Structured Sparsity: from Mixed Norms to Structured Shrinkage

Matthieu Kowalski, Bruno Torrésani

Structured Sparsity: from Mixed Norms to Structured Shrinkage

M. Kowalski and B. Torrésani
LATP, CMI, 39 rue Joliot-Curie, 13453 Marseille Cedex 13, France.
{kowalski,bruno.torresani}@cmi.univ-mrs.fr

***Abstract*—Sparse and structured signal expansions on dictionaries can be obtained through explicit modeling in the coefficient domain. The originality of the present contribution lies in the construction and the study of generalized shrinkage operators, whose goal is to identify structured significance maps. These generalize Group LASSO and the previously introduced Elitist LASSO by introducing more flexibility in the coefficient domain modeling. We study experimentally the performances of corresponding shrinkage operators in terms of significance map estimation in the orthogonal basis case. We also study their performance in the overcomplete situation, using iterative thresholding.**

I. Introduction

Standard variational approaches to the sparse regression problem, such as LASSO/Basis Pursuit Denoising, generally rely on the implicit assumption that the regression coefficients are *i.i.d.*. Departure from independence can also be implemented into the problem by suitable choices of the regularization term. Examples of these are given by the *joint sparsity* approaches [1], the *Elastic net* [2] type methods, Group LASSO [3] or Elitist LASSO [4].

We study in this paper some generalizations of the mixed-norm based dependent coefficient sparse regression methods, designed to overcome some limitations of some of G-LASSO and E-LASSO.

In the latter, described in Section II (together with the corresponding generalized shrinkage operators), dependences are introduced through a prior rigid hierarchy (defining groups and members). Variants were proposed in [4] that avoid such a rigidity and allow for adaptive groupings of coefficients. After recalling in Section III the construction of these variants (termed Windowed Group LASSO, and persistent Elitist LASSO), in a somewhat simplified way, we describe corresponding iterative shrinkage algorithms. We describe a numerical simulation setup designed to study experimentally their performances in regression problems. Numerical results, displayed and discussed in Section IV, show that even though the proposed methods do not compare very well with more classical LASSO in terms of SNR, they drastically outperform the latter in terms of significance map estimation.

II. Structured regression using mixed norms

We start here from the popular regression problem known as the Basis Pursuit Denoising (BPDN for short) [5], or LASSO [6]. Given an observation $y \in \mathbb{C}^M$, BPDN aims to identify a sparse expansion into an overcomplete dictionary $\{\varphi_k\}_{k=1}^N$ (with $M \leq N$), by minimizing the functional

$$\Psi(x) = \left\| y - \sum_k x_k \varphi_k \right\|_2^2 + \lambda \|x\|_1 \ . \tag{1}$$

The ℓ_1 penalty on the synthesis coefficients x_k yields sparsity. In a Bayesian interpretation, this corresponds to a prior with *i.i.d.* synthesis coefficients, distributed according to a Laplacian distribution. This remark is well illustrated when the dictionary is an orthonormal basis: denoting by $\underline{y}_k = \langle y, \varphi_k \rangle$ the analysis coefficients, the minimizer of the functional is obtained through the well known soft-thresholding operator, in which all analysis coefficients $\underline{y}_k$ are compared to a global threshold λ:

$$x_k = \arg(\underline{y}_k) \left(|\underline{y}_k| - \lambda \right)^+ \ . \tag{2}$$

In many cases, the synthesis coefficients can be labelled with two indices. Various example of such two levels indexing can be naturally found in signal processing, for example in time-frequency or time-scale decomposition, or multichannel signals. Using these two indices, a hierarchy may be introduced in the synthesis coefficients, in such a way that they can be gathered into groups. Such a hierarchy can be exploited by a mixed norm penalty [7], [8], the functional to minimize taking the form:

$$\begin{aligned} \Psi(x) &= \left\| y - \sum_g \sum_m x_{g,m} \varphi_{g,m} \right\|_2^2 + \lambda \sum_g \left(\sum_m |x_{g,m}|^p \right)^{q/p} \\ &= \left\| y - \sum_g \sum_m x_{g,m} \varphi_{g,m} \right\|_2^2 + \lambda \|x\|_{p,q}^q \ . \end{aligned} \tag{3}$$

The index g is the "group" index, and m is the "member" index. The relationship between a hierarchical model and mixed norms is well illustrated in [9].

For simplicity, let us shall limit ourselves to the cases $p, q \in \{1, 2\}$. When the atoms $\varphi_{g,m}$ are orthonormal, the minimizer of the functional is obtained through "generalized" soft thresholding operators. Before recalling them, let us remark that generalized soft thresholding (with threshold $\tau_{g,m}$) can

also be written in the general form (supposing that $\underline{y}_{g,m} \neq 0$)

$$\begin{aligned} x_{g,m} &= \arg(\underline{y}_{g,m}) \left(|\underline{y}_{g,m}| - \tau_{g,m}\right)^+ \\ &= \underline{y}_{g,m} \left(1 - \frac{\tau_{g,m}}{|\underline{y}_{g,m}|}\right)^+ \\ &= \underline{y}_{g,m} \left(1 - \nu_{g,m}(\underline{y})\right)^+ . \end{aligned} \quad (4)$$

In the following, we will use the notation $\underline{y}_g = \{\underline{y}_{g,1}, \ldots, \underline{y}_{g,m}, \ldots\}$ to represent the vector of coefficients that belong to the group g. If the φ_k are orthonormal, the minimizers of Ψ are given by the following shrinkage coefficients $\nu_{g,m}$

- if $p = 2$ and $q = 1$,

$$\nu_{g,m}(\underline{y}) = \frac{\lambda}{\|\underline{y}_g\|_2} . \quad (5)$$

- If $p = 1$ and $q = 2$, denoting by $\{\check{\underline{y}}_{g,m}\}$ the coefficients $|\underline{y}_{g,m}|$ ordered by descending order within each group g, and choosing M_g so that

$$\check{\underline{y}}_{g,M_g+1} \leq \lambda \sum_{m=1}^{M_g+1} (\check{\underline{y}}_{g,m} - \check{\underline{y}}_{g,M_g+1})$$

and

$$\check{\underline{y}}_{g,M_g} > \lambda \sum_{m=1}^{M_g(\lambda)} (\check{\underline{y}}_{g,m} - \check{\underline{y}}_{g,M_g}) ,$$

the shrinkage coefficient is given by

$$\nu_{g,m}(\underline{y}) = \frac{\lambda}{1 + M_g\lambda} \frac{\|\check{\underline{y}}_{g,m:M_g}\|_1}{|\underline{y}_{g,m}|} . \quad (6)$$

Actually, these shrinkage operators correspond to the proximity operators, used by Combettes *et al.* [10], associated with the mixed norms under consideration.

The regression problem with $\ell_{2,1}$ penalty is known as the Group-LASSO (G-LASSO) regression, or regression with multiple measurement vectors. Observing the shrinkage coefficient in (5), one can see that the selected groups are the ones with the biggest ℓ_2 norms. Then, when using this penalty, one keeps entire groups of coefficients, namely the most energetic groups. At the opposite, operator (6) selects coefficients with the biggest modulus within each group. Therefore, the regression based upon the $\ell_{1,2}$ penalty was called the Elitist-LASSO (E-LASSO) regression to illustrate the fact that only the "best" coefficients are chosen for each group.

These operators are used to minimise the convex, but non-differentiable, functional (3), when the atoms $\varphi_{g,m}$ form an overcomplete dictionary, inside iterative thresholding algorithms such as the thresholded Landweber iteration [11] or proximal algorithms [10].

III. Selection by shrinkage operators

The mixed norms allow one to introduce structure (organized in terms of groups and members) in regression problems, however groups are defined once for all. One main shortcoming is the independence of the groups: a given coefficient cannot belong to two different groups. However, in some tasks, one cannot design such independent groups, but would like to gather a coefficient and its neighborhood. Conversely, even when independent groups can be constructed, one may want to introduce persistence accross the groups. Motivated by these remarks, we present generalizations of the two above defined operators (G-LASSO and E-LASSO). We introduce two new shrinkage operators constructed directly on the analysis coefficients with respect to an orthonormal basis. The shrinkage will be defined by the corresponding $\nu_{g,m}(\underline{y})$ in equation (4).

A. Windowed Group-LASSO

In G-LASSO the hierarchy is fixed once for all, and a given member cannot belong to several groups. To relax this constraint we use here a single indice k to label the analysis coefficients $\underline{y}_k$. Then we associate with any index k a family of neighborhood indices $g(k) = \{m \ ; \ m \in \text{ Neighborood of } k\}$, and use the G-LASSO shrinkage (5). While the neighborhood system still has to be defined in advance, a given index can now belong to the neighborhood of several other indices. Figure 1 gives an illustration of such a neighborhood.

The corresponding shrinkage coefficient (WG-LASSO) takes the form

$$\nu_k(\underline{y}) = \frac{\lambda}{\sqrt{\sum_{m \in g(k)} |y_m|^2}} = \frac{\lambda}{\|\underline{y}_{g(k)}\|_2} . \quad (7)$$

When the neighborhoods are disjoint (ie. there is no overlap between the groups), WG-LASSO reduces to G-LASSO.

We stress that in the iteration, the shrinkage operates *only* on the original analysis coefficients $\{\underline{y}_k\}$ to give the corresponding synthesis coefficients $\{x_k\}$.

B. Persistent Elitist-LASSO

The Persistent Elitist-LASSO (PE-LASSO) is constructed in order to introduce persistence between groups inside E-LASSO. In this case, neighborhood systems have to be defined as in section II. Then, for a given group g, we denote by $\mathcal{N}(g)$ the index set of the groups which are defined as close to this group g. Then, we define the following coefficients

$$z_{g,m} = \left(\sum_{g' \in \mathcal{N}(g)} |\underline{y}_{g',m}|^2 \right)^{1/2} .$$

For each member of a group, we consider the energy of this member and its neighbors in the close groups $\mathcal{N}(g)$. Such a coupling is illustrated on Figure 2. Then, the PE-LASSO shrinkage is defined by

$$\nu_{g,m}(\underline{y}) = \frac{\lambda}{1 + \lambda M_g} \frac{\|\check{z}_{g,1:M_g}\|_1}{z_{g,m}} , \quad (8)$$

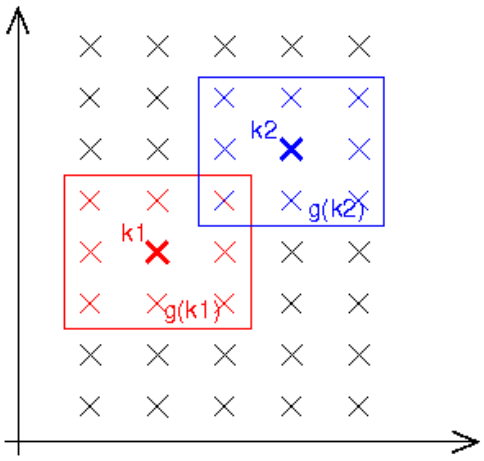

Fig. 1. Windowed Group-LASSO: two overlaping groups. The neighborhood of the coefficient k_1 is given by the red window, and the neighborhood of the coefficient k_2 by the blue one. These two neighbordhoods share one coefficient.

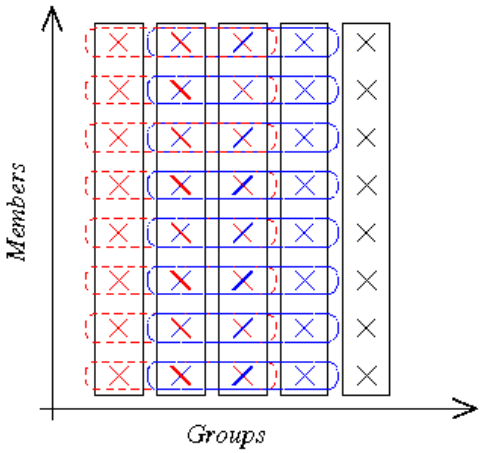

Fig. 2. Persistent Elitist-LASSO: persitence is introduced between the groups. The groups are defined by the black rectangle. Then, for each member of a group, we are considering the left neighbor ant the right neighbor. We compute then the energy of each red groups, and we apply a "E-LASSO like" shrinkage on these coefficients. We do the same for the next coefficients ans its neighborhood defined by blue groups.

with M_g defined by

$$\check{z}_{g,M_g+1} \leq \lambda \sum_{m=1}^{M_g+1} (\check{z}_{g,m} - \check{z}_{g,M_g+1})$$

and

$$\check{z}_{g,M_g} > \lambda \sum_{m=1}^{M_g} (\check{z}_{g,m} - \check{z}_{g,M_g}) \ ,$$

the $\{\check{z}_{g,m}\}$ being the coefficients $z_{g,m}$ ordered by descending order within each group g.

Here again, if the groups are independent, the PE-LASSO shrinkage coincides with the E-LASSO shrinkage. Moreover, this operator is also a generalization of WG-LASSO: if the neighborhood are only between the groups for a given member, then $M_g = 0$ and by setting $\|\check{z}_{g,1:M_g}\|_1 = 1$, the two shrinkage operators are the same. We stress again that, as for the WG-LASSO, the shrinkages operate only on the analysis coefficients.

Remark 1: The shrinkage (8) is in fact related to the proximity operator of the mixed norm defined on triply labelled coefficients $x_{i,j,k}$ (i.e. a two levels hierarchy – see [12]):

$$\|x\|_{2,1,2}^2 = \sum_i \left(\sum_j \sqrt{\sum_k |x_{i,j,k}|^2} \right)^2 \ ,$$

and can be viewed as a version of this operator with overlaped groups.

C. Algorithms and simulations

The above defined shrinkage operators have been used heuristically in iterative thresholding strategies. It is worth noticing that unlike G-LASSO and E-LASSO, for which convergence to a fixed point can be proven, the shrinkage operations in (7) and (8) are not associated with a simple (convex) functional such as (3).

Denoting by $\underline{y} \mapsto \mathbb{S}(\underline{y})$ the shrinkage operator defined coordinatewise in the previous section, the iterative algorithm we use is the following

Algorithm 1:

- Let $x^{(0)} = \mathbf{0}$, $\gamma < \frac{1}{\|\Phi^*\Phi\|}$ and $t_{max} \in \mathbb{N}$.
- **For** $t = 0$ to t_{max}

$$x^{(t+1/2)} = x^{(t)} + \gamma \Phi^*(y - \Phi x^{(t)})$$
$$x^{(t+1)} = \mathbb{S}(x^{(t+1/2)})$$

End For

One can remark that while the shrinkage operator $\mathbb{S}$ is the (proximity) operator (5) or (6) defined in section II, algorithm 1 is the iterative thresholded Landweber algorithm [11] applied to mixed norms [8], or proximal algorithm [10], and converges to a solution of (3). However, with the shrinkage operators (7) and (8), nothing is known about the convergence of this algorithm: neither its convergence, nor its limit when it converges. But, as we will see in section IV-B, the numerical results we obtain are nevertheless quite interesting.

In the numerical experiments presented below, we use a time-frequency dictionary $\{\varphi_{(t,f)}\}$, we consider signals of the form

$$y = \sum_{(t,f)\in\Delta} x_{(t,f)}\varphi_{(t,f)} + b \ ,$$

where b is an additive Gaussian noise, and Δ is the (structured sparse) significance map.

The latter is generated using fixed frequency Markov chains as introduced in [13], and the synthesis coefficients $x_{(t,f)}, (t,f) \in \Delta$ are generated from a standard normal distribution. An example of such a map is displayed in Figure 3. b is a white Gaussian noise (tuned so as to obtain a signal to noise ratio of about $5-6$ dB).

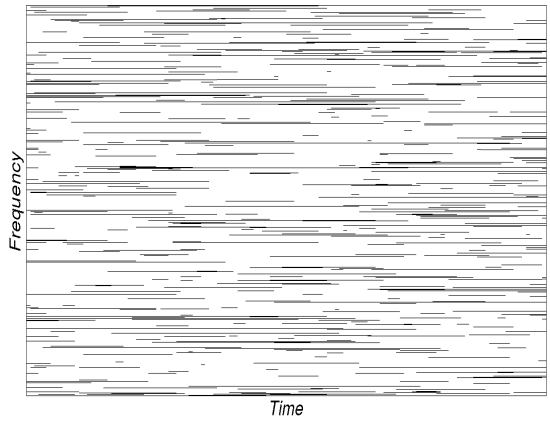

Fig. 3. A structured sparse significance map generated with fixed frequency Markov chains.

Used as such, the shrinkage operators introduced above turn out to perform quite poorly for denoising task compared to usual soft thresholding, except in some very specific situations. However, they also turn out to outperform very significantly standard soft thresholding if one limits oneself to another problem, namely the significance map estimation. We analyze below the results for this problem in terms of type one and two errors:

Type 1: $\pi_1 = \mathbb{P}\{(t,f) \notin \hat{\Delta} \mid (t,f) \in \Delta\}$;
Type 2: $\pi_2 = \mathbb{P}\{(t,f) \in \hat{\Delta} \mid (t,f) \notin \Delta\}$.

with Δ the true significance map, and $\hat{\Delta}$ the estimated one.

IV. Experimental Results

A. The orthogonal basis case

First, assume that $\{\varphi_{(t,f)}\}$ is an orthogonal basis such as a MDCT (Modified Discret Cosine Transform) basis. The group index is defined as the frequency index, and the member index is therefore the time index. We compared systematically all the previous shrinkage operators, with various values of the parameter λ (the bigger of λ, the sparser the $\hat{\Delta}$). The time-persistent groups were constructed as follows: $\mathcal{N}(g) = \{g-p, g-p+1, \ldots, g, g+1, \ldots, g+p\}$ with $p = 1$ or 2. We used the same neighborhood system for WG-LASSO. The results are shown in figure 4, where type one and two errors are represented as functions of the estimated significance map size. PE-LASSO and WG-LASSO obviously outperform LASSO and E-LASSO, in particular for type 1 error, but also for type 2 error. The PE-LASSO performs better than WG-LASSO. In addition, results are better for long persistence ($p = 2$) than short persistence ($p = 1$) as could be expected. All this clearly shows that taking into account explicitely the persistence properties of significance maps improves significantly their estimation.

B. The frame case

We also used the shrinkage operators inside algorithm 1. In this experiment, the dictionary $\{\varphi_{(t,f)}\}$ is a Gabor frame, and

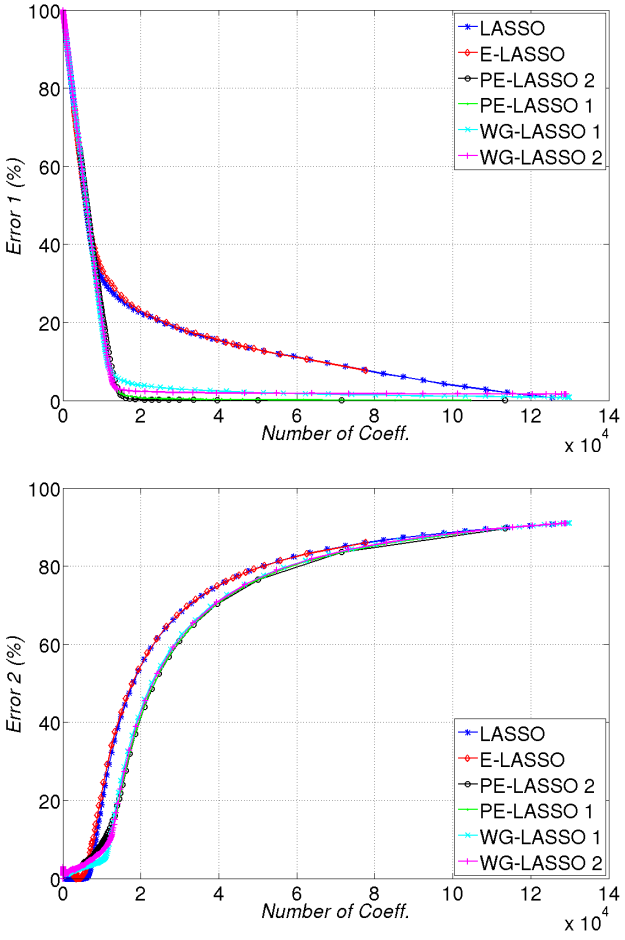

Fig. 4. Top: type 1 error. Bottom: type 2 error.

the significance map is generated as before. Results shown in Figure 5 are qualitatively similar to the previous ones. Regarding the convergence of the algorithm, we observed that the thresholded Landweber iteration seems to "oscillate" after a number of iterations for both PE-LASSO and WG-LASSO. However, this is not sufficient to undermine once for all the convergence of algorithm 1. However, both outperform significantly simple shrinkage strategies.

V. Conclusions, perspectives

The iterative shrinkage schemes described in this paper provide a flexible way for performing sparse approximation while maintaining some dependence relationships between the regression coefficients.

Numerical results show that such approaches perform extremely well (in comparison with more standard sparse regression) if one limits oneself to the estimation of significance maps, i.e. the locations of significant coefficients. This suggests to develop a different, "two-stage" regression approach, in which significance map estimation would be followed by a more standard regression, limited to the identified atoms. This new approach will be discussed in a forthcoming contribution.

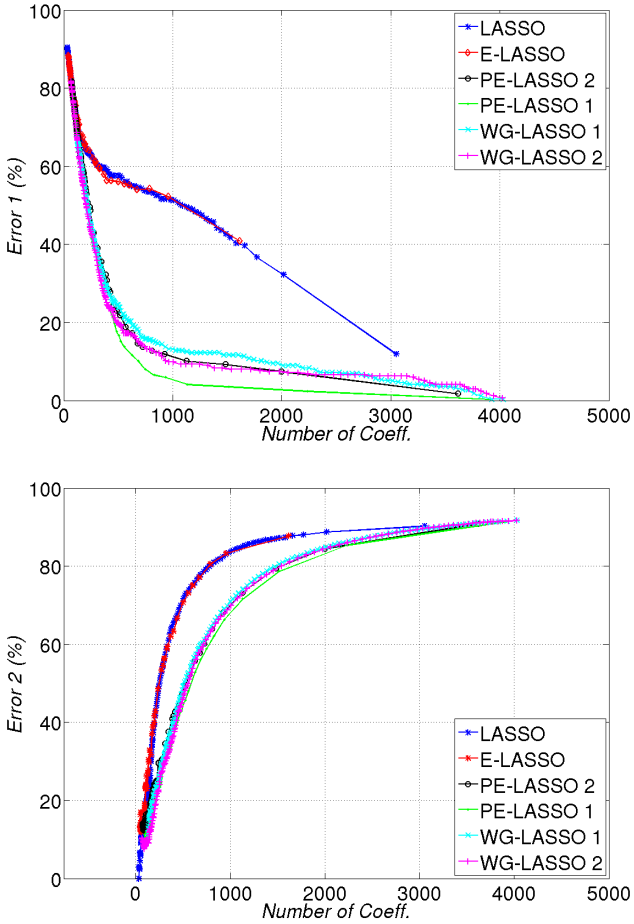

Fig. 5. Top: type 1 error. Bottom: type 2 error.

The convergence of algorithm 1 with shrinkage operations such as the WG-LASSO or PE-LASSO ones should also be studied. The difficulty comes from that these shrinkages are not the solution of the variational equation associated with the minimization of a simple functional.

REFERENCES

[1] M. Fornasier and H. Rauhut, "Recovery algorithm for vector-valued data with joint sparsity constraints," *SIAM Journal on Numerical Analysis*, vol. 46, no. 2, pp. 577–613, 2008.

[2] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of Royal Statistical Society Serie B*, vol. 67, no. 2, pp. 301–320, 2005.

[3] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society Serie B*, vol. 68, no. 1, pp. 49–67, 2006.

[4] M. Kowalski and B. Torrésani, "Random models for sparse signals expansion on unions of basis with application to audio signals," *IEEE Transaction On Signal Processing*, vol. 56, no. 8, pp. 3468–3481, Aug. 2008.

[5] S. S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Serie B*, vol. 58, no. 1, pp. 267–288, 1996.

[7] M. Kowalski, "Sparse regression using mixed norms," 2008. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00202904/

[8] M. Kowalski and B. Torrésani, "Sparsity and persistence: mixed norms provide simple signals models with dependent coefficients," *Signal, Image and Video Processing*, 2008, doi:10.1007/s11760-008-0076-1.

[9] M. Szafranski, Y. Grandvalet, and P. Morizet-Mahoudeaux, "Hierarchical penalization," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008.

[10] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200, Nov. 2005.

[11] 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, August 2004.

[12] A. Gramfort and M. Kowalski, "Improving m/eeg source localization with an inter-condition sparse prior," 2009, submitted.

[13] S. Molla and B. Torrésani, "An hybrid audio scheme using hidden Markov models of waveforms," *Applied and Computational Harmonic Analysis*, vol. 18, no. 2, pp. 137–166, 2005.