

Inversion of second-difference operators with application to infrared astronomy

M. Bertero \S , P. Boccacci \S and M. Robberto $\dagger\ddagger$

\S INFN and DISI, Università di Genova, Via Dodecaneso 35, I-16146 Genova, Italy

\dagger Space Telescope Science Institute, 3700 San Martin Dr., Baltimore, MD, 21218, USA

Abstract. Ground-based astronomical imaging at thermal-infrared wavelengths requires a differential technique, known as *chopping & nodding*, to extract the weak astronomical signal from the huge background due to atmosphere and telescope emission. The resulting image is the second difference of the intensity distribution of the astronomical target, and leads to an image restoration problem that can be formulated as the inversion of a second-difference operator. In general, the problem is affected by a huge non-uniqueness, but the degeneracy is reduced when convenient boundary conditions can be used. In particular, if the target field is surrounded by empty sky, it is natural to require that the solution is zero at the boundary of the image. In this paper we investigate the problem of inverting a second-difference operator with the addition of Dirichlet boundary conditions. We show that the related discrete problem can be reduced to the inversion of a non-singular positive definite matrix whose eigenvalues and eigenvectors can be explicitly given. We also give an inversion formula and we investigate the numerical stability of the solution. Since in most practical situations, the inversion problem is ill-conditioned, we give a reformulation as a least-squares problem. The advantage is that it is possible to introduce additional constraints such as the non-negativity of the solution. Moreover, we introduce an iterative algorithm converging to the unique non-negative least-squares solution. Since the latter can be still affected by numerical instability, we show that early stopping of the iterations has a regularization effect. We conclude with a discussion of the observational implications of our analysis.

1. Introduction

Observations of astronomical targets at wavelengths $\sim 10-20 \mu\text{m}$ (*thermal infrared*), corresponding to the peak of blackbody emission at temperatures $\sim 300 - 150 \text{ K}$) are hampered by the strong thermal emission from the atmosphere and the telescope itself. In order to extract the weak astronomical signal from the large and rapidly varying

\ddagger on assignment from the Space Telescopes Division of the European Space Agency

background, a differential technique, known as *chopping & nodding* (see, for instance, [1]), is used in many circumstances.

Chopping refers to rapid modulation (a few Hz) of the secondary mirror of the telescope, in order to illuminate the detector alternatively with the source and with a nearby sky region. The difference between the two images is the so-called chopped image (first-difference operator). However, the use of two slightly different optical paths adds an offset to the signal that can be subtracted by switching the two beams, i.e. pointing the telescope so to have the source on the previous sky beam and another sky region in the original source beam. In this way a second chopped image is obtained. Beam switching is done at lower frequency (of the order of once per minute), and the characteristic oscillatory movement of the entire telescope is called nodding. The difference between the two chopped images is just the chopped & nodded one (second-difference operator).

Let us denote by $f(\xi, \eta)$ the brightness distribution of the astronomical target, where ξ, η are angular variables in the sky; this function is non-negative and can be assumed to vanish in regions of empty sky. Moreover, let us denote by $g(\xi, \eta)$ the image of $f(\xi, \eta)$ obtained by means of the chopping & nodding technique. If chopping and nodding are in the direction of the angular variable η , then the relationship between g and f is:

$$g(\xi, \eta) = -f(\xi, \eta - \Delta) + 2f(\xi, \eta) - f(\xi, \eta + \Delta) \quad , \quad (1.1)$$

where Δ is the chopping amplitude. Therefore, if we neglect noise contamination, for a fixed value of ξ the chopped & nodded image is just the second difference of the unknown brightness distribution.

A chopped & nodded image in general contains negative values due to the negative counterparts of the sources. Therefore, except for compact sources whose angular size is smaller than the chopping amplitude, this technique does not provide a reliable image of the astronomical target. An image restoration problem arises, which consists in estimating the target f from the knowledge of the detected image g .

From Eq. 1.1 it follows that, for any given ξ , we have to solve the same one-dimensional problem which can be formulated as the inversion of the second-difference operator (we omit here the ξ variable):

$$(D^{(2)}f)(\eta) = -f(\eta - \Delta) + 2f(\eta) - f(\eta + \Delta) \quad . \quad (1.2)$$

This operator is bounded and injective in $L^2(\mathbb{R})$. However, the injectivity does not hold in practical situations. Any image g has a finite extent corresponding, in a first instance, to the field imaged by the panoramic detector or, possibly, to a strip of sky obtained as a mosaic of individual chopped & nodded images. Therefore, the domain of g is a bounded interval $\mathcal{I} = [\eta_{min}, \eta_{max}] \subset \mathbb{R}$ and, as follows from Eq. 1.1, the values of g in this interval receive contributions from the values of f in the broader interval $\mathcal{D} = [-\Delta + \eta_{min}, \eta_{max} + \Delta]$. If we consider $D^{(2)}$ as an operator from $L^2(\mathcal{D})$ into $L^2(\mathcal{I})$, then

$D^{(2)}$ has an infinite dimensional null space, whose complete characterization is given in [9]. It follows that the solution of the restoration problem is not unique.

A first attempt to restore chopped & nodded images is proposed in [2], where it is suggested to use three images of the same target taken with different optimized chopping amplitudes. The inversion strategy, essentially an inverse Fourier filter without further constraints, does not produce in general satisfactory results [12]. Our group has taken a different approach, proposing an iterative method, the so-called projected Landweber method, for approximating non-negative solutions of the discrete restoration problem [4]. The mathematical structure of the imaging matrix is analysed in [5]; the performance, and the limits, of the restoration algorithm are demonstrated in [6] on real chopped & nodded images taken at the United Kingdom Infrared Telescope (UKIRT). In many cases, the method provides excellent results, especially if data acquisition is carefully optimized. On the other hand, it also produces annoying artifacts. In [5] it is shown that the structure of the artifacts is somehow predictable, as it is related to the mathematical properties of the imaging matrix. In other words, artifacts are intrinsically related to the problem of extracting a particular solution from the huge set of all possible ones, i. e. to the non-uniqueness of the solution of the restoration problem.

In order to reduce, or remove, the non-uniqueness, the use of boundary conditions is fundamental. In [7] we investigate the case of periodic boundary conditions. These, wherever applicable, lead to a new iterative method which, in general, provides better restorations than those obtained with the previous one. However, if it is possible to assume empty sky in the regions above and below the image field (this condition can be easily satisfied by considering suitable mosaics of chopped & nodded images) the most natural conditions are of the Dirichlet type. The purpose of this paper is just to investigate the restoration problem in such a case.

In Section 2 we discretize the imaging problem and we show that, when Dirichlet boundary conditions are met, it can be reduced to the inversion of a non-singular matrix. The spectral properties of this matrix are investigated in Section 3 in a particular case that we call the *integer case*, i.e. when the chopping amplitude is an integer multiple of the size of the detector pixel. In Section 4 we give an explicit inversion formula of this matrix, which is in general ill-conditioned. In Section 5 we reformulate the image restoration as a least-squares problem, so that regularizing approaches can be introduced; in particular, we propose an iterative method converging to the unique non-negative least-squares solution. Since also this solution can be strongly affected by noise propagation, regularization can be obtained by early stopping of the iterations. In Section 6 we present the results of a few numerical experiments and in Section 7 we discuss the observational implications of our analysis.

2. Discretization of the problem and the boundary conditions

Let us denote by δ the angular size of the detector pixels and by M and N the numbers of rows and columns of the discrete image formed by the pixel values $g_{m,n}$. We assume that the indices m, n take the values $m = 1, 2, \dots, M$ and $n = 1, 2, \dots, N$, respectively.

We first assume, for simplicity, that the chopping amplitude Δ is an integer multiple K of the pixel size δ , i. e.

$$\Delta = K\delta \quad . \quad (2.1)$$

At the end of this Section we drop this assumption which is never exactly satisfied in practice.

If we denote by ξ_1, η_1 the coordinates of the central point of a reference pixel $\{1,1\}$, then the coordinates of the central point of the pixel $\{m, n\}$ are given by:

$$\xi_n = \xi_1 + \delta (n - 1) , \quad \eta_m = \eta_1 + \delta (m - 1) , \quad (2.2)$$

and, by assuming that the response function of the pixel is uniform, the corresponding pixel value $g_{m,n}$ can be modeled as the integral of $g(\xi, \eta)$ over the pixel domain:

$$g_{m,n} = \int_{-\delta/2}^{\delta/2} d\xi' \int_{-\delta/2}^{\delta/2} d\eta' \quad g(\xi_n + \xi', \eta_m + \eta') \quad . \quad (2.3)$$

In a similar way we can model the pixel values $f_{m,n}$ of the object f . The indexes m, n can take the same values used in the case of $g_{m,n}$ for the pixels which belong to \mathcal{I} . However the index m must be extended to take also the values $-K + 1, -K + 2, \dots, 0$ and $M + 1, M + 2, \dots, M + K$ for characterizing the rows of the image which belong to $\mathcal{D} \setminus \mathcal{I}$.

By inserting Eq. 1.1 and Eq. 2.1 into Eq. 2.3 and observing that $\eta_m \pm \Delta = \eta_{m \pm K}$, we obtain the following relationship between $g_{m,n}$ and $f_{m,n}$:

$$g_{m,n} = -f_{m-K,n} + 2f_{m,n} - f_{m+K,n} \quad . \quad (2.4)$$

In general, the measured values of $g_{m,n}$ are contaminated by noise.

Since the relationship between $g_{m,n}$ and $f_{m,n}$ does not depend on n , we can drop this index. We will denote by \mathbf{g} the vector, of length M , corresponding to a generic column of the image array and by \mathbf{f} the vector, of length $M + 2K$, corresponding to a generic column of the object array. The restoration problem, to be solved column by column, consists in the estimation of \mathbf{f} being given \mathbf{g} . Since we have to determine $M + 2K$ unknowns with M data, the solution is not unique.

If we assume now that the regions above and below the image domain are regions of empty sky, then the vector \mathbf{f} corresponding to a generic column of the object satisfies the boundary conditions:

$$\begin{aligned} f_m &= 0 \quad ; & (2.5) \\ m &= -K + 1, -K + 2, \dots, 0; M + 1, M + 2, \dots, M + K. \end{aligned}$$

By combining Eq. 2.4 and Eq. 2.5 we find that the restoration problem is equivalent to solve the linear system:

$$A_{M,K} \mathbf{f} = \mathbf{g} \quad , \quad (2.6)$$

where $A_{M,K}$ is the $M \times M$ matrix given by:

$$\begin{aligned} (A_{M,K} \mathbf{f})_m &= 2f_m - f_{m+K} \quad , \quad 1 \leq m \leq K \\ &= -f_{m-K} + 2f_m - f_{m+K} \quad , \quad K+1 \leq m \leq M-K \\ &= -f_{m-K} + 2f_m \quad , \quad M-K+1 \leq m \leq M . \end{aligned} \quad (2.7)$$

If Δ is not an integer multiple of the pixel size δ , we can write Δ as follows:

$$\Delta = (K + r) \delta \quad , \quad (2.8)$$

with K integer and $0 \leq r \leq 1$. Inserting Eq. 1.1 and Eq. 2.8 into Eq. 2.3, a simple change of variable shows that the pixel value $g_{m,n}$ receives not only the complete contribution of the values of f in the pixel $\{m, n\}$ but also partial contributions of its values in the pixels $\{m \mp K, n\}$ and $\{m \mp (K+1), n\}$. More precisely, the partial contributions of the pixels $\{m \mp K, n\}$ are given by integrals extended over intervals of length $(1-r)\delta$ in the variable η' , while the partial contributions of the pixels $\{m \mp (K+1), n\}$ are given by integrals extended over intervals of length $r\delta$, also in the variable η' . As long as f is not rapidly varying inside the pixel domain, these integrals can be approximated respectively by a fraction $1-r$ and a fraction r of the corresponding pixel values. As a consequence, the chopped and nodded image is approximately given by:

$$\begin{aligned} g_{m,n} &= (1-r)(-f_{m-K,n} + 2f_{m,n} - f_{m+K,n}) \\ &\quad + r(-f_{m-(K+1),n} + 2f_{m,n} - f_{m+(K+1),n}). \end{aligned} \quad (2.9)$$

If $f(\xi, \eta) = 0$ outside the image domain $[\eta_{min}, \eta_{max}]$, then the imaging process is described by the matrix:

$$A_{M,K+r} = (1-r) A_{M,K} + r A_{M,K+1} \quad (2.10)$$

and the image restoration problem is the solution of the linear equation:

$$A_{M,K+r} \mathbf{f} = \mathbf{g} \quad . \quad (2.11)$$

In the following, the matrices $A_{M,K}$ or $A_{M,K+r}$ will be called the *imaging matrices* (in the *integer case* when Eq. 2.1 holds true).

3. Spectral properties of the imaging matrix in the integer case

Let us consider the $N \times N$ matrix which, for $N = M$, coincides with the imaging matrix $A_{M,1}$:

$$A_N = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdot & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdot & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdot & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdot & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdot & 0 & -1 & 2 \end{pmatrix} . \quad (3.1)$$

As it is known, this matrix, which derives from the discretization of the second derivative with the addition of Dirichlet boundary conditions, is an element of the τ_N algebra; it is a non-singular matrix whose eigenvalues and eigenvectors are given by:

$$\begin{aligned} \lambda_k^{(N)} &= 4 \sin^2 \left(\frac{\pi k}{2(N+1)} \right) , & (3.2) \\ (\mathbf{v}_k^{(N)})_m &= \sqrt{\frac{2}{N+1}} \sin \left(\frac{\pi k m}{N+1} \right) ; \quad m = 1, 2, \dots, N , \\ k &= 1, 2, \dots, N . \end{aligned}$$

Since $\sin^2(x)$ is an increasing function of x in the interval $[0, \pi/2]$, the eigenvalues form an increasing sequence. Therefore the condition number of the matrix is given by:

$$\text{cond}(A_N) = \frac{\sin^2 \left(\frac{\pi N}{2(N+1)} \right)}{\sin^2 \left(\frac{\pi}{2(N+1)} \right)} \simeq \frac{4N^2}{\pi^2} , \quad (3.3)$$

and it grows quadratically for large values of N . In this Section we show that the eigenvalues and eigenvectors of the imaging matrix $A_{M,K}$ are given in terms of the eigenvalues and eigenvectors of matrices of type A_N , with suitable values of N .

We denote by q the integer part of M/K , $q = [M/K]$, so that, if $R < K$ is the remainder, we have:

$$M = qK + R . \quad (3.4)$$

Moreover, we group the M values of the index m into the following K subgroups:

$$J_j = \{m \in (1, 2, \dots, M) \mid m \equiv j \pmod{K}\} , \quad j = 1, \dots, K ; \quad (3.5)$$

for $j = 1, \dots, R$, J_j contains $q + 1$ values while, for $j = R + 1, \dots, K$, it contains q values. Finally, we introduce the permutation matrix:

$$\Pi_M = ((\mathbf{e}_l)_{l \in J_1} \mid (\mathbf{e}_l)_{l \in J_2} \mid \dots \mid (\mathbf{e}_l)_{l \in J_K}) , \quad (3.6)$$

where $\{\mathbf{e}_l\}_{l=1}^M$ is the set of the canonical basis vectors. We remark that the columns of Π_M are partitioned into R blocks of $q + 1$ columns and $K - R$ blocks of q columns.

Then it is easy to prove the following result:

Theorem 3.1. The matrix $T_{M,K} = \Pi_M^T A_{M,K} \Pi_M$ has a block-diagonal structure:

$$T_{M,K} = \begin{pmatrix} I_R \otimes A_{q+1} & 0 \\ 0 & I_{K-R} \otimes A_q \end{pmatrix}, \quad (3.7)$$

where I_N denotes the $N \times N$ identity matrix and \otimes is the Kronecker product.

From this result and the explicit expression of the eigenvalues and eigenvectors of the matrix A_N , as given in Eq. 3.2, it is also easy to prove that:

Theorem 3.2. The eigenvalues $\lambda_j^{(M,K)}$ of $A_{M,K}$ are the eigenvalues $\lambda_k^{(q+1)}$ of A_{q+1} , counted with multiplicity R , and the eigenvalues $\lambda_k^{(q)}$ of A_q , counted with multiplicity $K - R$. The ordering is as follows:

$$\lambda_1^{(q+1)} < \lambda_1^{(q)} < \lambda_2^{(q+1)} < \lambda_2^{(q)} < \dots < \lambda_q^{(q)} < \lambda_{q+1}^{(q+1)}, \quad (3.8)$$

so that the multiplicity is alternately R and $K - R$. Moreover, if we denote by $\mathbf{v}_{k,j}^{(q+1)}$, $j = 1, \dots, R$, the eigenvectors associated with the eigenvalue $\lambda_k^{(q+1)}$ and by $\mathbf{v}_{k,j}^{(q)}$, $j = R + 1, \dots, K$, the eigenvectors associated with $\lambda_k^{(q)}$, we have:

$$(\mathbf{v}_{k,j}^{(q+1)})_m = \begin{cases} \sqrt{\frac{2}{q+2}} \sin\left(\frac{\pi km}{q+2}\right) & \text{if } m \equiv j \pmod{K} \\ 0, & \text{otherwise,} \end{cases} \quad (3.9)$$

$$(\mathbf{v}_{k,j}^{(q)})_m = \begin{cases} \sqrt{\frac{2}{q+1}} \sin\left(\frac{\pi km}{q+1}\right) & \text{if } m \equiv j \pmod{K} \\ 0, & \text{otherwise.} \end{cases} \quad (3.10)$$

These eigenvectors become sparse when K becomes large.

From Eq. 3.8 it follows that the condition number of the matrix $A_{M,K}$ is given by:

$$\text{cond}(A_{M,K}) = \frac{\sin^2\left(\frac{\pi(q+1)}{2(q+2)}\right)}{\sin^2\left(\frac{\pi}{2(q+2)}\right)} \simeq \frac{4q^2}{\pi^2}, \quad (3.11)$$

and therefore it grows quadratically for large q .

In Fig. 1a) we plot the behaviour of the logarithm of the condition number as a function of q ; the initial value for $q = 1$ is 3, but the condition number grows rapidly and is already of the order of 10^2 for $q = 14$. Moreover, in order to clarify the dependence

of the condition number on K when the size M of the image is fixed, in Fig. 1b) we plot its behaviour as a function of K in the case $M = 128$. It decreases for increasing K and, for a sufficiently large value of K one can have a mild ill-conditioning of the problem. The jumps in the plot occur at the values of K corresponding to variations in the value of q .

We conclude this Section with the following remarks:

- in most circumstances, such as in the case of large mosaics produced to satisfy the condition of empty sky at the edges of the image, or in the case of small chopping amplitude (a limitation probably intrinsic to observations with 8-10 m class telescopes), the value of q can be quite large so that the corresponding restoration problem is very ill-conditioned;
- the ill-conditioning of the problem is due to the small eigenvalues associated with the low-frequency eigenvectors.

This feature is opposite to that of more traditional restoration problems, such as image deconvolution, where the small eigenvalues are associated with the high frequency eigenvectors. The basic reason is that the problem we are investigating is related to the inversion of a differential operator, while standard deconvolution problems imply the inversion of an integral operator. This remark must be taken into account when considering the application of regularization methods to the restoration of chopped & nodded images; it will be discussed in Section 5.

4. The inversion formula in the integer case

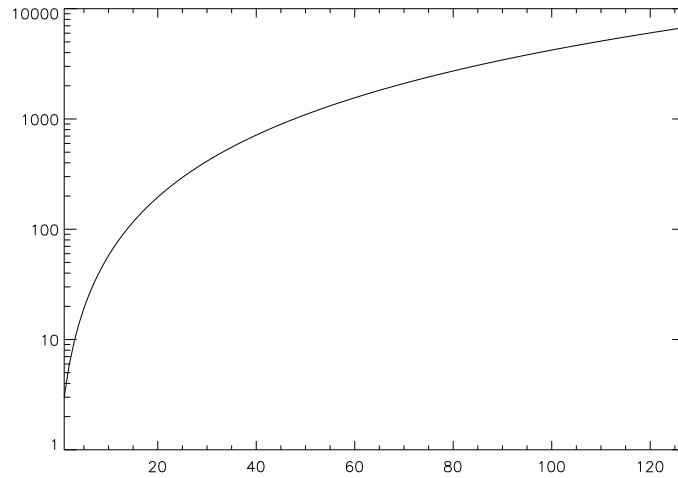
Let us consider first the problem of determining an explicit inversion formula for the matrix of Eq. 3.1. This is equivalent to solve the problem:

$$\begin{aligned} f_{m-1} - 2 f_m + f_{m+1} &= - g_m ; \quad m = 1, 2, \dots, N ; \\ f_0 &= f_{N+1} = 0 . \end{aligned} \quad (4.1)$$

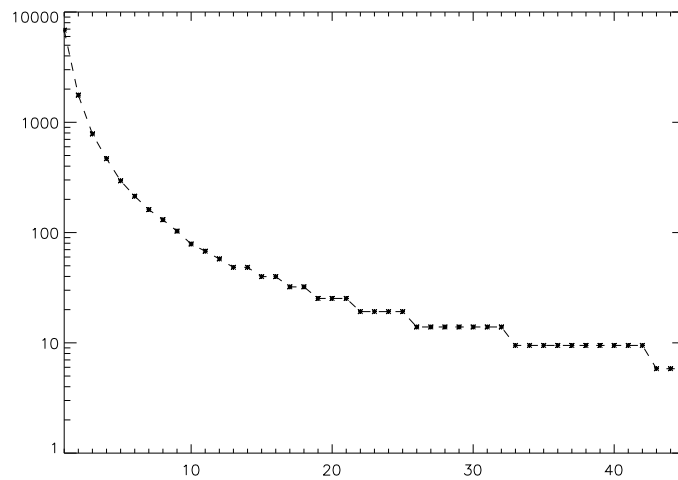
The solution can be easily obtained by the method of variation of constants: first, we write f_m as a linear combination of the two solutions of the homogeneous problem satisfying separately the two boundary conditions, namely $f_m^{(+)} = N + 1 - m$ and $f_m^{(-)} = m$; then, we replace the constants with vectors satisfying an additional condition which implies that they can be determined by solving first-difference problems. The result is:

$$f_m = \sum_{l=1}^m \frac{(N+1-m)l}{N+1} g_l + \sum_{l=m+1}^N \frac{m(N+1-l)}{N+1} g_l , \quad (4.2)$$

the second term being dropped in the case $m = N$.



a)



b)

Figure 1. Behaviour of the condition number of the imaging matrix $A_{M,K}$: a) semi-logarithmic plot of the condition number as a function of $q = [M/K]$ (the limit for $q = 1$ is 3); b) plot of the condition number as a function of K in the case $M = 128$. The last plotted value is 5.828, which is constant up to $K = 64$ while the value is 3 ($q = 1$) from $K = 64$ to 127.

We can consider now the problem of determining the inversion formula for the imaging matrix $A_{M,K}$. This is equivalent to solve the problem

$$\begin{aligned} f_{m-K} - 2 f_m + f_{m+K} &= - g_m ; \quad m = 1, \dots, M ; \\ f_0 &= f_{-1} = \dots = f_{-K+1} = 0 , \\ f_{M+1} &= f_{M+2} = \dots = f_{M+K} = 0 , \end{aligned} \quad (4.3)$$

and it is easy to see that it can be reduced to the solution of K problems of the type described in Eq. 4.1, one for each of the groups of indices defined in Eq.3.5. Therefore, if we write the indices of J_j in the form $j + K(m - 1)$, with m running from 1 to M_j ($M_j = q + 1$ for $j = 1, \dots, R$; $M_j = q$ for $j = R + 1, \dots, K$), we have:

$$\begin{aligned} f_{j+K(m-1)} &= \sum_{l=1}^m \frac{(M_j + 1 - m) l}{M_j + 1} g_{j+K(l-1)} + \\ &+ \sum_{l=m+1}^{M_j} \frac{m (M_j + 1 - l)}{M_j + 1} g_{j+K(l-1)} ; \\ j &= 1, \dots, K ; \quad m = 1, \dots, M_j , \end{aligned} \quad (4.4)$$

the second sum at the r.h.s. being dropped when $m = M_j$.

The analysis of the previous Section implies that, in the presence of data noise, this formula can be affected by strong noise propagation, the amount of the effect being dependent on the value of q . To illustrate this point, we have performed a few numerical experiments in the following way:

- we define a target object \mathbf{f}_0 , of length M , satisfying the boundary conditions;
- we compute the noise-free image defined as $A_{M,K}\mathbf{f}_0$;
- we compute a noisy image by adding Gaussian white noise, with zero mean and variance σ , to $A_{M,K}\mathbf{f}_0$, i. e. the model image is given by:

$$\mathbf{g} = A_{M,K}\mathbf{f}_0 + \mathbf{w} , \quad (4.5)$$

where \mathbf{w} denotes the noise term;

- finally we solve Eq. 2.6 with \mathbf{g} as given by Eq. 4.5.

In our tests on the numerical stability of the inversion formula we use a vector of length 128 obtained from a Gaussian pulse, centered at $m = 64$, with $\sigma_0 = 7$ and a maximum value of about 5×10^6 ; moreover, we consider two values of chopping amplitude: $K = 3$ and 37. In the first case the condition number is 784 and in the second one 9.47. In both cases the noise-free images are perturbed by means of Gaussian noise, as indicated above, with two values of σ : 5×10^3 and 5×10^4 (respectively 0.1 and 1% of the maximum value of the Gaussian pulse). These figures are adequate for observations of relatively bright stars.

In Figure 2 we give the results obtained by applying to these four cases the exact inversion formula of Eq. 4.5. Noise propagation is evident in the case $K = 3$ (Figure 2c)-

2e)), while in the case $K = 37$ the result is satisfactory with 0.1% noise (Figure 2d). We point out that the oscillations observable in the case $K = 3$ presumably are not due to the propagation of high frequency noise but to the amplification of the noise affecting the low frequency components and to the particular structure of the corresponding eigenvectors.

Indeed, in the case $K = 3$, the smallest eigenvector is $\lambda_1^{(43)} = 5.096 \times 10^{-3}$ with multiplicity 2 while the next one is $\lambda_1^{(42)} = 5.335 \times 10^{-3}$ with multiplicity 1. Division by these eigenvalues amplifies the corresponding noisy components of the chopped & noded image, whose eigenvectors have alternately one non-zero and two zero components. This approximate ‘‘periodicity’’ appears in the restored image. In the case of 1% noise it is also evident the effect of the incorrect restoration of the components corresponding to the subsequent eigenvalues, whose eigenvectors have one change of sign in the image domain.

This property explains the strong asymmetry observed in Figure 2e) as well as in the subsequent Figures 3c) and 4c). In all these simulations we have used exactly the same data, i.e. the same noise realization. If we change the noise realization, then also the asymmetry of the solution will change.

5. Constrained iterative regularization method

The ill-conditioning of the imaging matrix $A_{M,K}$ implies that, in general, the solution of the linear system of Eq. 2.6 does not provide sensible reconstructions. The same remark applies to the linear system of Eq. 2.11 because $A_{M,K+r}$ is also ill-conditioned. The results of this Section apply to both cases and therefore, in order to simplify the notations, we will indicate by A the generic imaging matrix: it coincides with $A_{M,K}$ in the integer case, otherwise with $A_{M,K+r}$.

Since the imaging matrix A is non-singular and positive definite, its square root is well defined and non-singular; if we write:

$$B = A^{1/2} \quad , \quad \mathbf{h} = A^{-1/2} \mathbf{g} \quad , \quad (5.1)$$

then the solution \mathbf{f} of the linear system $A\mathbf{f} = \mathbf{g}$ is also the solution of the least-squares problem:

$$\mathbf{f} = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \| B \mathbf{x} - \mathbf{h} \|_2 : \mathbf{x} \in \mathbb{R}^M \right\} \quad , \quad (5.2)$$

where $\| \cdot \|_2$ denotes the usual Euclidean norm. The advantage of this reformulation is that it leads quite naturally to the introduction of regularization methods. The most simple example is the well-known Tikhonov regularized solution, which is given by:

$$\mathbf{f}^\mu = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \| B \mathbf{x} - \mathbf{h} \|_2^2 + \mu \| \mathbf{x} \|_2^2 : \mathbf{x} \in \mathbb{R}^M \right\} \quad . \quad (5.3)$$

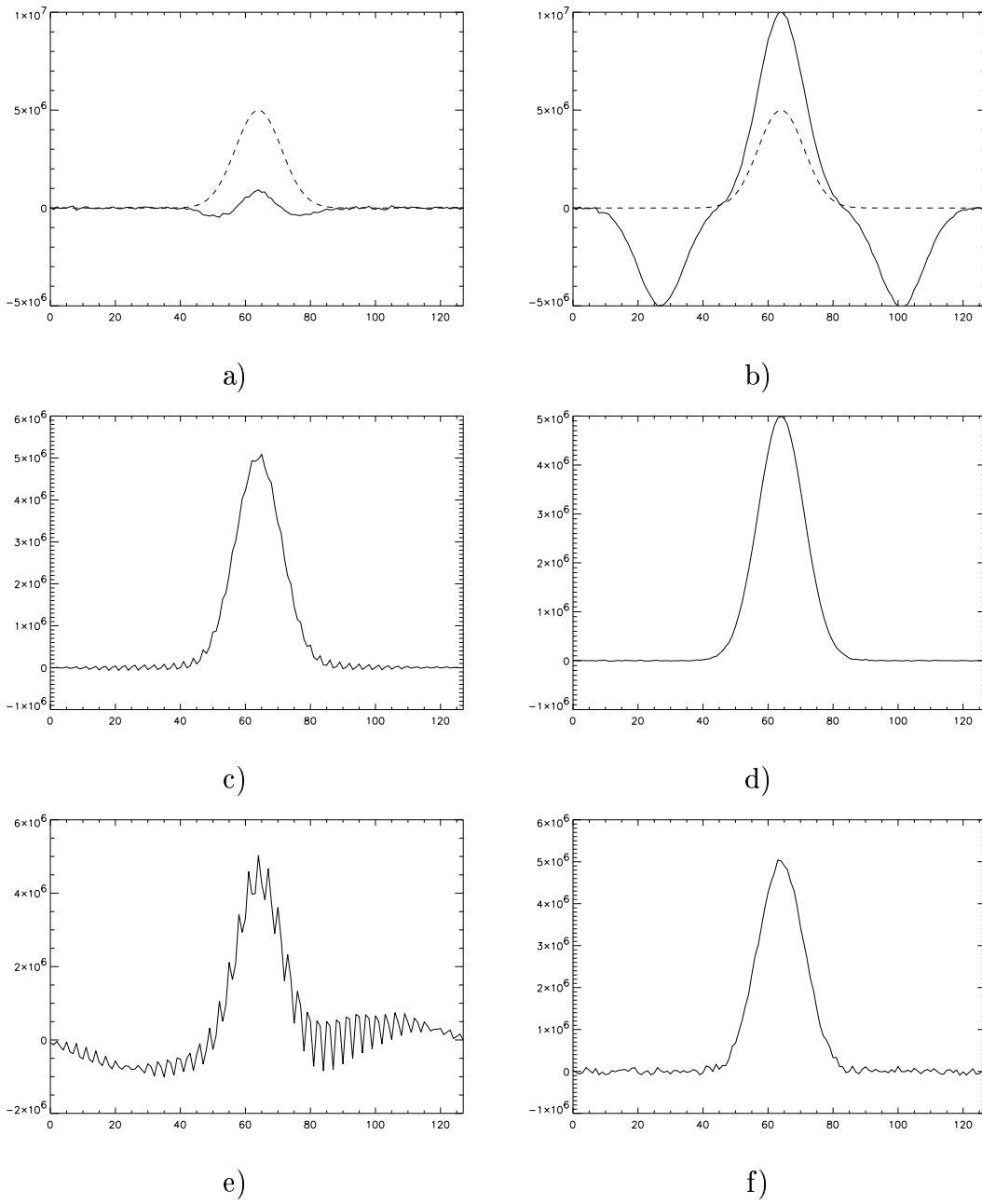


Figure 2. Examples of reconstructions of a Gaussian pulse: a)-b) the Gaussian pulse and its noisy images respectively with $K = 3$ and $K = 37$ (1% noise, as described in the text); c)-d) the restorations of the Gaussian pulse from the images with $K = 3$ and $K = 37$ and 0.1% noise; e)-f) the restorations of the Gaussian pulse from the images with $K = 3$ and $K = 37$ and 1% noise.

In the limit $\mu \rightarrow 0$, it tends to the solution \mathbf{f} of the linear system. Indeed it is the solution of the linear equation:

$$(A + \mu I) \mathbf{f}^\mu = \mathbf{g} \quad ; \quad (5.4)$$

this approach, which applies to any positive semi-definite matrix, is also known as *Laurentev regularization method*.

However, as a consequence of the remark at the end of Section 3, the application of standard regularization methods must be considered with great caution. For instance, regularization by means of a penalty on derivatives (finite differences) of the solution, does not work. This is obvious in the case $K = 1$: a penalty on the second difference, as implied by a penalty on $\|B\mathbf{x}\|_2^2$, does not regularize at all the problem. In order to get regularization one should rather introduce a penalty on some "integral" transform of the object such as, for instance, $B^{-1}\mathbf{x}$. In a sense, there is a sort of duality between the regularization of the inversion of differential operators and that of the inversion of integral operators. Maximum Entropy could be a good compromise for both kinds of problems.

There is another point which must be stressed. Regularization methods, in general, recover completely the components of the object associated with the largest eigenvalues. In the present case, these eigenvalues correspond to the high-frequency components so that the high-frequency noise is not filtered out by the inversion method. On the other hand, the recovery of the low-frequency components and of the corresponding low-frequency noise can generate the artifacts discussed just at the end of the previous Section.

In conclusion, an accurate investigation of the existing regularization methods is required for deciding what method is appropriate to the solution of the present problem and what is not. Such an investigation is beyond the scope of this paper and we limit the analysis to a modified version of the method proposed in our previous papers.

As shown in our previous work, it is important to look for non-negative solutions. The following result gives a possible approach to the use of this constraint.

Theorem 5.1. There exists a unique solution \mathbf{f}_+ of the least-squares problem with the non-negativity constraint:

$$\mathbf{f}_+ = \underset{\mathbf{x}}{\operatorname{argmin}} \{ \|B\mathbf{x} - \mathbf{h}\|_2 : \mathbf{x} \geq 0 \} \quad ; \quad (5.5)$$

then, for any initial non-negative guess $\mathbf{f}^{(0)}$ the following iterative algorithm converges to \mathbf{f}_+ :

$$\mathbf{f}^{(k+1)} = P_+ \left\{ \mathbf{f}^{(k)} + \tau (\mathbf{g} - A \mathbf{f}^{(k)}) \right\} \quad , \quad (5.6)$$

where P_+ is the projection onto the cone of the non-negative vectors and τ is a relaxation parameter satisfying the conditions:

$$0 < \tau \leq 0.25 . \quad (5.7)$$

Proof: The existence and uniqueness of the non-negative least-squares solution \mathbf{f}_+ follows from the strict convexity of the least-squares functional which is implied by the non-singularity of the matrix B . Moreover from the results proved in [10] it follows that the projected Landweber method converges, for any initial guess $\mathbf{f}^{(0)}$, to \mathbf{f}_+ . This method leads to the iterative algorithm:

$$\mathbf{f}^{(k+1)} = P_+ \left\{ \mathbf{f}^{(k)} + \tau B^T (\mathbf{h} - B \mathbf{f}^{(k)}) \right\} , \quad (5.8)$$

with τ satisfying the conditions:

$$0 < \tau < \frac{2}{\lambda_{max}} , \quad (5.9)$$

where λ_{max} is the maximum eigenvalue of the matrix $B^T B$. From Eq. 5.1, it follows that $B^T B = A$ and $B^T \mathbf{h} = \mathbf{g}$ so that Eq. 5.8 implies Eq. 5.6. Moreover, since the maximum eigenvalue of A is smaller than 4, Eq. 5.9 implies Eq. 5.7. ■

The previous result shows that the limit of the iterations does not depend on the initial guess. However it may be convenient to initialize with the null vector because in such a case we have:

$$\mathbf{f}^{(0)} = 0 , \quad \mathbf{f}^{(1)} = \tau P_+ \mathbf{g} , \quad (5.10)$$

so that this choice is equivalent to initialize with a suitably scaled version of the detected image where the negative counterparts of the sources are set to zero.

There is another reason for choosing this initialization. If we consider the particular case of a ‘‘compact’’ source, namely a source which extends over a number of pixels smaller than the chopping amplitude K , then, as we will show, all the iterates are essentially scaled version of the unknown source. This result is, in a sense, a consistency result because, in the case of a ‘‘compact’’ source, the chopping & nodding technique already provides a reliable image. The proof is based on arguments introduced in [7], that are repeated here for the convenience of the reader.

Theorem 5.2. Let the chopped & noded image satisfy the conditions

$$\mathbf{g} = A \mathbf{f} \quad , \quad P_+ \mathbf{g} = 2 \mathbf{f} ; \quad (5.11)$$

the first one means that the data are noise-free, while the second one is a consequence of the non-superposition of the source and of its negative counterparts. Then the result of the k -th iteration of the algorithm of Eq. 5.6 is given by:

$$\mathbf{f}^{(k)} = \left\{ 1 - (1 - 2\tau)^k \right\} \mathbf{f} . \quad (5.12)$$

Proof: The proof can be obtained by induction. Indeed the relation is trivially true for $k = 0$. We assume that it is true for a given k and, in order to prove that it is also true for $k + 1$, we insert Eq. 5.12 into Eq. 5.6; if we use the first relation of Eq. 5.11, we obtain:

$$\mathbf{f}^{(k+1)} = P_+ \left\{ [1 - (1 - 2\tau)^k] \mathbf{f} + \tau (1 - 2\tau)^k \mathbf{g} \right\} . \quad (5.13)$$

If we remark that negative values appear only in the second term at r.h.s. of this equation and that the first term is zero where the second one is negative (the source and its negative counterparts do not overlap), from the second relation of Eq. 5.11, we get:

$$\mathbf{f}^{(k+1)} = [1 - (1 - 2\tau)^k] \mathbf{f} + 2\tau (1 - 2\tau)^k \mathbf{f}. \quad (5.14)$$

and the result is proved. ■

From the arguments used in the proof of this result it follows that artifacts may appear in the restored images if we do not use the correct value of the chopping amplitude. Indeed, if the image \mathbf{g} satisfies the first relation of Eq. 5.11 but we use in the iterative algorithm of Eq. 5.6 a matrix A' corresponding to a slightly different chopping amplitude, then the negative counterparts of the source in \mathbf{g} do not overlap exactly to the negative counterparts of the source in $A' \mathbf{f}$. As a consequence of this incomplete overlapping, artifacts appear in the first iterations. They are approximately spaced by K with respect to the source and they propagate in the subsequent iterations, producing other weaker artifacts spaced by $2K$, $3K$ etc. The result looks similar to the Type A artifacts produced by our previous method and discussed in [6]. However they are due now to the inexact knowledge of K and are not intrinsic to the restoration method. Their presence could be used to “tune” the chopping amplitude.

6. Validation of the iterative method

In this Section we present the results of a few numerical experiments intended to validate the iterative method introduced in the previous Section. It is worth to point out that its implementation is quite easy and that the computational cost of each iteration is low: it essentially requires a matrix-vector product and the imaging matrix has only three diagonals different from zero in the integer case (five in the fractional case). Therefore, since the number of iterations required is, in general, not very high, the method is fast. In our experience, a 128×128 image can be usually processed in a few seconds.

Before discussing the results we wish to point out an important property of this method. As proved in the previous Section, when the number of iterations tends to infinity, the limit is the non-negative least-squares solution. However, due to the ill-conditioning of the imaging matrix, this solution can also be affected by strong

noise propagation (we will show examples in the following), so that the introduction of regularization terms may be needed. One can avoid such an approach using the regularizing effect of early stopping of the iterations. This property is proved for unconstrained iterative methods such as Landweber, steepest decent, conjugate gradients, etc. [3], [11].

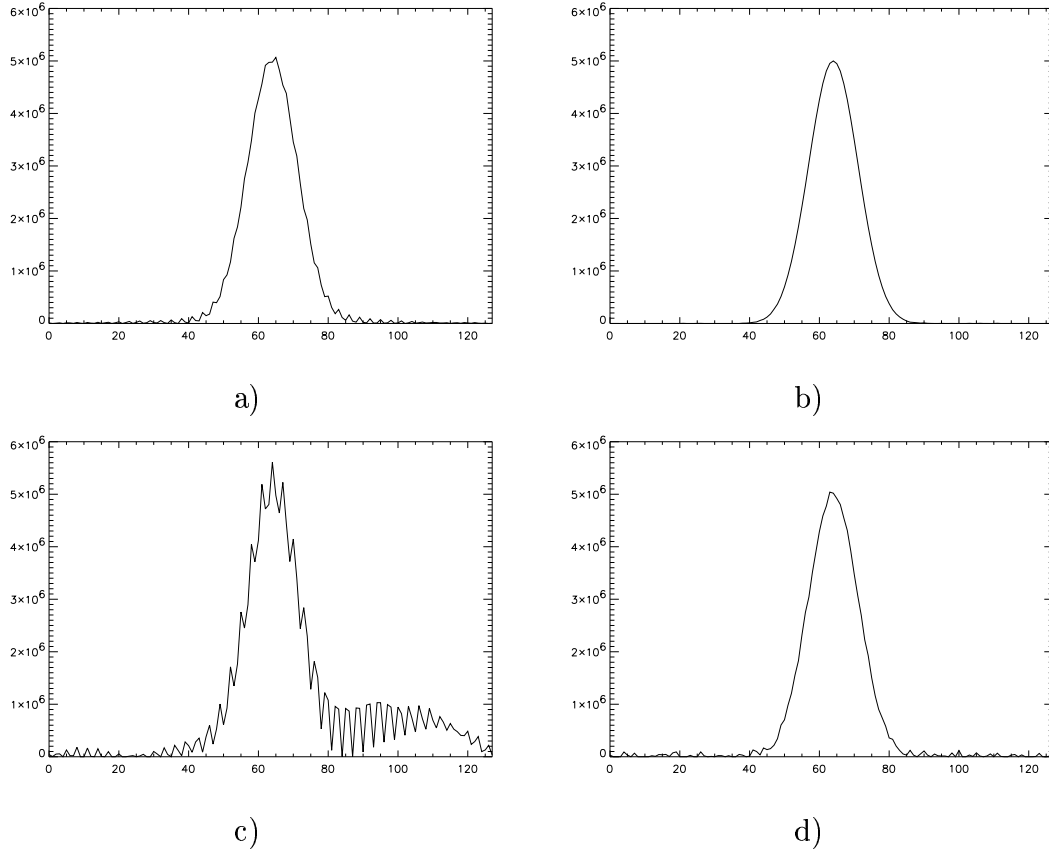


Figure 3. Reconstructions of the Gaussian pulse of Fig. 2 obtained after 1000 iterations of the method of Eq. 5.6: a)-b) reconstructions obtained from the images with $K = 3$ and $K = 37$ in the case of 0.1% noise; c)-d) reconstructions obtained from the images with $K = 3$ and $K = 37$ in the case of 1% noise.

A basic consequence is that these methods have the so-called *semiconvergence property*. Let us consider an image model as given in Eq. 4.5 and let us define the restoration error as follows:

$$\rho^{(k)} = \frac{\|\mathbf{f}^{(k)} - \mathbf{f}_0\|_2}{\|\mathbf{f}_0\|_2}, \quad (6.1)$$

where \mathbf{f}_0 is the target model and $\mathbf{f}^{(k)}$ the k -th iterate; then, for growing values of k ,

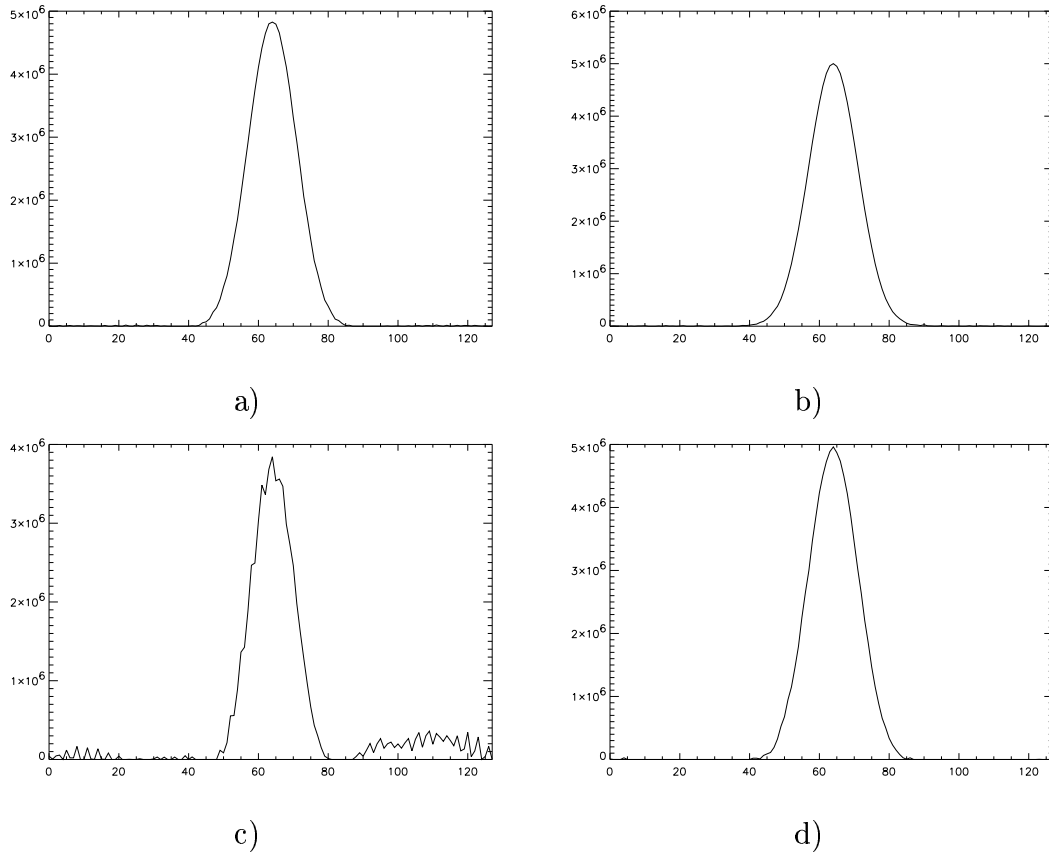


Figure 4. Reconstructions of the Gaussian pulse of Fig. 2 obtained by stopping the iterations of the method of Eq. 5.6 at the minimum of the restoration error: a)-b) reconstructions obtained from the images with $K = 3$ (141 iterations) and $K = 37$ (4 iterations) in the case of 0.1% noise; c)-d) reconstructions obtained from the images with $K = 3$ (38 iterations) and $K = 37$ (2 iterations) in the case of 1% noise.

$\rho^{(k)}$ first decreases, passes through a minimum and then increases; the minimum of the restoration error defines, in a sense, an optimal number of iterations.

The iterative method of Eq. 5.6, without the projection on the cone of non-negative vectors, has such a property, as it can be easily proved by means of standard methods [3]. We do not have yet a proof of this result for the projected method; however, from the many numerical experiments we have performed, we conjecture that it should be true. In the case of real data, since it is not possible to compute the restoration error, this property implies that one needs stopping criteria. This problem is beyond the scope of this paper. In the numerical experiments, we stop the iterations when we reach the minimum of the restoration error.

In order to justify our previous remarks on the numerical instability of the non-

negative least-squares solution and on the regularization effect of early stopping of the iterations, we first apply the method to the examples of Figure 2. For the four images of a Gaussian pulse (corresponding to two values of K and two noise levels) we give the results obtained both by pushing the iterations to convergence ($k = 1000$) and by stopping the iterations at the minimum of the restoration error.

In Fig. 3 we plot the results obtained after 1000 iterations. It is evident that in the domain where the non-negativity constraint is not active (the central part of the pulse) the effect of the noise is exactly the same that can be observed in the reconstructions of Fig. 2. In the external part of the pulse the effect of the noise is reduced by the constraint but the quality of the restorations obtained in the cases $K = 3$, 0.1% noise and $K = 37$, 1% noise, is not very good while it is definitely bad in the case $K = 3$, 1% noise.

In Fig. 4 we plot the results obtained by stopping the iterations at the minimum of the restoration error. In the caption we indicate, for each case, the number of iterations corresponding to this minimum. Incidentally we remark that, as expected, the number of iterations decreases when the noise increases. We obtain excellent results in all cases except $K = 3$ with 1% noise. In this case it is still evident the effect of incorrect evaluation of the components corresponding to the lowest eigenvalues.

We also give an example of the application of the method to the restoration of an object with a complex structure. We use a 321×128 image of the region of the *Becklin – Neugenbauer* (BN) source in the Great Orion Nebula. This image is part of a large image obtained at the 3.8m United Kingdom Infrared Telescope (UKIRT) on Mauna Kea, Hawaii, using the Mid-IR camera MAX developed by the Max Planck Institute für Astronomie. The image has been obtained from a mosaic of chopped & nodded images restored by means of the method proposed in our previous papers. Artifacts have been removed by combining images taken with different chopping throws and orientations (Robberto et al. 2003, in preparation). The field shown in Fig. 5a) is a severe test, as it contains complex structures with a very high dynamic range. For instance, the peak value of the main source is about 5×10^4 while the values of the diffuse emission around this source fluctuates around 5×10^2 , with peaks of about 1000 in correspondence of the other sources. From this field we obtain a (noise free) chopped & nodded image with $K = 37$, corresponding to $q = 8$. Therefore the condition number is about 40. Two noisy images are obtained by adding zero-mean white Gaussian noise with $\sigma = 10.3$ and $\sigma = 103$. The second value corresponds to a rather large noise compared with the number of counts per pixel of the structures around the main source.

The two noisy images are restored by means of the proposed method and iterations are stopped at the minimum of the restoration error. The “optimal” number of iterations is 43 in the case of low noise and 5 in the other one, while the corresponding restoration errors are 4.1 % and 12.7 %. The restored images are shown in Fig. 5b) and Fig. 5c),

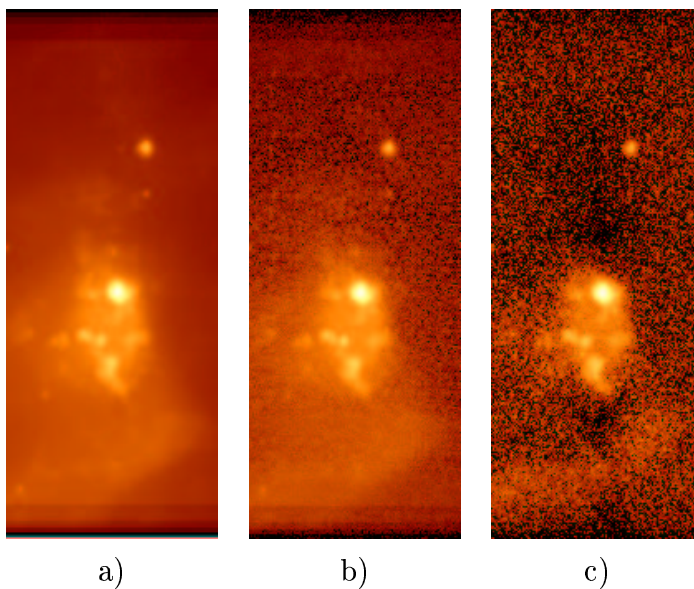
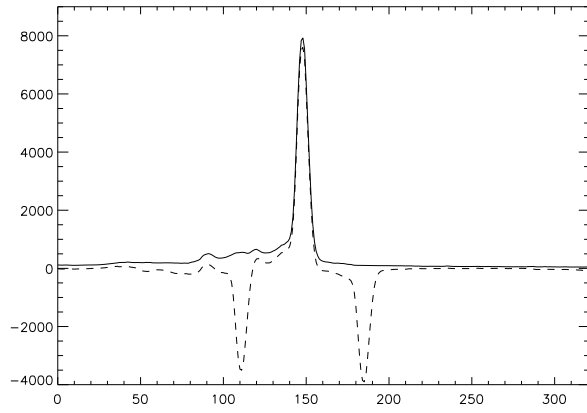


Figure 5. Restoration of the image of the BN system in the Orion Nebula: a) the original image; b) the restored image in the case of low noise; c) the restored image in the case of high noise. The values of the parameters which are relevant for the interpretation of these results are given in the text.

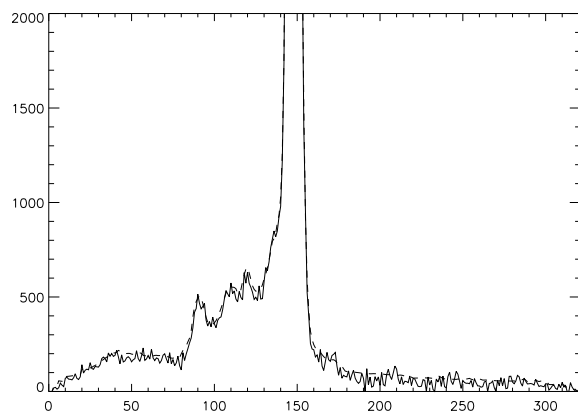
respectively. Especially in the case of higher noise the restoration looks rather corrupted by noise propagation.

In order to have a clear picture of the effect of noise propagation we give in Fig. 6 the cuts of the original and of the restored images corresponding to column 70 (the peak of the main source is on column 67). The noise corrupting these cuts looks as a high-frequency noise, especially in the case of Figure 6c). However, it may be due both to the high-frequency noise of the image (completely transmitted to the solution, as associated to the largest eigenvalues of the imaging matrix) and to the artifacts generated by the propagation of the low-frequency noise (discussed at the end of Section 3). It is evident that, in the case of high noise, important features are almost entirely lost. They are more correctly reproduced in the case of low noise.

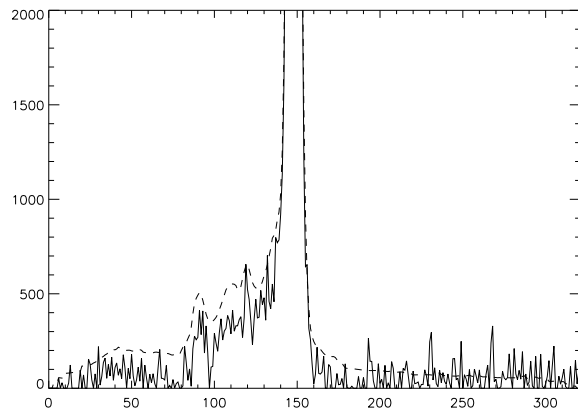
We also performed a few tests with the model developed for images with a chopping amplitude which is not an integer multiple of the pixel size. In the first experiment a compact Gaussian source is created over a number of pixels equal to twice the number used in the previous simulations. A chopped and noded image of this object is then formed with $K = 37$ and the result rebinned by taking the arithmetic mean of two adjacent rows, producing a chopped and noded image of the Gaussian with $\Delta = 18.5$. The iterative method is applied to this image using the matrix of Eq. 2.8 with $K = 18$



a)



b)



c)

Figure 6. Cuts of the images of the BN system corresponding to column 70: a) the original image (full line) and the corresponding chopped & nodded image with $K=37$ (dashed line); b) the restored image in the case of low noise (full line) compared with the original image (dashed line); c) the restored image in the case of high noise (full line) compared with the original image (dashed line).

and $r = 0.5$. A restoration error of about 2.5 % is reached after 27 iterations. No visible artifacts are produced: the two small ghosts, replacing the two negative counterparts, are fainter than the maximum value of the Gaussian by a factor $\approx 10^{-3}$.

This promising result suggests to apply the method to real images. As a first test we use a very simple image already investigated in Bertero et al. 2000, namely that of the bright star *BS1370* obtained with a chopping amplitude corresponding to about $K = 36$. As discussed in that paper, our previous method reproduces the stellar profile with great accuracy; however, the restored image contains also two small ghosts, similar to spikes, at a distance of approximately ± 36 pixels from the centre of the star and with an integrated flux of about 3.5 % of the stellar flux. An inspection of the structure of these spikes suggests that the chopping amplitude may be a bit greater than 36.

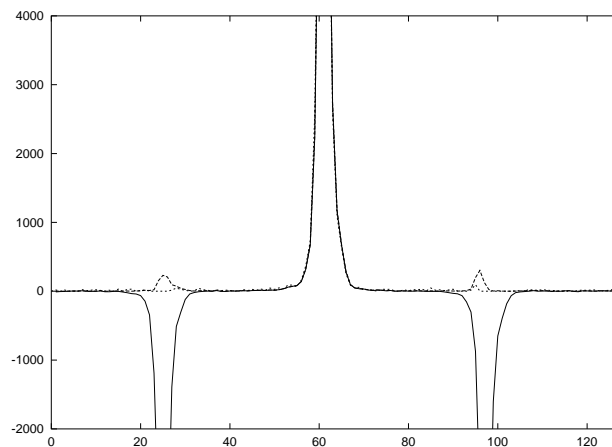


Figure 7. Vertical cuts passing through the maximum of *BS1370*: original image (full line), restored image obtained with our previous method (dashed line), restored image obtained with the new proposed method (dotted line). For comparison the profile of the original image has been divided by 2.

The application of the new method to the image of *BS1370* provides a restored image of the star with a photometric accuracy as good as that provided by the previous one. Concerning the two spikes they are already smaller in the case $K = 36$. However the best result is obtained assuming a chopping amplitude $K = 36.1$: the height of the spikes is reduced by a factor of 10 with respect to those generated by our previous method. These results are shown in Figure 7 where we plot the profiles of the original and of the restored images along the column passing through the star maximum. The profile of the original image has been divided by 2 for comparison (full line), while the dashed line and the dotted line correspond respectively to the restorations provided by the old and the new method. It is evident that new method provides a reconstruction

which is essentially free from artifacts.

7. Concluding remarks

In this paper we have investigated the effect of the use of Dirichlet boundary conditions in the restoration of chopped & nodded images and we have shown that they lead to a new reconstruction method which does not produce the artifacts generated by our previous one. Such a result however holds true only in the case of high quality of the original data. There are various pitfalls of data acquisition that may limit the accuracy of the method. If they are not accurately calibrated and corrected, artifacts may appear in the restored image, due this time not to the restoration method but to errors in image acquisition. In addition it must be observed that the new method is much more affected by noise propagation than our previous one. Such a feature suggests that it may be convenient to combine the restoration method with a suitable denoising of the original image. A recently proposed method [8] is just going in this direction: roughly speaking, it is an iterative Landweber method with a wavelets-based denoising at each iteration step.

Acknowledgements

We warmly thank Giovanni Alessandrini for stimulating discussions on this problem and for suggesting the nice inversion formula of Section 4, which has been the starting point of this work.

References

- [1] D. A. Allen, 1975, *Infrared, the New Astronomy*, Halsted Press, New York
- [2] J. M. Beckers, 1994, Imaging with array detectors using chopping and other forms of differential detection, in *Instrumentation in Astronomy VIII*, D. L. Crawford, and E. R. Craine eds., *Proc SPIE* **2198**, 1432-1437
- [3] M. Bertero, and P. Boccacci, 1998, *Introduction to Inverse Problems in Imaging*, IOP Press, Bristol
- [4] M. Bertero, P. Boccacci and M. Robberto, 1998, An inversion method for the restoration of chopped and nodded images, in *Infrared Astronomical Instrumentation*, A. M. Fowler ed., *Proc SPIE* **3354**, 877-886
- [5] M. Bertero, P. Boccacci, F. Di Benedetto, and M. Robberto, 1999, Restoration of chopped and nodded images in infrared astronomy, *Inverse Problems*, **15**, 345-372
- [6] M. Bertero, P. Boccacci, and M. Robberto, 2000, Wide-field imaging at mid-infrared wavelengths: Reconstruction of chopped and nodded data, *Pub. Astr. Soc. Pac.*, **112**, 1121-1137

- [7] M. Bertero, P. Boccacci, A. Custo, C. De Mol, and M. Robberto, 2003, A Fourier-based method for the restoration of chopped and nodded images, *Astron. Astrophis.*, **400**, 765-772
- [8] R. H. Chan, L.X. Shen, and Z.W. Shen, 2003, Restoration of Chopped and Nodded Images by Wavelet Frames, *Research report*, <http://www.math.cuhk.edu.hk/~rchan/>
- [9] F. Di Benedetto, 2003, The m-th difference operator applied to L^2 functions on a finite interval, *Linear Algebra Appl.*, **366**, 173-198
- [10] B. Eicke, 1992, Iteration methods for convexly constrained ill-posed problems in Hilbert space, *Num Funct Anal Opt*, **13** 413-429
- [11] H. W. Engl, M. Hanke, and A. Neubauer, 1996, *Regularization of Inverse Problems*, Kluwer, Dordrecht
- [12] H. U. Käufel, 1995, Observing extended objects with chopping restrictions on 8 m class telescopes in the thermal infrared, in *Calibrating and understanding HST and ESO instruments*, P. Benvenuti ed., ESO Conference and Workshop Proceedings, **50**, 159-163