# REGULARIZED ESTIMATION OF LINE SPECTRA FROM IRREGULARLY SAMPLED ASTROPHYSICAL DATA

*Sébastien Bourguignon, Hervé Carfantan, Loïc Jahan*

Laboratoire d'Astrophysique de Toulouse - Tarbes, UMR 5572 CNRS/UPS
14 avenue Edouard Belin, 31400 TOULOUSE, France
bourgui@ast.obs-mip.fr, herve.carfantan@obs-mip.fr, loic.jahan@obs-mip.fr

## ABSTRACT

We adress the problem of the estimation of spectral lines from irregularly sampled time series. As astrophysical data often present periodic time gaps, many undesirable peaks may appear in the Fourier spectrum. In this paper, the problem is addressed through the Bayesian regularization framework. The work by Sacchi *et al.*[1] and Ciuciu *et al.*[2] is extended to the irregular sampling case. Both quadratic and non quadratic regularizations are studied. Quadratic regularization is shown to be inappropriate for line spectra estimation, and non quadratic regularization is shown to give satisfactory simulation results compared to the classical CLEAN method. Finally, an application to astrophysical data processing is presented.

## 1. INTRODUCTION

The search for periodicities is a very important topic in astronomical time series analysis. It is the key tool for stellar oscillation analysis, as the pulsation modes of variable stars are determined from the frequency content of their light curves. Other fields of interest are the study of multiple stars systems and exoplanet detection, which are also based on the detection of periodicities in observational data.

Because of observation constraints, astronomical data generally suffer incomplete sampling. First, the day/night alternation, the non visibility of the object or erroneous measurements generate gaps in the time series. Therefore, the corresponding spectral window may have high secondary lobes, especially if the time gaps are periodic. Secondly, long time observations are generally fully irregularly sampled, that is, the time spacing between consecutive data points is not related to a sampling period. Multisite observation campaigns aim at improving the time covering, which consequently lessen the secondary lobes of the spectral window. As the instruments and local observation conditions may have different characteristics, this specificity may also be accounted for.

The Lomb-Scargle periodogram [3] (LSP) is an extension of the classical periodogram suited to irregular sampling. The complementary CLEAN algorithm [4], which remove iteratively the peaks in the Fourier spectrum as single frequency components, is widely used in astrophysics. This method, however, lacks theoretical background and is shown to fail in some cases.

We address the problem of line spectra estimation from the regularized estimation framework, following the work by [1, 2, 5]. After studying the properties and limits of the classical LSP and CLEAN techniques, we state the problem of line spectra restoration within the regularization framework. Then, both quadratic and non quadratic regularizations are studied. We examine two kinds of incomplete data that lead to different properties: *missing data*, which are regularly sampled but with missing samples and fully *irregularly sampled* data. We show on simulated data that non quadratic regularization gives satisfactory results compared to the LSP and the CLEAN algorithm. Finally, we present an application to real data with the study of Herbig Ae star HD 104237.

## 2. A CLASSICAL APPROACH

An ideal approach to the line spectra estimation from the data points $\{t_n, y(t_n)\}_{n=1\cdot\cdot N}$ is to consider a noise-corrupted sinusoidal model:

$$y(t_n) = \sum_{k=1}^{K} a_k sin(2\pi f_k t_n + \phi_k) + \epsilon_n \qquad (1)$$

whose parameters $\boldsymbol{\theta}_K = (a_k, f_k, \phi_k)_{k=1\cdot\cdot K}$ and order $K$ are to be identified. This is a difficult task, however. For independant and identically distributed (*i.i.d.*) centered gaussian noise $\epsilon_n$, the maximum likelihood (ML) estimation of $\boldsymbol{\theta}_K$ corresponds to the best fit to model (1) in a least-squares sense. In a regular sampling case, the shape of the corresponding likelihood function is investigated by Stoica *et al.*[6] and is shown to be multimodal, which make the estimation of $\boldsymbol{\theta}_K$ critical.

The classical periodogram of evenly spaced data is a spectral estimator whose maximum frequency identifies with the ML estimation of a single sinusoid model. If the data are irregularly sampled, the Fourier spectrum $Y(f) \triangleq \sum_{n=1}^{N} y(t_n) e^{-j2\pi f t_n}$ and $I_S(f) \triangleq |Y(f)|^2$, known as the Schuster periodogram, do not provide statistically significant spectral estimation. The extension to the case of irregular sampling is the so-called Lomb-Scargle periodogram

Authors would like to thank Nassim Seghouani for discussions on the subject of this paper and Torsten Böhm for the comments on the HD 104237 data.

or LSP [3] defined as:

$$\begin{aligned}\mathcal{I}_{\text{LS}}(f) &\triangleq \frac{\left(\sum_n y(t_n)\cos 2\pi f(t_n-\tau)\right)^2}{\sum_n \cos^2 2\pi f(t_n-\tau)} \\ &\quad + \frac{\left(\sum_n y(t_n)\sin 2\pi f(t_n-\tau)\right)^2}{\sum_n \sin^2 2\pi f(t_n-\tau)} \\ \text{with } \tau &= \frac{1}{4\pi f}\tan^{-1}\frac{\sum_n \sin 2\pi f t_n}{\sum_n \cos 2\pi f t_n}.\end{aligned}$$

Its maximum is shown to provide the least-squares fit of a single-sinusoid model to the data.

In the case of a multiple sinusoid model, however, the use of periodograms is not appropriate. The Fourier spectrum of the data $Y(f)$ is the convolution of the theoretical spectrum of the signal by the frequency response of the observation window $W(f) \triangleq \sum_{n=1}^{N} e^{j2\pi f t_n}$. As the spectral window $W(f)$ may have high secondary lobes because of the irregularities in the sampling scheme (especially with periodic gaps), the resulting periodogram may have many false peaks.

CLEAN deconvolution techniques [4] are widely used to process astrophysical data. The key point is to remove iteratively the peaks from the Fourier spectrum and their contribution in terms of sidelobes until some stopping condition. The resulting frequential components are then identified with the parameters of model (1). In order to produce a more readable spectrum, the CLEAN components are convolved by the *clean beam* which is a gaussian fitting of the central lobe of the spectral window. In addition, the residual spectrum is added in to form the *clean spectrum*. This method, however, is not satisfactory as a general spectral analysis tool and may give inaccurate results. For example, the mix of close spectral peaks can shift the location of the maximum frequencies to be removed in the Fourier spectrum, as will be shown on the simulations in section 6.1. The use of a *clean factor* to stabilize the algorithm is an ad-hoc solution, but it makes of CLEAN an ambiguous mathematical procedure. Because of the lack of theoretical background, the choice of a stopping condition is also problematic.

As the problems of these techniques are inherent to the direct identification of the parameters of model (1), we propose to consider a more general model with a large number of *fixed* frequencies. Thus, the problem becomes that of estimating the corresponding spectral amplitudes, in which zero values are encouraged through the regularization framework.

## 3. REGULARIZATION FRAMEWORK

We consider a linear model with an arbitrary large number of frequencies equispaced on the grid $\{\pm\frac{k}{P}f_{\max}\}_{k=0\cdot\cdot P}$:

$$\begin{aligned} y(t_n) &= \sum_{k=-P}^{P} x_k \exp\left(j2\pi\frac{k}{P}f_{\max}t_n\right) + \epsilon_n \\ \Leftrightarrow \quad \boldsymbol{y} &= \mathbf{W}\boldsymbol{x} + \boldsymbol{\epsilon} \end{aligned} \tag{2}$$

where $t_n$ are the measurement times, $\boldsymbol{x} \in \mathbb{C}^{2P+1}$ the spectral amplitudes, $N\times(2P+1)$ matrix $\mathbf{W} = (e^{j2\pi\frac{k}{P}f_{\max}t_n})_{k=-P\cdot\cdot P}^{n=1\cdot\cdot N}$ and $\boldsymbol{\epsilon}$ the perturbations, that may be complex. Of course, *real* data can be specifically accounted for by considering $x_{-k} = x_{2P-k}^*$ and *real* noise $\boldsymbol{\epsilon}$ in the previous model, but for the sake of clarity we use the complex formulation (2). Note that physical prior information on the signal bandwidth can be accounted for. If the frequency content is known to be in $[f_0, f_1]$, one can consider $\mathbf{W} = (e^{j2\pi\frac{k}{P}f_{\max}t_n})_{|k|=k_0\cdot\cdot k_1}^{n=1\cdot\cdot N}$ with $f_{0/1} = \frac{k_{0/1}}{P}f_{\max}$.

The objective is to estimate the parameters $\boldsymbol{x}$ from the noisy data $\boldsymbol{y}$. High resolution is achieved for $2P+1 \gg N$, which leads to an underdetermined inverse problem. The generalized inverse of the problem, defined as the minimum norm solution of $\boldsymbol{y} = \mathbf{W}\boldsymbol{x}$, leads to an unacceptable solution: in the *missing data* case, we show that it is proportional to the zero-substituted Discrete Fourier Transform (DFT) of the data $\boldsymbol{y}$ (see section 4.2). Thus, it corresponds to the convolution of the theoretical line spectrum with the observation window, which may have its own periodicities. The resulting estimate can then present many false peaks.

A classical regularization approach consists in accounting for prior information on the expected shape of the spectrum, in our case the *sparseness* information. Then, Maximum *A Posteriori* (MAP) estimation of $\boldsymbol{x}$ can be performed within the statistical Bayesian framework. If $p(\boldsymbol{x})$ is a prior law on the spectral amplitudes, the posterior law writes:

$$p(\boldsymbol{x}|\boldsymbol{y},\sigma^2) \propto \mathcal{L}(\boldsymbol{y};\boldsymbol{x},\sigma^2)\times p(\boldsymbol{x})$$

where the likelihood function of model (2) writes: $\mathcal{L}(\boldsymbol{y};\boldsymbol{x},\sigma^2) \propto \frac{1}{\sigma^N}\exp\left(-Q(\boldsymbol{x})/2\sigma^2\right)$ for complex circular gaussian noise $|\epsilon_n| \sim \mathcal{N}(0,\sigma^2)$, with $Q(\boldsymbol{x}) = ||\boldsymbol{y}-\mathbf{W}\boldsymbol{x}||^2$. Thus, maximizing the posterior law $p(\boldsymbol{x}|\boldsymbol{y},\sigma^2)$ is equivalent to minimizing a penalized least-squares criterion:

$$\hat{\boldsymbol{x}} = \arg\min J(\boldsymbol{x}), J(\boldsymbol{x}) \triangleq Q(\boldsymbol{x}) + \lambda R(\boldsymbol{x}) \tag{3}$$

where the penalization function $R(\boldsymbol{x})$ has to express the sparseness of the solution, that is, take its lowest values for solutions $\boldsymbol{x}$ with only a few number of non zero components. Hyperparameter $\lambda > 0$ balances then between fidelity to the data and confidence in prior information. This formulation has already been used for the spectral analysis of evenly spaced data [1, 2, 5] and we propose to study its extension to the irregular sampling case.

Using the previous formulation, the multisite observation case can be accounted for within the same statistical framework. Consider $|\epsilon_n| \sim \mathcal{N}(0,\sigma_n^2)$ where $\sigma_n$ is related to the signal-to-noise ratio (SNR) of the n$^{\text{th}}$ observation. The corresponding likelihood writes:

$$\mathcal{L}(\boldsymbol{y};\boldsymbol{x},\sigma_n^2) = \frac{1}{\prod_n \sigma_n}\exp\left(-Q_\Sigma(\boldsymbol{x})/2\right)$$

where $Q_\Sigma(\boldsymbol{x}) = ||\boldsymbol{y}-\mathbf{W}\boldsymbol{x}||^2_{\Sigma^{-1}} = (\boldsymbol{y}-\mathbf{W}\boldsymbol{x})^\dagger\Sigma^{-1}(\boldsymbol{y}-\mathbf{W}\boldsymbol{x})$ and $\Sigma$ is the $N\times N$ diagonal matrix of elements $\sigma_n^2$. Then, criterion (3) should be adapted by considering $Q_\Sigma$ instead of $Q$. Note that the Lomb-Scargle periodogram can also be modified to account for different noise levels $\epsilon_n$ by similarly

adapting the least-squares fitting of the sine wave parameters. We can show that it results in weighting every sum in $\mathcal{I}_{\text{LS}}(f)$ by the corresponding noise variance $\sigma_n^2$ and adequately modifying parameter $\tau$.

## 4. QUADRATIC REGULARIZATION

Quadratic regularization $R_2(\boldsymbol{x}) = ||\boldsymbol{x}||^2$, or more generally $R_2^{\Pi}(\boldsymbol{x}) = \boldsymbol{x}^\dagger \Pi \boldsymbol{x}$, is attractive as it leads to an explicit solution to problem (3). Several properties are established in [5], where the solution is interpreted in terms of data windowing. We study the extension of these properties to the case of unevenly spaced data. As they depend on specific structures of matrices $\mathbf{W}^\dagger\mathbf{W}$ and $\mathbf{W}\mathbf{W}^\dagger$, these operators are studied for both missing data and irregular sampling cases.

### 4.1. Regular sampling

Consider $t_n = nT_s$ and $f_{\text{max}}$ is set at the Nyquist limit $f_{\text{max}} = 1/2T_s$. Operator $\mathbf{W}$ writes in this case[1] $\mathbf{Z}_{N,2P} = (\exp j\pi \frac{kn}{P})_{n=1\cdot\cdot N}^{k=-P+1\cdot\cdot P}$. The case $2P = N$ corresponds to the Fourier matrix $\mathbf{Z}_{2P,2P} \triangleq \mathbf{F}_{2P}$ and $\mathbf{Z}_{2P,2P}\,\boldsymbol{x}$ is the inverse DFT of spectrum $\boldsymbol{x}$. In the high-resolution underdetermined case $2P > N$ we note $\mathbf{Z} = \mathbf{Z}_{N,2P}$ for the sake of clarity. It is straightforward to show that $\mathbf{Z}\mathbf{Z}^\dagger = 2P\mathbf{I}_N$ and that $\mathbf{Z}^\dagger\mathbf{Z} = \mathbf{F}_{2P}^\dagger \mathbf{D}_Z \mathbf{F}_{2P}$ is *circulant*, where $\mathbf{D}_Z$ is a diagonal matrix with $N$ ones and $2P-N$ zeros. Using such notations, the following properties hold [5]:

P.1 The regularized solution for the quadratic function $R_2$ is:

$$\widehat{\boldsymbol{x}}_2 = (\mathbf{Z}^\dagger\mathbf{Z} + \lambda \mathbf{I}_{2P})^{-1}\mathbf{Z}^\dagger\boldsymbol{y} = \tfrac{1}{2P+\lambda}\mathbf{F}_{2P}^\dagger \widetilde{\boldsymbol{y}}$$

where $\widetilde{\boldsymbol{y}}[n] = \boldsymbol{y}[n]$ for $n = 1\cdot\cdot N$ and $\widetilde{\boldsymbol{y}}[n] = 0$ for $n = N+1\cdot\cdot 2P$. It is proportional to the DFT of the zero-padded data. Therefore, this kind of regularization is of no interest as the regularization parameter only controls the proportionality coefficient of the DFT of the zero-padded data.

P.2 The generalized inverse solution (*i.e.* the minimum norm minimizer of least-squares criterion $Q$) is the limit of $\widehat{\boldsymbol{x}}_1$ when $\lambda$ tends to zero, which is the DFT of the zero-padded data:

$$\widehat{\boldsymbol{x}}_{\text{GI}} = \tfrac{1}{2P}\mathbf{F}_{2P}^\dagger \widetilde{\boldsymbol{y}}$$

P.3 The regularized solution for the quadratic function $R_2^{\Pi}$ (with a circulant matrix $\Pi = \mathbf{F}_{2P}^\dagger \mathbf{D}_\Pi \mathbf{F}_{2P}$ where $\mathbf{D}_\Pi$ is a diagonal matrix $\mathbf{D}_\Pi = \text{diag}\{\boldsymbol{d}_\Pi[k]\}_{k=-P+1\cdot\cdot P}$) corresponds to the DFT of the windowed[2] zero-padded data:

$$\widehat{\boldsymbol{x}}_2^{\Pi} = (\mathbf{Z}^\dagger\mathbf{Z} + \lambda\Pi)^{-1}\mathbf{Z}^\dagger\boldsymbol{y} = \mathbf{F}_{2P}^\dagger \overline{\boldsymbol{y}}$$

with $\overline{\boldsymbol{y}}[n] = \frac{1/2P}{1+\lambda \boldsymbol{d}_\Pi[n]}\widetilde{\boldsymbol{y}}[n]$ for $n = 1\cdot\cdot N$.

[1] Line $k = -P$ has been removed as it gives the same spectral contribution as $k = P$ at frequency $f_{\text{max}}$.

[2] In [5] this kind of windowing is compared to the classical windowing approach.

### 4.2. Missing data

The missing data case writes $t_n = i_n T_s, i_n \in \mathbb{N}$. If parameter $f_{\text{max}}$ is set to its maximum limit $f_{\text{max}} = 1/2T_s$, operator $\mathbf{W}$ writes $\mathbf{M}_{N,2P} = (\exp j\pi \frac{ki_n}{P})_{k,n}$ and is a Fourier matrix with missing lines. We show (we note hereafter $\mathbf{M} = \mathbf{M}_{N,2P}$) that $\mathbf{M}\mathbf{M}^\dagger = 2P\mathbf{I}_N$ and that $\mathbf{M}^\dagger\mathbf{M} = \mathbf{F}_{2P}^\dagger \mathbf{D}_M \mathbf{F}_{2P}$ is still *circulant*, where $\mathbf{D}_M$ is a diagonal matrix with $N$ ones for indexes $\{i_n\}_{n=1\cdot\cdot N}$ and zeros elsewhere. Thus, similar properties (P.1-3) hold by considering $\widetilde{\boldsymbol{y}}_M$ instead of $\widetilde{\boldsymbol{y}}$, where $\widetilde{\boldsymbol{y}}_M$ is the zero-padded and zero-substituted data: $\widetilde{\boldsymbol{y}}_M[i_n] = \boldsymbol{y}[n]$ for $n = 1\cdot\cdot N$ and 0 elsewhere. Once again, the DFT of the zero-padded and zero-substituted data corresponds to a frequency sampling of the theoretical signal spectrum convolved by the frequency response of the observation window. However, the latter has no more a $\text{sinc}$-like shape because of the missing data and may have many undesirable secondary lobes. The use of windowing in order to lessen these lobes is not as simple as in the regular sampling case, which reduces the interest of quadratic regularization for such incomplete sampling.

### 4.3. Irregular sampling

In the general irregular sampling case, operator $\mathbf{W}$ no longer has a Fourier-like structure and matrices $\mathbf{W}\mathbf{W}^\dagger$ and $\mathbf{W}^\dagger\mathbf{W}$ loose the former properties, *i.e.* proportional to identity and circulant, respectively. Thus, the interpretation of $\widehat{\boldsymbol{x}}_{\text{GI}}$, $\widehat{\boldsymbol{x}}_2$ and $\widehat{\boldsymbol{x}}_2^{\Pi}$ in terms of DFT of windowed and zero-padded data cannot be generalized. The regularized solution, however, is expected to present similar characteristics to that obtained for missing data, as will be shown in section 6.2.

Note that in this case frequency $f_{\text{max}}$ has to be set according to the underlying physics. As there is no Nyquist limit, the largest spectral window free of aliases can be much wider than with regular sampling. Thus, the irregularities in the sampling scheme make it theoretically possible to detect very short periods [7].

## 5. NON QUADRATIC REGULARIZATION

Non quadratic regularization has already been used for spectral analysis. Sacchi *et al.* [1] choose a non convex regularization function corresponding to a Cauchy prior law on the modulus $|x_k|$. It is appropriate for line spectra restoration, but criterion $J$ is not convex and may have local minima. The $l_1$ norm penalization ($R_1(\boldsymbol{x}) = ||\boldsymbol{x}||_1 = \sum_k |x_k|$) aims to find the sparsest solution to the inverse problem $\boldsymbol{y} = \mathbf{W}\boldsymbol{x}$ [8]. That is, under certain assumptions[3], the minimizer of criterion $J_1(\boldsymbol{x}) = Q(\boldsymbol{x}) + \lambda R_1(\boldsymbol{x})$ corresponds to the least-squares identification of the parameters of model (1). As criterion $J_1$ is not differentiable at the 0 point, however, its minimization requires sophisticated numerical tools, such as quadratic programming, which are very computationally expensive for high dimension problems.

[3] Note that the assumptions presented in [8] cannot be checked in this case.

We choose hereafter the strictly convex and differentiable hyperbolic function $R_{\text{hyp}}(\boldsymbol{x}) = \sum_k \sqrt{s^2 + |x_k|^2}$ proposed by [2]. Since $R_{\text{hyp}} \to R_1$ as $s \to 0$, the solution is similar to that of $l_1$ regularization for low $s$ and leads to a strictly convex and differentiable criterion. Thus, the solution can be computed easily by standard descent algorithms.

Although specific algorithms have been designed for line spectra estimation [1, 2], their attractive computational efficiency is based on the Fourier-like shape of operator $\mathbf{W}$ due to the regularity of the sampling scheme. In the irregular sampling case, we propose to compute an approximation to the minimizer of criterion (3) by minimizing the following modified criterion:

$$J_{\text{irreg}}(\boldsymbol{x}) = ||\mathbf{W}^\dagger \boldsymbol{y} - \mathbf{W}^\dagger \mathbf{W} \boldsymbol{x}||^2 + \lambda R(\boldsymbol{x})$$

As matrix $\mathbf{W}^\dagger \mathbf{W}$ is Toeplitz, the minimizer of $J_{\text{irreg}}$ can be computed at a low cost using FFT algorithms. In [9], differences between both minimizers of criteria $J$ and $J_{\text{irreg}}$ – adapting the optimal value of hyperparameter $\lambda$ – are shown to be not significant, while minimization of $J_{\text{irreg}}$ is performed at a much lower cost.

## 6. SIMULATION RESULTS

### 6.1. Lomb-Scargle periodogram and CLEAN

We consider a sum of 3 sinusoids with close periods around 100 day, corrupted by 15dB white Gaussian noise. This kind of spectrum simulates the effect of rotational splitting on variable stars. Frequencies are set off the reconstruction grid $\{\frac{k}{P} f_{\max}\}$ (here $P = 500$ and $f_{\max} = 0.02\,\text{day}^{-1}$). Sampling is irregular, spanning 2000 days with additional periodic gaps of 10 days every month and 200 days every year. Signal is $N = 38$ points long and is shown on figure 1. Because of the gaps in the sampling scheme, the frequency

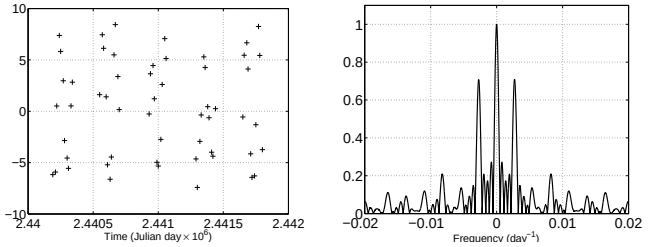



**Fig. 1**. Simulated data (left) and the corresponding spectral window (right).

response of the observation window presents high secondary lobes. Thus, the Lomb-Scargle periodogram presents false peaks, and the peaks corresponding to the model frequencies are slightly shifted (figure 2). Results obtained by the CLEAN algorithm are presented on figure 3. The solution was computed with a *clean factor* $g = 0.2$, and algorithm stops when the maximum amplitude in the residual spectrum is less than 10% of the maximum amplitude of the initial Fourier

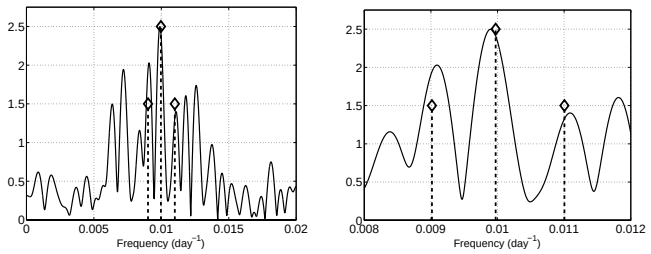
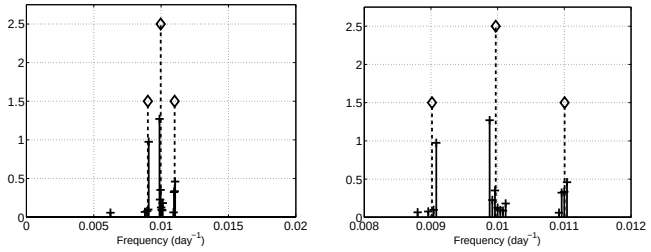
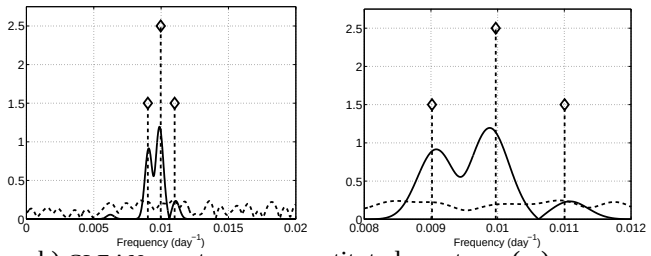



**Fig. 2**. Lomb-Scargle periodogram of the data (left) and zoom around the model frequencies (right). $\diamond$ correspond to the true spectral lines.




a) CLEAN components ($+$) and true spectral lines ($\diamond$). Total frequency range (left) and zoom around the model frequencies (right).




b) CLEAN spectrum: reconstituted spectrum ($-$), residual spectrum ($--$) and true spectral lines ($\diamond$).

**Fig. 3**. CLEAN algorithm simulation results.

spectrum. Because of the shifts in the maximum frequencies of the LSP, the procedure produces several components around the true spectral lines and the maximum components are slightly shifted. The posterior convolution by the *clean beam* shows that only two of the three peaks are significantly retrieved, while the amplitude of the third one is at the same level that the residual noise spectrum.

### 6.2. Quadratic penalization

The regularized solution for quadratic penalization $R_2$ is computed. Although properties of section 4 do not extend to the general irregular sampling case, the same practical conclusion hold, as shown on figure 4. The solution is almost proportional to the Fourier spectrum, *i.e.* $\boldsymbol{x}_2 \simeq \frac{1}{2P+1+\lambda} \mathbf{W}^\dagger \boldsymbol{y}$ using previous notations, and hyperparameter $\lambda$ only controls the proportionality coefficient. If $\lambda$ is too low, however, matrix $\mathbf{W}^\dagger \mathbf{W} + \lambda \mathbf{I}_{2P+1}$ becomes ill-conditioned and the regularized solution diverges. One can understand the fact that property (P.1) of section 4 is almost satisfied by considering that matrix

$\mathbf{W}\mathbf{W}^\dagger$ is almost diagonal, which is studied in [9]. Then, the regularized solution in the irregular sampling case is close to that expected in a regular sampling or missing data case and cannot be used for line spectra restoration.

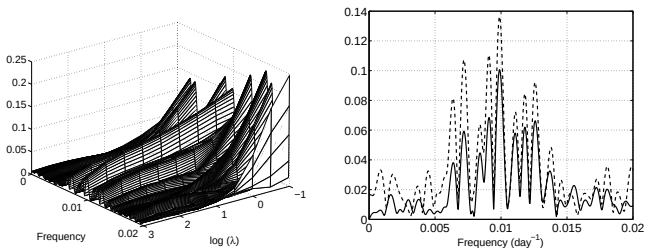



**Fig. 4**. Quadratic regularization. Left: $|\widehat{\boldsymbol{x}}_2|$ as a function of $\lambda$. Right: Solution $|\widehat{\boldsymbol{x}}_2|$ for $\lambda = 10$ $(-)$ and « expected » solution $\frac{1}{2P+1+\lambda}|\mathbf{W}^\dagger \boldsymbol{y}|$ $(--)$.

### 6.3. Non quadratic penalization

Minimizer $\widehat{\boldsymbol{x}}_{\text{hyp}}$ of criterion $J_{\text{irreg}}$ is performed for the penalization function $R_{\text{hyp}}$ with $s = 10^{-4}$. Optimal value of parameter $\lambda$ is tuned using Hansen's L-curve criterion. Although there is no theoretical background in the case of non quadratic regularization, this method has shown to give satisfactory results in practice. Figure 5 shows the estimate sensibility to hyperparameter $\lambda$. The « L » shape of curve $(Q(\boldsymbol{x}), R(\boldsymbol{x}))$ is not as pronounced as it is in the quadratic case. We can see on figure 5 right that the three model spectral lines are correctly estimated for $\log\lambda \in [1.5\,;2.5]$. Smaller $\lambda$ values produce artefacts in the estimate and for $\lambda \geq 2.5$ only the main line is detected.

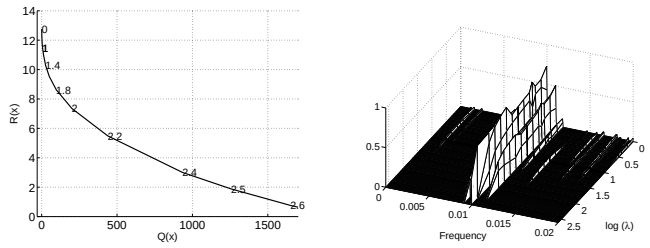



**Fig. 5**. Non quadratic penalization with $R_{\text{hyp}}$. Left: L-curve $(Q(\boldsymbol{x}), R(\boldsymbol{x}))$ parameterized by $\log\lambda$. Right: $|\widehat{\boldsymbol{x}}_{\text{hyp}}|$ as a function of $\lambda$.

Results on figure 6 show the minimizer of criterion $J_{\text{irreg}}$ for $\log\lambda = 1.8$. The maxima of the regularized solution correspond to the best approximation of the model frequencies on the grid $\{\frac{k}{P} f_{\max}\}$. We note a loss in the amplitude estimation, which is inherent to any regularization process. However, thanks to the accurate frequency localization, the amplitudes can be correctly estimated *a posteriori* by least-squares.

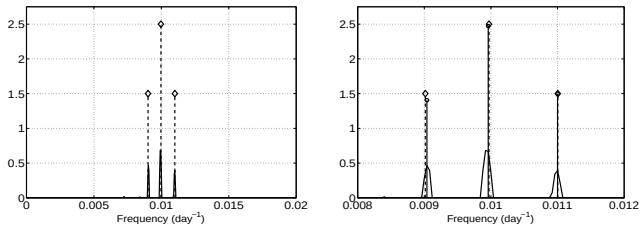



**Fig. 6**. Non quadratic regularization with $R_{\text{hyp}}$. Left: estimate $|\widehat{\boldsymbol{x}}_{\text{hyp}}|$ on the whole frequency grid $(-)$ and true spectral lines $(\diamond)$. Right: zoom around the model frequencies and posterior least-squares amplitude estimation $(\bullet)$.

## 7. APPLICATION TO REAL DATA

### 7.1. HD 104237 raw data

Data presented on figure 7 are radial velocity values of the pulsating Herbig Ae star HD 104237 obtained from spectroscopic observations in 1999 [10]. This is a complex multiple system whose primary component is pulsating, so that several frequencies are expected to be found in the data near $30$ day$^{-1}$. Data consist of $5$ observing nights. The $N = 514$

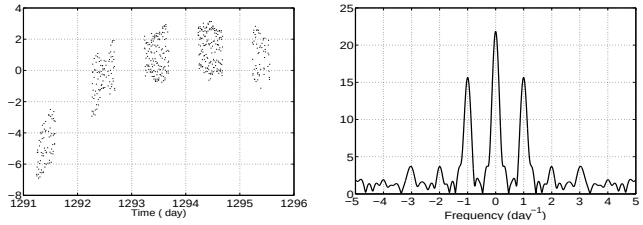



**Fig. 7**. HD 104237 data and spectral window.

samples are irregularly spaced with additional periodic gaps due to the day/night alternation. Because of these gaps, the spectral window on figure 7 presents high secondary lobes at frequencies $\pm 1$day$^{-1}$. The Lomb-Scargle periodogram of the centered data is shown on figure 8. It emphasizes the low frequency behaviour and one spectral peak near $33$ day$^{-1}$, with secondary lobes at $1$ day$^{-1}$ intervals due to the spectral window. Because of the low frequency perturbations, the CLEAN algorithm only finds the $33$ day$^{-1}$ frequency before stopping. (If the stopping condition is pushed farther the algorithm diverges.)

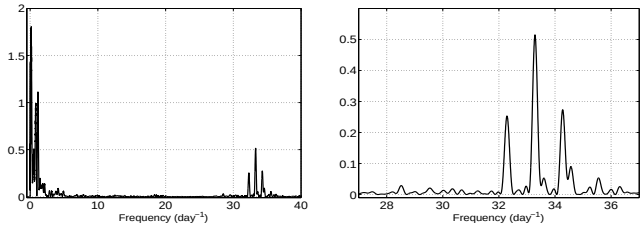



**Fig. 8**. Lomb-Scargle periodogram of HD 104237 data.

Figure 9 shows the regularized solution $\widehat{\boldsymbol{x}}_{\text{hyp}}$ for optimal hyperparameter value $\log\lambda = 2.4$. In the high frequency range we retrieve the $33$ day$^{-1}$ frequency, but also three other frequencies with smaller amplitudes.

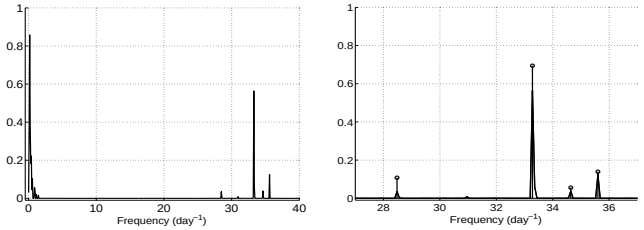



**Fig. 9**. Regularized estimation. Left: $|\widehat{\boldsymbol{x}}_{\text{hyp}}|$ for $\log\lambda = 2.8$. Right: zoom around $30$ day$^{-1}$ and posterior least-squares estimation of the spectral amplitudes (•).

### 7.2. HD 104237 « corrected » data

In [10] the main orbital movement of HD 104237 multiple system is fitted with a binary approximation and substracted from the initial data, which results in the data shown on figure 10. In that case, the CLEAN algorithm is no more disturbed

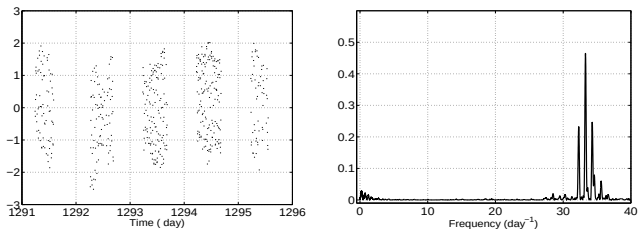



**Fig. 10**. Left: corrected HD 104237 data. Right: Lomb-Scargle periodogram.

by the low frequency components. Results are shown on figure 11: the same frequencies are detected as those obtained by the regularized estimation on the raw data (see section 7.1). Of course, estimator $\widehat{\boldsymbol{x}}_{\text{hyp}}$ obtained with these corrected data gives the same « high frequency » estimation as in section 7.1. Consequently, the regularized solution is not sensitive to the low frequency perturbations as the CLEAN method is.

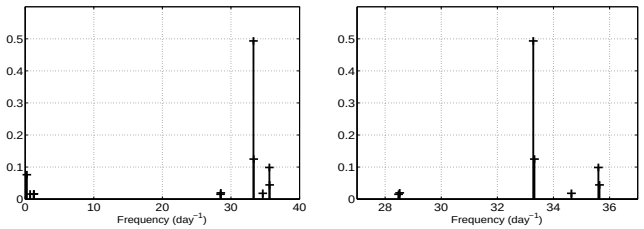



**Fig. 11**. CLEAN results on the corrected data. Left: total frequency range. Right: zoom around $30$ day$^{-1}$.

## 8. CONCLUSION

The regularization framework makes it possible to define a spectral estimator that can account for typical astrophysical data specificities (various instruments, irregular sampling with large time gaps). In the *missing data* case, the quadratic regularization leads to the DFT of the windowed zero-substituted data, and thus is not appropriate for line spectra estimation. Simulations show that the same practical conclusions hold in the general *irregular sampling* case.

The use of an hyperbolic penalization function produces a sparse solution with accurate frequency localization, and its computational cost is reduced by taking advantage of the specific shape of the operator. Simulations show the ability of the regularized estimation to retrieve closely spaced spectral lines from little data, whereas the CLEAN method does not provide such accurate estimation. An application to real data analysis finally shows that the regularized solution is not sensitive to low frequency perturbations, so that no data preprocessing is needed to retrieve the « high frequency » spectral lines.

## 9. REFERENCES


[1] M.D. Sacchi, T.J. Ulrych, and C.J. Walker, Interpolation and extrapolation using a high-resolution discrete Fourier transform," *IEEE Transactions on Signal Processing*, vol. 46, no. 1, pp. 31–38, January 1998.

[2] P. Ciuciu, J. Idier, and J.-F. Giovannelli, Regularized estimation of mixed spectra using a circular Gibbs-Markov model," *IEEE Transactions on Signal Processing*, vol. 49, no. 10, pp. 2201–2213, October 2001.

[3] J.D. Scargle, Studies in astronomical time series analysis. *II* - Statistical aspects of spectral analysis of unevenly spaced data," *Astrophysical Journal, Part 1*, vol. 263, pp. 835–853, December 1982.

[4] D. H. Roberts, J. Lehar, and J. W. Dreher, Time series analysis with CLEAN. I. Derivation of a spectrum," *The Astronomical Journal*, vol. 93, no. 4, pp. 968–989, April 1987.

[5] J.-F. Giovannelli and J. Idier, Bayesian interpretation of periodograms," *IEEE Transactions on Signal Processing*, vol. 49, no. 7, pp. 1988–1996, July 2001.

[6] P. Stoica, R. L. Moses, B. Friedlander, and T. Söderström, Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 3, pp. 378–392, March 1989.

[7] L. Eyer and P. Bartholdi, Variable stars: which Nyquist frequency?," *Astronomy and Astrophysics Supplement Series*, vol. 135, pp. 1–3, February 1999.

[8] J.-J. Fuchs, On sparse representations in arbitrary redundant bases," *IEEE Transactions on Information Theory*, vol. 50, pp. 1341–1344, June 2004.

[9] S. Bourguignon, H. Carfantan, and L. Jahan, Regularized spectral analysis of unevenly spaced data," in *Proc. ICASSP*, Philadelphia, USA, March 2005.

[10] T. Böhm, C. Catala, L. Balona, and B. Carter, Spectroscopic monitoring of the Herbig Ae star HD 104237. I. Multiperiodic stellar oscillations," *Astronomy and Astrophysics*, vol. 427, pp. 907–922, December 2004.