Simultaneous sparse approximation : insights and algorithms

Alain Rakotomamonjy

Simultaneous Sparse Approximation : insights and algorithms

Alain Rakotomamonjy

Abstract

This paper addresses the problem of simultaneous sparse approximation of signals given an overcomplete dictionary of elementary functions. At first, we propose a simple algorithm for solving the multiple signals extension of the Basis Pursuit Denoising problem. Then, we consider the M-FOCUSS problem which performs sparse approximation by using non-convex sparsity-inducing penalties and show that M-FOCUSS is actually equivalent to an automatic relevance determination problem. Based on this novel insight, we introduce an iterative reweighted Multiple-Basis Pursuit for solving M-FOCUSS; we trade the non-convexity of M-FOCUSS against several resolutions of the convex M-BP problem. Relations between our reweighted algorithm and the Multiple-Sparse Bayesian Learning are also highlighted. Experimental results show how our algorithms behave and how they compare to previous approaches for solving simultaneous sparse approximation problem.

EDICS: DSP-TFSR, MLR-LEAR

I. Introduction

Since several years now, there has been a lot of interest about sparse signal approximation. This large interest comes from frequent wishes of practitioners to represent data in the most parsimonious way. According to this objective, in signal analysis, one usually wants to approximate a signal by using a linear combination of elementary functions called a dictionary. Mathematically, such a problem can be formulated as the following optimization problem

$$\min_{\mathbf{c}} \|\mathbf{c}\|_0 \quad \text{st} \;\; \mathbf{s} = \mathbf{\Phi c}$$

where $\mathbf{s} \in \mathbb{R}^N$ is the signal vector to be approximated, $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$ is a matrix of unit-norm elementary functions, $\mathbf{c}$ a weight vector and $\|\cdot\|_0$ the ℓ_0 pseudo-norm that counts the number of non-zero components

A. Rakotomamonjy is with the LITIS EA4108, University of Rouen, France.

in its vector parameter. Solving this problem of finding the sparsest approximation over a dictionary $\mathbf{\Phi}$ is a hard problem, and it is usual to relax the problem in order to make it more tractable. For instance, Chen et al. [6] have posed the problem as a convex optimization problem by replacing the ℓ_0 pseudo-norm with a ℓ_1 norm and proposed the so-called Basis Pursuit algorithm. Greedy algorithms are also available for solving this sparse approximation problem [22], [31]. Such a family of algorithms known as Matching Pursuit is simply based on iterative selection of dictionary elements. Although the original sparse approximation problem has been relaxed, both Basis Pursuit and Matching Pursuit algorithms can be provided with some conditions whereby they are guaranteed to produce the sparsest approximation of the signal vector [9], [30].

A natural extension of sparse approximation problem is the problem of finding jointly sparse representations of multiple signal vectors. This problem is also known as simultaneous sparse approximation and it can be stated as follows. Suppose we have several signals describing the same phenomenon, and each signal is contaminated by noise. We want to find the sparsest approximation of each signal by using the same set of elementary functions. Hence, the problem consists in finding the best approximation of each signal while controlling the number of functions involved in all the approximations. Such a situation arises in many different application domains such as sensor networks signal processing [20], neuroelectromagnetic imaging [14], [24] and source localization [21].

A. Problem formalization

Formally the problem of simultaneous sparse approximation can be stated as follows. Suppose that we have measured L signals $\{\mathbf{s}_i\}_{i=1}^{L}$ where each signal is of the form

$$\mathbf{s}_i = \mathbf{\Phi}\mathbf{c}_i + \epsilon$$

where $\mathbf{s}_i \in \mathbb{R}^N$, $\mathbf{\Phi} \in \mathbb{R}^{N\times M}$ is a matrix of unit-norm elementary functions, $\mathbf{c}_i \in \mathbb{R}^M$ a weighting vector and ϵ is a noise vector. $\mathbf{\Phi}$ will be denoted in the sequel as the dictionary matrix. Since we have several signals, the overall measurements can be written as

$$\mathbf{S} = \mathbf{\Phi}\mathbf{C} + \mathcal{E} \tag{1}$$

with $\mathbf{S} = [\mathbf{s}_1\ \mathbf{s}_2\ \cdots\ \mathbf{s}_L]$ a signal matrix, $\mathbf{C} = [\mathbf{c}_1\ \mathbf{c}_2\ \cdots\ \mathbf{c}_L]$ and $\mathcal{E}$ a noise matrix. Note that in the sequel, we have adopted the following notations. $c_{i,\cdot}$ and $c_{\cdot,j}$ respectively denote the ith row and jth column of matrix $\mathbf{C}$. $c_{i,j}$ is the ith element in the jth column of $\mathbf{C}$.

For the sparse simultaneous approximation problem, the goal is then to recover the matrix $\mathbf{C}$ given the signal matrix $\mathbf{S}$ and the dictionary $\mathbf{\Phi}$ under the hypothesis that all signals $\mathbf{s}_i$ share the same sparsity profile. This latter hypothesis can also be translated into the coefficient matrix $\mathbf{C}$ having a minimal number of non-zero rows. In order to measure the number of non-zero rows of $\mathbf{C}$, a frequent measure is the so-called *row-support* or *row diversity measure* of a coefficient matrix defined as

$$\text{rowsupp}(\mathbf{C}) = \{i \in [1 \cdots M] : c_{i,k} \neq 0 \text{ for some } k\}$$

The row-support of $\mathbf{C}$ tells us which atoms of the dictionary have been used for building the signal matrix. Hence, if the cardinality of the row-support is lower than the dictionary cardinality, it means that at least one atom of the dictionary has not been used for synthetizing the signal matrix. Then, the row-ℓ_0 pseudo-norm of a coefficient matrix can be defined as

$$\|\mathbf{C}\|_{row-0} = |\text{rowsupp}(\mathbf{C})|$$

According to this definition, the sparse simultaneous approximation problem can be stated as

$$\begin{array}{ll} \min_{\mathbf{C}} & \frac{1}{2}\|\mathbf{S} - \mathbf{\Phi}\mathbf{C}\|_F^2 \\ \text{st.} & \|\mathbf{C}\|_{row-0} \leq T \end{array} \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm and T a user-defined parameter that controls the sparsity of the solution. Note that the problem can also take a different form

$$\begin{array}{ll} \min_{\mathbf{C}} & \|\mathbf{C}\|_{row-0} \\ \text{st.} & \frac{1}{2}\|\mathbf{S} - \mathbf{\Phi}\mathbf{C}\|_F \leq \epsilon \end{array} \tag{3}$$

For this latter formulation, the problem translates in minimizing the number of non-zero rows in the coefficient matrix $\mathbf{C}$ while keeping control on the approximation error. Both problems (2) and (3) are appealing for their formulation clarity. However, similarly to the single signal approximation case, solving these optimization problems are notably intractable because $\|\cdot\|_{row-0}$ is a discrete-valued function. Hence, some relaxed versions of these problems have been proposed in the literature.

B. Related works

Two ways of relaxing problems (2) and (3) are possible : by replacing the $\|\cdot\|_{row-0}$ function with a more tractable row-diversity measure or by using some suboptimal algorithms. We details these two approaches in the sequel.

A large class of relaxed versions of $\|\cdot\|_{row-0}$ proposed in the literature are emcompassed into the following form

$$J_{p,q}(\mathbf{C}) = \sum_i \|c_{i,\cdot}\|_q^p$$

where typically $p \leq 1$ and $q \geq 1$. This novel penalty term can be interpreted as the ℓ_p quasi-norm of the sequence $\{\|c_{i,\cdot}\|_q\}_i$. Note that as p converges to 0, $J_{p,q}(\mathbf{C})$ provably converges towards $\sum_i \log(\|c_{i,\cdot}\|)$. According to this relaxed version of the row diversity measure, most of the algorithms proposed in the literature try to solve the relaxed problem

$$\min_{\mathbf{C}} \frac{1}{2}\|\mathbf{S} - \mathbf{\Phi C}\|_F^2 + \lambda J_{p,q}(\mathbf{C}) \tag{4}$$

where λ is another user-defined parameter that balances the approximation error and the sparsity-inducing penalty $J_{p,q}(\mathbf{C})$. The choice of p and q results in a compromise between the row-support sparsity and the convexity of the optimization problem. Indeed, problem (4) is known to be convex when $p, q \geq 1$ while it is known to produce a row-sparse matrix $\mathbf{C}$ if $p \leq 1$ (due to the penalty function singularity at $\mathbf{C} = 0$ [11]).

Several authors have proposed methods for solving problem (4). For instance, Cotter el al. [7] have developed an algorithm for solving problem (4) when $p \leq 1$ and $q = 2$, known as M-FOCUSS. Such an algorithm based on factored gradient descent have been proven to converge towards a local or global (when $p = 1$) minimum of problem (4) if it does not get stuck in a fixed-point.

The case $p = 1, q = 2$, named as M-BP for Multiple Basis Pursuit in the following, is a special case that deserves special attention. Indeed, it seems to be the most natural extension of the so-called Lasso problem [28] or Basis Pursuit Denoising [6], since for $L = 1$, problem (4) reduced to the Lasso problem. The key point of this case is that it yields to a convex optimization problem and thus it can benefit from all properties resulting from convexity *e.g* global minimum. Malioutov et al. [21] have proposed an algorithm based on a second-order cone programming formulation for solving the resulting M-BP convex problem which at the contrary of M-FOCUSS, always converges to the problem global solution.

When $p = 1$ and $q = 1$, again we fall within a very particular case that has been studied by Chen et al. [5]. In this case, the sparse simultaneous problem can be decoupled in L independent problems. In such a situation, the hypothesis of the L signals having the same sparsity profile is no more guaranteed, thus the problem can not be considered as a simultaneous sparse approximation problem. However, in this case, one can use efficient algorithms that solve the well-known *Lasso* problem [27], [10].

The approach proposed by Wipf et al. [35] for solving the sparse simultaneous approximation is somewhat related to the optimization problem in equation (4) but from a very different perspective. Indeed, if we consider that the above described approaches are equivalent to a MAP-estimation procedures, then Wipf et al. have explored a Bayesian model which prior encourages sparsity. In this sense, their approach is related to the relevance vector machine of Tipping et al. [29]. Algorithmically they proposed an empirical bayesian learning approach based on Automatic Relevance Determination (ARD). The ARD prior over each row they have introduced is

$$p(c_{i,\cdot}; d_i) = \mathcal{N}(0, d_i \mathbf{I}) \qquad \forall i$$

where $\mathbf{d}$ is a vector of non-negative hyperparameters that govern the prior variance of each coefficient matrix row. Hence, these hyperparamaters aim at catching the sparsity profile of the approximation. Mathematically, the resulting optimization problem is to minimize according to $\mathbf{d}$ the following cost function

$$L \log |\Sigma_t| + \sum_{j=1}^{L} \mathbf{s}_j^t \Sigma_t^{-1} \mathbf{s}_j \tag{5}$$

where $\Sigma_t = \sigma^2 \mathbf{I} + \mathbf{\Phi} \mathbf{D} \mathbf{\Phi}^t$, $\mathbf{D} = \text{diag}(\mathbf{d})$ and σ^2 a parameter of the algorithm related to the noise level presented in the signals to be approximated. The algorithm is then based on a likelihood maximization which is performed through an Expectation-Minimization approach. Very recently, a very efficient algorithm for solving this problem has been proposed [19]. However, the main drawback of this latter approach is that due to its greedy nature, the algorithm can be easily stucked in local minima.

The second family of methods for solving the simultaneous sparse approximation is to use a suboptimal forward sequential selection of a dictionary element. These algorithms denoted as M-OMP in the sequel [32], are a simple extension of the well-known Matching Pursuit technique to simultaneous approximation. They provide a solution of problem (2) or (3) which corresponds to a local minimum of the cost function. While the algorithms are relatively simple, their main advantage is their efficiency and some theoretical guarantees about the correctness of the approximation can be provided [17], [32].

C. Our contributions

At the present time, the most interesting approach for simultaneous sparse approximation is the Bayesian approach introduced by Wipf et al. [35] and further improved by Ji et al. [19] in terms of speed efficiency. However, in this paper, we depart from this route and instead consider a (frequentist) regularized

empirical minimization approach. Indeed, in view of the very flourishing literature on ℓ_p minimization algorithms and subsequent theoretical results (*e.g* consistency of estimator, convergence rate, ...) related to single signal sparse approximation, we think that many of these results can be transposed to the multiple signal approximation case and we hope with this paper to give our dime to reach that objective. Hence, we follow the steps of Cotter et al. and Malioutov et al. in considering using ℓ_p minimization problem for simultaneous sparse approximation, but propose a different way of solving the minimization problem (4). Our contributions here are essentially on novel insights on the problem and algorithms for solving it.

At first, we develop a simple and efficient algorithm for solving the M-Basis Pursuit problem. We show that by using results from non-smooth optimization theory, we are able to propose an iterative method which only needs some matrix multiplications.

Then, we focus on the more general situation where $p \leq 1$ and $q = 2$ in $J_{p,q}$. We show that such a row diversity measure is actually related to automatic relevance determination (ARD). Indeed, we show that for any $p \leq 1$ and $q = 2$, $J_{p,q}$ can be interpreted as a weighted row 2-norm measure, and these weights measure the relevance of a given row in the approximation. Owing to that interpretation, we clarify the relation between M-FOCUSS and M-SBL (which also uses ARD) for any value of $p \leq 1$. Afterwards, instead of directly deriving a proper algorithm for solving the non-convex optimization problem when $p < 1$ and $q = 2$, we introduce an iterative reweighted M-Basis pursuit (IrM-BP) algorithm. We then show that depending on the chosen weights, such an iterative scheme can actually solve problem (4). Our main contribution at this point is then to have translated the non-convex problem (4) into a series of convex problems which are easy to solve with our iterative method for M-BP. Furthermore, by choosing a different weighting scheme, we show that our iterative reweighted approach is strongly related to M-SBL.

The paper is organized as follows. Section II introduces the iterative shrinking algorithm for solving M-BP. After having discussed the ARD formulation of M-FOCUSS in Section III, we propose in Section IV a reweighted M-BP algorithm for addressing the optimization problem related to M-FOCUSS. Experimental results presenting performance of our algorithms are in Section V while conclusion and perspectives in Section VI close the paper. For a sake of reproducibility, the code used in this paper is available on http://asi.insa-rouen.fr/enseignants/~arakotom/code/SSAindex.html

II. Simple Algorithm for M-Basis Pursuit

The algorithm we propose in this section addresses the particular case of $p = 1$ and $q = 2$, denoted as the M-BP problem. We show in the sequel that the specific structure of the problem leads to a very simple iterative shrinking algorithm. Furthermore if the dictionary is under-complete then it can be shown that the solution of the problem is equivalent to a simple shrinkage of the coefficient matrix.

The M-BP optimization problem is the following

$$\min_{\mathbf{C}} W(\mathbf{C}) = \frac{1}{2}\|\mathbf{S} - \mathbf{\Phi}\mathbf{C}\|_F^2 + \lambda \sum_i \|c_{i,\cdot}\|_2 \tag{6}$$

where the objective function $W(\mathbf{C})$ is a non-smooth but convex function. Since the problem is unconstrained a necessary and sufficient condition for a matrix $\mathbf{C}^\star$ to be a minimizer of (6) is that $\mathbf{0} \in \partial W(\mathbf{C}^\star)$ where $\partial W(\mathbf{C})$ denotes the subdifferential of our objective value $W(\mathbf{C})$ [1]. By computing the subdifferential of $W(\mathbf{C})$ with respect to each row $c_{i,\cdot}$ of $\mathbf{C}$, the optimality condition of problem (6) is then

$$-\mathbf{r}_i + \lambda g_{i,\cdot} = 0 \qquad \forall i$$

where $\mathbf{r}_i = \phi_i^t(\mathbf{S} - \mathbf{\Phi}\mathbf{C})$ and $g_{i,\cdot}$ is the i-th row of a subdifferential matrix $\mathbf{G}$ of $J_{1,2}(\mathbf{C}) = \sum_i \|c_{i,\cdot}\|_2$. The following lemma which proof has been postponed to the appendix, characterizes this subdifferential $\mathbf{G}$ of $J_{1,2}(\mathbf{C})$.

Lemma 1: A matrix $\mathbf{G}$ is a subdifferential of $J_{1,2}(\mathbf{C}) = \sum_i \|c_{i,\cdot}\|_2$ if and only if the j-th row of $\mathbf{G}$ satisfies

$$\mathbf{e}_j^t\mathbf{G} \in \begin{cases} \{\mathbf{g} \in \mathbb{R}^L : \|\mathbf{g}\|_2 \leq 1\} & \text{if } \ \forall k, c_{j,k} = 0 \\ \frac{c_{j,\cdot}}{\|c_{j,\cdot}\|_2} & \text{otherwise} \end{cases}$$

where $\mathbf{e}_j$ is a canonical vector of $\mathbb{R}^M$.

According to this definition of $J_{1,2}$'s subdifferential, the optimality condition can be rewritten as

$$\begin{aligned} -\mathbf{r}_i + \lambda\frac{c_{i,\cdot}}{\|c_{i,\cdot}\|_2} &= \mathbf{0} \quad \forall i, \quad c_{i,\cdot} \neq \mathbf{0} \\ \|\mathbf{r}_i\|_2 &\leq \lambda \quad \forall i, \quad c_{i,\cdot} = \mathbf{0} \end{aligned} \tag{7}$$

A matrix $\mathbf{C}$ satisfying these equations can be obtained after the following algebra. Let us expand each $\mathbf{r}_i$ so that

$$\begin{aligned} \mathbf{r}_i &= \phi_i^t(\mathbf{S} - \mathbf{\Phi}\mathbf{C}_{-i}) - \phi_i^t\phi_i c_{i,\cdot} \\ &= T_i - c_{i,\cdot} \end{aligned} \tag{8}$$

Algorithm 1 Solving M-BP through iterative shrinking

$\mathbf{C} = 0$, Loop = 1

while Loop **do**

for $i = 1, 2, \cdots, M$ **do**

if $c_{i,\cdot}$ KKT condition is not satisfied **then**

$c_{i,\cdot} = \left(1 - \frac{\lambda}{\|T_i\|}\right)_+ T_i$

end if

end for

if all KKT Conditions are satisfied **then**

Loop = 0

end if

end while

where $\mathbf{C}_{-i}$ is the matrix $\mathbf{C}$ with the i-th row being set to 0 and $T_i = \phi_i^t(\mathbf{S} - \mathbf{\Phi}\mathbf{C}_{-i})$. The second equality is obtained by remembering that $\phi_i^t\phi_i$=1. Then, equation (7) tells us that if $c_{i,\cdot}$ is non-zero, T_i and $c_{i,\cdot}$ have to be collinear. Plugging all these points into equation (7) yields to an optimal solution that can be obtained as :

$$c_{i,\cdot} = \left(1 - \frac{\lambda}{\|T_i\|}\right)_+ T_i \qquad \forall i \tag{9}$$

From this update equation, we can derive a simple algorithm which consists in iteratively applying the update (9) to each row of $\mathbf{C}$. Such an iterative scheme actually performs a block-coordinate optimization. Although, block-coordinate optimization does not converge in general for non-smooth optimization problem, Tseng [33] has shown that for an optimization problem which objective value is the sum of a smooth and convex function and a non-smooth but block-separable convex function, block-coordinate optimization converges towards the global minimum of the problem. Since for M-BP we are considering a quadratic function and a row-separable penalty function, Tseng's results can be directly applied in order to prove convergence of our algorithm.

Our approach, detailed in Algorithm (1), is a simple and efficient algorithm for solving M-BP especially when the dictionary size is large. A similar approach has also been proposed for solving the lasso [12], the group lasso [36] and the elastic net [37]. Intuitively, we can understand this algorithm as an algorithm which tends to shrink to zero rows of the coefficient matrix that contribute poorly to the approximation. Indeed, T_i can be interpreted as the correlation between the residual when row i has been removed and

ϕ_i. Hence the smaller the norm of T_i is, the less ϕ_i is relevant in the approximation. And according to equation (9), the smaller the resulting $c_{i,\cdot}$ is. Insight into this iterative shrinking algorithm can be further obtained by supposing that $M \leq N$ and that $\mathbf{\Phi}$ is composed of orthonormal elements of $\mathbb{R}^N$, hence $\mathbf{\Phi}^t\mathbf{\Phi} = \mathbf{I}$. In such situation, we have

$$T_i = \phi_i^t \mathbf{S} \quad \text{and} \quad \|T_i\|_2^2 = \sum_{k=1}^{L} (\phi_i^t s_k)^2$$

and thus

$$c_{i,\cdot} = \left(1 - \frac{\lambda}{\sqrt{\sum_k^L (\phi_i^t s_k)^2}}\right)_+ \phi_i^t \mathbf{S}$$

This last equation highlights the relation between the single Basis Pursuit (when $L = 1$) and the Multiple-Basis Pursuit algorithm presented here. Both algorithms lead to a shrinkage of the coefficient projection. With the inclusion of multiple signals, the shrinking factor becomes more robust to noise since it depends on the correlation of the atom ϕ_i to all signals.

As we stated previously, the M-BP problem is equivalent to the M-FOCUSS problem with $p = 1$ and $q = 2$. For solving such a problem Cotter et al. [7] have proposed a factored gradient algorithm. That algorithm is related to iterative reweighted least-squares, which at each iteration updates the coefficient matrix $\mathbf{C}$. However, their factored gradient algorithm presents a important issue. Indeed, the updates they propose are not guaranteed to converge to a local minima of the problem (if the problem is not convex $p < 1$) or to the global minimum of the convex problem ($p = 1$). Indeed, their algorithm presents several fixed-points since when a row of $\mathbf{C}$ is equal to 0, it stays at 0 at the next iteration. Although such a point may be harmless if the algorithm is initialized with a "good" starting point, it is nonetheless an undesiderable point when solving a convex problem. At the contrary, our iterative shrinking algorithm does not suffer from the presence of such fixed-points. Thus, it can benefit from a good initialization like $\mathbf{C} = \mathbf{0}$ since most of rows would stay at zero.

From a computational complexity point of view, it is not possible to evaluate the exact number of iterations that will be needed before convergence of our algorithm. However, we can analyze the computational cost per each iteration. We can note that each shrinking operation,in the worst case scenario, has to be done M times and the dominating cost for each shrinking is the computation of T_i. This computation involves the matrix multiplication $\mathbf{\Phi}\mathbf{C}_{-i}$ and a matrix-vector multiplication which respectively need $\mathcal{O}(NML)$ and $\mathcal{O}(NL)$ operations. On the overall, if we assume that at each iteration,

all $c_{i,\cdot}$ are updated, we can consider that the computational cost of our algorithm is about $\mathcal{O}(M^2NL)$. This cost per iteration can be compared to the one of M-FOCUSS algorithm and second-order code programming of Malioutov et. al [21] which are respectively $\mathcal{O}(MN^2)$ and $\mathcal{O}(M^3L^3)$. Theoretically, it seems that our algorithm suffers more than M-FOCUSS from large dictionary size but it is far more efficient than the SOC programming.

Illustrations of how our algorithm behaves and empirical computational complexity evaluation are given in section V.

III. ARD Formulation of Simultaneous Sparse Approximation

In this section, we now focus on the relaxed optimization problem given in (4) for the general where $p \leq 1$ and $q = 2$ in $J_{p,q}(\mathbf{C})$. Our objective here is to clarify the connection between such a form of penalization and the automatic relevance determination of $\mathbf{C}$'s rows, which has been the keystone of the Bayesian approach of Wipf et al [35].

For this purpose, we first consider the following formulation of the simultaneous sparse approximation problem

$$\begin{array}{ll} \min_{\mathbf{C}} & \frac{1}{2}\|\mathbf{S} - \mathbf{\Phi}\mathbf{C}\|_F^2 \\ \text{st} & J_{p,q}(\mathbf{C}) \leq T \end{array} \tag{10}$$

Due to the non-convexity of $J_{p,q}$, equivalence between this formulation and the one given in equation (4) is not strict in the sense that it is possible that for some values of T, there exists no λ so that solution of problem 4 is a solution of problem (10) (for more details on this issue, one can refer to [8]). However, such a formulation is useful due to the nature of the transformation we apply to problem (10). Indeed, since the power function is strictly monotonically increasing, we can equivalently replace the constraint of that problem with the constraint

$$(J_{p,q}(\mathbf{C}))^{\frac{1}{p}} = \left(\sum_i \|c_{i,\cdot}\|_q^p\right)^{\frac{1}{p}} \leq T^{\frac{1}{p}}$$

Now, let us introduce the key lemma that allows us to derive the ARD-based formulation of the problem. This lemma gives a variational form of the $\ell_{2/s}$ norm of a sequence $\{a_i\}$ for any $s > 1$.

Lemma 2: if $r > 0$ and $\{a_i\}_i^I$ so that $I \in \mathbb{N}$ and $\forall i = 1, \cdots, I, a_i \in \mathbb{R}$, let us define $s = 1 + \frac{1}{r}$ then,

$$\min_{\mathbf{d}} \left\{ \sum_i \frac{a_i^2}{d_i} : d_i \geq 0, \sum_i d_i^r \leq 1 \right\} = \left(\sum_i |a_i|^{\frac{2}{s}} \right)^{s/2} \tag{11}$$

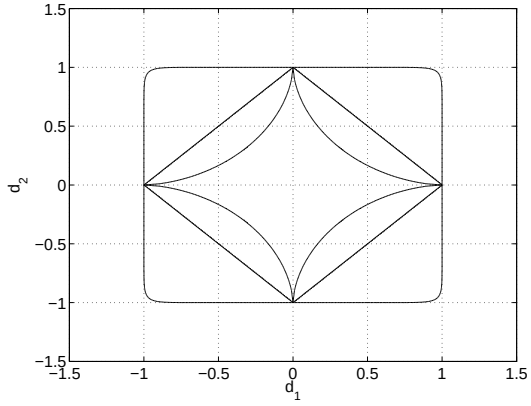

Fig. 1. Plots of $d_1^r + d_2^r = 1$ for different values of r. From the inner to the outer curves we have $r = \{0.6, 1, 19\}$ which corresponds to p=$\{0.75,\ 1,\ 1.9\ \}$. We can note that for $r = p < 1$, the plot delimitates a non-convex region which presents singularities at $d_1 = 0$ or $d_2 = 0$. At the contrary, when p tends towards 2, r goes to ∞ and thus the constraints becomes equivalent to a ℓ_∞ and defines a convex feasibility domain.

and equality occurs for $\sum_i |a_i| > 0$ at

$$d_i^\star = \frac{|a_i|^{\frac{2}{r+1}}}{\left(\sum_i |a_i|^{\frac{2r}{r+1}}\right)^{1/r}}$$

The proof of this lemma, which is simply based on a Holder inequality can be found for instance in [23]. According to this lemma, the $\ell_{\frac{2}{s}}$ norm of a sequence can be computed through a minimization problem. Hence, applying this lemma to $(J_{p,q}(\mathbf{C}))^{\frac{1}{p}}$ by defining $a_i = \|c_{i,\cdot}\|_q$ gives for $p < 2$ and $q = 2$

$$\min\left\{\sum_i \frac{\|c_{i,\cdot}\|_2^2}{d_i} : d_i \geq 0, \sum_i d_i^r \leq 1\right\} = \left(\sum_i \|c_{i,\cdot}\|_2^p\right)^{1/p} \tag{12}$$

with $s = \frac{2}{p}$ and $r = \frac{p}{2-p}$.

Now, we can go back to a regularized form of the problem by replacing in (4) $J_{p,q}(\mathbf{C})$ with $J_{p,q}(\mathbf{C})^{1/p}$. Using the above lemma yields to the following equivalent problem

$$\begin{array}{ll} \min\limits_{\mathbf{C},\mathbf{d}} & \frac{1}{2}\|\mathbf{S} - \mathbf{\Phi C}\|_F^2 + \lambda \sum_i \frac{\|c_{i,\cdot}\|_2^2}{d_i} \\ \text{s.t.} & \sum_i d_i^r \leq 1 \\ & d_i \geq 0 \qquad \forall i \end{array} \tag{13}$$

This final problem is the one which makes clear the automatic relevance determination interpretation of the original formulation of M-FOCUSS problem. We can see that we have transformed problem (4) into a problem with a smooth objective function at the expense of adding some additional variables d_i. These parameters d_i actually aim at determining the relevance of each coefficient matrix row like in

problem (5). Indeed, in the objective function, each row square norm is now inversely weighted by a coefficient d_i. By taking the convention that $\frac{x}{0} = \infty$ if $x \neq 0$ and 0 otherwise, the objective value of the optimization problem becomes finite only if $\|c_{i,\cdot}\|_2^2 = 0$ for $d_i = 0$ Then the smaller d_i is, the smaller the $c_{i,\cdot}$ norm will be. Furthermore, optimization problem (13) also involves some constraints on $\{d_i\}$. These constraints impose the vector $\mathbf{d}$ to be in the positive orthant of $\mathbb{R}^M$ and so that its ℓ_r quasi-norm is smaller than 1. Examples of the feasibility domain imposed by these constraints related to ℓ_r quasi-norm, are given in Figure 1. According to the relation between p and r, for $p < 1$, we also have $r < 1$, and we can see in the figure that the ℓ_r norm delimits a non-convex region with singularities at $d_i = 0$. Such singularities favor sparsity of the vector $\mathbf{d}$ at optimality. As we have noted above, when a d_i is equal to 0, the corresponding row norm should be equal to 0 which means that the corresponding element of the dictionary is "irrelevant" for the approximation of all signals simultaneously.

The problem (13) proposes an equivalent formulation of the M-FOCUSS problem for which the row-diversity measure has been transformed in a easily interpretable penalty function owing to an ARD formulation. The trade-off between convexity of the problem and the sparsity of the solution has been transfered from p, q to r.

From a Bayesian perspective, we can interpret the d_i as the diagonal term of the covariance matrix of a Gaussian prior over the row norm distribution. This is typically the classical Bayesian Automatic Relevance Determination approach as proposed for instance in the following works [25], [29]. This novel insight on the ARD interpretation of $J_{p,q}(\mathbf{C})$ clarifies the connection between the M-FOCUSS of Cotter et al. [7] and the M-SBL of Wipf et al. [35] for any value of $p < 1$. In their previous works, Wipf et al. have proved that these two algorithms were related when $p \approx 0$. Here, we refine their result by enlarging the connection for other values of p. In a frequentist framework, we can also note that Grandvalet et al. has proposed a similar approach for feature selection in generalized linear models and SVM [15], [16].

This particular ARD-based formulation of the problem (4) can still be simplified by exploiting the specific relation between the Frobenius norm and the ℓ_2 norm. Indeed, we can further expand equation (13), to yield

$$\begin{array}{ll} \min\limits_{\mathbf{C},d} & \sum_j \frac{1}{2} \left(\|\mathbf{s}_j - \mathbf{\Phi} c_{\cdot,j}\|_2^2 + \lambda \sum_i \frac{c_{i,j}^2}{d_i} \right) \\ \text{s.t.} & \sum_i d_i^r \leq 1 \\ & d_i \geq 0, \qquad \forall i \end{array} \tag{14}$$

From this formulation, we can exhibit the relation between each single signal approximation problem and the sparsity-inducing ARD parameters. Indeed, if we consider fixed parameters d_i then the problem

is equivalent to L least-square regression problems where the relevance of each atoms ϕ_i is weighted by d_i. Then, we can interpret again the problem as several least-square problems which are tied together by the sparsity profile induced by the ARD parameters.

IV. Reweighted M-Basis Pursuit

This section introduces an iterative reweighted M-Basis Pursuit (IrM-BP) algorithm and proposes two ways of setting these weights. By using the first weighting scheme, we are able to provide an iterative algorithm which solves problem (4) when $p < 1$ and $q = 2$. The second weighting scheme makes clear the strong relation between M-SBL and our work.

A. Reweighted algorithm

Recently, several works have advocated that sparse approximations can be recovered through iterative algorithms based on a reweighted ℓ_1 minimization [38], [3], [4]. Typically, for a single signal case, the idea consists in iteratively solving the following problem

$$\min_{\mathbf{c}} \frac{1}{2}\|\mathbf{s} - \mathbf{\Phi c}\|_2^2 + \lambda \sum_i z_i |c_i|$$

where z_i are some positive weights, and then to update the positive weights z_i according to the solution $\mathbf{c}^\star$. Besides, providing empirical evidences that reweighted ℓ_1 minimization yields to sparser solutions than a simple ℓ_1 minimization, the above cited works theoretically support such claims. These results for the single signal approximation case suggest that in the simultaneous sparse approximation problem, reweighted M-Basis Pursuit would lead to sparser solutions than the classical M-Basis Pursuit.

Our iterative reweighted M-Basis Pursuit is defined as follows. We iteratively solve until convergence the optimization problem

$$\min_{\mathbf{C}} \frac{1}{2}\|\mathbf{S} - \mathbf{\Phi C}\|_F^2 + \lambda \sum_i z_i \|c_{i,\cdot}\|_2 \tag{15}$$

where the positive weight vector $\mathbf{z}$ depends on the previous iterate $\mathbf{C}^{(n-1)}$. In our case, we will consider the following weighting scheme

$$z_i = \frac{1}{(\|c_{i,\cdot}^{(n-1)}\|_2 + \varepsilon)^r} \quad \forall i \tag{16}$$

where $\{c_{i,\cdot}^{(n-1)}\}$ is the i-th row of $\mathbf{C}^{(n-1)}$, r a user-defined positive constant and ε a small regularization term that avoids numerical instabilities and prevents from having an infinite regularization term for $c_{i,\cdot}$as soon as $c_{i,\cdot}^{(n-1)}$ vanishes. This is a classical trick that has been used for instance by Candès et al. [3] . Note that for any positive weight vector $\mathbf{z}$, problem (15) is a convex problem that does not present local minima. Furthermore, it can be solved using our iterative shrinking algorithm by simply replacing λ with $\lambda_i = \lambda \cdot z_i$. Such a scheme is similar to the *adaptive lasso* algorithm of Zou et al. [38] but uses several iterations and addresses the simultaneous approximation problem.

B. Relation with M-FOCUSS

The IrM-BP algorithm we proposed above can also be interpreted as an algorithm for solving problem (4) when $0 < p < 1$. Indeed, similarly to the reweighted ℓ_1 scheme of Candès et al. [3] or the one-step reweighted lasso of Zou et al. [39], our algorithm falls in the class of majorize-minimize (MM) algorithms [18]. MM algorithms consists in replacing a difficult optimization problem with a more easier one, for instance by linearizing the objective function, by solving the resulting optimization problem and by iterating such a procedure.

The connection between MM algorithms and our reweighted scheme can be made through linearization. In effect, in our case, since $J_{p,2}$ is concave in $c_{i,\cdot}$ for $0 < p < 1$, a linear approximation of $J_{p,2}(\mathbf{C})$ around $\mathbf{C}^{(n-1)}$ yields to the following majorizing inequality

$$J_{p,2}(\mathbf{C}) \leq J_{p,2}(\mathbf{C}^{(n-1)}) + \sum_i \frac{p}{\|c_{i,\cdot}^{(n-1)}\|_2^{1-p}}(\|c_{i,\cdot}\| - \|c_{i,\cdot}^{(n-1)}\|)$$

then for the minimization step, replacing in problem (4) $J_{p,2}$ with the above inequality and dropping constant terms lead to our optimization problem (15) with appropriately chosen z_i and r. Note that for the weights given in equation (16), $r = 1$ corresponds to the linearization of a log penalty $\sum_i \log(\|c_{i,\cdot}\|)$ whereas setting $r = 1-p$ corresponds to a ℓ_p penalty ($0 < p < 1$). According to the convergence properties for MM algorithms towards a local minimum of their objective function [18], we can state that our IrM-BP algorithm converges towards a local minimum of problem (4) with p and r being appropriately related.

Note that problem (15) can also be interpreted as a manual relevance determination of matrix $\mathbf{C}$'s rows. Indeed, compared to the ARD formulation given in equation (13), the weights z_i can be considered as pre-fixed weights that determine the importance of row $c_{i,\cdot}$ in the approximation. Furthermore, it is clear that for a pair $\mathbf{C}^\star, d^\star$ that minimizes problem (14), $\mathbf{C}^\star$ also minimizes problem (15) for

$$z_i = \frac{\|c_{i,\cdot}^\star\|_2}{d_i^\star}$$

C. Relation with M-SBL

Very recently, Wipf et al. [34] have proposed some new insights on Automatic Relevance Determination and on Sparse Bayesian Learning. They have shown that, for the vector regression case, ARD can be achieved by means of iterative reweighted ℓ_1 minimization. Furthermore, in that paper, they have sketched an extension of such results for matrix regression in which ARD is used for automatically selecting the most relevant covariance components in a dictionary of covariance matrices. Such an extension is more related to learning with multiple kernels in regression as introduced by Girolami et al. [13] or Rakotomamonjy et al. [26] although some connections with simultaneous sparse approximation can be made. Here, we build on the works on Wipf et al. [34] and give all the details about how M-SBL and reweighted M-BP are related.

Recall that the cost function minimized by the M-SBL of Wipf et al. [35] is

$$\mathcal{L}(\mathbf{d}) = L \log |\Sigma_t| + \sum_{j=1}^{L} \mathbf{s}_j^t \Sigma_t^{-1} \mathbf{s}_j \tag{17}$$

where $\Sigma_t = \sigma^2 \mathbf{I} + \mathbf{\Phi} \mathbf{D} \mathbf{\Phi}^t$ and $\mathbf{D} = \text{diag}(\mathbf{d})$. Now, let us define $g^\star(z)$ as the conjugate function of the concave $\log |\Sigma_t|$. Since, that log function is concave and continuous on $\mathbb{R}_+^M$, according to the scaling property of conjugate functions we have [2]

$$L \cdot \log |\Sigma_t| = \min_{\mathbf{z} \in \mathbb{R}^M} \mathbf{z}^t \mathbf{d} - L g^\star \left(\frac{\mathbf{z}}{L} \right)$$

Thus, the cost function $\mathcal{L}(\mathbf{d})$ in equation (17) can then be upper-bounded by

$$\mathcal{L}(\mathbf{d}, \mathbf{z}) \triangleq \mathbf{z}^t \mathbf{d} - L g^\star \left(\frac{\mathbf{z}}{L} \right) + \sum_{j=1}^{L} \mathbf{s}_j^t \Sigma_t^{-1} \mathbf{s}_j \tag{18}$$

Hence when optimized over all its parameters, $\mathcal{L}(\mathbf{d}, \mathbf{z})$ converges to a local minima or a saddle point of (17). However, for any fixed $\mathbf{d}$, one can optimize over $\mathbf{z}$ and get the tight optimal upper bound. If we denote as $\mathbf{z}^\star$ such an optimal $\mathbf{z}$ for any fixed $\mathbf{d}^\dagger$, since $L \cdot \log |\Sigma_t|$ is differentiable, we have, according to conjugate function properties, the following closed form of $\mathbf{z}^\star$

$$\mathbf{z}^\star = L \cdot \nabla \log |\Sigma_t| (\mathbf{d}^\dagger) = \text{diag}(\mathbf{\Phi}^t \Sigma_t^{-1} \mathbf{\Phi}) \tag{19}$$

Similarly to what proposed by Wipf et al., Equations (18) and (19) suggest an alternate optimization scheme for minimizing $\mathcal{L}(\mathbf{d}, \mathbf{z})$. Such a scheme would consist, after initialization of $\mathbf{z}$ to some arbitrary

vector, in keeping $\mathbf{z}$ fixed and in computing

$$\mathbf{d}^{\dagger} = \arg\min_{\mathbf{d}} \mathcal{L}_z(\mathbf{d}) \triangleq \mathbf{z}^t\mathbf{d} + \sum_{j=1}^{L} \mathbf{s}_j^t \Sigma_t^{-1} \mathbf{s}_j \tag{20}$$

then to minimize $\mathcal{L}(\mathbf{d}^{\dagger}, \mathbf{z})$ for fixed $\mathbf{d}^{\dagger}$, which can be analytically done according to equation (19). This alternate scheme is then performed until convergence to some $\mathbf{d}^{\star}$.

Owing to this iterative scheme proposed for solving M-SBL, we can now make clear the connection between M-SBL and our iterative reweighted M-BP according to the following lemma. Again this is an extension to the multiple signals case of a Wipf's lemma.

Lemma 3: The objective function in equation (20) is convex and can be equivalently solved by computing

$$\mathbf{C}^{\star} = \arg\min_{\mathbf{C}} \mathcal{L}_z(\mathbf{C}) = \frac{1}{2}\|\mathbf{S} - \mathbf{\Phi}\mathbf{C}\|_F^2 + \sigma^2 \sum_i z_i^{1/2} \|c_{i,\cdot}\| \tag{21}$$

and then by setting

$$d_i = z_i^{-1/2} \|c_{i,\cdot}^{\star}\| \qquad \forall i$$

Proof: Convexity of the objective function in equation (20) is straightforward since it is just a sum of convex functions [2]. The key point of the proof is based on the equality

$$\mathbf{s}_j^t \Sigma_t^{-1} \mathbf{s}_j = \frac{1}{\sigma^2} \min_{c_{\cdot,j}} \|\mathbf{s}_j - \mathbf{\Phi} c_{\cdot,j}\|_2^2 + \sum_i \frac{c_{i,j}^2}{d_i} \tag{22}$$

which proof is given in appendix. According to this equality, we can upper-bound $\mathcal{L}_z(\mathbf{d})$ with

$$\mathcal{L}_z(\mathbf{d}, \mathbf{C}) = \mathbf{z}^t\mathbf{d} + \sum_j \frac{1}{\sigma^2} \|\mathbf{s}_j - \mathbf{\Phi} c_{\cdot,j}\|_2^2 + \sum_{i,j} \frac{c_{i,j}^2}{d_i} \tag{23}$$

The problem of minimizing $\mathcal{L}_z(\mathbf{d}, \mathbf{C})$ is smooth and jointly convex in its parameters $\mathbf{C}$ and $\mathbf{d}$ and thus an iterative coordinatewise optimization scheme (iteratively optimizing over $\mathbf{d}$ with fixed $\mathbf{C}$ and then optimizing over $\mathbf{C}$ with fixed $\mathbf{d}$) yields to the global minimum. It is easy to show that for any fixed $\mathbf{C}$, the minimal value of $\mathcal{L}_z(\mathbf{d}, \mathbf{C})$ with respects to $\mathbf{d}$ is achieved when

$$d_i = z_i^{-1/2} \|c_{i,\cdot}\| \qquad \forall i$$

Plugging these solutions back into (23) and multiplying the the resulting objective function with $\sigma^2/2$ yields to

$$\mathcal{L}_z(\mathbf{C}) = \frac{1}{2} \sum_j \|\mathbf{s}_j - \mathbf{\Phi} c_{\cdot,j}\|_2^2 + \sigma^2 \sum_i z_i^{1/2} \|c_{i,\cdot}\| \tag{24}$$

Making the relation between ℓ_2 and Frobenius norms concludes the proof. ■

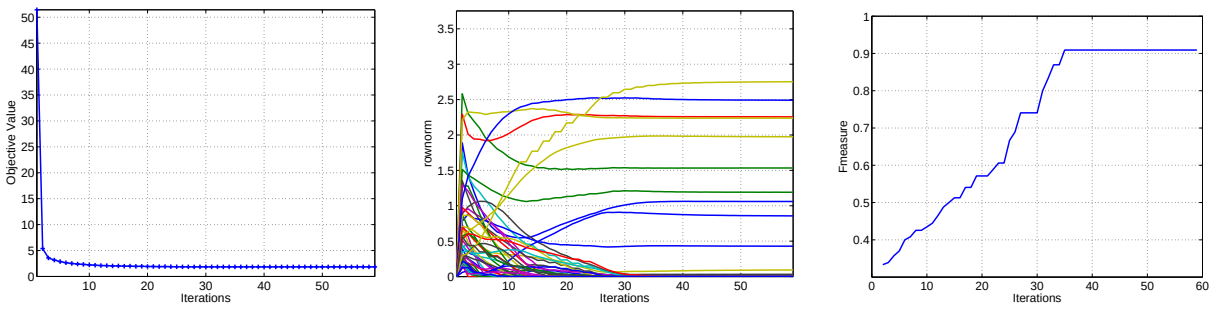

Fig. 2. Illustration of the Iterative shrinking algorithm for M-BP. Example of variation along the iterations of : left) Objective value, middle) rownorm $\|c_{i,\cdot}\|$, right) F-measure. For this example, the dictionary size is 50 while 10 active elements have been considered in the true sparsity profile.

Minimizing $\mathcal{L}_z(\mathbf{C})$ boils down to minimize the M-BP problem with an adaptive penalty $\lambda_i = \sigma^2 \cdot z_i^{1/2}$ on each row-norm. This latter point makes the alternate optimization scheme based on equation (19) and (20) equivalent to our iterative reweighted M-BP for which weights z_i would be given by equation (19).

The impact of this relation between M-SBL and reweighted M-BP is essentially methodological. Indeed, its main advantage is that it turns the original M-SBL optimization problem into a serie of convex optimization problems. In this sense, our iterative reweighted algorithm can again be viewed as an application of MM approach for solving problem (17). Indeed, we are actually iteratively minimizing a proxy function which has been obtained by majorizing each term of equation (17). This MM point of view offers us the convergence of our iterative algorithm towards a local minimum of equation (17). Convergence for the single signal case using other arguments has also been shown by Wipf et al. [34]. Note that similarly to M-FOCUSS, the original M-SBL algorithm based on EM approach suffers from presence of fixed-points (when $d_i = 0$). Hence, such an algorithm is not guaranteed to converge towards a local minimum of (17). This is then another argument for preferring IrM-BP.

V. Numerical experiments

Some computer simulations have been carried out in order to evaluate the algorithms proposed in the above sections. Results that have been obtained from these numerical studies are detailed in this section.

A. Experimental protocol

In order to quantify the performance of our algorithms and compare them to other approaches, we have used simulated datasets with different redundancies $\frac{M}{N}$, number k of active elements and number L

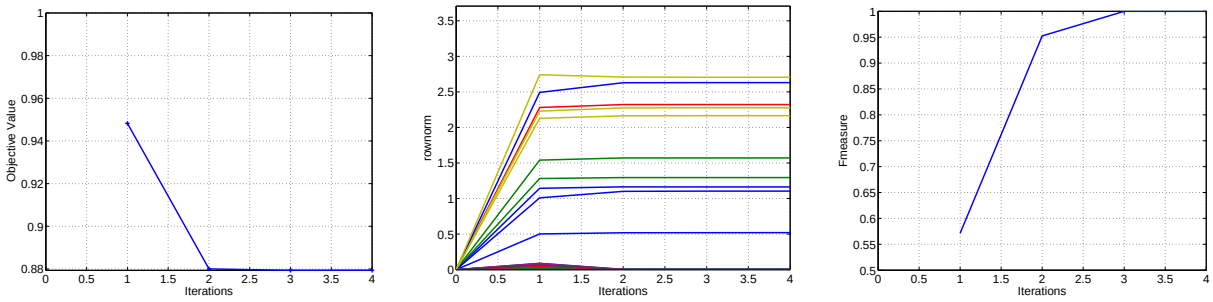

Fig. 3. Illustration of the Iterative reweighted M-BP applied for $J_{\frac{1}{2},2}$ penalty. Example of variation along the iterations of : left) Objective value, middle) rownorm $\|c_{i,\cdot}\|$, right) F-measure

of signals to approximate. The dictionary is based on M vectors sampled from the unit hypersphere of $\mathbb{R}^N$. The true coefficient matrix $\mathbf{C}$ has been obtained as follows. The positions of the k non-zero rows in the matrix are randomly drawn. The non-zero coefficients of $\mathbf{C}$ are then drawn from a zero-mean unit variance Gaussian distribution. The signal matrix $\mathbf{S}$ is obtained as in equation (1) with the noise matrix being drawn i.i.d from a zero-mean Gaussian distribution and variance so that the signal-to-noise ratio of each single signal is 10 dB.

Each algorithm is provided with the signal matrix $\mathbf{S}$ and the dictionary $\mathbf{\Phi}$ and will output an estimate of $\mathbf{C}$. Since our objective is to evaluate whether the sparsity profile of the coefficient matrix has been recovered, we use as a performance criterion the F-measure between the row support of the true $\mathbf{C}^\star$ and the estimate $\hat{\mathbf{C}}$. In order to take into account numerical precisions, we overload the row support definition as

$$\text{rowsupp}(\mathbf{C}) = \{i \in [1 \cdots M] : \|c_{i,\cdot}\| < \mu\}$$

where μ is a threshold coefficient that has been set by default to $1e^{-16}$ in our experiments. From $\text{rowsupp}(\hat{\mathbf{C}})$ and $\text{rowsupp}(\mathbf{C}^\star)$ respectively the estimated and true sparsity profile, Precision, Recall and F-measure are computed as

$$\text{Prec} = \frac{|\text{rowsupp}(\hat{\mathbf{C}}) \cap \text{rowsupp}(\mathbf{C})|}{|\text{rowsupp}(\hat{\mathbf{C}})|}$$

$$\text{Recall} = \frac{|\text{rowsupp}(\hat{\mathbf{C}}) \cap \text{rowsupp}(\mathbf{C})|}{|\text{rowsupp}(\mathbf{C})|}$$

$$\text{F-measure} = \frac{2 \cdot \text{Prec} \cdot \text{Recall}}{(\text{Prec+Recall})}$$

Note that Precision and Recall are equal to 1 respectively when the estimated sparsity profile is a subset of the true sparsity profile and when the true sparsity profile is a subset of the estimated one. F-measure is equal to 1 when the estimated sparsity profile coincides exactly with the true one.

In the experiments presented below, we consider empirical convergence of the iterative shrinking algorithm when the KKT conditions given in equation (7) are satisfied up to a tolerance of 0.001. For IrM-BP, we have used $r = 1$ and $r = 0.5$ which respectively corresponds to a log penalty and M-FOCUSS with $p = 0.5$. ϵ has been set to 0.01 and we stop our Iterative reweighted M-BP when $\|\mathbf{C}^{(n)} - \mathbf{C}^{(n-1)}\|_\infty \leq 1e^{-5}$ or when the maximal number of fifty iterations are reached. The M-SBL algorithm we used is the one proposed by Wipf et al. [35] and available on his website. We have used the fast EM updates which according to Wipf et al. provides a good trade-off between speed and precision. The M-OMP is the one described by Tropp et al. [32].

B. Illustrating our M-BP and IrM-BP algorithms

This first experimental results aim at illustrating how our M-BP and Ir-MBP algorithms work. As an experimental set-up, we have used $M = 50$, $N = 25$, $L = 3$ and the number k of active elements in the dictionary is equal to 10. λ has been chosen so as to optimize the sparsity profile recovery. Since we just want to illustrate how the algorithm works, we think that such a default value of λ is sufficient for making our point.

Figure 2 respectively plots the variations of the objective value, the row norms $\|c_{i,\cdot}\|$ and the F-measure for our iterative shrinking algorithm. For this example, many iterations are needed for achieving convergence. However, we can note that the objective value decreases rapidly whereas the row-support (middle plot) of $\hat{\mathbf{C}}$ first increases then many of these row norms get shrinked to zero. Following this trend, the F-measure slowly increases before yielding to its maximal value. In this example, we can see that we have more non-zero rows than expected. Figure 3 shows the same plots resulting from the same approximation problem but using Iterative reweighted M-BP with a penalty $J_{\frac{1}{2},2}$. The first iteration corresponds to a single pass of M-BP. After the second iteration, the objective value already seems to have reached its optimal value. However the next iterations still help in shrinking to zero some undesired coefficients and thus in improving sparsity recovery. For this problem, IrM-BP is able to perfectly recover the sparsity profile while M-BP does not.

We have also empirically assessed the computational complexity of our algorithms (we used $r = 1$ for IrM-BP). We varied one of the different parameters (dictionary size M, signal dimensionality N,

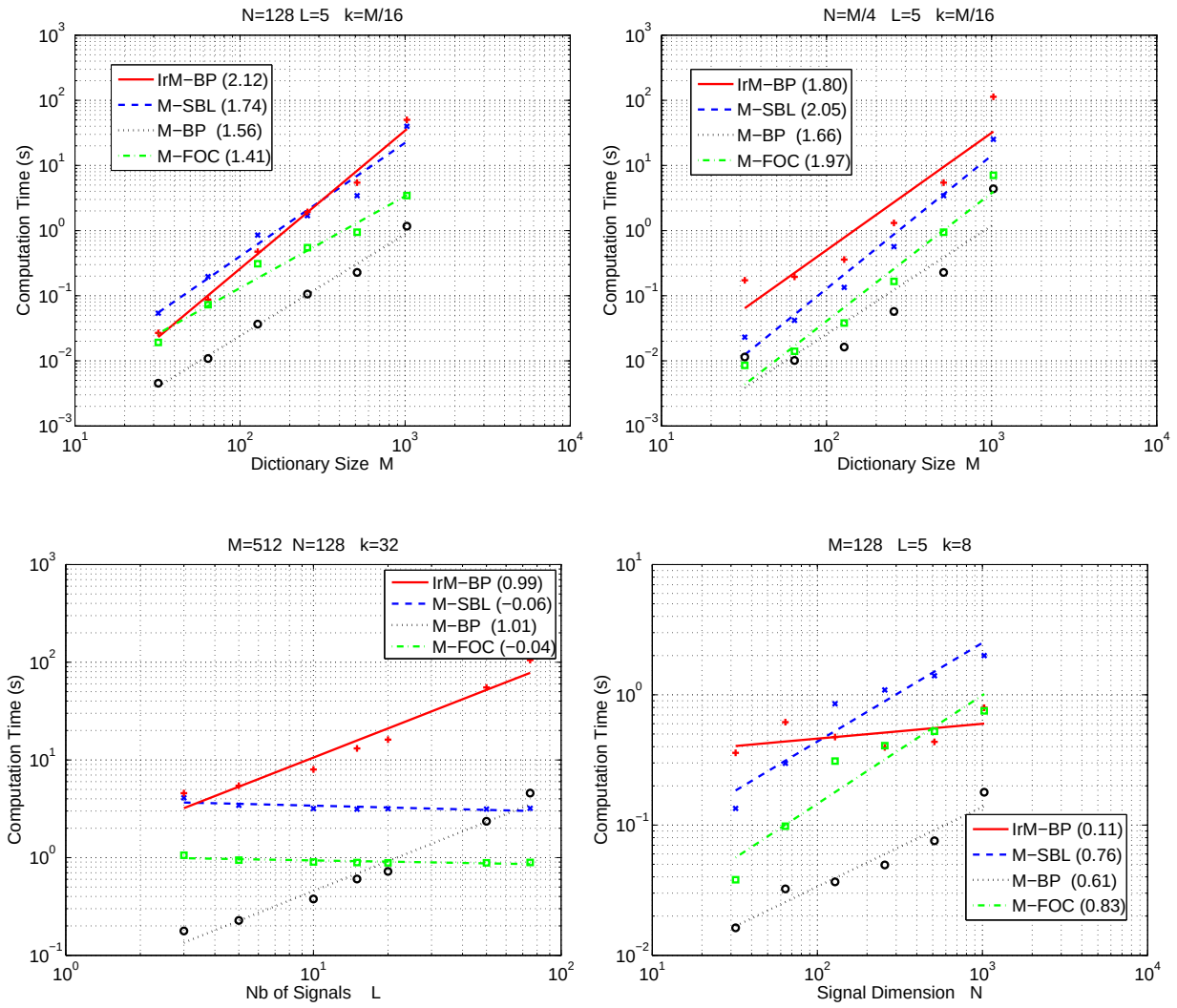

Fig. 4. Estimation of the empirical exponent of the computational complexity of different algorithms (M-BP, IrM-BP, M-SBL and M-FOCUSS). For the two last algorithms, we have used the code available from Wipf's website. The top plots give the computation time of the algorithms with respects to the dictionary size. On the top left, the signal dimension N has been kept fixed and equal to 128 whereas on the right one, the signal dimension is related to the dictionary size. The bottom plots respectively depict the computational complexity with respects to the number of signals to approximate and the dimensionality of these signals.

number of signals L) while keeping the others fixed. All matrices $\mathbf{\Phi}$, $\mathbf{C}$ and $\mathbf{S}$ are created as described above. Experiments have been run on a Pentium D-3 GHz with 4 GB of RAM using Matlab code. The results, averaged over 50 trials, in Figure 4 show the computational complexity of the different

algorithms for different experimental settings. Note that we have also experimented on the M-SBL and M-FOCUSS computational performances owing to the code of Wipf et al. [35]. All algorithms need one hyperparameter to be set, for M-SBL and M-FOCUSS, we were able to choose the optimal one since the hyperparameter is dependent on a known noise level. For our algorithms, the choice of λ is more critical and has been manually set so as to achieve optimal performances. Note that our aim here is not give an exact comparison of computational complexity of the algorithms but just to give an order of magnitude of these complexities. Indeed, careful comparisons are difficult since the different algorithms do not solve the same problem and do not use the same stopping criterion.

We can remark in Figure 4 that with respects to the dictionary size, all algorithms present an empirical exponent between 1.4 and 2.4. Interestingly, we have theoretically evaluate the complexity of M-BP as quadratic whereas we measure a sub-quadratic complexity. We suppose that this happens because at each iteration, only the non-optimal $c_{i,.}$'s are updated and thus the number of updates drastically reduces along iterations. We can note that among all approaches, M-BP is the less demanding algorithm while IrM-BP is the less efficient one. This is clearly the cost to be paid for trading the resolution of a non-convex problem against several convex ones. Note however, that this complexity can be controlled by reducing the number of iterations while preserving good sparsity recovery. This is the case of many weighted Lasso algorithms which use only two iterations [39], [38].

The difference between the two top plots in Figure 4 shows that algorithm complexities not only depend on the dictionary size but also on the redundancy of the dictionary. Indeed, on the right plot, signal dimensionality is related to the dictionary size (redundancy is kept fixed) while on the left plot, the signal size is fixed. This results in a non-uniform variation of the complexities which is difficult to understand. It is not clear if it is related to the problem difficulty or is intrinsic to algorithms. Further researches are still needed to clarify this point.

Bottom left plot of Figure 4 depicts the complexity dependency of all algorithms with respects to the number of signal to approximate. The results we obtain is in agreement with theoretical exponents since for M-BP and IrM-BP we have exponents of approximately 1 while the other algorithm complexities do not depend on L. On the bottom right, we have evaluated these exponents with respects to signal dimension. Here again, we have results in accordance to theoretical expectations : M-BP and IrM-BP have lower complexities than M-SBL and M-FOCUSS. Furthermore, we note that IrM-BP has unexpectedly a very low exponent complexity. We assume that this is due to the fact that as dimension increases, the approximation problem becomes easier and thus needs less M-BP iterations.

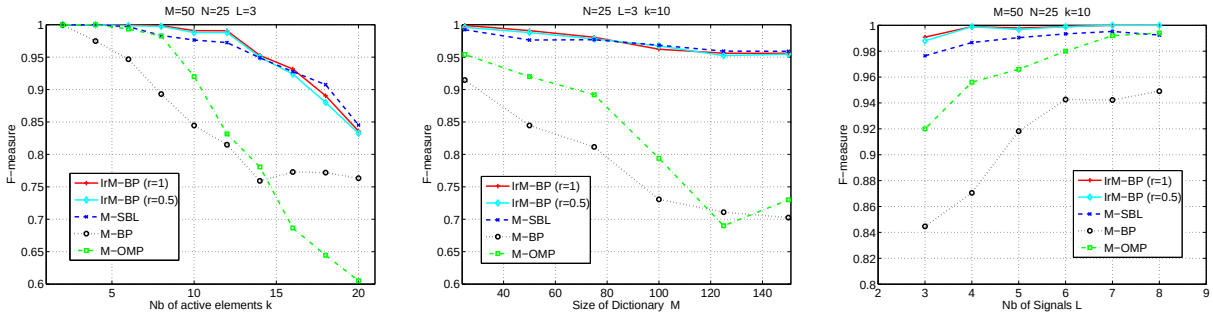

Fig. 5. Results comparing performances of different simultaneous sparse algorithms. from left to right, we have varied the number of generating functions k, the dictionary size M and the number of signal to approximate L. The default setting is $M = 50$, $N = 25$, $k = 10$ and $L = 3$.

C. Comparing performances

The objective of the next empirical study is to compare performances of M-BP, IrM-BP (with $r = 0.5$ and $r = 1$), M-SBL and M-OMP for different experimental situations. The baseline context is $M = 50$, $N = 25$, $k = 10$ and $L = 3$. Note that for the M-OMP, we stop the algorithm after exactly k iterations. For this experiment, we did not performed model selection but instead tried several values of λ and σ and chosen the ones that maximize performances.

Figure 5 shows, from left to right, the performance averaged over 50 trials, on sparsity recovery when k increases from 2 to 20, when M goes from 25 to 150 and when $L = 2, \cdots, 8$. We can note that, M-BP performs worse that IrM-BP. This is a result that we could expected in views of the literature [38], [3] which compare Lasso and reweighted Lasso, the single signal approximation counterpart of M-BP and IrM-BP.

For all experimental situations, we remark that IrM-BP and M-SBL perform equally well. Again, this similar performances can easily be understood because of the strong relation between reweighted M-BP and M-SBL as explained in Subsection IV-C. When considering M-OMP, although we suppose that k is known, we can see that the M-OMP performance is not as good as those of M-SBL and IrM-BP.

VI. Conclusions and perspectives

This paper aimed at contributing to simultaneous sparse signal approximation problems on several points. Firstly, we have proposed an algorithm for solving the multiple signal counterpart of Basis Pursuit

Denoising named M-BP. The algorithm we introduced is rather efficient and simple and it is based on a soft-threshold operator which only needs matrix multiplications. Then, we have considered the more general non-convex M-FOCUSS problem for which M-BP is a special case. We have shown that M-FOCUSS can also been understood as an ARD approach. Indeed, we have transformed the M-FOCUSS penalty in order to exhibit some weights that automatically influence the importance of each dictionary elements in the approximation. Finally, we have introduced an iterative reweighted M-BP algorithm for solving M-FOCUSS. We also made clear the relationship between M-SBL and such a reweighted algorithm. We also provided some experimental results that show how our algorithms behave and how they compare to other methods dedicated to simultaneous sparse approximation. In terms of performances for sparsity profile recovery, our algorithms does not necessarily perform better than others approaches but they are provided with interesting features such as convexity and convergence guarantees.

Owing to this clear formulation of the problem and its numerically reproducible solution (due to convexity), our perspective on this work is now to theoretically investigate the properties of the M-BP and IrM-BP solutions. We believe that the recent works on the Lasso and related methods can be extended in order to make clear in which situations M-BP and Ir-MBP achieve consistency. Further improvements of algorithm speed can also be interesting so that tackling very large-scale approximation becomes tractable.

ACKNOWLEDGMENTS

This work was partly supported by the KernSig project grant from the Agence Nationale de la Recherche.

VII. APPENDIX

A. Proof of Lemma 2

By definition, a matrix $\mathbf{G}$ lies in $\partial J_{1,2}(\mathbf{B})$ if and only if for every matrix $\mathbf{Z}$, we have

$$J_{1,2}(\mathbf{Z}) \geq J_{1,2}(\mathbf{B}) + \langle \mathbf{Z} - \mathbf{B}, \mathbf{G} \rangle_F \tag{25}$$

If we expand this equation we have the following equivalent expression

$$\sum_i \|z_{i,\cdot}\|_2 \geq \sum_i \|b_{i,\cdot}\|_2 + \sum_i \langle z_{i,\cdot} - b_{i,\cdot}, g_{i,\cdot} \rangle \tag{26}$$

From this latter equation, we understand that, since both $J_{1,2}$ and the Frobenius inner product are row-separable, a matrix $\mathbf{G} \in \partial J_{1,2}(\mathbf{B})$ if and only if each row of $\mathbf{G}$ belongs to the subdifferential of the ℓ_2 norm of the corresponding row of $\mathbf{B}$.

Indeed, suppose that $\mathbf{G}$ is so that any row of $\mathbf{G}$ belongs to the subdifferential of the ℓ_2 norm of the corresponding row of $\mathbf{B}$. We thus have for any row i

$$\forall \mathbf{z}, \quad \|\mathbf{z}\|_2 \geq \|b_{i,\cdot}\|_2 + \langle z - b_{i,\cdot}, g_{i,\cdot} \rangle \tag{27}$$

A summation over all the rows then proves that $\mathbf{G}$ satisfies equation (26) and thus belongs to the subdifferential of $J_{1,2}(\mathbf{B})$.

Now, let us show that a matrix $\mathbf{G}$ for which there exists a row that does not belong to the subdifferential of the ℓ_2 norm of the corresponding row of $\mathbf{B}$ can not belong to the subdifferential of $J_{1,2}(\mathbf{B})$. Let us consider $g_{i,\cdot}$ the i-th row of $\mathbf{G}$, since we have supposed that $g_{i,\cdot} \notin \partial \|b_{i,\cdot}\|_2$, the following equation holds

$$\exists \mathbf{z}_0 \text{ st.} \quad \|\mathbf{z}_0\|_2 < \|b_{i,\cdot}\|_2 + \langle \mathbf{z} - b_{i,\cdot}, g_i \rangle$$

Now let us construct $\mathbf{Z}$ so that $\mathbf{Z} = \mathbf{B}$ except for the i-th row where $z_{i,\cdot} = \mathbf{z}_0$. Then it is easy to show that this matrix $\mathbf{Z}$ does not satisfy equation (26), which means that $\mathbf{G}$ does not belong to $\partial J_{1,2}(\mathbf{B})$. In conclusion, we get $\partial J_{1,2}(\mathbf{B})$ by applying the ℓ_2 norm subdifferential to each row of $\mathbf{B}$. And it is well known [1] that

$$\partial \|\mathbf{b}\|_2 = \begin{cases} \{\mathbf{g} \in \mathbb{R}^L : \|\mathbf{g}\|_2 \leq 1\} & \text{if } \mathbf{b} = \mathbf{0} \\ \frac{\mathbf{b}}{\|\mathbf{b}\|_2} & \text{otherwise} \end{cases} \tag{28}$$

B. Proof of equation (22)

We want to show that at optimality which occurs at $\mathbf{C}^\star$, we have

$$\mathbf{s}_j^t \Sigma_t^{-1} \mathbf{s}_j = \frac{1}{\sigma^2} \mathbf{s}_j^t (\mathbf{s}_j - \mathbf{\Phi} \mathbf{C}^\star)$$

which is equivalent, after factorizing with $\mathbf{s}^t$, to show that

$$\sigma^2 \mathbf{s}_j = \Sigma_t \mathbf{s}_j - \Sigma_t \mathbf{\Phi} \mathbf{C}^\star$$

This last equation can be proved using simple algebra

$$\begin{aligned} \Sigma_t \mathbf{s}_j - \Sigma_t \mathbf{\Phi} \mathbf{C} &= \sigma^2 \mathbf{s}_j + \mathbf{\Phi} \mathbf{D} \mathbf{\Phi}^t \mathbf{s} - (\sigma^2 I + \mathbf{\Phi} D \mathbf{\Phi}^t) \mathbf{\Phi} \mathbf{C}^\star \\ &= \sigma^2 \mathbf{s}_j + \mathbf{\Phi} \mathbf{D} \mathbf{\Phi}^t \mathbf{s} - \mathbf{\Phi} (\sigma^2 I + D \mathbf{\Phi}^t \mathbf{\Phi}) \mathbf{C}^\star \\ &= \sigma^2 \mathbf{s}_j + \mathbf{\Phi} \mathbf{D} \mathbf{\Phi}^t \mathbf{s} - \mathbf{\Phi} \mathbf{D} \mathbf{\Phi}^t \mathbf{s} \\ &= \sigma^2 \mathbf{s}_j \end{aligned}$$

REFERENCES

[1] D. Bertsekas, A. Nedic, and A. Ozdaglar, *Convex Analysis and Optimization*. Athena Scientific, 2003.

[2] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[3] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Analysis and Applications*, vol. To appear, 2008.

[4] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *33rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[5] J. Chen and X. Huo, "Sparse representations for multiple measurements vectors (mmv) in an overcomplete dictionary," in *Proc IEEE Int. Conf Acoustics, Speech Signal Processing*, vol. 4, 2005, pp. 257–260.

[6] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal Scientific Comput.*, vol. 20, no. 1, pp. 33–61, 1999.

[7] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[8] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley, 2001.

[9] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1-norm minimization," *Proceedings of the National Academy of Sciences USA*, vol. 1005, pp. 2197–2202, 2002.

[10] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression (with discussion)," *Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[11] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[12] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[13] M. Girolami and S. Rogers, "Hierarchic bayesian models for kernel learning," in *Proc. of 22nd International Conference on Machine Learning*, 2005, pp. 241–248.

[14] I. Gorodnitsky, J. George, and B. Rao, "Neuromagnetic source imaging with FOCUSS : a recursive weighted minimum norm algorithm," *J. Electroencephalogr. Clin. Neurophysiol.*, vol. 95, no. 4, pp. 231–251, 1995.

[15] Y. Grandvalet, "Least absolute shrinkage is equivalent to quadratic penalization," in *ICANN'98*, ser. Perspectives in Neural Computing, L. Niklasson, M. Bodén, and T. Ziemske, Eds., vol. 1. Springer, 1998, pp. 201–206.

[16] Y. Grandvalet and S. Canu, "Adaptive scaling for feature selection in svms," in *Advances in Neural Information Processing Systems*, vol. 15. MIT Press, 2003.

[17] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms," IRISA N1848, Tech. Rep., 2007.

[18] D. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, pp. 30–37, 2004.

[19] S. Ji, D. Dunson, and L. Carin, "Multi-task compressive sensing," *IEEE Trans. Signal Processing, to appear*, 2008.

[20] Z. Luo, M. Gaspar, J. Liu, and A. Swami, "Distributed signal processing in sensor networks," *IEEE Signal Processing magazine*, vol. 23, no. 4, pp. 14–15, 2006.

[21] D. Malioutov, M. Cetin, and A. Willsky, "Sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 3010–3022, 2005.

[22] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[23] C. Micchelli and M. Pontil, "Learning the kernel function via regularization," *Journal of Machine Learning Research*, vol. 6, pp. 1099–1125, 2005.

[24] C. Phillips, J. Mattout, M. Rugg, P. Maquet, and K. Friston, "An empirical Bayesian solution to the source reconstruction problem in EEG," *NeuroImage*, vol. 24, pp. 997–1011, 2005.

[25] Y. Qi, T. Minka, R. Picard, and Z. Ghahramani., "Predictive Automatic Relevance Determination by Expectation Propagation," in *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[26] A. Rakotomamonjy, F. Bach, Y. Grandvalet, and S. Canu, "SimpleMKL," *Journal of Machine Learning Research, to appear*, 2008.

[27] R. Tibshirani, "Regression selection and shrinkage via the lasso," *Journal of the Royal Statistical Society*, pp. 267–288, 1995.

[28] ——, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 46, pp. 431–439, 1996.

[29] M. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[30] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Info. Theory*, vol. 51, no. 3, pp. 1030–1051, 2006.

[31] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[32] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Journal of Signal Processing*, vol. 86, pp. 572–588, 2006.

[33] P. Tseng, "Convergence of block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Application*, vol. 109, pp. 475–494, 2001.

[34] D. Wipf and S. Nagarajan, "A new view of automatic relevance determination,," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2008, vol. 20.

[35] D. Wipf and B. Rao, "An empirical bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans on Signal Processing*, vol. 55, no. 7, pp. 3704–3716, July 2007.

[36] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of Royal Statisctics Society B*, vol. 68, pp. 49–67, 2006.

[37] H. Zhou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistics Society Ser. B*, vol. 67, pp. 301–320, 2005.

[38] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[39] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *The Annals of Statistics*, vol. 36, no. 4, pp. 1509–1533, 2008.