Restoration of Astrophysical Spectra With Sparsity Constraints: Models and Algorithms

Sébastien Bourguignon, David Mary, and Éric Slezak

***Abstract*—We address the problem of joint signal restoration and parameter estimation in the context of the forthcoming MUSE instrument, which will provide spectroscopic measurements of light emitted by very distant galaxies. Restoration of spectra is formulated as a linear inverse problem, accounting for the instrument response and the noise spectral variability. Estimation is considered in the setting of sparse approximation, where restoration is performed jointly with the detection of relevant patterns in the spectra. To this aim, a dictionary of elementary spectral features is designed according to astrophysical spectroscopy. Sparse estimation is considered through the minimization of a quadratic data misfit criterion with an ℓ^1-norm penalization, where nonzero components are associated to the detected features. An efficient optimization strategy is proposed, based on the Iterative Coordinate Descent (ICD) principle, with accelerations that dramatically reduce the computational cost. The algorithm does not rely on fast transforms and can be applied to a wide variety of criteria if the sparsity constraint is separable. Results on simulated MUSE-like data reveal satisfactory performance in terms of denoising and detection of physically relevant spectral features. On such data, the proposed algorithm is shown to outperform both state-of-the-art gradient-based and homotopy continuation methods. Simulations with a compressed sensing-like random matrix also reveal better performance compared with usual algorithms, showing that ICD can be a powerful strategy for sparse optimization.**

***Index Terms*—ℓ^1-norm penalization, deconvolution, denoising, iterative coordinate descent, sparse approximation, sparse optimization.**

I. Introduction

MUSE (Multi-Unit Spectroscopic Explorer) is a second-generation instrument under construction for the European Southern Observatory, which will be installed at the Very Large Telescope in Chile in 2012. It is a very powerful integral field spectrograph [1], which will provide massive hyperspectral data cubes with images of 300×300 pixels, with 0.2 arcsec angular resolution, at up to 4000 wavelengths, covering the visible and the near infrared parts of the electromagnetic spectrum. One of the main science cases of MUSE, which is of interest here, concerns the detection of very distant galaxies and their characterization by their spectra.

Because of their large distance to the observer and of other disturbances, data will be collected in a very noisy environment, with spectrally variable characteristics. In particular, noise is not expected to be identically distributed along the wavelength axis. The Line Spread Function (LSF) of MUSE—the instrument impulse response in the spectral domain—is also spectrally variable, producing more degradations as wavelength increases. In this paper, we consider the restoration of MUSE-like spectra within the setting of inverse problems regularization [2], which provides a robust framework for taking into account the latter observational specificities. Regularization is also an efficient approach for adding prior information on the solution. This is absolutely necessary in our case where, given the very high level of noise contaminating the data, estimation has to be constrained with prior assumptions.

Spectra of galaxies have been studied for a long time [3]. Several spectral components can be identified in the wavelength range covered by MUSE, which depend mainly on the star formation history of each object, on its chemical composition and dust content. In particular, such spectra contain emission and absorption lines and a continuum, which can exhibit a blueward break due to the absorption of photons by intergalactic hydrogen clouds along the line of sight. Hence, we consider that a galaxy spectrum can be modeled as the superposition of these three basic components. Since the characteristics of the features of each component are unknown, we build a catalog containing a high number of possible ones with discretized parameters, some of which will be selected in order to fit the spectrum. In the works that we present next, restoration is consequently combined with the detection of physically relevant parameters.

Such an approach can be viewed as a sparse approximation problem, where only *a few* components are selected in the catalog. Sparsity-based methods have been widely used for signal and image denoising problems in the past 20 years [4]. They rely on the assumption that, for a given class of signals, most information concentrates into a small number of significant coefficients, expressed in some appropriate space. Many contributions initiated by the work of Donoho and Johnstone [5] considered transforms based on multi-scale representations such as wavelets, curvelets, or other XX-let transforms that have been introduced for specific problems [4]. Additionally to the fact that most information in natural signals (and especially, images) can be efficiently represented by a few number of decomposition coefficients—wavelets are good generic *sparsifying* transforms—two "technical" reasons also drove the design of such transforms.

Manuscript received October 17, 2010; revised February 22, 2011; accepted April 11, 2011. Date of publication April 25, 2011; date of current version August 17, 2011. This work was supported in part by ANR project 08-BLAN-0253-01 DAHLIA—Dedicated Algorithms for HyperspectraL Imaging in Astronomy and in part by PPF-ISSO, University of Nice Sophia Antipolis. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jean-Luc Starck.

S. Bourguignon and É. Slezak are with the Cassiopée Laboratory, University of Nice Sophia Antipolis, CNRS, Côte d'Azur Observatory, F-06304 Nice, France (e-mail: sebastien.bourguignon@oca.eu; eric.slezak@oca.eu).

D. Mary is with the Fizeau Laboratory, University of Nice Sophia Antipolis, CNRS, Côte d'Azur Observatory, F-06103 Nice, France (e-mail: david.mary@unice.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2011.2147278

Orthogonality: If the transform is orthogonal, then sparse estimation amounts to thresholding the transform coefficients, under the assumption of white independent and identically distributed (i.i.d.) noise contaminating the data [5]. This provides a simple framework for both computation and analysis of estimator properties [6], [7]. In contrast, numerous recent works explored the use of *redundant* transforms, such as overcomplete families of homogeneous functions [8], [9], or unions of structurally different bases [8], [10]—a concept sometimes called morphological diversity [11]. In these cases, near-orthogonality is also a crucial property in the transform choice, in order to guarantee the ability of algorithms to retrieve the sparsest solution [10], [12]. This property is also a cornerstone of the powerful emergent theory of compressed sensing (CS) [13].

Fast Transforms: if fast algorithms are available, then large data sets can be processed efficiently. In the redundant case especially, where simple thresholding generally does not provide the best sparse estimates, efficient optimization is needed. Many recent convex optimization algorithms in this field rely on intensive computations of gradient-like functionals which exploit fast transforms [14]–[18].

Although the methodology considered in this paper is based on a sparse decomposition of data, fundamental differences exist with such usual setting. Considered data are one-dimensional and with "reasonable" size, so that we can afford the construction of a specific dictionary (with no associated fast operator), which better models sparse prior information than generic transforms. We suppose that galaxy spectra can be synthesized as the superposition of astrophysically meaningful features [3]. Doing so, a physical interpretation can be associated to the active coefficients in the decomposition, indicating for example the *detection* of spectral lines or breaks. In other words, we are interested in both restoring the spectrum and estimating associated parameters by means of the synthesis coefficients, whereas usual denoising mainly focuses on reconstruction in data space. Moreover, sparsity should be better expressed in such an adapted dictionary than in generic ones; hence better denoising performance is expected. This approach has a double price to pay, however. First, dictionary atoms are highly correlated. Hence, theoretical properties about the resulting sparse approximation cannot be obtained. Nevertheless, as shown below, satisfactory results are achieved in practice. A second disadvantage concerns optimization, since no fast transform can be used to compute matrix-vector products. Note that such sparse representation problem with "constrained" dictionary arises in many applications, e.g., sparse linear regression in statistics, source separation, or when the dictionary is learned from the data. Hence, the optimization context of this paper is a rather generic one.

The methodological objectives of this paper are mainly twofold. First, attention is given to precisely modeling both data formation – in particular, noise variability and instrumental characteristics – and prior information. A dictionary of elementary spectral features is built based on prior physical knowledge, so that sparse estimation is combined with the detection of relevant spectral information in the data. Estimation is set under the usual sparsity-promoting ℓ^1-penalization framework [8], [19]. Observational specificities are shown to modify the equivalent dictionary, whose properties are studied.

The second part of our paper concerns sparse optimization when no fast transform can be used. Alternatives to gradient-based methods are considered which exploit sparsity, namely Homotopy Continuation (HC) and Iterative Coordinate Descent (ICD). HC or Least Angle Regression [20]–[23] is specifically designed for sparse solutions, where most computations are concentrated towards identifying the support of the solution. ICD [24], [25] performs successive componentwise optimization steps and can be interpreted similarly. Whereas HC is often considered as the most efficient strategy in this context [21], [23], [26], the efficiency of ICD for sparse estimation was recently exhibited, e.g., in [27]–[29]. In particular, Friedman *et al.* [27] showed that ICD can outperform HC in large size problems, and conclude their work by stating that "coordinate-wise descent algorithms deserve more attention in convex optimization." In this paper, we build an ICD-based strategy, with accelerations specifically designed for sparse problems. It can be efficiently applied to any penalized least-squares criterion with separable penalization, provided that coordinatewise optimizations can be performed at low cost and that the solution is sparse. Hence, we also consider a "compressed sensing like" scenario with high-dimensional random matrix, similarly to an example proposed in [17].

The paper is organized as follows. Section II introduces a model for data observation, with variable LSF and non identically distributed noise. In Section III, a prior model is proposed, based on a sparse representation of the data, and a specific dictionary is built. Estimation is formulated in Section IV as an ℓ^1-norm penalized optimization problem. Hyperparameter tuning is addressed and the resulting equivalent dictionary is studied. Section IV ends with the description of a posterior amplitude re-estimation step. Optimization is studied in Section V. The use of HC and ICD strategies is motivated, and improvements on standard ICD are introduced by exploiting the sparsity of the solution. Simulation results are given in Section VI. An application to deconvolution and noise reduction for a MUSE-like simulated spectrum is presented. Then, the behavior and performance of several optimization strategies are compared, revealing the efficiency of the proposed algorithm, both on MUSE-like data and on a CS-like example with random and noisy measurements of a sparse process. Conclusions and directions for further work are finally given in Section VII.

II. Data Formation Model

Let $\boldsymbol{y} = [y_1, \ldots, y_N]^T \in \mathbb{R}^N$ denote a spectrum as observed by MUSE, discretized at equispaced wavelengths $\lambda_1, \ldots, \lambda_N$. The spectral sampling step here is $\Delta\lambda = 0.13$ nm, ranging from $\lambda_1 = 450$ nm to $\lambda_N = 900$ nm. This yields $\mathrm{N} = 3\,463$. We consider the following linear observational model:

$$\boldsymbol{y} = \mathbf{H}\boldsymbol{s} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{s} \in \mathbb{R}^N$ is the spectrum to be restored, $\mathbf{H}$ is the $N \times N$ matrix form of the LSF, and $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^T$ is an additive perturbation term. Model (1) supposes that $\boldsymbol{s}$ is reconstructed at the

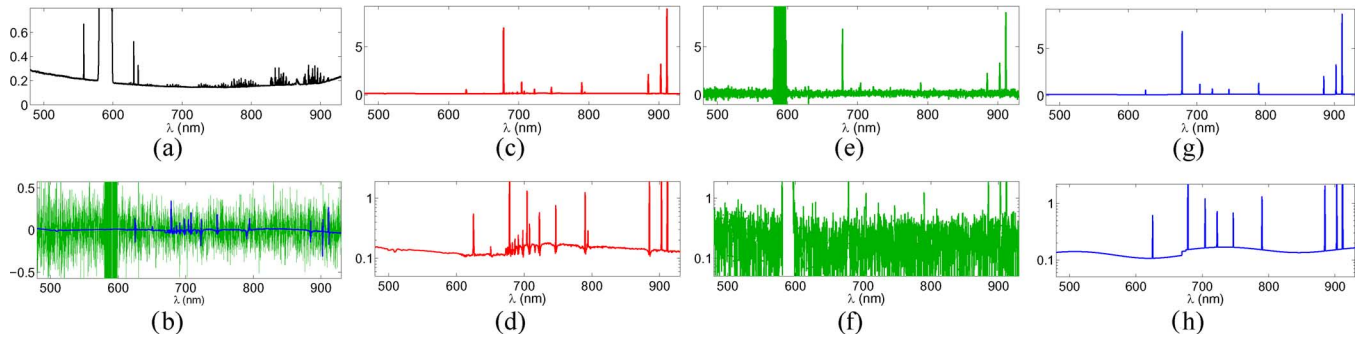

Fig. 1. An example of MUSE-like spectrum and corresponding estimation results. (a) Standard deviation of the noise affecting each wavelength. (c)–(h) Noise-free (red), noisy (green), and restored (blue) spectra at two amplitude scales (bottom panels (d)-(h) are in log scale). (b) Reconstruction error (blue) and original noise (green). Amplitudes are light fluxes, in $\mathrm{erg} \cdot \mathrm{s}^{-1} \cdot \mathrm{cm}^{-2}$ ($\times 10^{-20}$).

resolution of the data $\boldsymbol{y}$. Actually, the true spectrum is a continuous function of wavelength, which may have much higher *frequencies*[1] than the Nyquist limit $1/2\Delta\lambda$, hence model (1) can be viewed as a band-limited approximation of the true reconstruction problem. Restoration at a higher resolution could be addressed by considering a smaller sampling step for $\boldsymbol{s}$. However, this would increase the numerical complexity and the ill-posedness of problem (1)[2, Ch. 2].

A. Variable Line Spread Function

The LSF represents the instrumental spreading affecting every point in the spectrum: each column $\boldsymbol{h}_p = [h_{1,p}, \ldots, h_{N,p}]^T$ of $\mathbf{H}$ is the instrument's impulse response at wavelength λ_n, such that

$$y_n = \sum_{p=1}^{N} h_{n,p} s_p + \epsilon_n.$$

The LSF of MUSE is variable: as wavelength increases, it becomes more spiky and the spectral blurring effect decreases. Consequently, $h_{n,p}$ cannot be written under the typical form h_{n-p}, $\mathbf{H}$ is not a Toeplitz matrix and product $\mathbf{H}\boldsymbol{s}$ cannot be written as a convolution. Note that, at the instrument's spectral resolution, each $\boldsymbol{h}_p$ has a support of 11 points (that is, a spectral extension of 1.43 nm), which is small compared with the size of the spectra, so that matrix $\mathbf{H}$ is very sparse.

B. Non-Identically Distributed Noise

Noise affecting MUSE-like observations is also expected to strongly vary with wavelength, as shown in Fig. 1(a). Three main factors contribute to such variability. First, as in all ground-based astronomical observations, data will suffer from a parasite atmospheric emission. Indeed, some chemical components in the atmosphere emit light at specific wavelengths. In practice, such emission can be supposed spatially constant across the 1 arcmin $\times$ 1 arcmin field of view of the instrument. Hence, it can be estimated from the related 300×300 pixels at each wavelength, and then subtracted from the data. Note that current MUSE simulations suppose perfect estimation of such background emission, whereas estimation residuals will for sure affect observational data. Even so, however, fluctuations remain in the background-subtracted spectrum, which are proportional to the emission level because light emission is a Poisson process: the higher the emission, the higher the associated noise variance. Consequently, in model (1), $\boldsymbol{\epsilon}$ contains a noise source whose power varies with wavelength. In particular, one can see in Fig. 1(a) the signature of powerful parasite emission lines at 528 nm and around 600 nm, which correspond to [OI] emission, and of line packets at higher wavelengths, caused by the presence of water in the atmosphere [3]. MUSE quantum efficiency—the number of produced electrons over the number of photons hitting the detector—is also variable with wavelength, with decreasing performance at each extremity of the spectral range, where consequently the noise influence is stronger. Least, a laser reference star system will be implemented for adaptive optics, generating a very powerful parasite emission in a 20-nm-wide spectral band around the sodium line $\lambda_{\mathrm{Na}} = 589.2$ nm, where no signal can be detected. These three effects are visible in Fig. 1(a), which shows a typical model for the variation of the noise level, jointly to the noise-free and noisy simulated spectra of a moderately bright galaxy in MUSE-like data (panels c)-d) and e)-f), respectively). MUSE data for a deep extragalactic exposure will contain thousands of such spectra, and for the vast majority of them the noise level is expected to be as high, or even higher, as that of Fig. 1.

Note that MUSE is still under construction. Hence, only simulated data generated by the MUSE consortium are available at this time: astronomical scenes are first computed from high-complexity astrophysical simulations, and a mock observation is then generated by applying the MUSE Instrument Numerical Model [30]. In the following, we will suppose that both LSF and noise variance are known at each wavelength.

III. Prior Model With Sparse Representations

A. Motivation for Sparsity Priors

Estimating spectrum $\boldsymbol{s}$ from data $\boldsymbol{y}$ in (1) can be viewed as a denoising and deconvolution inverse problem. Given the high level of noise contamination in the data, only poor results can be reached without any additional assumption on the spectrum. To improve the restoration quality, a typical approach incorporates prior information on the searched solution [2]. Widespread sparsity-based methods can be viewed within this setting. The unknown spectrum $\boldsymbol{s}$ is supposed to have a sparse representation in an appropriate dictionary, say $\boldsymbol{s} = \mathbf{W}\boldsymbol{x}$ with $\boldsymbol{x}$ *sparse*, that

[1] *Frequency* here denotes the inverse (dual) dimension of wavelength.

is, only a few coefficients in $\boldsymbol{x}$ are significant. Since the signal energy is concentrated in a few points, nonzero coefficients in $\boldsymbol{x}$ take higher values than coefficients in $\boldsymbol{s}$, so that coefficients are relatively less affected by noise in the first case. *Denoising* then consists in estimating such sparse vector, say $\widehat{\boldsymbol{x}}$, and restoring the signal by $\widehat{\boldsymbol{s}} = \mathbf{W}\widehat{\boldsymbol{x}}$.

As far as astrophysical spectra are concerned, we consider sparsity in a twofold objective. Denoising and deconvolution is a first one, but we also aim at associating a physical interpretation to the detected atoms. The latter aspect relies on a dictionary design specially adapted to such spectra, which is addressed hereafter.

B. Modeling Galaxies' Spectra: The Dictionary

In most signal and image denoising applications with sparsity assumptions, the dictionary is generally chosen according to the following guidelines:

- the considered data must, of course, contain information that is sparse in some transform domain(s);
- fast transforms must be available to compute products $\mathbf{W}\cdot$ (*synthesis* operator) and $\mathbf{W}^T\cdot$ (*analysis* operator), enabling algorithms to handle efficiently large-size data;
- orthogonality (or near-orthogonality for redundant dictionaries) is another property of interest, which is crucial for theoretical analysis such as uniqueness of the sparse solution and ℓ^0-ℓ^1 equivalence.

In this paper, we adopt a rather different approach, where we prefer using a specific, highly redundant and correlated dictionary, which is more adapted to the morphological features of the considered spectra than generic transforms. We consider that a spectrum is made of three components, each of which is supposed to have a sparse decomposition in an appropriate dictionary: a line spectrum, a step-like spectrum and a continuous spectrum.

1) Complex Line Spectrum: Most relevant information in astrophysical spectroscopy is contained in emission and absorption lines. Each line is characterized by its central wavelength, amplitude and width—we do not consider here the line spreading produced by the LSF, already included in the model (1). Lines can be either *resolved* if their profile spans several points in the wavelength axis, or *unresolved* if the linewidth is smaller than the spectral sampling step. Hence, a dictionary $\mathbf{W}_\ell$ is built, composed of spectral lines with variable widths and central wavelengths. Positive and negative coefficients in the decomposition then characterize, respectively, emission and absorption lines. Unresolved lines are modeled by N delta functions[2] at wavelengths $\lambda_{n,\,n=1\ldots N}$. The shapes of resolved lines are modeled with spline functions of variable size, corresponding to different linewidths. Because resolved lines cannot have arbitrarily large widths, we consider spline widths ranging from 0.39 nm to 18 nm, corresponding to supports varying from 3 to 138 points. In order to limit the dictionary size and to avoid too high correlation between spline atoms, the spacing between adjacent splines with same support size S was set to $S/8$ (up to rounding error). Empirical studies led us to the following selection of parameters, written under the form (support size, translation step) and given in number of points: (3, 1), (5, 1), (9, 1), (11, 1), (17, 2), (25, 3), (35, 4), (49, 6), (69, 9), (97, 12), (138, 17). By doing so, the number of spline atoms is approximately reduced by half compared with the "full" version with unitary translation steps, while the loss in precision is small.

Note that although the rest-frame spectrum of almost all chemical components are well characterized by laboratory measurements, the locations in frequency of these lines in the observed spectrum are unknown, due to the cosmological *redshift* [3] of the data: for an object measured at a redshift $z > 0$ (the larger z, the farther the object), the observed wavelengths λ_{obs} are shifted with respect to the known rest-frame wavelengths λ_{lab} according to the famous equation $1 + z = \lambda_{\mathrm{obs}}/\lambda_{\mathrm{lab}}$. MUSE should be sensitive to galaxies with redshifts as high as 6 or more, that is to objects as far away as 10 billion light-years.

2) Step Spectrum: One major science case of MUSE is the detection of distant galaxies exhibiting a strong discontinuity in their observed spectra [1], [3]. This discontinuity can be modeled by a break in the spectrum. We consider for this purpose the dictionary $\mathbf{W}_s$ made of step functions, which we initially centered at wavelengths $\lambda_{n,\,n=1\ldots N}$. Since the step functions centered at the very first wavelengths possess very high correlations with the "continuous component" (the mean value, which is always nonzero), they may be selected by approximation algorithms although no significant break is present in the data. In order to avoid this effect, only steps centered at wavelengths $\lambda_{n,\,n=50\ldots N}$ were considered.

3) Continuous Spectrum: Several works in the field of sparse representations [8], [10] (see also [31] for a previous work of the authors on astrophysical spectra) have considered a sparse decomposition in the discrete cosine transform (DCT) basis (or, similarly, in the discrete Fourier transform basis) for smoothly varying signals. An N-point signal is implicitly modeled as the sum of a few sinusoids, taken in a dictionary of N atoms with distinct *frequencies* between 0 and the Nyquist limit. In our case, the continuous spectrum is supposed to show very smooth variations, so that high frequencies are unnecessary. On the other hand, more accuracy on the low frequency model is desired. Hence, dictionary $\mathbf{W}_c$ is built by considering all sinusoids with:

- reduced frequencies $f_k = k/N$ with $k = 0 \ldots 8$;
- for $f_k \neq 0, 8$ equispaced phases $\varphi_\ell = \ell\pi/8$, $\ell = 0 \ldots 7$.

The whole dictionary $\mathbf{W} = [\mathbf{W}_\ell \;\; \mathbf{W}_s \;\; \mathbf{W}_c]$ has $N = 3\,463$ lines (number of data) and $M = 26\,015$ columns (number of atoms). Note that $\mathbf{W}$ shows obvious redundancies: a delta function in $\mathbf{W}_c$ is the difference of two adjacent step functions in $\mathbf{W}_s$, and three sine functions with same frequency and different phases are linearly dependent. In both cases, however, the sparsity constraint should remove the ambiguity by favoring a combination with the fewest atoms. Estimation accounting for MUSE LSF and noise statistics is discussed in Section IV. This is shown to lead to an *equivalent dictionary*, whose structure is studied in Section IV-D.

Note that alternatives exist for designing dictionaries adapted to specific data. In particular, approaches based on dictionary learning (e.g., [32], [33]) build a dictionary which sparsifies a

[2]Note that locating unresolved lines with higher precision than that imposed by the instrument would require a high-resolution formulation of problem (1).

given set of training data. The objective of this paper is different, where attention is given in priority to the physical meaning of the dictionary elements. An interesting and more related recent work [34] considers parametric dictionary design, where parameters of elementary components are discretized in order to satisfy a minimal coherence criterion. In our case, we prefer building specific sets of atoms according to precise prior information, even if the resulting dictionary is very coherent.

Of course, our dictionary design only performs an *approximation* of galaxies spectra with simple atoms, and true data do not exactly correspond to a superposition a such atoms. In particular, emission and absorption line profiles are rarely symmetric. Neither does the parametrization of the continuous spectrum with sine functions correspond to physical reality—except for the imposed low-frequency trend. Indeed, such dictionary results from a compromise between sufficiently rich and physically meaningful modeling and limited complexity.

IV. Estimation Setting

A. ℓ^1-Norm Penalization

Let $\mathbf{W}$ denote the dictionary built in Section III-B. One wants to find a sparse approximation of spectrum $\boldsymbol{s}$ using dictionary $\mathbf{W}$ under the observation model (1), that is, a sparse vector $\boldsymbol{x} \in \mathbb{R}^M$ to problem

$$\boldsymbol{y} = \mathbf{H}\mathbf{W}\boldsymbol{x} + \boldsymbol{\epsilon}. \tag{2}$$

In the literature of sparse representations, two well-known approaches coexist. The greedy approach was formalized as Matching Pursuit (MP) by Mallat and Zhang [35]—even if the principle can be traced back to much earlier works such as [36]. MP is an iterative procedure, which removes the most correlated component between the data and the dictionary, and repeats the process on the residual until some stopping condition is met. Although it benefits from a low computational cost, it is known to propagate erroneous atom selections in cases where atoms of $\mathbf{W}$ are too much correlated ([37], see also examples in [8]). Improvements such as Orthogonal Matching Pursuit [38] and Orthogonal Least Squares [39] try to overcome this problem by performing more complex iterations, at the price of a higher computational cost. They remain, however, sensitive to error propagation and are discarded in the rest of this paper, where the dictionary is highly correlated—see Section IV-D.

We consider the usual alternative to greedy approaches, which turns the problem into convex optimization, by defining estimate $\widehat{\boldsymbol{x}}$ as the minimizer of the following criterion [19]:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \frac{1}{2}\left\|\boldsymbol{y} - \mathbf{H}\mathbf{W}\boldsymbol{x}\right\|^2 + \gamma \left\|\boldsymbol{x}\right\|_1 \text{ with } \gamma > 0$$

which is a compromise between data fidelity and the sparsity measure $\|\boldsymbol{x}\|_1 = \sum_m |x_m|$. Such an approach, named *Basis Pursuit De-Noising* after the work of Chen *et al.* [8], is known to yield a sparse solution for well chosen values of γ [10], [12]. Compared with greedy methods, convexity of the optimization problem brings less sensitivity of the estimate toward high correlations between atoms of $\mathbf{W}$ [8], [23], [40].

Note that other penalizations can be used to enforce sparsity. Strictly convex approximations of the ℓ^1-norm (e.g., [41]) yield a strictly convex criterion, for which uniqueness of the minimizer is ensured and optimization can be tackled by a wide range of algorithms. However, corresponding estimates are not strictly sparse, which is not in accordance with our detection purpose. On the other hand, other sparsity-promoting functions than the ℓ^1-norm are non-convex, so that optimization can be trapped in local minima. Hence, we select the ℓ^1-norm as the limiting case of a convex (but not strictly) penalization function that yields strict sparsity, which is also a key point for the efficiency of the optimization procedure proposed in Section V.

The former generic ℓ^1-based criterion has to be adapted, however, in order to integrate noise statistics and the necessary dictionary normalization. These topics are addressed in Sections IV-B and IV-C, respectively, that yield the final optimization criterion given in (5).

B. Adjustment for Non-i.i.d. Noise: Equivalent Dictionary

Data-misfit term $(1/2)\left\|\boldsymbol{y} - \mathbf{H}\mathbf{W}\boldsymbol{x}\right\|^2$ can be viewed as the neg-log-likelihood of model (2) under the assumption that perturbations ϵ_n are i.i.d. centered Gaussian [2]. The specific structure of the noise affecting MUSE spectra was described in Section II-B. We suppose in the following that data are contaminated by zero-mean Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_N^2)$ and σ_n^2 is the variance of the noise contaminating the spectrum at wavelength λ_n. Hence, the neg-log-likelihood of model (2) reads (up to an additive constant):

$$\begin{aligned}\frac{1}{2}\left\|\boldsymbol{y} - \mathbf{H}\mathbf{W}\boldsymbol{x}\right\|_{\boldsymbol{\Sigma}}^2 &\triangleq \frac{1}{2}\left(\boldsymbol{y} - \mathbf{H}\mathbf{W}\boldsymbol{x}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y} - \mathbf{H}\mathbf{W}\boldsymbol{x}\right) \\ &= \frac{1}{2}\left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y} - \boldsymbol{\Sigma}^{-1/2}\mathbf{H}\mathbf{W}\boldsymbol{x}\right\|^2\end{aligned}$$

with $\boldsymbol{\Sigma}^{-1/2} = \operatorname{diag}(\sigma_1^{-1}, \ldots, \sigma_N^{-1})$. This expresses the correct data-misfit measurement to be considered in order to account for MUSE noise statistics. Note that non-diagonal noise covariance could also be taken into account similarly. The optimization criterion then reads

$$\frac{1}{2}\left\|\boldsymbol{z} - \mathbf{A}\boldsymbol{x}\right\|^2 + \gamma\left\|\boldsymbol{x}\right\|_1 \tag{3}$$

where

- $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}$ are weighted data;
- $\mathbf{A} = \boldsymbol{\Sigma}^{-1/2}\mathbf{H}\mathbf{W}$ is the weighted and convolved dictionary.

C. Hyperparameter Tuning and Dictionary Normalization

Tuning the weight γ of the ℓ^1-norm penalization term in (3) is a crucial issue. For $\gamma \geq \left\|\mathbf{A}^T\boldsymbol{z}\right\|_\infty$, the minimizer is identically zero [12]. Conversely, for too small γ, the solution may not be sparse. In this section, a statistical interpretation of γ is given and the need for dictionary normalization is evidenced, which yields a practical rule for hyperparameter tuning.

Consider a slightly more general penalization function of the form $\sum_m \gamma_m |x_m|$. Karush–Kuhn–Tucker (KKT) conditions stipulate [19] that $\widehat{\boldsymbol{x}}$ minimizes $\left\|\boldsymbol{z} - \mathbf{A}\boldsymbol{x}\right\|^2/2 + \sum_m \gamma_m |x_m|$ if and only if

$$\begin{cases} \text{for } \widehat{x}_m = 0: & |\widehat{e}_m| < \gamma_m \\ \text{for } \widehat{x}_m \neq 0: & \widehat{e}_m = \gamma_m \operatorname{sign}(\widehat{x}_m) \end{cases} \tag{4}$$

where $\widehat{e}_m$ is the mth component of projected residual $\widehat{\boldsymbol{e}} = \mathbf{A}^T(\boldsymbol{z} - \mathbf{A}\widehat{\boldsymbol{x}})$ and the sign function equals 1, -1 and 0 for positive, negative and zero arguments, respectively. Thus, γ_m can be viewed as a threshold on $\widehat{e}_m$ under which the mth component is not detected.

Suppose that data contain only noise. One then wants $\widehat{\boldsymbol{x}} = \mathbf{0}$ so that residual $\boldsymbol{z} - \mathbf{A}\widehat{\boldsymbol{x}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_N)$ and $\widehat{e}_m \sim \mathcal{N}(0, ||\boldsymbol{a}_m||)$, where $\boldsymbol{a}_m$ is the mth column of $\mathbf{A}$. From (4), a detection test with false alarm rate τ_{FA} is achieved by choosing the threshold on $\widehat{e}_m$ at $\gamma_m = q\,||\boldsymbol{a}_m||$, so that $\tau_{\mathrm{FA}} = Pr(\widehat{x}_m \neq 0) = 1 - \mathrm{erf}(q/\sqrt{2})$, where erf is the Gaussian error function. That is, the weight γ_m on $|x_m|$ should be proportional to $||\boldsymbol{a}_m||$. Equivalently, $\mathbf{A}$ should have normalized columns so that a unique hyperparameter $\gamma_m = \gamma$ yields a uniform false detection rate on each component. A practical illustration of such a dependence between hyperparameter values, dictionary normalization and false alarms can be found in [31].

Let $\mathbf{N}_A$ denote the diagonal matrix with elements $\{||\boldsymbol{a}_m||\}_{m=1\ldots M}$, so that the columns of $\mathbf{B} \triangleq \mathbf{A}\mathbf{N}_A^{-1}$ have unit norm. The problem we want to solve reads finally

$$\begin{aligned} \widehat{\boldsymbol{x}} &= \arg\min_{\boldsymbol{x}} \frac{1}{2}||\boldsymbol{z} - \mathbf{A}\boldsymbol{x}||^2 + q\sum_m ||\boldsymbol{a}_m||\,|x_m| \\ \Leftrightarrow \widehat{\boldsymbol{u}} &= \arg\min_{\boldsymbol{u}} J(\boldsymbol{u}), \text{ where} \\ J(\boldsymbol{u}) &= \frac{1}{2}||\boldsymbol{z} - \mathbf{B}\boldsymbol{u}||^2 + q\sum_m |u_m| \end{aligned} \tag{5}$$

with $\boldsymbol{u} = \mathbf{N}_A\boldsymbol{x}$, that is $u_m = ||\boldsymbol{a}_m||\,x_m$. Typical values of q are 3 or 4 for which one has, respectively, $\tau_{\mathrm{FA}} \simeq 2.7\,10^{-3}$ and $\tau_{\mathrm{FA}} \simeq 6.3\,10^{-5}$. The latter formulation (5) is used in the rest of the paper.

D. Equivalent Dictionary Analysis and Solution Uniqueness

Equation (5) defines a classical *Basis Pursuit De-Noising* problem, where relevant atoms are searched in the equivalent dictionary $\mathbf{B} = \boldsymbol{\Sigma}^{-1/2}\mathbf{H}\mathbf{W}\mathbf{N}_A^{-1}$. Such a dictionary is imposed by observational constraints, prior information on the spectra of galaxies, and a desired uniform false alarm rate. It is certainly not designed in order to satisfy usual recovery conditions such as low mutual coherence [10], [12], positive Exact Recovery Coefficient [42] or the Restricted Isometry Property [43], that can be used to prove uniqueness of the minimizer of (5). The Unique Representation Property (URP, [44]) is another sufficient condition for uniqueness, and supposes that any N columns of the dictionary are independent, where N is the number of data. In our case, even small sets of columns in $\mathbf{W}$ are obviously linearly dependent—see Section III-B. Thus, the same columns in $\mathbf{B}$ are also linearly dependent. Hence, $\mathbf{B}$ does not satisfy the URP. Addressing uniqueness in our case thus seems intractable. Consequently, we can only claim that the minimizer is a convex set of solutions, possibly not reduced to a single element, but where all members are equivalent solutions: in the Bayesian maximum *a posteriori* estimation framework [2], all of them maximize the posterior distribution under a Laplacian prior assumption on each coefficient: $p(u_m) \propto \exp(-q|u_m|)$. Let us remark that, in our experiments, all tested optimization algorithms always converged to the same solution, suggesting that uniqueness may be obtained in practice.

Let us note here that in the literature of compressed sensing, the first theoretical results regarding dictionaries with high correlations were, to the best of our knowledge, established only very recently in [45]. These results cannot be used in the present case, however, because they concern a sparse analysis approach (for which $\mathbf{W}^T\boldsymbol{s}$ is sparse) rather than the synthesis approach considered here (for which $\boldsymbol{s} = \mathbf{W}\boldsymbol{x}$ with $\boldsymbol{x}$ sparse).

E. Amplitude Re-Estimation

The ℓ^1-norm penalization is known to introduce bias on the amplitudes of the nonzero components in $\widehat{\boldsymbol{u}}$ [12]. In practice, once such components have been identified by solving (5), their best fit to the data can be obtained in the least-squares sense. Let $\widehat{\boldsymbol{u}}_\star$ collect only the nonzero components in $\widehat{\boldsymbol{u}}$, and $\mathbf{B}_\star$ collect the corresponding columns of $\mathbf{B}$. Amplitude re-estimation is performed by computing

$$\widehat{\boldsymbol{u}}_\star^{\mathrm{reest}} = \arg\min_{\boldsymbol{u}_\star} ||\boldsymbol{z} - \mathbf{B}_\star\boldsymbol{u}_\star||^2 = \left(\mathbf{B}_\star^T\mathbf{B}_\star\right)^{-1}\mathbf{B}_\star^T\boldsymbol{z} \tag{6}$$

where the matrix inversion in the latter equation is properly defined if $\mathbf{B}_\star$ is full-rank, that is, if optimal atoms are linearly independent. Despite the high number of linear dependencies in our dictionary and its high coherence, we expect the sparsity constraint to be strong enough in order to select linearly independent atoms. In practice, given the very low signal-to-noise ratio, no more than a few tens of active atoms are selected, so that in all our experiments, we never had to face non-invertible or ill-conditioned matrices $\mathbf{B}_\star^T\mathbf{B}_\star$. Equation (6) can be efficiently computed by conjugate gradients [16], [17], or by direct matrix inversion if the number of nonzero components is small enough. Note that such re-estimation is applied to the output of the normalized problem (5). Spectral components are then finally retrieved by multiplying the coefficients of $\widehat{\boldsymbol{u}}_\star^{\mathrm{reest}}$ by the corresponding atoms of $\mathbf{W}\mathbf{N}_A^{-1}$.

V. Optimization Without Fast Transforms

The growing interest for ℓ^1-based regularization in signal and image processing gave birth to an abundant literature about optimization of criteria such as (5). In many applications, priority is given to the ability of processing large-size data—in particular, images—and thus to the computational efficiency of the transforms. In practice, dictionaries based on DCT or wavelets are not built explicitly, and optimization efficiency relies on fast transforms to compute products $\mathbf{B}\cdot$ and $\mathbf{B}^T\cdot$ (or $\mathbf{B}^T\mathbf{B}\cdot$). In this category fall all gradient-based methods such as Iterative Thresholding [14], [15], Fast Iterative Shrinkage-Thresholding (FISTA, [18]), Gradient Projection for Sparse Reconstruction (GPSR, [16]) or Sparse Reconstruction by Separable Approximation (SpaRSA, [17]).

The problem tackled in this paper is different, where priority is given to the design of an adapted dictionary. The price to pay is the impossibility to compute high-dimensional operations at low cost, causing the inefficiency of gradient-based methods. Algorithmic solutions exist, though, that are suited to this problem. In particular, homotopy continuation (HC) methods [20], [21], introduced for sparse regression in statistics, do not

rely on the use of fast transforms. HC is known to perform very efficiently, in particular it outperforms usual quadratic programming methods, especially when the solution is highly sparse [21], [22]. Iterative Coordinate Descent (ICD) is another alternative, whose efficiency was recently exhibited for large-scale sparse regression problems [27], [29] and for sparse spectral analysis [28].

A. Support Exploration Algorithms

Both HC and ICD can be viewed as *support exploration* algorithms, because they are particularly efficient at quickly finding the correct support of the solution if it is sparse enough. Identifying the *signed support* of the solution, that is, the location of the nonzero components and their corresponding signs, is the hardest task in ℓ^1 minimization. Indeed, once the signed support is found, amplitude estimation is straightforward. Let $\widehat{\boldsymbol{u}}$ denote a minimizer of (5), let $\widehat{\boldsymbol{u}}_\star$ collect its nonzero components and $\mathbf{B}_\star$ collect the corresponding columns of $\mathbf{B}$. One has $\nabla \|\widehat{\boldsymbol{u}}_\star\|_1 = \operatorname{sign}(\widehat{\boldsymbol{u}}_\star)$ and minimization of (5) in $\widehat{\boldsymbol{u}}_\star$—other components of $\widehat{\boldsymbol{u}}$ are zero—can be written as the gradient cancellation

$$-\mathbf{B}_\star^T(\boldsymbol{z}-\mathbf{B}_\star\widehat{\boldsymbol{u}}_\star)+q\operatorname{sign}(\widehat{\boldsymbol{u}}_\star)=\mathbf{0} \Leftrightarrow \widehat{\boldsymbol{u}}_\star=\left(\mathbf{B}_\star^T\mathbf{B}_\star\right)^{-1}\left(\mathbf{B}_\star^T\boldsymbol{z}-q\operatorname{sign}(\widehat{\boldsymbol{u}}_\star)\right) \quad (7)$$

where the matrix inversion in the latter equation is properly defined if $\mathbf{B}_\star$ is full-rank—see the former discussion in Section IV-E.

The efficiency of HC and ICD depends on their ability to quickly identify the correct signed support. Basically, each loop of HC tends to add nonzero components to the iterates starting from the zero vector. Although the principle of ICD is essentially different, it was found that most of the changes in the iterates operated similarly, where only a few changes in the support are performed at each iteration. Consequently, algorithmic efficiency is directly related to the degree of sparsity: the sparser the solution, the faster the support is identified. In that sense, such algorithms can be linked with greedy support exploration techniques, but operate on the convex formulation of the sparse representation problem (see [23] for a detailed study of the parallelism between HC and Orthogonal Matching Pursuit). Hence, HC and ICD fundamentally differ from gradient-based strategies, which do also produce sparse solutions, but do little exploit sparsity for computational efficiency. On the other hand, support exploration algorithms would probably not be appropriate for very large size problems or less sparse solutions.

B. Homotopy Continuation in Practice

We recall the principle of HC, first proposed in [22] for the specific formulation of (5). Let $\widehat{\boldsymbol{u}}(q)$ denote the minimizer of (5) for a given value of q, and let $\overline{q}$ denote the desired value of q. HC is based on the observation that the signed support of $\widehat{\boldsymbol{u}}$ is a piecewise constant function of q. Starting at $q_0=\left\|\mathbf{B}^T\boldsymbol{u}\right\|_\infty$, for which $\widehat{\boldsymbol{u}}(q_0)=\mathbf{0}$, HC works by decreasing q until a change appears in the signed support of $\widehat{\boldsymbol{u}}$. Indeed, for a given support, the value $q^{(m)}$ at which a change would appear on the sign of the mth component by decreasing q can be computed analytically (see [22] for details). The largest value among all $q^{(m)}$ is selected as q_1 and the signed support is updated, that is, one component switches from zero to ±1 or from ±1 to zero. The procedure is repeated and a non-increasing sequence of values $\{q_t\}_{t=1\ldots T}$ is built at which the support of $\widehat{\boldsymbol{u}}(q)$ changes—then, the support is updated. Algorithm stops when $\overline{q}\in[q_T,q_{T-1}]$. The signed support of $\widehat{\boldsymbol{u}}(\overline{q})$ is then identified and amplitudes are obtained by (7).

The most time-consuming part of this procedure is the computation, at each iteration t, of all $q^{(m)}$, which involve M system inversions whose size generally increase with t. Since every change in the support only operates on one component, inversions can be performed by recursively building inverse matrices. Note that HC provides all supports for q varying from q_0 to $\overline{q}$: this could help selecting *a posteriori* the regularization parameter, for example by choosing a fixed number of nonzero components. On the contrary, in our case, $\overline{q}$ is tuned according to the simple statistical rule given in Section IV-C, and fixing a certain number of components in the solution is to be avoided here. Indeed, depending on their position in the field of view, the spectra may be rich in spectral features (bright galaxies), or present only one or a few emission lines (faintest galaxies)—or even, for regions of the sky with no detectable source, contain only noise.

C. Iterative Coordinate Descent and Accelerations

1) Basic Version: ICD consists in performing componentwise minimizations of (5), which have the analytical solution

$$\arg\min_{u_m} J(\boldsymbol{u})=\phi_q^{\mathrm{st}}\left(\boldsymbol{b}_m^T\left(\boldsymbol{z}-\sum_{p\neq m}u_p\boldsymbol{b}_p\right)\right) \quad (8)$$

where ϕ_q^{st} is the soft-thresholding function [5]:

$$\begin{cases}\phi_q^{\mathrm{st}}(u)=0, & \text{if } |u|\leq q\\ \phi_q^{\mathrm{st}}(u)=u-\operatorname{sign}(u)q, & \text{if } |u|>q.\end{cases}$$

Basic ICD works by starting at any point in $\mathbb{R}^M$ and then repeatedly updating all coordinates successively, until some stopping rule is met. KKT conditions in (4) provides an explicit characterization of the minimizer that can be used as a strong convergence test, which reads for criterion (5):

$$\begin{cases}\text{for } \widehat{u}_m=0: & |\boldsymbol{b}_m^T(\boldsymbol{z}-\mathbf{B}\widehat{\boldsymbol{u}})|<q\\ \text{for } \widehat{u}_m\neq0: & \boldsymbol{b}_m^T(\boldsymbol{z}-\mathbf{B}\widehat{\boldsymbol{u}})=q\operatorname{sign}(\widehat{u}_m).\end{cases} \quad (9)$$

In the following, we will denote by $M_{\mathrm{KKT}}(\boldsymbol{u})$ the number of coordinates in $\boldsymbol{u}$ that satisfy (9). Convergence is then declared when $M_{\mathrm{KKT}}(\boldsymbol{u})=M$.

Standard ICD is ensured to converge toward the minimum of (5) [25]. Its efficiency for minimizing ℓ^1-penalized functionals was shown in particular in [27]–[29], and has two main reasons:

- residuals $\boldsymbol{z}-\sum_{p\neq m}u_p\boldsymbol{b}_p$ can be updated recursively and each update (8) can be performed at low cost;
- if $\widehat{\boldsymbol{u}}$ is sparse, then most updates concern zero values, which do not require any computation.

We propose the following improvements to this standard algorithm.

2) Selective Cycling [28], [46]: The most straightforward improvement consists in cycling only through the components that *need* updating. Indeed, each update (8) has a nonzero cost

due to the inner product $\boldsymbol{b}_m^T\cdot$, even if no update is performed. If the solution is sparse, however, then most zero components are quickly identified by ICD and do not need to be updated. Consequently, we consider a cycling rule (which we refer to as NZ cycling) where updates (8) are only applied to the nonzero components of the current iterate. Cycling is also periodically performed through all components, which ensures convergence: including cycling steps on nonzero components only, all of which decrease the value of J, can be viewed as the addition of *spacer steps* [47, Ch. 7] in the standard ICD, and yields a convergent procedure.

3) Support Testing: Recall that, if the signed support of the solution is known, then corresponding amplitudes can be found analytically by (7). Consequently, any signed support can be tested as the optimal one: corresponding amplitudes are computed according to (7), then optimality condition (9) is checked. Implementing support testing in ICD aims at performing a shortcut in last iterations, which generally only work at estimating amplitudes while the correct support has already been identified.

4) Local Tricks: When implementing ICD, two efficiency-limiting factors were identified: 1) some zero components are sometimes reached slowly; and 2) many successive iterates may contain two highly correlated atoms, whereas only one is active at the optimum. We propose to test the optimality of the support obtained by setting to zero each nonzero coefficient [for 1)], or each coefficient in pairs of active atoms whose correlation exceeds some threshold [for 2)]: we call this step *atom disambiguation*. In both cases, the component is set to zero in the current iterate if the number of optimal coordinates M_{KKT} is increased.

Support testing requires the computation of both (7) and (9). Hence, in practice, it is only performed periodically, and only if the current support has not been already tested before. Similarly, local accelerations are only performed when the iterate is close to convergence. Table I details the implementation of the accelerated ICD algorithm, where:

- $\mathbf{S}$ collects all signed supports for which optimality has already been tested;
- T_{all} controls the period at which all coordinates are swept in one ICD iteration;
- T_{test} controls the period at which support testing and local exploration are performed;
- $M_{\mathrm{KKT}}^{\min}$ defines the minimum number of components satisfying KKT conditions (9) required for performing optimality testing. In the following, it is set to $M-10$;
- $\mu^{\max}$ defines the correlation value between two active atoms above which zeros are tested in the corresponding support. In the following, it is set to $1-10^{-4}$;
- `tol` is a numerical tolerance parameter, used in checking KKT equalities. In the following, it is set to 10^{-4}.

TABLE I
ACCELERATED ICD ALGORITHM FOR ℓ^1-NORM-CONSTRAINED OPTIMIZATION

Initialize $t=0$, $\boldsymbol{u}^{(0)}\in\mathbb{R}^M$ and $\mathbf{S}=\emptyset$. Then: i)

1) Selection of sweep indexes.
   - If $t\equiv 0 \pmod{T_{\mathrm{all}}}$, then set $\mathcal{M}=[1,\dots,M]$;
   - else, set $\mathcal{M}=\mathcal{M}_\star=\{m|u_m^{(t)}\neq 0\}$.
2) ICD sweep. For $m\in\mathcal{M}$, update successively $u_m^{(t)}$ with soft-thresholding:
$$u_m^{(t)} = \phi_q^{\mathrm{st}}\Big(\boldsymbol{b}_m^T\big(\boldsymbol{z}-\sum_{p\in\mathcal{M},\,p<m}u_p^{(t)}\boldsymbol{b}_p\ \dots \ \dots-\sum_{p\in\mathcal{M},\,p>m}u_p^{(t-1)}\boldsymbol{b}_p\big)\Big).$$
3) If $t\equiv 0 \pmod{T_{\mathrm{test}}}$ and $\mathrm{sign}(\boldsymbol{u}^{(t)})\notin\mathbf{S}$, then test optimality of $\mathrm{sign}(\boldsymbol{u}^{(t)})$:
   - with $\star$ indexing elements in $\mathcal{M}_\star$, compute
$$\widehat{\boldsymbol{u}}_\star^{(t)}=\left(\mathbf{B}_\star^T\mathbf{B}_\star\right)^{-1}\left(\mathbf{B}_\star^T\boldsymbol{z}-q\ \mathrm{sign}(\boldsymbol{u}^{(t)})\right)$$
     and form $\widehat{\boldsymbol{u}}^{(t)}$ by $\widehat{\boldsymbol{u}}_\star^{(t)}=\widehat{\boldsymbol{u}}_\star^{(t)}$ and $\widehat{u}_m^{(t)}=0$ for $m\notin\mathcal{M}_\star$;
   - check optimality of $\widehat{\boldsymbol{u}}^{(t)}$ by computing the number $M_{\mathrm{KKT}}(\widehat{\boldsymbol{u}}^{(t)})$ of components satisfying:
$$\begin{cases}\text{for } \widehat{u}_m^{(t)}=0: & |\boldsymbol{b}_m^T(\boldsymbol{z}-\mathbf{B}\widehat{\boldsymbol{u}}^{(t)})|<q\\ \text{for } \widehat{u}_m^{(t)}\neq 0: & \boldsymbol{b}_m^T(\boldsymbol{z}-\mathbf{B}\widehat{\boldsymbol{u}}^{(t)})=q\,\mathrm{sign}(\widehat{u}_m^{(t)})\end{cases}$$
     where the equality condition is tested up to numerical tolerance given by `tol`;
   - if $M_{\mathrm{KKT}}(\widehat{\boldsymbol{u}}^{(t)})=M$, then stop. Otherwise, add $\mathrm{sign}(\boldsymbol{u}^{(t)})$ to $\mathbf{S}$.
4) If $t\equiv 0 \pmod{T_{\mathrm{test}}}$ and $M_{\mathrm{KKT}}\geq M_{\mathrm{KKT}}^{\min}$, then try local tricks: $\forall m\in\mathcal{M}_\star$ (alternately $\forall m,p\in\mathcal{M}_\star$ such that $|\boldsymbol{b}_m^T\boldsymbol{b}_p|>\mu_{\max}$),
   - set the m^{th} (p^{th}) component to 0: form $\boldsymbol{u}^{\mathrm{test}}$ by $\boldsymbol{u}^{\mathrm{test}}=\boldsymbol{u}^{(t)}$ except $u_m^{\mathrm{test}}=0$ (except $u_p^{\mathrm{test}}=0$);
   - perform optimality test 3) with $\boldsymbol{u}^{\mathrm{test}}$ instead of $\boldsymbol{u}^{(t)}$ and add $\mathrm{sign}(\boldsymbol{u}^{\mathrm{test}})$ to $\mathbf{S}$;
   - if $M_{\mathrm{KKT}}(\boldsymbol{u}^{\mathrm{test}})>M_{\mathrm{KKT}}(\boldsymbol{u}^{(t)})$, then keep $u_m^{(t)}=0$ ($u_p^{(t)}=0$).
5) If $M_{\mathrm{KKT}}(\widehat{\boldsymbol{u}}^{(t)})=M$, then stop. Otherwise, set $t=t+1$ and go back to i).

VI. Simulation Results

A. Restoration of MUSE-Like Spectra

Estimation results are given for the data in Fig. 1. We recall that such data are the result of complex simulations. In particular, they are not built in accordance with model (2), which is used for reconstruction. Criterion (5) is minimized for $q=4$ (see Section IV-C) and amplitudes are re-estimated according to (6). For spectrum $\boldsymbol{s}$ and corresponding noisy or reconstructed $\widetilde{\boldsymbol{s}}$, we define the signal-to-noise ratio (SNR) and the spectral angle (SA), respectively, by

$$\mathrm{SNR}_{\mathrm{dB}}=10\log_{10}\frac{\|\boldsymbol{s}\|^2}{\|\boldsymbol{s}-\widetilde{\boldsymbol{s}}\|^2}$$
$$\text{and}\quad \mathrm{SA}_{\mathrm{deg}}=\mathrm{acos}\frac{\widetilde{\boldsymbol{s}}^T\boldsymbol{s}}{\|\widetilde{\boldsymbol{s}}\|\,\|\boldsymbol{s}\|}$$

where the latter is a common measurement of spectral similarity in hyperspectral imaging [48]. Restoration yields[3] SNR $=12.1$ dB and SA $=3.5\,\mathrm{deg}$, whereas for noisy data we have respectively SNR $=3.1$ dB and SA $=26.7\,\mathrm{deg}$. Note that before amplitude re-estimation with (6), we have SNR $=9.2$ dB and

[3]Current MUSE simulations only provide noiseless but already *convolved* spectra: $\boldsymbol{s}$ is thus unknown. The SNRs that would be obtained with respect to $\boldsymbol{s}$ instead of its convolved version, are thus expected to be slightly higher than the SNRs presented here.

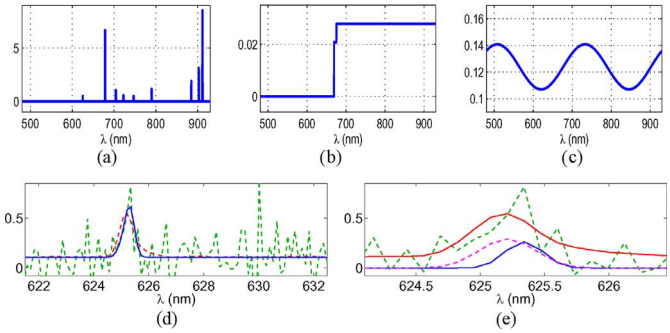

Fig. 2. Decomposition results on the MUSE-like spectrum of Fig. 1. (a)–(c) Estimated spectral components: (a) line spectrum; (b) step spectrum; and (c) continuous spectrum. (d) Zoom on the spectral line at 625 nm, with noise-free (red), noisy (green), and restored (blue) data. (e) Noise-free (red) and noisy (green) data, together with the two selected atoms (magenta dashed and blue solid curves). The noise-free data serve here as ground truth, but are already convolved by the LSF. Amplitudes are light fluxes, in $\mathrm{erg}\cdot\mathrm{s}^{-1}\cdot\mathrm{cm}^{-2}$ ($\times 10^{-20}$).

SA $= 4.8\,\mathrm{deg}$. Right panels in Fig. 1 show the restored spectrum, and associated spectral components are plotted in Fig. 2. The estimated line spectrum correctly locates the main lines. In particular, a faint emission line is detected at 625 nm in a very noisy environment, as shown on the zooms in Fig. 2(d) and (e), where the asymmetric line profile is estimated by two spline atoms with different widths and slightly shifted centers. A break at 673 nm is detected by two close step atoms, and the continuous spectrum is estimated with two components (the mean value and one low frequency oscillation). The estimated continuous spectrum fits the noiseless spectrum quite well, except at the highest wavelengths, where the oscillation produced by the sine atom generates a more important difference with the reference spectrum. However, given the low SNR (see the noisy data in green), this is still a rather satisfactory result. The noise reduction is clearly visible in Fig. 1(b), which also shows undetected faint lines around 700 nm and side effects caused by line profile approximation errors. Such errors, however, remain relatively small compared with the signal amplitude. In the case shown here, the number of synthesis coefficients is 23. This is a favorable case in the sense that for most spectra, the SNR will be lower than shown here—as will also be the number of detectable spectral features.

We note that all selected columns in $\mathbf{B}$ are linearly independent, so that inversion of $\mathbf{B}_\star^T\mathbf{B}_\star$ in (6) is possible. In this example, the Exact Recovery Coefficient (ERC, [42]) equals -2.81; hence, it cannot be used to claim uniqueness of the solution[4]. However, all algorithms that have been tested (see next Section VI-B) converged to the same solution.

B. Optimization Efficiency

In this section, the behavior and performance of several algorithms are compared. Minimization of (5) is performed on the example of Section VI-A by previously described HC and ICD methods and by recent gradient-based techniques: GPSR, SpaRSA, and FISTA. Many algorithms recently appeared in the literature; hence, this comparison is far from exhaustive. However, the former gradient methods were shown to outperform several other algorithms in their corresponding papers, and HC was shown to be much more efficient than standard quadratic programming solvers on several problems of similar (or smaller) sizes than the problem considered here [26], [27].

1) Implementation Details: All algorithms are implemented in Matlab and run on an AMD Opteron 2356 Dual Quad-Core processor under Linux with Central Processing Units (CPUs) clocked at 2.30 GHz. Matrix-vector products are computed only on nonzero components and Matlab's sparse data structure for matrices[5] was used if the computational cost was reduced. Parameter q was set to 4. HC provides in addition the solution paths for higher values of q, but this gain is not considered here.

GPSR and SpaRSA were implemented from the codes available at M. Figueiredo's web page,[6] and only the most favorable results are presented among four tested options (with/without "warm start" and with/without enforcing monotonicity). For FISTA, the gradient step size depends on the spectral norm of matrix $\mathbf{B}$ [15], [18], whose computation is extremely cost consuming in our case. In the following, the maximum step size that yields convergence was determined empirically for each simulation. For HC, both recursive and direct matrix inversions were implemented and only the most favorable results are presented. In general, since the solution is very sparse, direct computations were more efficient.

All algorithms are initialized at $\mathbf{0}$, and are stopped when KKT conditions (4) are satisfied with a numerical tolerance of 10^{-4} for equalities. In all tests that were performed, all algorithms yielded the same support and similar amplitudes up to numerical tolerance.

2) CPU Costs for Astrophysical Data: Fig. 3 plots the number of components with correct sign in the current iterate versus CPU time for all tested algorithms, jointly with the corresponding criterion. For HC, at each change in the support, amplitudes were computed with (7) and criterion (5) was evaluated. The ICD version described in Table I was implemented with $T_{\mathrm{all}} = 250$, $T_{\mathrm{test}} = 500$.

ICD with full cycling is not efficient, because most time is lost by unnecessary cycling along zero components. Gradient-based strategies also perform very poorly. One can see in particular that they are very slow at finding a support close to the true one. On the contrary, HC and ICD with selective cycling are much more efficient. Their performance is shown for two realizations of noise affecting the noise-free data in Fig. 1. In the first one (center row), all versions of ICD with NZ cycling (hereafter, ICD-NZ, that is, with step 1) in Table I) outperform HC. One can see in particular that mostly the half of the computational time of ICD-NZ is dedicated to the estimation of amplitudes—this can be evaluated by comparing solid and dotted vertical blue lines in Fig. 3, center left, respectively indicating convergence and correct support identification (the three dotted lines for the three ICD versions are almost superimposed). Performing support testing (step 3) in Table I) then reduces the cost almost by half. In this example, local tricks [step 4)] do not bring any significant improvement. In the second simulation (Fig. 3

[4]A positive ERC value is actually a sufficient condition for uniqueness.

[5]Indeed, for MUSE-like data, $\mathbf{H}$ is sparse and most atoms in the dictionary are support-limited; hence, $\mathbf{B}$ also has a sparse structure.

[6]http://www.lx.it.pt/~mtf

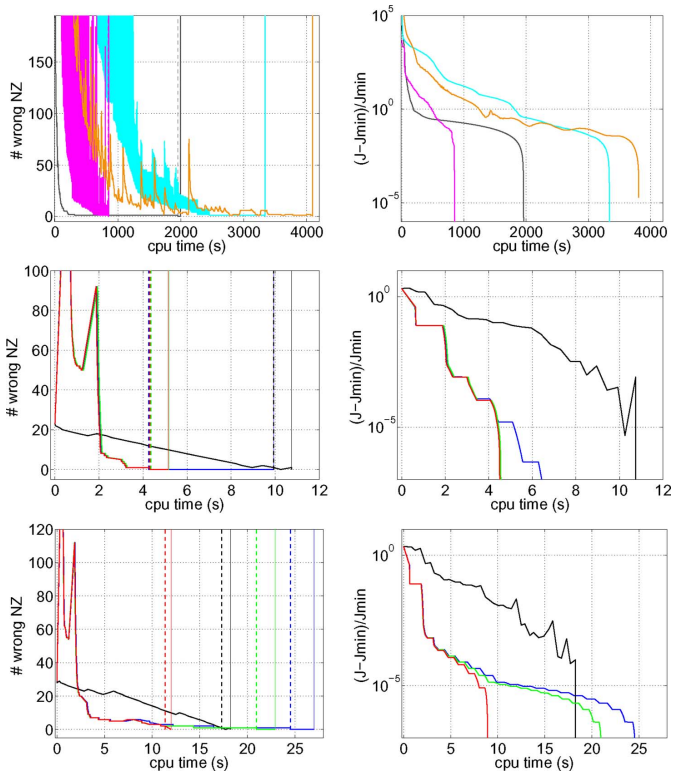

Fig. 3. Number of correctly identified components (left) and criterion value (right) versus CPU time. Top: FISTA (orange), GPSR (cyan), basic ICD (gray), and SpaRSA (magenta). Center and bottom: HC (black), ICD with NZ cycling (blue), ICD with NZ cycling and support testing (green), ICD with NZ cycling, support testing and local tricks (red). Top and center panels correspond to the same data set. Bottom panels correspond to a critical data set for ICD with slow atom disambiguation (see text). Vertical full (resp., dotted) lines on left panels indicate CPU time for convergence (resp., for correct signed support identification).

TABLE II
OPTIMIZATION COSTS FOR DIFFERENT ALGORITHMS ON ASTROPHYSICAL SPECTRA AND CS-LIKE SCENARIOS (IN SECONDS). VALUES IN PARENTHESES INDICATE TIMES FOR CORRECT SUPPORT IDENTIFICATION

	MUSE-like data NZ $\simeq$ 23		CS scenario CS-30		CS-300	
ICD						
full cycling	1992	(1955)	3.23	(1.71)	5.58	(3.54)
NZ cycling	25.0	(11.6)	1.20	(1.15)	2.53	(2.40)
NZ cycling + support testing	11.7	(10.5)	1.35	(1.30)	2.53	(2.40)
NZ cycling + support testing + local tricks	7.2	(6.6)	1.31	(1.26)	2.51	(2.39)
HC	15.5	(15.0)	17.98	(17.52)	459	(457)
FISTA	4090	(4090)	3.69	(3.69)	7.31	(7.31)
GPSR	3340	(3337)	1.78	(1.19)	3.92	(3.08)
SpaRSA	850.9	(849.4)	1.78	(1.15)	2.98	(2.20)

bottom), critical configurations are yielded by ICD-NZ iterations caused by slow disambiguation between correlated atoms (see Section V-C4), so that 28 s are needed for ICD-NZ convergence, whereas HC takes 18 s. Adding a support testing step reduces the cost by 4 s, and local tricks yield an additional gain of 11 s.

Table II (left column) shows the computational costs for all algorithms, averaged on 15 realizations of the noise process. Costs are given for both convergence and correct support identification. Confirming the example in Fig. 3, gradient methods and ICD with full cycling yield excessive costs. Cycling only nonzero components reduces the cost of ICD, but the average cost still exceeds the one of HC. Note in particular that the correct support is obtained, in average, at less that half the total duration of the algorithm. Performing support testing then divides the cost by more than half, and adding local exploration steps still saves approximately 40% of the CPU time. The final version of our ICD algorithm then reduces by most by half the cost of HC.

3) CPU Costs for an Artificial Compressed Sensing Example: Such poor performance of gradient methods are partly due to the absence of fast operators, but this is not the only reason. Algorithms were also run on an artificial example with the same operator size than before ($\mathbf{B}$ is $3463 \times 26\,015$), but whose coefficients were randomly drawn from a Gaussian distribution with unit variance, in a "compressed-sensing" (CS) philosophy—see [17] for a similar example. Contrary to the astrophysical spectra case, the dictionary is almost *incoherent*. Data were generated by applying $\mathbf{B}$ to a sparse vector with randomly chosen locations and amplitudes of nonzero components, drawn from uniform and unit variance Gaussian distributions, respectively. Then, 5-dB white Gaussian noise was added, to reach a similar SNR to that of former MUSE-like data. Two different sparsity levels with 30 and 300 nonzero components were generated, that we denote CS-30 and CS-300. Here again, all algorithms always converged to the same solution. ICD versions were implemented with $T_{\text{all}} = 100$ and $T_{\text{test}} = 20$ for CS-30, and $T_{\text{all}} = 10$, $T_{\text{test}} = 5$ for CS-300.

CPU times averaged on 15 random realizations are given in Table II, center and right columns. All algorithms except HC run much more quickly on CS problems, especially ICD with full cycling and gradient methods, among which SpaRSA yields the lowest CPU times. We explain such differences between the two problems by the very different structures of the two dictionaries. Whereas atoms in the astrophysical dictionary are highly correlated, randomly drawn atoms are almost orthogonal. Consequently, descent directions in the CS case are much better separated, and much deeper descent steps are performed by both gradient methods and ICD. Indeed, the spectral norm of the dictionary (the square root of the maximum eigenvalue of $\mathbf{B}^t\mathbf{B}$, which equals 1 for orthonormal matrices) equals 2411 for the normalized astrophysical dictionary and 14 in the CS case. Recall that the maximum step size ensuring convergence of gradient methods like FISTA is inversely proportional to the spectral norm of the dictionary [15], [18].

By construction, the cost of HC is globally proportional to the number of nonzero components in the solution, and does not strongly depend on the dictionary. In both CS examples, HC is not competitive with ICD and gradient strategies, and its cost becomes prohibitive for CS-300. We note that the cost of all algorithms increases when switching from CS-30 to CS-300 data. This is a logical result for ICD because a larger support is searched (in particular, NZ cycling is performed on more components) and also for gradient-based algorithms, because products $\mathbf{B}\boldsymbol{u}$ operate in higher dimension.

ICD with NZ cycling is still the most efficient algorithm among all tested methods, but additional support testing and local exploration steps do not improve efficiency. This is an expected result because such tests are designed for coherent dictionaries, and their contribution in this case does not compensate for their additional cost. We also note that, on both CS examples, SpaRSA is the most efficient strategy for support identification. This suggests that gradient-based methods could also benefit from support testing steps.

VII. Conclusion and Further Work

Restoration of astrophysical spectra was addressed as a sparse approximation problem. A data formation model was constructed, which accounts for observational constraints. A specific dictionary was designed in accordance with astrophysical spectroscopy, where each atom corresponds to an elementary spectral feature. This allowed us to combine denoising and deconvolution with the detection of physically relevant features, such as emission or absorption lines and discontinuities. Sparse estimation was considered through the minimization of an ℓ^1-norm based criterion, where specificities were shown to require normalization of the equivalent dictionary. Results on an artificial spectrum extracted from MUSE simulations were presented, where detections of lines and of a discontinuity were achieved; however, restoration quality of the continuous spectrum was limited by the too constraining oscillating atoms present in the dictionary. Optimization was studied, and an algorithm based on the ICD scheme was proposed, with accelerations that exploit the sparsity of the solution, substantially reducing the computational cost. The procedure was shown to outperform other recent algorithms for ℓ^1-penalized optimization with our designed dictionary, and also on simulations with random matrices. Consequently, ICD-based methods can be considered as a powerful alternative to gradient-based optimization techniques, especially in cases where no fast transform algorithm can be implemented.

In terms of data modeling, refinements in the design of the dictionary should be investigated. In particular, the parametrization of the continuous spectrum is not fully satisfactory. In continuity with our sparsity-based approach, using a dictionary with low-frequency splines is currently under study. Other parametric models such as polynomials, or non-parametric models based on a Markovian smoothness-promoting penalization such as in [49] are other possibilities. Both require, however, the tuning of additional parameters compared with the "full-sparse" approach. Smoother breaks could also be added in the dictionary, e.g., with several values of a discretized slope parameter.

A crucial extension of this work for MUSE data concerns the three-dimensional restoration of hyperspectral cubes. Formulating a three-dimensional problem, accounting for the instrument spatial point spread function, would perform joint spatial and spectral deconvolution and would certainly improve the restoration quality. Specific problems include the characterization of spatially extended sources (where several pixels have proportional spectra) and spectral *unmixing* [48]. Indeed, in the case of overlapping objects, the spectrum in a given pixel is a mixture of the corresponding spectra. By processing all pixels independently, the method presented in this paper does not allow to separate such spectral components. The main issue for addressing such problems obviously concerns the related computational complexity. The implicit dimension reduction operated by a sparse decomposition in the spectral domain is consequently a key point that should be exploited in order to efficiently tackle three-dimensional problems.

From the algorithmic point of view, automatic rules for fixing the parameters of the accelerated ICD procedure should be studied, based for example on the size of the problem and the expected degree of sparsity of solutions. We have shown that the structure of the dictionary deeply impacts the behavior of algorithms, especially for gradient methods. Precise characterizations of such a dependence deserve more attention. Including acceleration steps as performed with ICD could also be studied for other optimization methods. More generally, the merging of different algorithmic structures with complementary properties should be investigated for high-dimensional and complex sparse optimization problems.

References

[1] R. Bacon *et al.*, "Probing unexplored territories with MUSE: A second generation instrument for the VLT," in *Proc. SPIE*, Jul. 2006, vol. 6269, Ground-based and Airborne Instrumentation for Astronomy.

[2] *Bayesian Approach to Inverse Problems*, J. Idier, Ed.. New York: STE and Wiley.

[3] J. Tennyson, *Astronomical Spectroscopy*. London, U.K.: Imperial College Press, 2005.

[4] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*. New York: Academic, 2008.

[5] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[6] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 909–919, Apr. 1999.

[7] A. Antoniadis and J. Fan, "Regularization of wavelet approximations," *J. Amer. Statist. Assoc.*, vol. 96, pp. 939–967, 2001.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[9] S. Sardy, A. G. Bruce, and P. Tseng, "Block coordinate relaxation methods for nonparametric wavelet denoising," *J. Comput. Graph. Statist.*, vol. 9, pp. 361–379, 2000.

[10] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov. 2001.

[11] J.-L. Starck, M. Elad, and D. L. Donoho, "Redundant multiscale transforms and their application for morphological component analysis," *Adv. Electron. El. Phys.*, vol. 132, pp. 287–348, 2004.

[12] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1341–1344, Jun. 2004.

[13] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[14] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, Aug. 2003.

[15] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[16] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2980–2991, Dec. 2007.

[17] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, Jul. 2009.

[18] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[19] S. Alliney and S. A. Ruzinsky, "An algorithm for the minimization of mixed ℓ^1 and ℓ^2 norms with application to Bayesian estimation," *IEEE Trans. Signal Process.*, vol. 42, no. 3, pp. 618–627, Mar. 1994.

[20] M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *IMA J. Numer. Anal.*, vol. 20, no. 3, pp. 389–403, Jul. 2000.

[21] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.
[22] D. M. Malioutov, M. Çetin, and A. S. Willsky, "Homotopy continuation for sparse signal representation," in *Proc. IEEE ICASSP*, 2005, vol. 5, pp. 733–736.
[23] D. L. Donoho and Y. Tsaig, "Fast solution of ℓ_1-norm minimization problems when the solution may be sparse," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.
[24] W. J. Fu, "Penalized regressions: The bridge versus the lasso," *J. Comput. Graph. Stat.*, vol. 7, no. 3, pp. 397–416, Sep. 1998.
[25] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, pp. 475–494, 2001.
[26] M. Friedlander and M. Saunders, "Discussion: The Dantzig selector: Statistical estimation when p is much larger than n, by E. Candès and T. Tao," *Ann. Statist.*, vol. 35, no. 6, pp. 2385–2391, 2007.
[27] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *Ann. Appl. Stat.*, vol. 1, no. 2, pp. 302–332, 2007.
[28] S. Bourguignon, H. Carfantan, and J. Idier, "A sparsity-based method for the estimation of spectral lines from irregularly sampled data," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 575–585, Dec. 2007.
[29] T. T. Wu and K. Lange, "Coordinate descent algorithms for lasso penalized regression," *Ann. Appl. Statist.*, vol. 2, no. 1, pp. 224–244, 2008.
[30] A. Jarno, R. Bacon, P. Ferruit, and A. Pécontal-Rousset, "Numerical simulation of the VLT/MUSE instrument," in *Proc. SPIE*, 2008, vol. 7017, Modeling, Systems Engineering, and Project Management for Astronomy III, pp. 701710–701710–8.
[31] S. Bourguignon, D. Mary, and E. Slezak, "Sparsity-based denoising of hyperspectral astrophysical data with colored noise: Application to the MUSE instrument," in *Proc. IEEE WHISPERS*, Jun. 2010, pp. 1–4, DOI 10.1109/WHISPERS.2010.5594902.
[32] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vis. Res.*, vol. 37, no. 23, pp. 3311–3332, 1997.
[33] M. Aharon, E. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[34] M. Yaghoobi, L. Daudet, and M. E. Davies, "Parametric dictionary design for sparse coding," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 3311–3332, Dec. 2009.
[35] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[36] J. A. Högbom, "Aperture synthesis with a non-regular distribution of interferometer baselines," *Astron. Astrophys. Suppl.*, vol. 15, pp. 417–426, Jun. 1974.
[37] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
[38] G. M. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency decompositions," *SPIE J. Opt. Eng.*, vol. 33, no. 7, pp. 2183–2189, Jul. 1994.
[39] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, vol. 50, no. 5, pp. 1873–1896, 1989.
[40] J.-J. Fuchs, "On the application of the global matched filter to DOA estimation with uniform circular arrays," *IEEE Trans. Signal Process*, vol. 49, no. 4, pp. 702–709, Apr. 2001.
[41] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. Image Process.*, vol. 2, no. 3, pp. 296–310, Mar. 1993.
[42] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.
[43] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec.. 2005.
[44] I. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A recursive weighted norm minimization algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.
[45] E. J. Candès, Y. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Analysis*, vol. 31, no. 1, pp. 59–73, Jul. 2011.
[46] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statist. Software*, vol. 33, no. 1, pp. 1–22, Feb. 2010.
[47] D. G. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1989.
[48] C.-I. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. New York: Kluwer/Plenum, 2003.
[49] P. Ciuciu, J. Idier, and J.-F. Giovannelli, "Regularized estimation of mixed spectra using a circular Gibbs-Markov model," *IEEE Trans. Signal Process.*, vol. 49, no. 10, pp. 2201–2213, Oct. 2001.

Sébastien Bourguignon was born in Dijon, France, in 1977. He received the diploma degree in electrical engineering from École Supérieure d'Électricité, Gif-sur-Yvette, France, the engineer degree from ETSIT, Universidad Politécnica de Madrid, Spain, in 2001, and the Ph.D. degree in signal processing from the University of Toulouse, France, in 2005.

From 2002 to 2007, he was with the Astrophysics Laboratory of Toulouse-Tarbes, France. From 2007 to 2008, he was with IFREMER, the French research institute for exploitation of the sea. He is currently a Postdoctoral Fellow at the Cassiopée Laboratory, Côte d'Azur Observatory, Nice, France. His research interests include statistical inference and estimation, sparse approximation, optimization and MCMC algorithms, and applications to multi-dimensional observational data.

David Mary received the Ph.D. degree in signal processing from the École Nationale Supérieure des Télécommunications, Paris, France, in 2003.

In 2004, he joined the Aryabhatta Research Institute, Observational Sciences, Nainital, India, and the Astronomisches Rechen Institut, Heidelberg, Germany, in 2006. Since 2007, he has been an Assistant Professor in the Laboratoire Fizeau, University of Nice Sophia Antipolis, Nice, France. His research interests include statistical estimation and detection, approximation theory, and their applications.

Éric Slezak was born in France in 1961. He received the M.S. degree in signal processing from the University of Nice Sophia Antipolis, Nice, France, in 1984 and the Ph.D. degree in physics from the University of Nice Sophia Antipolis in 1988. His thesis research was carried out at the Nice Astronomical Observatory and dealt with the development of algorithms for object detection and classification in wide-field optical images.

He was among the first to introduce multiscale techniques in observational cosmology, developping wavelet-based approaches to analyze images, describe complex objects, compute density probability functions, and perform spectro-imagery from low S/N X-ray data. Doing so, he contributed in quantifying the subclustering, segregation properties and dynamical status of several clusters of galaxies and in detecting large-scale diffuse emission related to their building histories. Since 2006, he has held a Full Astronomer position at the Côte d'Azur Observatory, Nice, France. His research interests include astronomical image segmentation and classification from multi-wavelength data analysis. Since 2009, he has led a four-year IT research project aiming to develop the signal processing methods required to analyze the massive hyperspectral datasets provided by the forthcoming integral field spectrographs in astronomy.