $\sigma = 4$. The parameter γ combining these two kernels, and the regularization coefficient η, were set to 10^{-1} and 10^{-3}.

The cardinality of the training data set was fixed to 200 in order to reach an appropriate compromise between the computational cost and the performance. The root mean square error (RMSE) between the true and the estimated abundance vectors $\boldsymbol{\alpha}_i$ and $\hat{\boldsymbol{\alpha}}_i$ was used to compare the performance of the five algorithms. Results for Scene 1 and Scene 2 unmixing, with three and five endmember materials, are reported in Table 1 and Table 2, respectively.

Table 1. Scene 1 (three materials): RMSE comparison.

	SNR = 30 dB			SNR = 15 dB		
	linear	bilinear	PNMM	linear	bilinear	PNMM
FCLS	0.0037	0.0758	0.0604	0.0212	0.0960	0.0886
KFCLS	0.0054	0.2711	0.2371	0.0296	0.2694	0.2372
BilBay	0.0384	0.0285	0.1158	0.1135	0.1059	0.1191
RBF-OLS	0.0144	0.0181	0.0170	0.0561	0.0695	0.0730
Pre-image method (P)	0.0139	0.0221	0.0129	0.0592	0.0601	0.0764
Pre-image method (G)	0.0086	0.0104	0.0103	0.0422	0.0561	0.0597
Pre-image method (PL)	0.0072	0.0096	0.0098	0.0372	0.0395	0.0514

Table 2. Scene 2 (five materials): RMSE comparison.

	SNR = 30 dB			SNR = 15 dB		
	linear	bilinear	PNMM	linear	bilinear	PNMM
FCLS	0.0134	0.1137	0.1428	0.0657	0.1444	0.1611
KFCLS	0.0200	0.2051	0.1955	0.0890	0.1884	0.1572
BilBay	0.0585	0.0441	0.1741	0.1465	0.1007	0.1609
RBF-OLS	0.0200	0.0236	0.0259	0.0777	0.0805	0.0839
Pre-image method (P)	0.025	0.0267	0.0348	0.0905	0.0903	0.1000
Pre-image method (G)	0.0186	0.0233	0.0245	0.0775	0.0778	0.0875
Pre-image method (PL)	0.0148	0.0184	0.0203	0.0636	0.0616	0.0763

Consider first the semi-supervised algorithms. The FCLS method achieves a very low RMSE for linearly-mixed images because it was initially derived for the linear mixing model. As a consequence, it produces a large RMSE with nonlinearly-mixed images. The KFCLS should have overcome this drawback. It however performs worse than FCLS, even with nonlinearly-mixed images as it does not clearly investigate nonlinear interactions between materials (Chen *et al.* 2013b). BilBay algorithm was derived for the bilinear mixing model, and thus achieves very good performance with bilinearly-mixed images. Nevertheless, its performance severely degrades when dealing with a nonlinear mixing model for which it was not originally designed. Consider now the supervised algorithms. The pre-image method and RBF-OLS outperforms all the semi-supervised algorithms when dealing with non-linearly mixed images. Of course, they make use of more information to achieve this performance. Our approach is however much more flexible than RBF-OLS since it can be associated with any reproducing kernel. In particular, as already observed in (Chen *et al.* 2013b), the experiments demonstrate the benefit of using a partially-linear kernel.

6.1.2 Experiment on synthetic images: Test with swiss-roll data

In order to highlight the flexibility of our approach with respect to kernel selection, we shall now show that kernels designed with manifold learning techniques can be advantageously used. Let us consider the well-known swiss-role artificial data set for illustration purpose. It consists of random samples in a two-dimensional

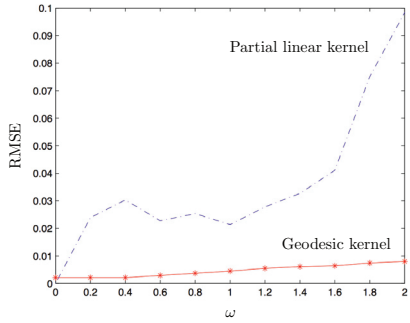


Fig. 3. Geodesic kernel *vs.* partially-linear kernel in the case where data lie in a manifold.

simplex, transformed into a three-dimensional nonlinear manifold by projection on a swiss-roll structure. The non-linearity of the swiss-roll data is parameterized by a variable ω. The coordinate of a data point $\boldsymbol{r}_i$ as a function of the local abundance $\boldsymbol{\alpha}_i$ are expressed by

$$\begin{cases} r_{i1} = \alpha_{i1} \sin(\omega \alpha_{i1}) + 1 \\ r_{i2} = \alpha_{i1} \cos(\omega \alpha_{i1}) + 1 \\ r_{i3} = \alpha_{i2} + 1. \end{cases} \tag{6.2}$$

Following the sum-to-one constraint, the abundance of the third endmembers can be generated by $\alpha_{i3} = 1 - (\alpha_{i1} + \alpha_{i2})$. By setting a single abundance equal to one, and the two others to zero, we obtain the endmember spectra

$$\begin{cases} \boldsymbol{m}_1 = [\sin(\omega) + 1,\ \cos(\omega) + 1,\ 1]^\top \\ \boldsymbol{m}_2 = [1,\ 1,\ 2]^\top \\ \boldsymbol{m}_3 = [1,\ 1,\ 1]^\top. \end{cases} \tag{6.3}$$

Swiss-roll data unmixing was performed with our pre-image algorithm, based on 100-sample training sets, for ω values in the interval $[0, 2]$. The partially-linear kernel with Gaussian kernel whose bandwidth was set to $\sigma = 4$, and the kernel based on geodesic distances provided by Isomap, were considered. The geodesic kernel was constructed using the geodesic distance matrix provided by Isomap and Djisktra algorithms. Note that this matrix was converted into a positive definite matrix using a technique described in (Muñoz & Diego 2006). Figure 3 clearly shows that the geodesic kernel is much more appropriate than the partially-linear kernel in the case where the data lie in a manifold, and the performance of the algorithm is quite steady even for large ω values.

6.2 Experiments with the spatially-regularized pre-image method

Two spatially correlated abundance maps were generated for the following experiments. The endmembers were randomly selected from the spectral library ASTER (Baldridge *et al.* 2009). Each signature of this library has reflectance

values measured over 224 spectral bands, uniformly distributed in the interval $3 - 12$ micrometers. Two synthetic abundance maps identical to (Iordache *et al.* 2011) were used.

The first data cube, denoted by IM1, and containing 50×50 pixels, was generated using five signatures randomly selected from the ASTER library. Pure regions and mixed regions involving between 2 and 5 endmembers, distributed spatially in the form of square regions, were generated. The background pixels were defined as mixtures of the same 5 endmembers with the abundance vector $[0.1149, 0.0741, 0.2003, 0.2055, 0.4051]^\top$. The first row in Figure 4 shows the true fractional abundances for each endmember. The reflectance samples were generated with the bilinear mixing model, based on the 5 endmembers, and corrupted by a zero-mean white Gaussian noise $\boldsymbol{v}_i$ with a SNR of 20 dB, namely,

$$\boldsymbol{r}_i = \boldsymbol{M}\boldsymbol{\alpha}_i + \sum_{p=1}^{R} \sum_{q=p+1}^{R} \alpha_{n,p}\, \alpha_{n,q}\, \boldsymbol{m}_p \otimes \boldsymbol{m}_q + \boldsymbol{v}_i \tag{6.4}$$

with $\otimes$ the Hadamard product.

The second data cube, denoted by IM2 and containing 100×100 mixed pixels, was generated using 5 endmember signatures. The abundance maps of the endmembers are the same as for the image DC2 in (Iordache *et al.* 2011). The first row of Figure 5 depicts the true distribution of these 5 materials. Spatially homogeneous areas with sharp transitions can be clearly observed. Based on these abundance maps, an hyperspectral data cube was generated with the bilinear model (6.4) applied to the 5 endmember spectral signatures. The scene was also corrupted by a zero-mean white Gaussian noise $\boldsymbol{v}_i$ with a SNR of 20 dB.

Algorithms with and without spatial regularization were compared in order to demonstrate the effectiveness of adding this type of information. Unsupervised algorithms that do not use spatial information, were also considered for comparison purpose. The tuning parameters of the algorithms were set using preliminary experiments on independent data, via a simple search over predefined grids.

1. The linear unmixing method FCLS (Heinz & Chang 2001): The regularization parameter λ was varied in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ in order to determine the best configuration.

2. The pre-image algorithm without spatial regularization: The partially-linear kernel with $\gamma = 0.1$ was considered. It was associated with the Gaussian kernel. The bandwidth of the latter was varied in [0.5, 5], and finally set to 4. The regularization parameter η of the pre-image algorithm was varied in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, and was finally set to 10^{-3}. The size of the training set was set to 200.

3. The pre-image algorithm with spatial regularization: The same parameter values as above were considered for this algorithm in order to clearly evaluate the interest of taking spatial information into account. The parameters ζ and ν, which are specifically related to the spatial regularization, were tuned as explained below.

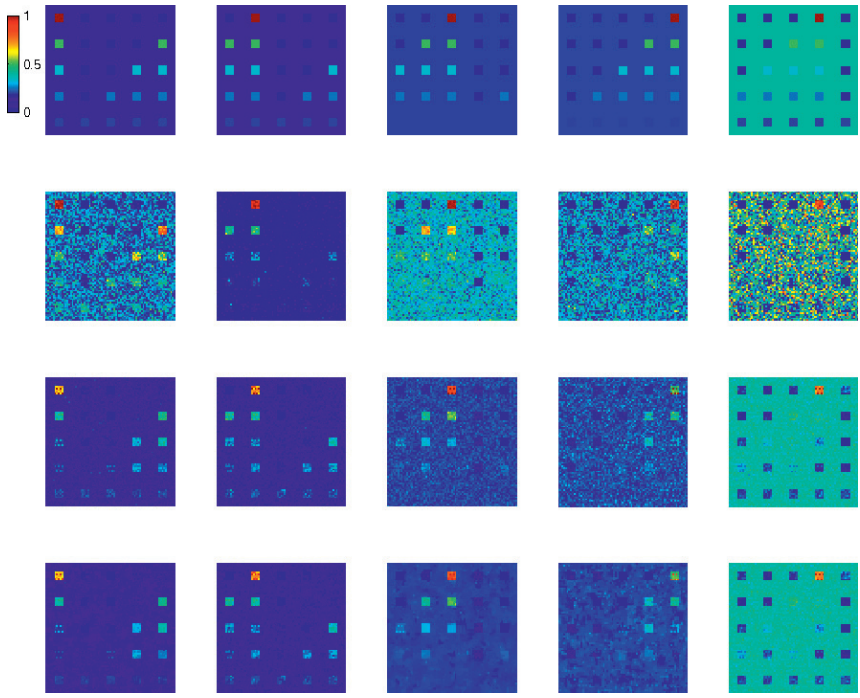


Fig. 4. Estimated abundance maps for IM1. From *top* to *bottom*: true abundance map, FCLS, pre-image method, pre-image method with spatial regularization.

With the image IM1, the preliminary tests led us to $\lambda = 10^{-2}$ for FCLS, and $\zeta = 1$, $\nu = 0.1$ for the proposed algorithm. With the image IM2, these tests led to $\lambda = 0.01$ for FCLS, and $\zeta = 20$, $\nu = 0.5$ for the proposed algorithm.

The estimated abundances are presented in Figures 4 and 5. The reconstruction errors (RMSE) are reported in Table 3. For both images IM1 and IM2, it can be observed that when applied on nonlinearly mixed data, the linear unmixing method FCLS has large reconstruction errors. The proposed pre-image method allows to notably reduce this error in the mean sense, but the estimated abundance maps are corrupted by a noise that partially masks spatial structures of the materials. Finally, the proposed spatially-regularized method has lower reconstruction error and clearer abundance maps. Using spatial information obviously brings advantages to the nonlinear unmixing process.

Table 3. Comparison of the RMSE for IM1 and IM2.

Algorithms	IM1	IM2
FCLS	0.1426	0.0984
pre-image	0.0546	0.0712
pre-image with reg.	0.0454	0.0603

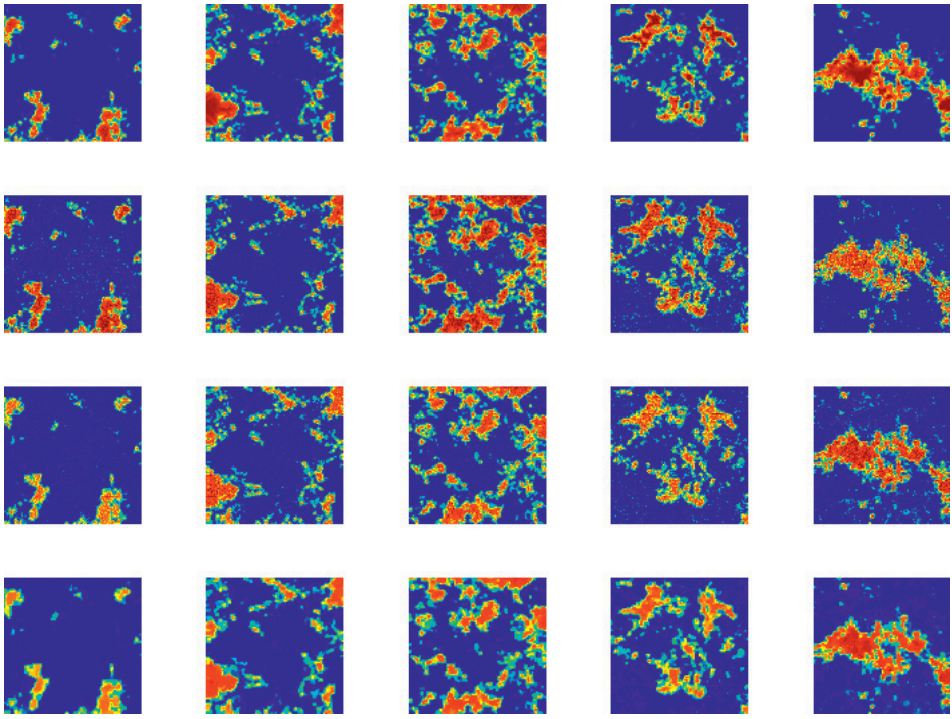


Fig. 5. Estimated abundance maps for IM2. From *top* to *bottom*: true abundance map, FCLS, pre-image method, pre-image method with spatial regularization.

7 Conclusion

In this chapter, we introduced an hyperspectral unmixing algorithm based on the pre-image principle, which is usually addressed by the community of machine learning. Our contribution is two-fold in the sense that the pre-image algorithm described here, and its spatially-regularized counterpart, are both original. We showed that these techniques can be advantageously applied for supervised unmixing provided that labeled pixel-vectors are available.

References

Altmann, Y., Dobigeon, N., McLaughlin, S., & Tourneret, J.-Y., 2011a, in Proc. IEEE IGARSS

Altmann, Y., Halimi, A., Dobigeon, N., & Tourneret, J.-Y., 2011b, in Proc. IEEE IGARSS

Arias, P., Randall, G., & Sapiro, G., 2007, in Proc. IEEE CVPR

Aronszajn, N., 1950, Trans. Amer. Math. Soc., 68, 337

Baldridge, A.M., Hook, S.J., Grove, C.I., & Rivera, G., 2009, Remote Sensing Env., 113, 711

Bengio, Y., Paiement, J.-F., Vincent, P., *et al.*, 2003, in Proc. NIPS

Boardman, J., 1993, in Proc. AVIRIS, 1, 11

Broadwater, J., Chellappa, R., Banerjee, A., & Burlina, P., 2007, in Proc. IEEE IGARSS, 4041

Chen, J., Richard, C., Bermudez, J.-C.M., & Honeine, P., 2011, IEEE Trans. Sig. Proc., 59, 5225

Chen, J., Richard, C., & Honeine, P., 2013a, IEEE Trans. Geosci. Remote Sens.

Chen, J., Richard, C., & Honeine, P., 2013b, IEEE Trans. Sig. Proc., 61, 480

Cucker, F., & Smale, S., 2002, Bull. Am. Math. Soc., 39, 1

Dobigeon, N., Moussaoui, S., Coulon, M., Tourneret, J.-Y., & Hero, A.O., 2009, IEEE Trans. Sig. Proc., 57, 4355

Eckstein, J., & Bertsekas, D., 1992, Math. Prog., 55, 293

Fauvel, M., Tarabalka, Y., Benediktsson, J.A., Chanussot, J., & Tilton, J., 2012, Proc. IEEE, to appear

Goldstein, T., & Osher, S., 2009, SIAM J. Imaging Sci., 2, 323

Halimi, A., Altmann, Y., Dobigeon, N., & Tourneret, J.-Y., 2011, IEEE Trans. Geosci. Remote Sens., 49, 4153

Ham, J., Lee, D. D., Mika, S., & Schölkopf, B., 2003, A kernel view of the dimensionality reduction of manifolds, Tech. Rep. TR-110 (Max-Planck-Institut für biologische Kybernetik)

Hapke, B., 1981, J. Geophys. Res., 86, 3039

Heinz, D.C., & Chang, C.-I., 2001, IEEE Trans. Geosci. Remote Sens., 39, 529

Honeine, P., & Richard, C., 2011, IEEE Signal Proc. Mag., 28, 77

Honeine, P., & Richard, C., 2012, IEEE Trans. Geosci. Remote Sens., 50, 2185

Iordache, M.-D., Bioucas-Dias, J.-M., & Plaza, A., 2011, in Proc. IEEE WHISPERS

Jutten, C., & Karhunen, J., 2003, in Proc. ICA, 245

Keshava, N., & Mustard, J.F., 2002, IEEE Signal Proc. Mag., 19, 44

Kwok, J., & Tsang, I., 2003, in Proc. ICML

Li, J., Bioucas-Dias, J.-M., & Plaza, A., 2011, IEEE Trans. Geosci. Remote Sens., 50, 809

Martin, G., & Plaza, A., 2011, IEEE Geosci. Remote Sens. Lett., 8, 745

Mika, S., Schölkopf, B., Smola, A., *et al.*, 1999, in Proc. NIPS

Muñoz, A., & Diego, I.M., 2006, in Lecture Notes in Computer Science, Structural, Syntactic, and Statistical Pattern Recognition, Vol. 4109, ed. D.-Y. Yeung, J. Kwok, A. Fred, F. Roli & D. Ridder (Springer), 764

Nascimento, J.M.P., & Bioucas-Dias, J.M., 2005, IEEE Trans. Geosci. Remote Sens., 43, 898

Nascimento, J.M.P., & Bioucas-Dias, J.-M., 2009, in Proc. SPIE, 7477

Nguyen, N.H., Richard, C., Honeine, P., & Theys, C., 2012, in Proc. IEEE IGARSS

Raksuntorn, N., & Du, Q., 2010, IEEE Geosci. Remote Sens. Lett., 7, 836

Rogge, D.M., Rivard, B., Zhang, J., *et al.*, 2007, Remote Sensing Env., 110, 287

Roweis, S., & Saul, L., 2000, Science, 2323

Schölkopf, B., Herbrich, R., & Williamson, R., 2000, A generalized representer theorem, Tech. Rep. NC2-TR-2000-81, NeuroCOLT, Royal Holloway College (University of London, UK)

Tenenbaum, J.B., de Silva, V., & Langford, J.C., 2000, Science, 290, 2319

Themelis, K., Rontogiannis, A.A., & Khoutroumbas, K., 2010, in Proc. IEEE ICASSP, 1194

Theys, C., Dobigeon, N., Tourneret, J.-Y., & Lanteri, H., 2009, in Proc. IEEE SSP

Tourneret, J.-Y., Dobigeon, N., & Chang, C.-I., 2008, IEEE Trans. Sig. Proc., 5, 2684

Winter, M.E., 1999, Proc. SPIE Spectrometry V, 3753, 266

Zortea, M., & Plaza, A., 2009, IEEE Trans. Geosci. Remote Sens., 47, 2679

Index