DICTIONARY LEARNING WITH *SPATIO-SPECTRAL* SPARSITY CONSTRAINTS

Y. Moudden[1], *J. Bobin*[1,2], *J.-L. Starck*[1] *and J. Fadili*[3]

[1] DSM /IRFU/SEDI, CEA/Saclay, F-91191 Gif-sur-Yvette, France
[2]Department of Applied & Computational Mathematics , California Institute of Technology, Pasadena, California 91125
[3]GREYC - CNRS UMR 6072, 10 Bd Maréchal Juin, 14050 Caen, France

ABSTRACT

Devising efficient sparse decomposition algorithms in large redundant dictionaries has attracted much attention recently. However, choosing the right dictionary for a given data set remains an issue. An interesting approach is to learn the *best* dictionary from the data itself. The purpose of this contribution is to describe a new dictionary learning algorithm for multichannel data analysis purposes under specific assumptions. We assume a large number of contiguous channels as in so-called *hyperspectral* data. In this case it makes sense to consider *a priori* that the collected data exhibits sparse spectral signatures and sparse spatial morphologies in specified dictionaries of spectral and spatial waveforms. Building on GMCA, the proposed algorithm gives a practical way to enforce the additional *a priori* spectral sparsity constraint on the dictionary space. Numerical experiments with synthetic and real hyperspectral data illustrate the efficiency of the proposed algorithm.

1. INTRODUCTION

Generalized Morphological Component Analysis (GMCA) is a recent algorithm for multichannel data analysis described in [1, 2]. It was derived as a multichannel extension to MCA [3] and as such it is a powerful algorithm for the sparse decomposition of multichannel data over multichannel dictionaries. A major feature of the GMCA algorithm, in comparison to other sparse decomposition algorithms, is that it alternates sparse decomposition steps with dictionary learning steps, in a fast iterative thresholding loop with a progressively decaying threshold leading to a very robust *salient to fine* estimation process. In fact, learning dictionaries for the sparse representation of given data sets is now a growing field of interest [4, 5] although it is worth noting that this problem has a long history [6, 7, 8].

Obviously, it is required to set *a priori* some constraints on the dictionary space in which the optimal dictionary is to be looked for. In the case of GMCA, it is assumed that the size of the multichannel dictionary Ω, *i.e.* the number of atoms $n \times t'$, is specified beforehand and that the multichannel atoms are all rank one matrices, product of a *spectral* signature $a^k \in \mathbb{R}^{m,1}$ and a *spatial* density profile $\phi_k \in \mathbb{R}^{1,t}$. The column vectors a^k are grouped into a matrix noted $A \in \mathbb{R}^{m,n}$ and the line vectors ϕ_k are taken from matrix $\Phi \in \mathbb{R}^{t',t}$. GMCA models the data $X \in \mathbb{R}^{m,t}$ as follows :

$$X = A\nu\Phi + N = \sum_k \sum_{k'} \nu_k^{k'} a^k \phi_{k'} + N \qquad (1)$$

where the entries of the sparse matrix of coefficients ν representing X in the multichannel dictionary $\Omega = A \otimes \Phi$ are noted $\nu_k^{k'}$ and $N \in \mathbb{R}^{m,t}$ is included to account for Gaussian instrumental noise or modeling errors. GMCA further assumes that the dictionary of spatial waveforms Φ is known beforhand while the spectral components A, also called the mixing matrix in blind source separation (BSS) applications, is learned from the data. The image from the p^{th} channel is represented here as the p^{th} row of $\mathbf{X}$, x_p.

The successful use of GMCA in a variety of multichannel data processing applications such as BSS [2], color image restoration and inpainting [1] motivated research to extend its applicability. In particular, there are instances where one is urged by additional *prior* knowledge to further constrain the dictionary space. For instance, one may want to enforce equality constraints on some atoms, or the positivity or the sparsity of the learned dictionary atoms.

Builiding on GMCA, the purpose of this contribution is to describe a new dictionary learning algorithm for so-called *hyperspectral* data processing. Hyperspectral imaging systems collect data in a large number (up to several hundreds) of contiguous regions of the spectrum so that it makes sense to consider for instance that some physical property will show some regularity from one channel to the next. In fact, the proposed algorithm, referred to as hypGMCA, assumes that the multichannel atoms to be learned from the collected data exhibit diversely sparse *spatial* morphologies as well as diversely sparse *spectral* signatures in specified dictionaries $\Phi \in \mathbb{R}^{t,t'}$ and $\Psi \in \mathbb{R}^{m,m'}$ of respectively spatial and spectral waveforms. The proposed algorithm is used to learn from the data rank one multichannel atoms which are diversely sparse [2] in a given larger multichannel dictio-

nary.

In what follows, regardless of other models living in other scientific communities, the term *hyperspectral* denotes multichannel data following model (1) with the above two specific properties *i.e.* that the number of channels is large and that these achieve a *regular* sampling of some additional and meaningful physical index (*e.g.* wavelength, space, time) which we refer to as the *spectral* dimension. We describe next the proposed modified GMCA which we devised to account for the *a priori* sparsity of columns a^k in Ψ, a given dictionary of spectral waveforms. Accounting for this prior requires a modified objective function, discussed in section 2. The resulting hypGMCA algorithm is given in section 3. Finally, numerical experiments in section 4 demonstrate the efficiency of the proposed method.

2. OBJECTIVE FUNCTION

With the above assumptions, equation (1) is rewritten as follows :

$$X = \sum_k X_k + N = \sum_k \Psi \gamma^k \nu_k \Phi + N = \Psi \alpha \Phi \qquad (2)$$

where $X_k = a^k s_k$ are rank one matrices sparse in $\Omega = \Psi \otimes \Phi$ such that a^k has a sparse representation γ^k in Ψ while s_k has a sparse representation ν_k in Φ. In a BSS context, the rows s_k of S are commonly referred to as sources. For the sake of simplicity, we assume that Ψ and Φ are orthonormal bases ant that the noise N is uncorrelated inter- and intra-channel with variance σ^2. Extensions to more general cases are readily derived. Denote $\alpha_k = \gamma^k \nu_k$ the rank one matrix of coefficients representing X_k in Ω .

Looking at the above matrix equation column-wise, each column of $\Psi\alpha = X\Phi^T$ is represented as a sparse linear combination of columns in A with coefficients in the corresponding column of $\nu = S\Phi^T$. Indeed, the assumption that the lines of ν are diversely sparse (*e.g.* sparse and independent, or sparse with disjoint support, etc.) results in the columns of ν also being sparse. Usual sparse decomposition algorithms work on each column of $X\Phi^T$ separately to seek its sparsest representation in A. GMCA comes into play when the dictionary is not fully known beforehand and A needs to be learned from the data. GMCA constrains the dictionary space assuming that matrix A is unknown yet of specified size $m \times n$ and that Φ is given. This leads to a joint sparse coding and dictionary learning objective which can be expressed as a minimization problem in augmented Lagrangian form using an ℓ_1 sparsity measure :

$$\min_{A,\nu} \sum_{k=1}^{K} \lambda_k \|\nu_k\|_1 + \left\| X - \sum_{k=1}^{K} a^k \nu_k \Phi \right\|_2^2 \qquad (3)$$

which is clearly a difficult non-convex optimization problem. Nonetheless, the GMCA algorithm is able to provide a practical approximate solution as reported in [1]. Problem (3) is readily interpreted as a MAP estimation of the model parameters A and ν where the ℓ_1 penalty terms imposing sparsity come from a Laplacian prior on the sparse coefficient matrix ν.

Building on GMCA, we want here to learn spectral dictionary A so that again on average the columns of $X\Phi^T$ are simultaneously sparse in A and, what is more, the columns of A are sparse in Ψ.

A well known property of the linear mixture model (2) is its *scale and permutation invariance*. A consequence is that unless *a priori* specified otherwise, information on the separate scales of γ^k and ν_k is lost, and only a joint scale parameter for γ^k, ν_k can be estimated. This needs to be translated into a *practical* prior on $\gamma^k \nu_k$. We propose here that the following p_π is a good and practical candidate *joint sparse prior* for γ^k and ν_k after the loss of information induced by multiplication :

$$\begin{aligned} p_\pi(\gamma^k, \nu_k) &\propto \exp(-\lambda_k \|\gamma^k \nu_k\|_1) \\ &\propto \exp(-\lambda_k \textstyle\sum_{i,j} |\gamma_i^k \nu_k^j|) \end{aligned} \qquad (4)$$

where γ_i^k is the i^{th} entry in γ^k and ν_k^j is the j^{th} entry in ν_k. The proposed distribution has the nice property, for subsequent derivations, that the conditional distributions of γ^k given ν_k and of ν_k given γ^k are Laplacian distributions which are commonly and conveniently used to model sparse distributions. Finally, inserting the latter prior distribution in a Bayesian MAP estimator leads to the following minimization problem, expressed in coefficient space :

$$\min_{\{\gamma^k, \nu_k\}} \frac{1}{2\sigma^2} \left\| \mathbf{X} - \sum_k \mathbf{\Psi} \gamma^k \nu_k \mathbf{\Phi} \right\|^2 + \sum_k \lambda_k \|\gamma^k \nu_k\|_1 \quad (5)$$

Let us first note that the above can be expressed slightly differently as follows :

$$\begin{gathered} \min_{\{\alpha_k\}} \tfrac{1}{2\sigma^2} \left\| \mathbf{X} - \textstyle\sum_k \mathbf{X}_k \right\|^2 + \textstyle\sum_k \lambda_k \|\alpha_k\|_1 \\ \text{with } \mathbf{X}_k = \mathbf{\Phi} \alpha_k \mathbf{\Psi} \text{ and } \forall k, \operatorname{rank}(\mathbf{X}_k) \leq 1 \end{gathered} \qquad (6)$$

which uncovers a nice interpretation of our problem as that of approximating the data $\mathbf{X}$ by a sum of rank one matrices $\mathbf{X}_k$ which are sparse in the specified dictionary of rank one matrices. This is the usual ℓ_1 minimization problem [9] but with the additional constraint that the $\mathbf{X}_k$ are all rank one at most. The latter constraint is enforced here mechanically through a proper parametric representation of $\mathbf{X}_k = a^k s_k$ or $\alpha_k = \gamma^k \nu_k$. A similar problem was previously investigated by [10] with a very different approach.

We also note that rescaling the columns of $\nu \leftarrow \rho\nu$ while applying the proper inverse scaling to the rows of $\mathbf{S} \leftarrow 1/\rho \mathbf{S}$, leaves both the quadratic measure of fit and the ℓ_1

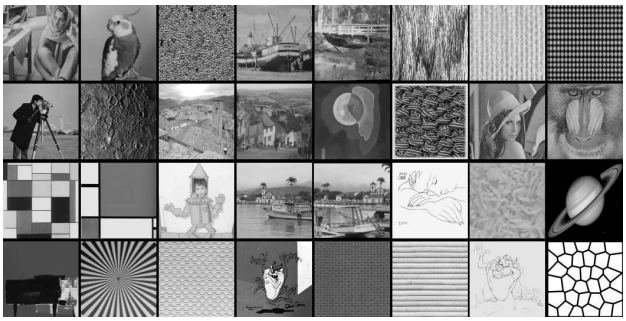

Fig. 1. Image data set used in the experiments.

sparsity measure in equation (5) unaltered. Although renormalizing is still worthwhile numerically, it is no longer dictated by the lack of scale invariance of the objective function and the need to stay away from trivial solutions, as in GMCA.

Finally, adopting a BSS point of view, the objective function (5) is fully symmetric in its treatment of the mixing matrix A and the source processes S. The great majority of BSS methods invoke a uniform *improper* prior distribution for the spectral parameters $\mathbf{A}$. Truly, $\mathbf{A}$ and $\mathbf{S}$ often have different roles in the model and very different sizes. However, dealing with so-called *hyperspectral* data, such an asymmetry is questionable. There have been previous reports of a symmetric treatment of $\mathbf{A}$ and $\mathbf{S}$ for BSS [11, 12] however in the noiseless case. We also note that very recently, the objective function (5) was proposed in [5]. However the algorithm used in [5] is very different from the method proposed here which benefits from all the good properties of GMCA, notably its speed and robustness which come along the iterative thresholding with a decreasing threshold.

3. ALGORITHM

Unfortunately, there is no obvious closed form solutions to optimization problem (5) which is again clearly non-convex. Similarly to the GMCA algorithm, we propose here a numerical approach by means of a block-coordinate relaxation iterative algorithm, alternately minimizing with respect to γ and ν. Indeed, thanks to the chosen prior, for fixed γ (resp. ν), the *marginal* minimization problem over ν (resp. γ) is convex and is readily solved using a variety of methods. Inspired by the iterative thresholding methods described in [13, 14, 15], akin to Projected Landweber algorithms, we obtain the following system of update rules :

$$\left\{ \begin{array}{rcl} \nu^{(+)} & = & \Delta_\eta \left(\nu^{(-)} + \boldsymbol{R}_\nu \left(\alpha - \gamma \nu^{(-)} \right) \right) \\ \gamma^{(+)} & = & \Delta_\zeta \left(\gamma^{(-)} + \left(\alpha - \gamma^{(-)} \nu \right) \boldsymbol{R}_\gamma \right) \end{array} \right. \quad (7)$$

where $\boldsymbol{R}_\nu$ and $\boldsymbol{R}_\gamma$ are appropriate relaxation matrices for the iterations to be non-expansive. Assume left invertibility of $\mathbf{A}$ and right invertibility of $\mathbf{S}$. Then, taking $\boldsymbol{R}_\nu = \left(\gamma^T\gamma\right)^{-1}\gamma^T$ and $\boldsymbol{R}_\gamma = \nu^T\left(\nu\nu^T\right)^{-1}$, the above are rewritten as follows :

$$\nu^{(+)} = \Delta_\eta \left(\left(\gamma^T\gamma\right)^{-1}\gamma^T\alpha \right) \quad (8)$$

$$\gamma^{(+)} = \Delta_\zeta \left(\alpha\nu^T \left(\nu\nu^T\right)^{-1} \right) \quad (9)$$

where vector η has length n and entries $\eta[k] = \lambda_k \|\gamma^k\|_1 / \|\gamma^k\|_2^2$, while ζ has length m and entries $\zeta[k] = \lambda_k \|\nu_k\|_1 / \|\nu_k\|_2^2$. The multichannel soft-thresholding operator Δ_η acts on each row k of ν with threshold $\eta[k]$ and Δ_ζ acts on each column k of γ with threshold $\zeta[k]$. Equations (8) and (9) rules are easily interpreted as thresholded alternate least squares solutions. Finally, in the spirit of the fast GMCA algorithm [2, 1], it is proposed that a solution to problem (5) can be approached efficiently using the following symmetric iterative thresholding scheme with a progressively decreasing threshold, which we refer to as hypGMCA :

1. Set the number of iterations $I_{\max}$ and initial thresholds $\lambda_k^{(0)}$
2. Transform the data $\mathbf{X}$ into α
3. While $\lambda_k^{(h)} \geq \lambda_k^{\min}$,
   - Update ν assuming γ is fixed using eq. (8).
   - Update γ assuming ν is fixed using eq. (9) .
   - Decrease the thresholds $\lambda_k^{(h)}$.
5. Transform back γ and ν to estimate $\mathbf{A}$ and $\mathbf{S}$.

The *salient to fine* estimation process is again the core of hypGMCA. With the threshold successively decaying towards zero along iterations, the current sparse approximations for γ and ν are progressively refined by including finer structures spatially and spectrally, alternatingly. The final threshold should vanish in the *noiseless* case or it may be set to a multiple of the noise standard deviation as in common detection or denoising methods. Soft thresholding results from the use of an ℓ_1 sparsity measure, which comes as an approximation to the ℓ_0 pseudo-norm. Applying a hard threshold instead towards the end of the iterative process, may lead to better results as was noted experimentally in [1, 2]. When non-unitary or redundant transforms are used, the above is no longer strictly valid. Nevertheless, simple shrinkage still gives satisfactory results in practice as studied in [16]. In the end, implementing the proposed update rules requires only a slight modification of the GMCA algorithm given in [1, 2]. Where a simple least squares linear regression was used in the GMCA update for a^k, the proposed update rule applies a thresholding operator to the least squares solution thus enforcing sparsity on the estimated spectral signatures as *a priori* desired.

4. NUMERICAL EXPERIMENTS

4.1. Toy experiment with synthetic data

In this section, we compare the ability of hypGMCA and GMCA to recover and learn the dictionary atoms that were used to synthesize a given data set, generated according to model (2). Using the language of BSS, we consider synthetic 2D data consisting of $m = 128$ mixtures of $n = 5$ image sources. The sources were drawn at random from a set of 128×128 structured images shown on Figure 1. These images provide us with 2D structured processes which are sparse enough in the curvelet domain [17]. The spectra were generated as sparse processes in some orthogonal wavelet domain given *a priori*. The wavelet coefficients of the spectra were sampled from a Laplacian probability density with scale parameter $\mu = 1$. Finally, white Gaussian noise with variance σ^2 was added to the pixels of the synthetic mixture data in the different channels. Figure 2 displays four typical noisy simulated mixture data with SNR $= 20$dB.

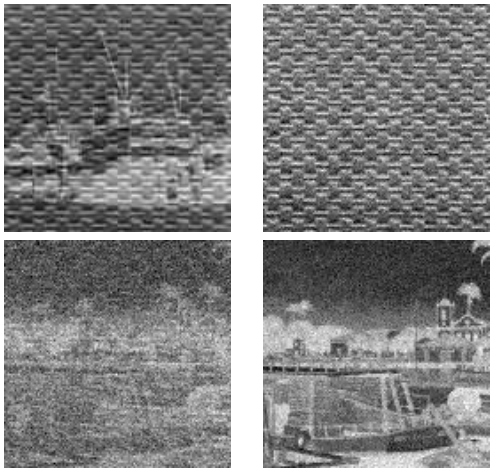

Fig. 2. Four 128×128 mixtures out of the 128 channels. The SNR is equal to 20dB.

The graph on figure 3 traces the evolution of $\mathcal{C}_{\mathbf{A}} = \|\mathbf{I}_n - \mathbf{P}\tilde{\mathbf{A}}^{\dagger}\mathbf{A}\|_1$, which we use to assess the recovery of the spectral dictionary A, as a function of the SNR which was varied from 0 to 40dB. Matrix P serves to reduce the scale and permutation indeterminacy inherent in model (2) and $\tilde{\mathbf{A}}^{\dagger}$ is the pseudo-inverse of the estimated spectral dictionary. In simulation, the true source and spectral matrices are known and so that $\mathbf{P}$ can be computed easily. Criterion $\mathcal{C}_{\mathbf{A}}$ is then strictly positive and null only if matrix A is correctly estimated up to scale and permutation. Finally, as we expected since it benefits from the added *a priori* spectral sparsity constraint it enforces, the proposed hypGMCA is clearly more robust to noise. A visual inspection of figure 4 allows

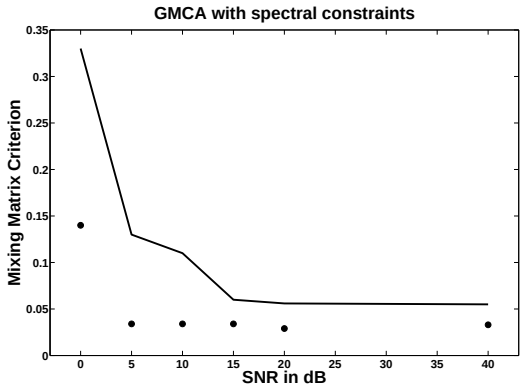

Fig. 3. **Evolution of the mixing matrix criterion $\mathcal{C}_{\mathbf{A}}$ as a function of the SNR in dB.** *Solid line :* recovery results with GMCA. • *:* recovery results with hypGMCA.

a further qualitative assessment of the improved dictionary and source recovery provided by correctly accounting for *a priori* spatial as well as spectral sparsity. The top images were obtained with GMCA while the bottom images, were obtained with hypGMCA. In all cases, both methods were run in the curvelet domain [17] with the same number of iterations.

4.2. Application to real data

We applied the proposed dictionary learning algorithm to hyperspectral data from the 128 channels of spectrometer OMEGA on Mars Express (*www.esa.int/marsexpress*), at wavelengths ranging from $0.93\mu m$ to $2.73\mu m$. Example maps collected in four different channels are shown on figure 5. Model 2 is clearly too simple to describe this hyperspectral reflectance data set. Non-linear instrumental and atmospheric effects are likely to contribute to the *true* generative process. In any case, we use hypGMCA to learn a dictionary of spectral atoms, sparse in some orthogonal wavelet dictionary, that well represent the data. The spectral dictionary was learned assuming it is sparse in an orthogonal wavelet representation. Preliminary results are shown on figure 6. Interestingly, exactly two atoms were found to be respectively strongly correlated with reference spectra for H_2O and CO_2 ice. Given the simplicity of the model, the close match between the learned spectra and the reference spectra is is remarkable. Figure 7 shows the corresponding spatial maps which were assumed sparse in an orthogonal wavelet basis. We note that the CO_2 ice appears spatially concentrated at the poles which is in close agreement with the results presented in [18].

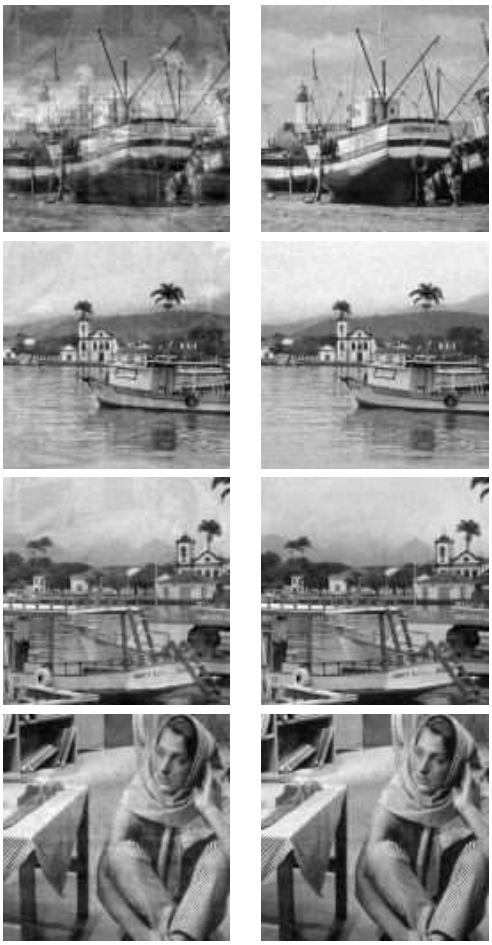

Fig. 4. **Left column :** Estimated sources using the original GMCA algorithm. **Right column :** Estimated sources using the new hypGMCA.

5. CONCLUSION

We described a new dictionary learning algorithm, hypGMCA, for sparse signal representation in the case where it is known *a priori* that the spatial and spectral features in the data have sparse representations in known dictionaries of template waveforms. The proposed method relies on an iterative thresholding procedure with a progressively decreasing threshold. This alone gives the method true robustness to noise. As expected, taking into account the additional prior knowledge of spectral sparsity leads to enhanced performance. Numerical experiments focus on a comparison between GMCA and hypGMCA. GMCA is compared to state-of-the-art dictionary learning and BSS algorithms in [2, 1].

Acknowledgments : The authors are grateful to Olivier Forni for providing the hyperspectral data from Omega on Mars Express.

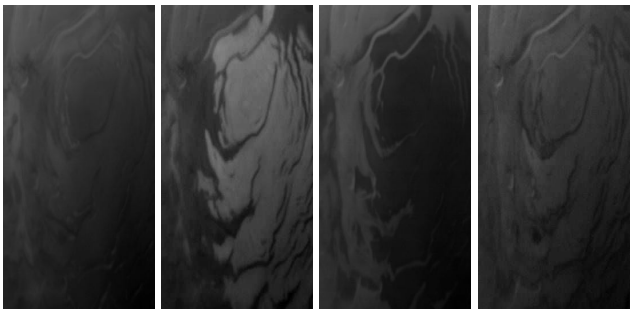

Fig. 5. **From left to right :** Mars Express observations at wavelengths = 1.38 - 1.75 - 1.94 and $2.41\mu m$.

6. REFERENCES

[1] J. Bobin, Y. Moudden, M. J. Fadili, and J.-L. Starck, "Morphological diversity and sparsity for multichannel data restoration," *Journal of Mathematical Imaging and Vision*, vol. 33, no. 2, pp. 149–168, 2008.

[2] J. Bobin, J.-L. Starck, Y. Moudden, and J. Fadili, *Blind Source Separation: the Sparsity Revolution*, vol. 152 of *Advances in Imaging and Electron Physics*, pp. 221–298, 2008.

[3] M. Elad, J.-L. Starck, D. L. Donoho, and P. Querre, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)," *ACHA*, vol. 19, no. 3, pp. 340–358, 2005.

[4] R. Gribonval and K. Schnass, "Dictionary Identifiability from Few Training Samples," in *EUSIPCO*, 2008.

[5] R. Rubinstein, M. Zibulevsky, and M. Elad, "Learning sparse dictionaries for sparse signal representations," *IEEE Transactions on signal processing*, 2008, submitted.

[6] B. A. Olshausen and D. J. Field, "Sparse coding of natural images produces localized, oriented, bandpass receptive fields," Tech. Rep. CCN-110-95, Department of Psychology, Cornell University, 1995.

[7] J.-F. Cardoso and D. L. Donoho, "Some experiments on independent component analysis of non-gaussian processes," in *Proc. IEEE SP Int. Workshop HOS '99*, 1999, pp. 74–77.

[8] A. Hyvärinen, P. Hoyer, and E. Oja, "Sparse code shrinkage: denoising by nonlinear maximum likelihood estimation," in *Advances in Neural Information Processing Systems 11 (NIPS*98*. 1999, pp. 473–479, MIT Press.

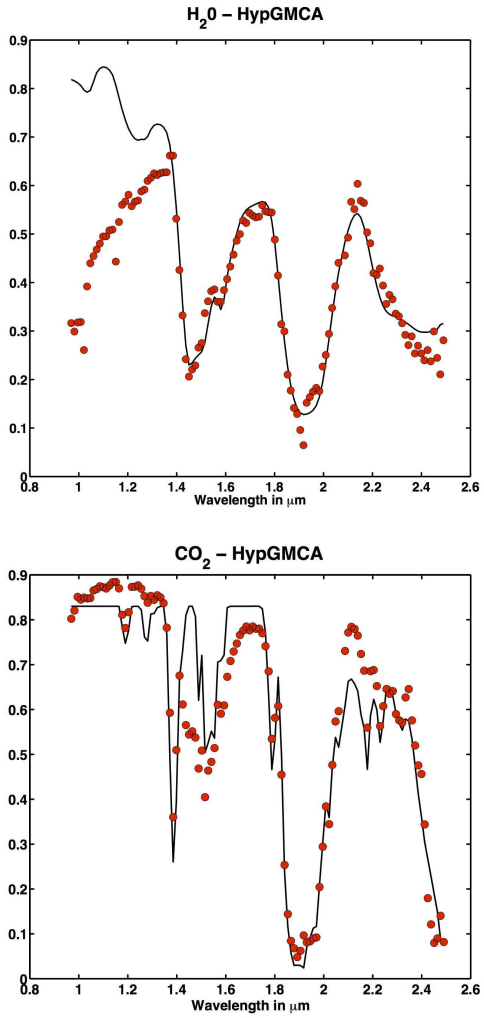

Fig. 6. **Left picture :** Reference (solid line) and estimated (•) spectra for H_2O ice. **Right picture :** Reference (solid line) and estimated (•) spectra for CO_2 ice.

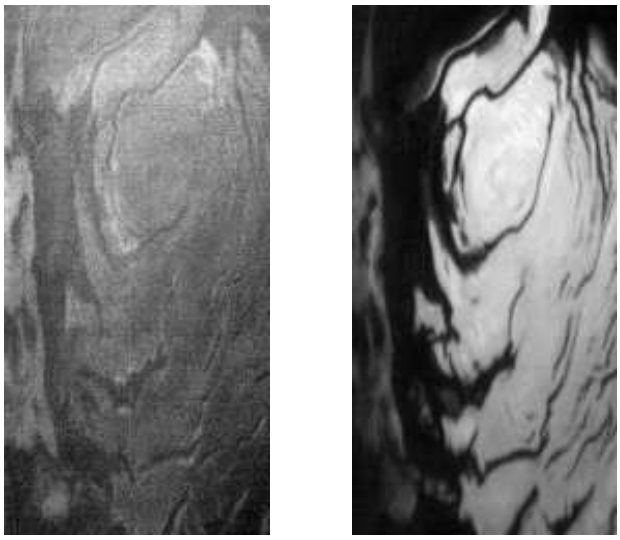

Fig. 7. Estimated spatial concentration maps of H_2O (**left**) and CO_2 (**right**) ice.

[9] D. L. Donoho and M. Elad, "Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Aca. Sci.*, vol. 100, pp. 2197–2202, 2003.

[10] Z. Zhang, H. Zha, and H. Simon, "Low-rank approximations with sparse factors I: Basic algorithms and error analysis," *SIAM Journal on Matrix Analysis and Applications*, vol. 23, no. 3, pp. 706–727, 2002.

[11] J. V. Stone, J. Porrill, N. R. Porter, and I. W. Wilkinson, "Spatiotemporal independent component analysis of event-related fmri data using skewed probability density functions," *NeuroImage*, vol. 15, no. 2, pp. 407–421, 2002.

[12] A. Hyvärinen and R. Karthikesh, "Imposing sparsity on the mixing matrix in independent component analysis," *Neurocomputing*, vol. 49, pp. 151–162, 2002.

[13] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, Aug 2004.

[14] Elaine T. Hale, Wotao Yin, and Yin Zhang, "A fixed-point continuation method for l1 -regularized minimization with applications to compressed sensing," Tech. Rep., Rice University, July 2007.

[15] M. A. Figueiredo, R. Nowak, and S.J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing - To appear*, 2007.

[16] M. Elad, "Why simple shrinkage is still relevant for redundant representations?," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5559–5569, 2006.

[17] J.-L. Starck, E. J. Candès, and D. L. Donoho, "The curvelet transform for image denoising.," *IEEE Transactions on Image Processing*, vol. 11, no. 6, pp. 670–684, 2002.

[18] S. Moussaoui, H. Hauksdottir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, S. Douté, and J.A. Benediktsson, "On the decomposition of mars hyperspectral data by ica and bayesian positive source separation," *Neurocomputing*, vol. in press, 2008.