Surveying and comparing simultaneous sparse approximation (or group lasso) 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 over-complete dictionary of elementary functions, with a joint sparsity profile induced by a $\ell_p - \ell_q$ mixed-norm. Our contributions are essentially two-fold i) making connections between such an approach and other methods available in the literature and ii) on providing algorithms for solving the problem with different values of p and q. At first, we introduce a simple algorithm for solving the multiple signals extension of the Basis Pursuit Denoising problem ($p = 1$ and $q = 2$). Then, we show that for general sparsity-inducing $\ell_p - \ell_q$ mixed-norm penalty, this optimization problem is actually equivalent to an automatic relevance determination problem. From this insight, we derive an simple EM-like algorithm for problems with $\ell_1 - \ell_{q\leq 2}$ penalty. For addressing approximation problem with non-convex penalty ($p < 1$), we propose an iterative reweighted Multiple-Basis Pursuit ; we trade the non-convexity of the problem against several resolutions of the convex multiple-basis pursuit problem. Relations between such a reweighted algorithm and the Multiple-Sparse Bayesian Learning are also highlighted. Experimental results show how our algorithms behave and how they compare to related approaches (such as CosAmp) 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

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

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 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. [7] 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 [37], [50]. 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 [11], [49].

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 [35], neuroelectromagnetic imaging [25], [40] and source localization [36].

A. Problem formalization

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

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

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 :

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

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 $\mathbf{C}$ and $c_{i,j}$ is the ith element in the jth column of $\mathbf{C}$.

For the simultaneous sparse 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 criterion 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 synthesizing 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 simultaneous sparse 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 the 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 addressing problems (2) and (3) are possible : relaxing the problem 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 encompassed 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 $\mathbf{C} = 0$ [16]).

Several authors have proposed methods for solving problem (4). For instance, Cotter et al. [8] 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 proved 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 attention. Indeed, it seems to be the most natural extension of the so-called Lasso problem [46] or Basis Pursuit Denoising [7], 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. [36] have proposed an algorithm based on a second-order cone programming formulation for solving the 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. [6]. In this case, the simultaneous sparse problem can be decoupled in L independent problems. In such a situation, estimations of the L true signals are no more guaranteed to have the same sparsity profile, thus the problem can hardly be considered as a simultaneous sparse approximation problem. However, in this case, one can use efficient algorithms that solve the well-known *Lasso* problem [46], [12], [18].

Another important piece of work belonging to the framework of convex relaxation is the one of Tropp [48]. In this latter work, Tropp proposed to relax problem (2), (3) and (4) by replacing $\|\mathbf{C}\|_0$ with a

$J_{1,\infty}(\mathbf{C})$ penalty. He then analyzed the theoretical properties of the different problem formulations and provided some conditions under which that convex relaxation produces good solutions.

Very recently, Baraniuk et al. [1] have extended the CosAmp algorithm of Needell et al. [39] in order to address simultaneous sparse approximation using the penalty $J_{0,2}(\mathbf{C})$. Besides providing a computationally efficient approach, they also provide theoretical guarantees on estimation robustness. In the sequel, we denote the algorithm of Baraniuk et al. as M-CosAmp.

The approach proposed by Wipf et al. [54], denoted in the sequel as Multiple-Sparse Bayesian Learning (M-SBL), 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. [47]. 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 [31]. However, the main drawback of this latter approach is that due to its greedy nature, and as any EM algorithm where the objective function is not convex, the algorithm can be easily stuck 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 S-OMP in the sequel [51], are a simple extension of the well-known Matching Pursuit technique to simultaneous approximation.

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 [28], [51].

Due to the recent interest around compressed sensing, the number of works addressing simultaneous sparse approximation problem have flourished. Many of these papers consider exactly sparse signal recovery which is not the case we are interested in since we assume that the signals we have are noisy. Nonetheless, they are of great interest since most of those papers propose theoretically guaranteed recovery schemes [38], [15], [14].

C. Our contributions

At the present time, one of the most interesting approach for simultaneous sparse approximation is the Bayesian approach introduced by Wipf et al. [54] and further improved by Ji et al. [31] 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, ... [32], [43], [56]) 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. [8] and Malioutov et al. [36] in considering using ℓ_p minimization problem for simultaneous sparse approximation, but propose a different way of solving the minimization problem (4). Our contributions in this paper are essentially on novel algorithms for solving that problem for different values of p and q and on some novel insights on its connection with different existing approaches for simultaneous sparse approximation.

At first, we develop a simple and efficient algorithm for solving the M-Basis Pursuit problem ($p = 1$ and $q = 2$). We show that by using results from non-smooth optimization theory, we are able to propose an block-coordinate descent method which only needs some matrix multiplications. In this sense, this first contribution is strongly related to the work of Sardy et al. [44] for single-signal approximation and thus it can be understood as an extension of their block-coordinate approach to simultaneous approximation. A proof of convergence and a discussion with related works such as those of Fornasier et al. [20], Elad et al. [13] and Sardy et al. [44] are also provided.

Then, we focus on the more general situation where $p > 0$ and $q \leq 2$ in $J_{p,q}(\mathbf{C})$. We show that such a row-diversity measure is actually related to automatic relevance determination (ARD). Indeed, we show that for any $p > 0$ and $q \leq 2$, $J_{p,q}$ can be interpreted as a weighted 2-norm row-measure and these

weights evaluate the relevance of a given entry of the matrix $\mathbf{C}$. The equivalence between $J_{p,q}(\mathbf{C})$ and the ARD needs the weights to be optimized according to certain constraints (which for instance induce some row-sparsity of $\mathbf{C}$). Owing to this insight, we clarify the relation between M-FOCUSS and the Multiple-Sparse Bayesian Learning of Wipf et al. [54] (which also uses ARD) for any value of $p > 0$. Then, from this ARD formulation, we derive an Iterative Reweighted Least-Square algorithm, which has the flavor of M-FOCUSS, for solving problem with $J_{1,1\leq q\leq 2}$. To the best of our knowledge, this is the first algorithm for sparse simultaneous approximation with $J_{1,1<q<2}(\mathbf{C})$.

Afterward, instead of directly deriving a proper algorithm for solving the non-convex optimization problem when $p < 1$ and $1 \leq q \leq 2$, we propose an iterative reweighted algorithm which reuses our algorithms that solve problem (4) with penalty $J_{1,1\leq q\leq 2}(\mathbf{C})$. Using a Majorize-Minimize optimization framework, we show that depending on the chosen weights, such an iterative reweighted scheme can actually solve problem (4). Our main contribution at this point is to have translated the non-convex problem (4) into a serie of convex problems. In this sense, it can be considered as an extension of the work of Candès et al. [4], Foucart et al. [21] and Gasso et al. [23] to simultaneous sparse approximation.

Furthermore, by choosing a different weighting scheme, we show that our iterative reweighted approach is strongly related to M-SBL. An experimental comparison of these two (and other) algorithms will make clear the benefits and disadvantages of using our iterative reweighted algorithm. For instance, experimental results show that although our iterative reweighted approach is slower than the M-CosAmp of Baraniuk et al. [1], it is able to provide better estimation of the coefficient matrix $\mathbf{C}$.

The paper is organized as follows. Section II introduces and discusses our block-coordinate algorithm for solving the M-BP problem. Section III deals with the equivalence between ARD and $J_{p,q}(\mathbf{C})$ optimization. An algorithm which addresses the general case where $p = 1$ and $1 \leq q \leq 2$ is then proposed. Then, the iterative reweighted algorithm for addressing the optimization problem when $p < 1$ is described and discussed in Section IV. 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 that the specific structure of the problem leads to a very simple block coordinate descent algorithm named as M-BCD. We also show that if the dictionary is under-complete then it can be proved that the solution of the problem is equivalent to a simple shrinkage of the coefficient

matrix. Proof of convergence of our M-BCD algorithm is also given in this section

Before delving into the algorithmic details, we should note that block-coordinate descent have already been considered by Sardy et al. [44] and Elad [13] for sparse signal approximations from orthogonal and redundant representations. Hence the numerical scheme we propose here can be seen as an extension of their works to vector-valued data.

A. Deriving optimality conditions

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 $W(\mathbf{C})$ [2]. By computing the subdifferential of $W(\mathbf{C})$ with respect to each row $c_{i,\cdot}$ of $\mathbf{C}$, the KKT 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 KKT optimality conditions 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

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

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 $\mathbf{C}$.

B. The algorithm and its convergence

Our block-coordinate descent algorithm is detailed in Algorithm (1). It is a simple and efficient algorithm for solving M-BP.

Basically, the idea consists in solving each row $c_{i\cdot}$ at a time. By starting from a sparse solution like, $\mathbf{C} = 0$, at each iteration, we check for a given i whether row $c_{i,\cdot}$ is optimal or not based on conditions (7). If not, $c_{i,\cdot}$ is then updated according to equation (9).

Although, such a block-coordinate algorithm does not converge in general for non-smooth optimization problem, Tseng [52] 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. Our proof of convergence is based on such properties and follows the same line as the one proposed by Sardy et al. [44].

Theorem 1: The M-BCD algorithm converges to a solution of the M-Basis Pursuit problem given in Equation (6), where convergence is understood as any accumulation point of the M-BCD algorithm is a minimum of problem (6) and the sequence of $\{\mathbf{C}_k\}$ generated by the algorithm is bounded.

Proof: Note that M-BP problem presents a particular structure with a smooth and differentiable convex function $\|\mathbf{S} - \mathbf{\Phi}\mathbf{C}\|_F^2$ and a row-separable penalty function $\sum_i h_i(c_{i,\cdot})$ where $h(\cdot)$ is a continuous and convex function with respects to $c_{i,\cdot}$.

Also note that our algorithm considers a cyclic rule where within each loop, for each $i \in [1, \cdots, M]$, each $c_{i,\cdot}$ is considered for optimization. The main particularity is that for some i, the $c_{i,\cdot}$ may be left unchanged by the block-coordinate descent if already optimal. This occurs especially for row $c_{i,\cdot}$ which are equal to 0.

Then according to the special structure of the problem and the use of a cyclic rule, the results of Tseng [52] prove that our M-BCD algorithm converges. ■

Algorithm 1 Solving M-BP through block-coordinate descent

$\mathbf{C} = 0$, Loop = 1
while Loop **do**
  for $i = 1, 2, \cdots, M$ **do**
    Compute $\|\mathbf{r}_i\|$
    if optimality condition of $c_{i,\cdot}$ according to equations (7) is not satisfied **then**
      $c_{i,\cdot} = \left(1 - \frac{\lambda}{\|T_i\|}\right)_+ T_i$
    end if
  end for
  if all optimality conditions are satisfied **then**
    Loop = 0
  end if
end while

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 block-coordinate descent 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.

C. Some relations with other works

As we have already stated, our M-BCD algorithm can be considered as an extension to simultaneous signal approximations of the works of Sardy et al. [44] and Elad [13]. However, here, we want to emphasize the importance of starting from a $\mathbf{C} = 0$. Indeed, since in the estimated $\hat{\mathbf{C}}$ is expected to be sparse, by doing so, only few updates are needed before convergence.

In addition to the works of Sardy et al. and Elad, many others authors have considered block-coordinate descent algorithm for related sparse approximation problems. For instance, it has also been used for solving the Lasso [22], and the elastic net [55]. Other works have also considered iterative thresholding algorithms for solving single signal sparse approximation problem [9], [19].

For recovering vector valued data with joint sparsity constraints, Fornasier et al. [20] have proposed an extension of the Landweber iterative approach of Daubechies et al. [9]. In their work, Fornasier et al. have also used an iterative shrinking algorithm (which has the flavor of a gradient projection approach) which is able to solve the general problem (4) with $p = 1$ and $q = \{1, 2, \infty\}$. For $q = 2$, the main difference between their algorithm and the one we propose here is that, by optimizing at each loop, only the $c_{i,\cdot}$'s that are not optimal yet, we have an algorithm that is more efficient than the one of Fornasier et al.

As we stated previously, the M-FOCUSS algorithm also solves the M-BP problem. In their M-FOCUSS approach, Cotter et al. [8] 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 undesirable point when solving a convex problem. At the contrary, our M-BCD algorithm does not suffer from the presence of such fixed-points. However, such fixed-points in the M-FOCUSS algorithm can be handled by introducing a smoothing term ε in the weight so that the updated weight (according to Cotter's notation and for $p = 1$) becomes

$$\mathbf{W} = \text{diag}\left(\sqrt{\|c_{i,\cdot}\| + \epsilon}\right)$$

where $\mathbf{W}$ is the diagonal weighting matrix and $\varepsilon > 0$. The use of ε avoids a given weight to be at zero and consequently it avoids the related $c_{i,\cdot}$ to stay permanently at zero. Then if we furthermore note that

M-FOCUSS is not more than an iterative reweighted least-square. According to the very recent works of Daubechies et al. [10] and Chartrand et al. [5], it seems justified to iterate the M-FOCUSS algorithm using decreasing value of ε. In our numerical experimental, we will consider the M-FOCUSS algorithm with fixed and decreasing value of ε.

D. Evaluating complexity

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. Although, this may not be relevant since the number of iterations needed for the considered algorithms to converge may be very different, such knowledge give an hint about the algorithm scaling with respects to parameters of the simultaneous sparse approximation problem.

For our M-BCD algorithm, we can note that each shrinking operation, in the worst case scenario, has to be done M times and the dominating cost for each update is the computation of T_i. This computation involves the matrix multiplication $\mathbf{\Phi C}_{-i}$ 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 [36] 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 evaluations are given in section V.

III. ARD Formulation of Simultaneous Sparse Approximation

In this section, we focus on the relaxed optimization problem given in (4) with the general penalization $J_{p,q}(\mathbf{C})$. 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 [54]. We will show that for a set of values of p and q, the mixed-norm $J_{p,q}(\mathbf{C})$ has an equivalent variational formulation. Then by using this novel formulation in problem (4), instead of $J_{p,q}(\mathbf{C})$, we exhibit the relation between our sparse approximation problem and ARD. We then propose an iterative reweighted least-square approach for solving the resulting ARD problem.

A. Exhibiting the relation with ARD

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

$$\min_{\mathbf{C}} \frac{1}{2}\|\mathbf{S}-\mathbf{\Phi C}\|_F^2 + \lambda' \left(J_{p,q}(\mathbf{C})\right)^{\frac{2}{p}}. \tag{10}$$

In the convex case (for $p \geq 1$ and $q \geq 1$), since the power function is strictly monotonically increasing, problems (4) and (10) are equivalent, in the sense that for a given value λ', there exists a λ so that solution of the two problems are equal. When $J_{p,q}$ is not convex, this equivalence does not strictly apply. However, due to the nature of the problem, the problem formulation (10) is more convenient for exhibiting the relation with ARD.

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 $\ell_{p,q}$ norm of a sequence $\{a_{t,k}\}$.

Lemma 2: if $s > 0$ and $\{a_{t,k} : k \in \mathbb{N}_n, t \in \mathbb{N}_T\} \in \mathbb{R}$ such that at least one $a_{t,k} > 0$, then

$$\min_{\mathbf{d}} \left\{ \sum_{t,k} \frac{|a_{t,k}|^2}{d_{t,k}} : d_{t,k} \geq 0, \sum_k \left(\sum_t d_{t,k}^{1/s} \right)^{\frac{s}{r+s}} \leq 1 \right\} = \left(\sum_k \left(\sum_t |a_{t,k}|^q \right)^{\frac{p}{q}} \right)^{\frac{2}{p}} \tag{11}$$

where $q = \frac{2}{s+1}$ and $p = \frac{2}{s+r+1}$. Furthermore, at optimality, we have:

$$d_{t,k}^\star = \frac{|a_{t,k}|^{\frac{2s}{s+1}} \left(\sum_u |a_{u,k}|^{\frac{2}{s+1}} \right)^{\frac{r}{s+r+1}}}{\left(\sum_v \left(\sum_u |a_{u,v}|^{\frac{2}{s+1}} \right)^{\frac{s+1}{s+r+1}} \right)^{r+s}} \tag{12}$$

Proof : See Appendix.

According to this lemma, the $\ell_{p,q}$ norm of a sequence can be computed through a minimization problem. Hence, applying this lemma to $\left(J_{p,q}(\mathbf{C})\right)^{\frac{2}{p}}$ by defining $a_{t,k} = c_{t,k}$, we get a variational form of the penalization term. We can also note that the mixed-norm on the matrix coefficients has been transformed to a mixed-norm on weight matrix $\mathbf{d}$.

Then plugging the above variational formulation of the penalization term in problem (10) yields to the following equivalent problem :

$$\begin{array}{ll} \min\limits_{\mathbf{C},\mathbf{d}} & \frac{1}{2}\|\mathbf{S}-\mathbf{\Phi C}\|_F^2 + \lambda \sum_{t,k} \frac{c_{t,k}^2}{d_{t,k}} \\ \text{s.t.} & \sum_k \left(\sum_t d_{t,k}^{1/s} \right)^{\frac{s}{r+s}} \leq 1 \\ & d_{t,k} \geq 0 \qquad \forall t,k \end{array} \tag{13}$$

This problem is the one which makes clear the automatic relevance determination interpretation of the original formulation (4). Indeed, we have transformed problem (4) into a problem with a smooth objective

	(r,s)		
	$(0,1)$	$(1,0)$	$(\frac{1}{2},\frac{1}{2})$
$\mathbf{d}$	$\ell_{1,1}$	$\ell_{1,\infty}$	$\ell_{1,2}$
$\mathbf{C}$	$\ell_{1,1}$	$\ell_{1,2}$	$\ell_{1,\frac{4}{3}}$

TABLE I

EQUIVALENCE BETWEEN MIXED-NORM ON $\mathbf{d}$ AND $\mathbf{C}$ FOR DIFFERENT VALUES OF r AND s.

function at the expense of some additional variables $d_{t,k}$. These parameters $d_{t,k}$ actually aim at determining the relevance of each element of $\mathbf{C}$. Indeed, in the objective function, each squared-value $c_{t,k}$ is now inversely weighted by a coefficient $d_{t,k}$. 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 $d_{t,k} = 0$ for $c_{t,k}^2 = 0$. Then the smaller $d_{t,k}$ is, the smaller the $c_{t,k}$ norm should be. Furthermore, optimization problem (13) also involves some constraints on $\{d_{t,k}\}$. These constraints impose the matrix $\mathbf{d}$ to have positive elements and to be so that its $\ell_{\frac{1}{r+s},\frac{1}{s}}$ mixed-norm is smaller than 1. Note that this mixed-norm on $\mathbf{d}$ plays an important role since it induces the row-norm sparsity on $\mathbf{C}$. According to the relation between p, r and s, for $p < 1$, we also have $r+s > 1$, making the $\ell_{\frac{1}{r+s},\frac{1}{s}}$ non-differentiable with respect to the first norm. Such singularities favor row-norm sparsity of the matrix $\mathbf{d}$ at optimality, inducing row-norm sparsity of $\mathbf{C}$. As we have noted above, when a row-norm of $\mathbf{d}$ is equal to 0, the corresponding row-norm of $\mathbf{C}$ should also be equal to 0 which means that the corresponding element of the dictionary is "irrelevant" for the approximation of all signals. Problem (13) proposes an equivalent formulation of problem (4) for which the row-diversity measure has been transformed in another penalty function owing to an ARD formulation. The trade-off between convexity of the problem and the sparsity of the solution has been transferred from p, q to r and s. Table I gives some examples of equivalence between the two mixed-norms on $\mathbf{d}$ and $\mathbf{C}$.

From a Bayesian perspective, we can interpret the row-norm on $\mathbf{d}$ as the diagonal term of the covariance matrix of a Gaussian prior over the row-norm on $\mathbf{C}$ distribution. This is typically the classical Bayesian Automatic Relevance Determination approach as proposed for instance in the following works [41], [47]. This novel insight on the ARD interpretation of $J_{p,q}(\mathbf{C})$ clarifies the connection between the M-FOCUSS algorithm of Cotter et al. [8] and the M-SBL algorithm of Wipf et al. [54] 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

Algorithm 2 Iterative Reweighted Least-Square for addressing $J_{1,1\leq q2}$ penalty

Initialize $\mathbf{d}^{(0)}$ to a strictly positive matrix
$t = 1$
while Loop **do**
  $\mathbf{C}^{(t)} \leftarrow$ solution of problem (14) with fixed $\mathbf{d} = \mathbf{d}^{(t-1)}$ as given by Equation (15)
  $\mathbf{d}^{(t)} \leftarrow$ solution of problem (14) with fixed $\mathbf{C} = \mathbf{C}^{(t)}$ as given by Equation (16)
  $t \leftarrow t + 1$
  if stopping condition is satisfied **then**
    Loop = 0
  end if
end while

linear models and SVM [26], [27].

B. Solving the ARD formulation for $p = 1$ and $1 \leq q \leq 2$

Here, we propose a simple iterative algorithm for solving problem (13) for $p = 1$ and $1 \leq q \leq 2$. Our algorithm, named as M-EM$_q$, is based on an iterative-reweighted least squares where the weights are updated according to equation (12). Thus, it can be seen as an extension of the M-FOCUSS algorithm of Cotter et al. for $q \leq 2$. Note that we have restricted ourselves to $p = 1$ since we will show in the next section that $p < 1$ can be handled using another reweighted scheme.

Since $p = 1$, thus $s + r = 1$, the problem we are considering is :

$$\begin{array}{ll} \min\limits_{\mathbf{C},\mathbf{d}} & \sum_k \frac{1}{2}\left(\|\mathbf{s}_k - \mathbf{\Phi} c_{.,k}\|_2^2 + \lambda \sum_t \frac{c_{t,k}^2}{d_{t,k}}\right) = J_{obj}(\mathbf{C},\mathbf{d}) \\ \text{s.t.} & \sum_k \left(\sum_t d_{t,k}^{1/s}\right)^s \leq 1 \\ & d_{t,k} \geq 0 \end{array} \tag{14}$$

Since, we consider that $1 \leq q \leq 2$ hence $0 \leq s \leq 1$ [1], this optimization problem is convex with a smooth objective function. We propose to address this problem through a block-coordinate algorithm which alternatively solves the problem with respects to $\mathbf{C}$ with the weight $\mathbf{d}$ being fixed and then keeping $\mathbf{C}$ fixed and computes the optimal weight $\mathbf{d}$. The resulting algorithm is detailed in Algorithm 2.

[1]for $s = 0$, we have explicitly used the sup norm of vector $d_{.,k}$ in the constraints.

Owing to the problem structure, step 4 and 5 of this algorithm has a simple closed form. Indeed, for fixed $\mathbf{d}$, each vector $c_{\cdot,k}^{(t)}$ at iteration t is given by :

$$c_{\cdot,k}^{(t)} = \left(\mathbf{\Phi}^t\mathbf{\Phi} + 2\lambda\mathbf{D}_k^{(t-1)}\right)^{-1}\mathbf{\Phi}^t\mathbf{s}_k \tag{15}$$

where $\mathbf{D}_k^{(t-1)}$ is a diagonal matrix of entries $d_{\cdot,k}^{(t-1)}$. In a similar way, for fixed $\mathbf{C}$, step 5 boils down in solving problem (11). Hence, by defining $a_{t,k} = c_{t,k}^{(t)}$, we also have a closed-form for $\mathbf{d}^{(t)}$ as

$$d_{t,k}^{(t)} = \frac{|a_{t,k}|^{\frac{2s}{s+1}}\left(\sum_u |a_{u,k}|^{\frac{2}{s+1}}\right)^{\frac{1-s}{2}}}{\sum_v\left(\sum_u |a_{u,v}|^{\frac{2}{s+1}}\right)^{\frac{s+1}{2}}} \tag{16}$$

Note that similarly to the M-FOCUSS algorithm, this algorithm can also be seen as an iterative reweighted least-square approach or as an Expectation-Minimization algorithm, where the weights are defined in equation (16). Furthermore, it can be shown that if the weights $\mathbf{d}$ are initialized to non-zero values then at each loop involving step 4 and 5, the objective value of problem (14) decreases. Hence, since the problem is convex, our algorithm should converge towards the global minimum of the problem.

Theorem 2: If the objective value of problem (14) is strictly convex (for instance when $\mathbf{\Phi}$ is full-rank), and if for the t-th loop, after the step 5, we have $\mathbf{d}^{(t)} \neq \mathbf{d}^{(t-1)}$, then the objective value has decreased, i.e :

$$J_{obj}(\mathbf{C}^{(t+1)}, \mathbf{d}^{(t)}) < J_{obj}(\mathbf{C}^{(t)}, \mathbf{d}^{(t)}) < J_{obj}(\mathbf{C}^{(t)}, \mathbf{d}^{(t-1)}).$$

Proof : The right inequality $J_{obj}(\mathbf{C}^{(t)}, \mathbf{d}^{(t)}) < J_{obj}(\mathbf{C}^{(t)}, \mathbf{d}^{(t-1)})$ comes from $\mathbf{d}^{(t)}$ being the optimal value of the optimization problem resulting from step 5 of algorithm (2). The strict inequality yields from the hypothesis that $\mathbf{d}^{(t)} \neq \mathbf{d}^{(t-1)}$ and from the strict convexity of the objective function. A similar reasoning allows us to derive the left inequality. Indeed, since $\mathbf{C}^{(t)}$ is not optimal with respects to $\mathbf{d}^{(t)}$ for the problem given by step (4), invoking the strict convexity of the associated objective function and optimality of $\mathbf{C}^{(t+1)}$ concludes the proof.

As stated by the above theorem, the decrease in objective value is actually guaranteed unless, the algorithm get stuck in some fixed points (*e.g* all the elements of $\mathbf{d}$ being zero expect for one entry $\{t_1, k_1\}$). In practice, we have experienced, by comparing for $q = 2$ with the M-BCD algorithm, that if $\mathbf{d}$ is initialized to non-zero entries, algorithm (2) converges to the global minimum of problem (14). Numerical experiments will illustrate this point.

IV. Reweighted M-Basis Pursuit

This section introduces an iterative reweighted M-Basis Pursuit (IrM-BP) algorithm and proposes two ways for 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 $1 \leq q \leq 2$. The second weighting scheme makes clear the strong relation between the empirical bayesian strategy of Wipf et al. [54] 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 [56], [4], [5]. 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 $\mathbf{c}^\star$ of the problem. 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 also lead to sparser solutions than the classical M-Basis Pursuit.

Our iterative reweighted M-Basis Pursuit algorithm is defined as follows. We iteratively construct a sequence $\mathbf{C}^{(m)}$ defined as

$$\mathbf{C}^{(m)} = \arg\min_{\mathbf{C}} \frac{1}{2}\|\mathbf{S} - \mathbf{\Phi C}\|_F^2 + \lambda \sum_i z_i^{(m)} \|c_{i,\cdot}\|_q \tag{17}$$

where the positive weight vector $\mathbf{z}^{(m)}$ depends on the previous iterate $\mathbf{C}^{(m-1)}$. For $m = 1$, we typically define $\mathbf{z}^{(1)} = 1$ and for $m > 1$, in our case, we will consider the following weighting scheme

$$z_i^{(m)} = \frac{1}{(\|c_{i,\cdot}^{(m-1)}\|_q + \varepsilon)^r} \quad \forall i \tag{18}$$

where $\{c_{i,\cdot}^{(m-1)}\}$ is the i-th row of $\mathbf{C}^{(m-1)}$, r a user-defined positive constant and ε a small regularization term that prevents from having an infinite regularization term for $c_{i,\cdot}$as soon as $c_{i,\cdot}^{(m-1)}$ vanishes. This is a classical trick that has been used for instance by Candès et al. [4] or Chartrand et al. [43]. Note that for any positive weight vector $\mathbf{z}$, problem (17) is a convex problem that does not present local minima. Furthermore, for $1 \leq q \leq 2$, it can be solved by our block-coordinate descent algorithm or by our M-EM$_q$

given in Algorithm 2, by simply replacing λ with $\lambda_i = \lambda \cdot z_i$. This reweighting scheme we propose is similar to the *adaptive lasso* algorithm of Zou et al. [56] but uses more than two iterations and addresses the simultaneous approximation problem.

B. Connections with Majorization-Minimization algorithm

The IrM-BP algorithm we proposed above can also be interpreted as an algorithm for solving an approximation of problem (4) when $0 < p < 1$ and $1 \leq q \leq 2$. Indeed, similarly to the reweighted ℓ_1 scheme of Candès et al. [4] or the one-step reweighted lasso of Zou et al. [57], our algorithm falls in the class of majorize-minimize (MM) algorithms [30]. 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. Let us first define $J_{p,q,\varepsilon}(\mathbf{C})$ as an approximation of the penalty term $J_{p,q}(\mathbf{C})$:

$$J_{p,q,\varepsilon}(\mathbf{C}) = \sum_i g\left(\|c_{i,\cdot}\|_q + \varepsilon\right)$$

where $g(\cdot) = |\cdot|^p$. Since $g(\cdot)$ is concave for $0 < p < 1$, a linear approximation of $J_{p,q,\varepsilon}(\mathbf{C})$ around $\mathbf{C}^{(m-1)}$ yields to the following majorizing inequality

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

then for the minimization step, replacing in problem (4) $J_{p,q}$ with the right part of the inequality and dropping constant terms lead to our optimization problem (17) with appropriately chosen z_i and r. Note that for the weights given in equation (18), $r = 1$ corresponds to the linearization of a log penalty $\sum_i \log(\|c_{i,\cdot}\| + \varepsilon)$ whereas setting $r = 1 - p$ corresponds to a ℓ_p penalty ($0 < p < 1$).

MM algorithms have already been considered in optimization problems with sparsity-inducing penalties. For instance, an MM approach have been used by Figueiredo et al. [17] for solving a least-square problem with a ℓ_p sparsity-inducing penalty, whereas Candès et al. [4] have addressed the problem for exact sparse signal recovery. In a context of simultaneous approximation, Simila [45] has also considered MM algorithms while approximating the non-convex penalty with a quadratic term. Hence our contribution here can be considered as an extension of their works to simultaneous sparse approximation using a mixed $\ell_p - \ell_q$ norm where at each iteration, this norm has been linearly approximated.

Algorithm 3 Reweighted M-Basis Pursuit with annealing setting of ε

1: Initialize $\varepsilon = 1$, $\mathbf{z} = 1$
2: **while** $\varepsilon > \varepsilon_{min}$ **do**
3: $(\hat{\mathbf{C}}, \hat{\mathbf{z}}) \leftarrow$ solution of problem (17) using $\mathbf{z}$
4: Update $\mathbf{z}$ according to equation (18) and $\hat{\mathbf{C}}$
5: $\varepsilon \leftarrow \varepsilon/10$
6: **end while**

Analyzing the convergence of the sequence $\mathbf{C}^{(m)}$ towards the global minimizer of problem (4) is a challenging issue. Indeed, several points make a formal proof of convergence difficult. At first, in order to avoid a row $c_{i,\cdot}^{(m)}$ to be permanently at zero, we have introduced a smoothing term ε, thus we are only solving a ε-approximation of problem (4). Furthermore, the penalty we use is non-convex, thus using a monotonic algorithm like a MM approach which decreases the objective value at each iteration, can not guarantee convergence to the global minimum of our ε-approximate problem. Hence, due to these two major obstacles, we have left the convergence proof for future works. Note however that few works have addressed the convergence issue of reweighted ℓ_1 or ℓ_2 algorithms for single sparse signal recovery. Notably, we can mention the recent work of Daubechies et al. [10] which provide a convergence proof of iterative reweighted least square for exact sparse recovery. In the same flavor, Foucart et al. [21] have proposed a tentative of rigorous convergence proof for reweighted ℓ_1 sparse signal recovery. Although, we do not have any rigorous proof of convergence, in practice, we will show that our reweighted algorithm provides good sparse approximations.

As already noted by several authors [4], [43], [10], ε plays a major role in the quality of the solution. In the experimental results presented below, we have investigated two methods for setting ε : the first one is to set it to a fixed value $\varepsilon = 0.001$, the other one is denoted as an annealing approach which consists in gradually decreasing ε after having solved problem (17). This annealing approach is detailed in Algorithm (3).

C. Relation with M-SBL

Recently, Wipf et al. [53] have proposed some new insights on Automatic Relevance Determination and 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. [24] or Rakotomamonjy et al. [42] although some connections with simultaneous sparse approximation can be made. Here, we build on the works on Wipf et al. [53] 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. [54] is

$$\mathcal{L}(\mathbf{d}) = L \log |\Sigma_t| + \sum_{j=1}^{L} \mathbf{s}_j^t \Sigma_t^{-1} \mathbf{s}_j \tag{19}$$

where $\Sigma_t = \sigma^2 \mathbf{I} + \mathbf{\Phi} \mathbf{D} \mathbf{\Phi}^t$ and $\mathbf{D} = \text{diag}(\mathbf{d})$, with $\mathbf{d}$ being a vector of hyperparameters that govern the prior variance of each coefficient matrix row. 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 [3]

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

Thus, the cost function $\mathcal{L}(\mathbf{d})$ in equation (19) can then be upper-bounded by

$$\mathcal{L}(\mathbf{d}, \mathbf{z}) \triangleq \mathbf{z}^t \mathbf{d} - L g^\star \left(\frac{\mathbf{z}}{L} \right) + \sum_{j=1}^{L} \mathbf{s}_j^t \Sigma_t^{-1} \mathbf{s}_j \tag{20}$$

Hence when optimized over all its parameters, $\mathcal{L}(\mathbf{d}, \mathbf{z})$ converges to a local minima or a saddle point of (19). 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 $\mathbf{z}^\star$

$$\mathbf{z}^\star = L \cdot \nabla \log |\Sigma_t| (\mathbf{d}^\dagger) = \text{diag}(\mathbf{\Phi}^t \Sigma_t^{-1} \mathbf{\Phi}) \tag{21}$$

Similarly to what proposed by Wipf et al., Equations (20) and (21) 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

$$\mathbf{d}^\dagger = \arg\min_{\mathbf{d}} \mathcal{L}_z(\mathbf{d}) \triangleq \mathbf{z}^t \mathbf{d} + \sum_{j=1}^{L} \mathbf{s}_j^t \Sigma_t^{-1} \mathbf{s}_j \tag{22}$$

then to minimize $\mathcal{L}(\mathbf{d}^\dagger, \mathbf{z})$ for fixed $\mathbf{d}^\dagger$, which can be analytically done according to equation (21). 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 an 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 (22) is convex and can be equivalently solved by computing

$$\mathbf{C}^{\star} = \arg\min_{\mathbf{C}} \mathcal{L}_z(\mathbf{C}) = \frac{1}{2}\|\mathbf{S} - \mathbf{\Phi}\mathbf{C}\|_F^2 + \sigma^2 \sum_i z_i^{1/2} \|c_{i,\cdot}\| \tag{23}$$

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 (22) is straightforward since it is just a sum of convex functions [3]. The key point of the proof is based on the equality

$$\mathbf{s}_j^t \Sigma_t^{-1} \mathbf{s}_j = \frac{1}{\sigma^2} \min_{c_{\cdot,j}} \|\mathbf{s}_j - \mathbf{\Phi} c_{\cdot,j}\|_2^2 + \sum_i \frac{c_{i,j}^2}{d_i} \tag{24}$$

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

$$\mathcal{L}_z(\mathbf{d}, \mathbf{C}) = \mathbf{z}^t \mathbf{d} + \sum_j \frac{1}{\sigma^2} \|\mathbf{s}_j - \mathbf{\Phi} c_{\cdot,j}\|_2^2 + \sum_{i,j} \frac{c_{i,j}^2}{d_i} \tag{25}$$

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 (25) and multiplying the the resulting objective function with $\sigma^2/2$ yields to

$$\mathcal{L}_z(\mathbf{C}) = \frac{1}{2} \sum_j \|\mathbf{s}_j - \mathbf{\Phi} c_{\cdot,j}\|_2^2 + \sigma^2 \sum_i z_i^{1/2} \|c_{i,\cdot}\| \tag{26}$$

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

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 (21) and (22) equivalent to our iterative reweighted M-BP for which weights z_i would be given by equation (21).

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 described here, can again be viewed as an application of MM approach for solving problem (19). Indeed, we are actually iteratively minimizing a proxy function which has been obtained by majorizing each term of equation (19). Owing to this MM point of view, convergence of our iterative algorithm towards a local minimum of equation (19) is guaranteed [30]. Convergence for the single signal case using other arguments has also been shown by Wipf et al. [53]. 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 (19). 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 set-up

In order to quantify the performance of our algorithms and compare them to other approaches, we have used simulated datasets with different redundancy factors $\frac{M}{N}$, number k of active elements and number L of signals to approximate. The dictionary $\mathbf{\Phi}$ is based on M vectors sampled from the unit hypersphere of $\mathbb{R}^N$. The true coefficient matrix $\mathbf{C}^\star$ has been obtained as follows. The positions of the k non-zero rows in the matrix are randomly drawn. The non-zero coefficients of $\mathbf{C}^\star$ 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. For a given experiment, when several trials are needed, we only resample the dictionary $\mathbf{\Phi}$ and the additive noise $\mathcal{E}$.

Each algorithm is provided with the signal matrix $\mathbf{S}$ and the dictionary $\mathbf{\Phi}$ and will output an estimate of $\mathbf{C}$. The performance criterion we have considered are the mean-square error between the true and the approximate signals and the sparsity profile of the coefficient matrix that has been recovered. For the latter, we use as a performance criterion the F-measure between the row-support of the true matrix $\mathbf{C}^\star$ and the estimate one $\hat{\mathbf{C}}$. In order to take into account numerical precisions, we have overloaded the row support definition as :

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

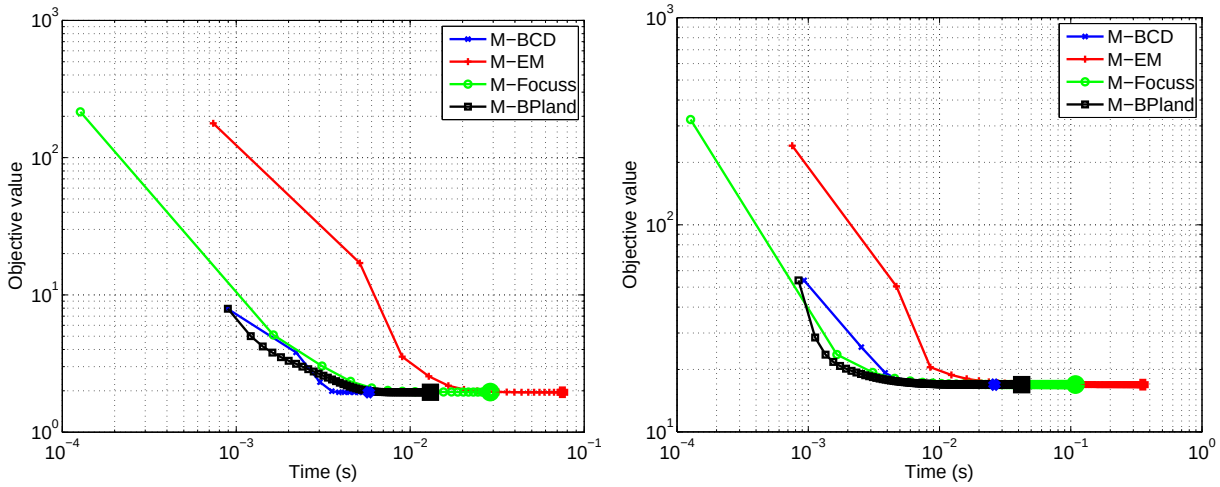

Fig. 1. Examples of objective value evolution with respects to computational time. Here we have, $M = 128$, $N = 64$, $L = 3$. The number of active elements is : left) $k = 5$. right) $k = 32$. For each curve, the large point corresponds to the objective value at convergence.

where μ is a threshold coefficient that has been set by default to 0.01 in our experiments. From rowsupp$(\hat{\mathbf{C}})$ and rowsupp$(\mathbf{C}^\star)$ respectively the estimated and true sparsity profile, we define :

$$\text{F-measure} = 2 \cdot \frac{|\text{rowsupp}(\hat{\mathbf{C}}) \cap \text{rowsupp}(\mathbf{C}^\star)|}{|\text{rowsupp}(\hat{\mathbf{C}})| + |\text{rowsupp}(\mathbf{C}^\star)|}.$$

Note that the F-measure is equal to 1 when the estimated sparsity profile coincides exactly with the true one.

Regarding the stopping criterion, in the experiments presented below, we have considered convergence of our M-BCD algorithm when the optimality conditions given in equation (7) are satisfied up to a tolerance of 0.001 and when all matrix coefficient $c_{i,j}$ variations are smaller than 0.001. This latter condition has also been used as a stopping criterion for our M-EM and IrM-BP algorithms.

B. Comparing $\ell_1 - \ell_2$ M-BP problem solvers

In this first experiment, we have compared different algorithms which solves the M-BP problem with $p = 1$ and $q = 2$. Besides our M-BCD and M-EM algorithms, we have also used the M-FOCUSS of Cotter et al. [8] and the approach of Fornasier et al. [20] based on Landweber iterations and denoted in the sequel as M-BPland. Note that for M-FOCUSS, we have modified the genuine algorithm by introducing a ε parameter, set to 0.001, which helps in avoiding a row-norm of $\mathbf{C}$ to be permanently at 0.

TABLE II

SUMMARY OF M-BP SOLVERS COMPARISON. COMPARISONS HAVE BEEN CARRIED OUT FOR TWO VALUES OF k, THE NUMBER OF ACTIVE ELEMENTS IN THE DICTIONARY AND HAVE BEEN AVERAGED OVER 100 TRIALS. COMPARISON MEASURES ARE THE TIME NEEDED BEFORE CONVERGENCE, THE DIFFERENCE IN OBJECTIVE VALUE AND THE LARGEST COEFFICIENT MATRIX DIFFERENCE. FOR THE TWO LATTER MEASURE, THE BASELINE ALGORITHM IS CONSIDERED TO BE THE M-BCD ONE.

	k=5			k=32		
	Time (ms)	Δ ObjVal (10^{-3})	$\|\Delta\mathbf{C}\|_\infty(10^{-3})$	Time (ms)	Δ ObjVal (10^{-3})	$\|\Delta\mathbf{C}\|_\infty(10^{-3})$
M-BCD	6.90 ± 3.13	-	-	29.2 ± 8.6	-	-
M-EM	58.87 ± 13.8	1.01 ± 0.36	2.54 ± 1.63	158.8 ± 7.1	8.02 ± 3.2	19.2 ± 5.3
M-Focuss	38.47 ± 9.97	9.75 ± 2.22	4.51 ± 1.57	74.6 ± 19.2	17.56 ± 3.2	25.3 ± 5.1
M-BPland	13.69 ± 3.62	0.04 ± 1.18	5.63 ± 1.19	24.7 ± 5.1	1.09 ± 6.9	31.1 ± 6.9

Figure 1 shows two examples of how the objective value of the different algorithms evolves with respects to computational time. We can note that the two iterative reweighted least-square algorithms (M-EM and M-FOCUSS) are the most computationally demanding. Furthermore, we also see that the Landweber iteration approach of Fornasier et al. quickly reduces its objective value but compared to our M-BCD method, it needs more time before properly converging. Table II summarizes more accurately the difference between the four algorithms. As comparison criteria, we have considered the computational time before convergence, the difference (compared to our M-BCD algorithm) in objective values and the maximal absolute difference in the coefficient matrix $c_{i,j}$. The table clearly shows that our M-BCD algorithm is clearly faster than M-BPland and the two iterative reweighted least-square approaches. We can also note from the table that, although M-FOCUSS and our M-EM are not provided with a formal convergence proof, these two algorithms seems to empirically converge to the problem global minimum.

C. Illustrating our M-BCD and IrM-BP algorithms

This other experiment illustrates the behavior of our M-BCD and Ir-MBP algorithms. As an experimental set-up, we have used $M = 128$, $N = 64$, $L = 3$ 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 recovered by our M-BCD algorithm. Since we just want to illustrate how the algorithms work, 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

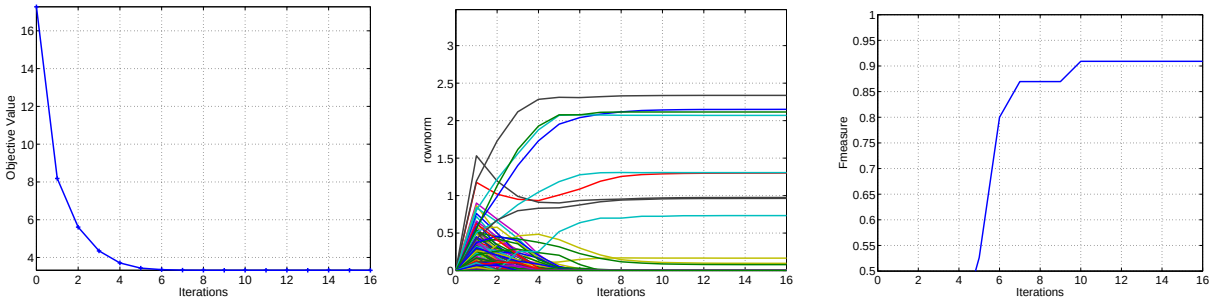

Fig. 2. Illustration of our block coordinate descent algorithm for solving M-BP. Example of variation along the iterations of : left) Objective value, middle) row-norm $\|c_{i,.}\|$, right) F-measure. For this example, the dictionary size is 128 while 10 active elements have been considered in the true sparsity profile.

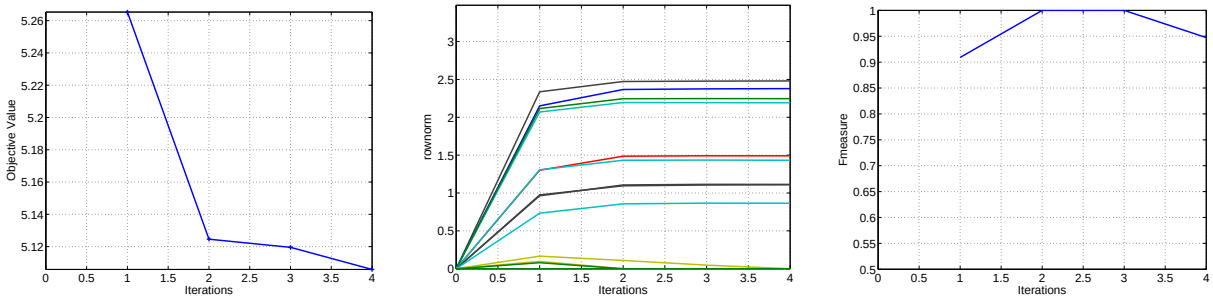

Fig. 3. Illustration of the iterative reweighted M-BP applied for $J_{\frac{1}{2},2}$ penalty. Example of variation along the iterations of : left) Objective value, middle) row-norm $\|c_{i,.}\|$, right) F-measure

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 shrunken 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 an Iterative reweighted M-BP with a penalty $J_{\frac{1}{2},2}$. The first iteration corresponds to a single pass of M-BCD. The next iterations still help in shrinking to zero some coefficients and thus in improving the sparsity profile of the estimate $\hat{\mathbf{C}}$ although some true non-zero row-norms have also been filtered out. For this problem, both algorithms are not able to perfectly recover the true sparsity profile, although for another value of λ, the Ir-MBP algorithm would.

Figure 4 compares the true row-norm of $\mathbf{C}$ with the ones obtained with our algorithms. On the left

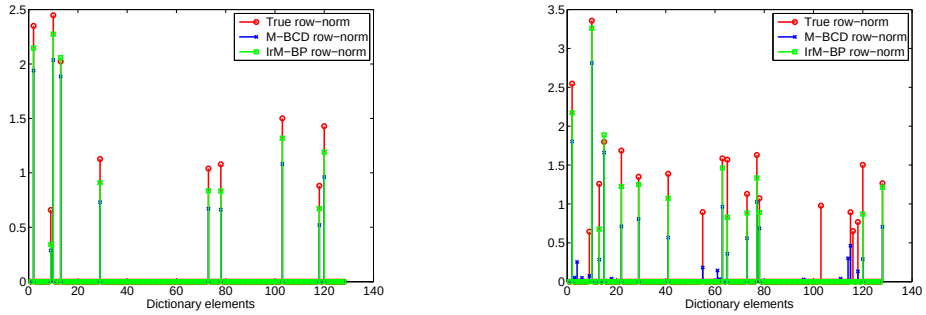

Fig. 4. Estimated row-norm obtained from the M-BCD and Ir-BCD (with $p = 0.5$) algorithms. left) k=10. right) k=20.

panel, the number of active elements is equal to 10 and we can see that both algorithms (with the same value of λ are able to recover the exact sparsity profile of $\mathbf{C}$. We can note that the concavity of the penalty yields to a better estimation of the row-norm values. The right panel illustrates an example, for $k = 20$, where the M-BCD algorithm tends to produce a solution which returns undesired non-zeros row-norms whereas the Ir-MBP approach tends to shrink to zero some true non-zero rows.

D. Computational performances

We have also empirically assessed the computational complexity of our algorithms (we used $s = 0.2$, thus $q = \frac{5}{3}$ for M-EM and $r = 1$ for IrM-BP). We varied one of the different parameters (dictionary size M, signal dimensionality N) 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 in Figure 5, averaged over 20 trials, 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. [54] and have implemented the CosAmp block-sparse approach of Baraniuk et al. [1] and the Landweber iteration method of Fornasier et al. [20] [2]. All algorithms need one hyperparameter to be set, for M-SBL and CosAmp, we were able to choose the optimal one since the hyperparameter respectively depends on a known noise level and a known number of active elements in the dictionary. For other algorithms, we have reported the computational complexity for the λ that yields to the best sparsity recovery. Note that our aim here is not give an exact comparison of computational complexity of the algorithms but just to give an order of

[2]All the implementations are included in the toolbox.

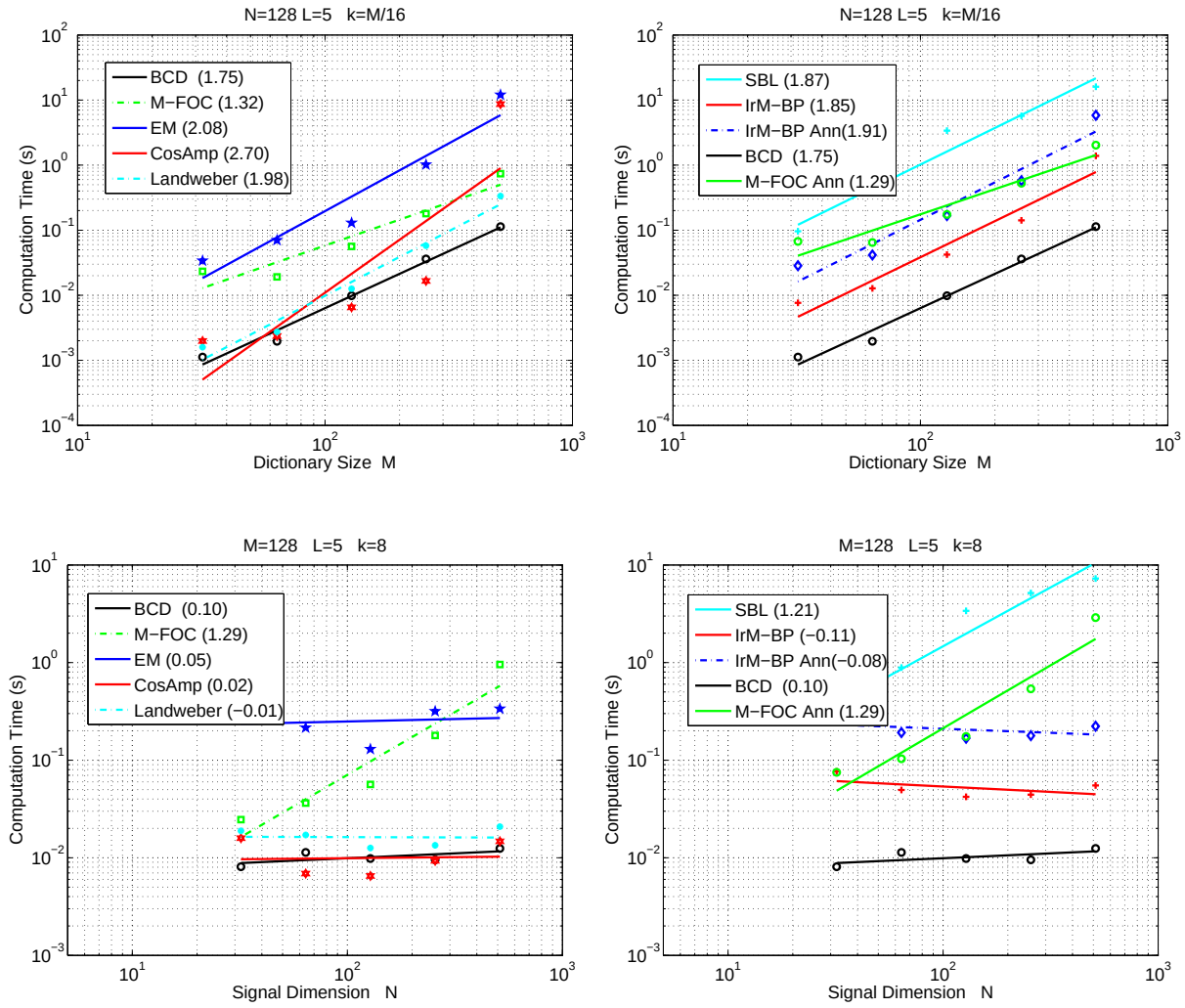

Fig. 5. Estimating the empirical exponent, given in parenthesis, of the computational complexity of different algorithms (M-BCD, IrM-BP, M-SBL, M-FOCUSS, CosAmp, Landweber iterations). The top plots give the computation time of the algorithms with respects to the dictionary size. The bottom plots respectively depict the computational complexity with respects to the signal dimensionality. For a sake of readability, we have separated the algorithms in two groups :(left) the ones that solve $\ell_1 - \ell_q$ problem. (right) the ones that solve $\ell_p - \ell_2$ problem (M-BCD result provided for baseline comparison). The "IrM-BP Ann" and "M-FOC Ann" refers to the Ir-MBP and M-FOCUSS algorithm using an annealing approach for iteratively decreasing ε as described in Algorithm (3).

magnitude of these complexities. Indeed, accurate 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 5 that with respects to the dictionary size, all algorithms present an empirical

exponent between 1.3 and 2.7. Interestingly, we have theoretically evaluate the complexity of our M-BCD algorithm 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 that solve the $\ell_1 - \ell_q$ problem (left plots), M-BCD, Landweber iteration approach and M-CosAmp have similar complexity with a slight advantage to M-BCD for large dictionary size. However, we have to note that the M-CosAmp algorithm sometimes suffers from lack of convergence and thus stop only when the maximal number of allowed iterations is reached. This is the reason why for large dictionary size CosAmp is computationally expensive. When considering the algorithms that solve the $\ell_p - \ell_2$ problem (right plots), they all have similar complexity, with a slightly better constant for IrM-BP while M-SBL seems to be the most demanding algorithm.

Bottom plots of Figure 5 depicts the complexity dependency of all algorithms with respects to signal dimension N. Interestingly, the results show that except for M-SBL and M-FOCUSS algorithms, all algorithms do not suffer from the signal dimension increase. We assume that this is due to the fact that as dimension increases, the approximation problem becomes easier and thus faster convergence of those algorithms occurs.

E. Comparing performances

The objective of the next empirical study is to compare the performances of the algorithms we propose with some of those proposed in the literature (M-SBL, CosAmp, Landweber iterations, S-OMP and M-FOCUSS with an annealing decreasing of ε). From our side, we have considered only our M-BCD algorithm and our IrM-BP with two values of p and an annealing decrease of ε.

The baseline experimental context is $M = 128$, $N = 64$, $k = 10$ and $L = 3$. For this experiment, we have considered an agnostic context with no prior knowledge about the noise level being available. Hence, for all models, we have performed model selection (either for selecting λ, the noise level σ for M-SBL or the number of elements for M-CosAmp and S-OMP). Model selection procedure is the following. Training signals $\mathbf{S}$ are randomly splitted in two parts of $N/2$ samples. Each algorithm is then trained on one part of the signal and the mean-square error of the resulting model is evaluated on the second part. This splitting and training is run 5 times and the hyperparameter yielding to the minimal averaged mean-square error is considered as optimal. Each method is then run on the full signals with that parameter. Performances, averaged over 50 trials of all methods have been evaluated according to the F-measure and a mean-square error computed on 10000 samples.

Figure 6 shows, from top to bottom, these performances when k increases from 2 to 40, when M goes

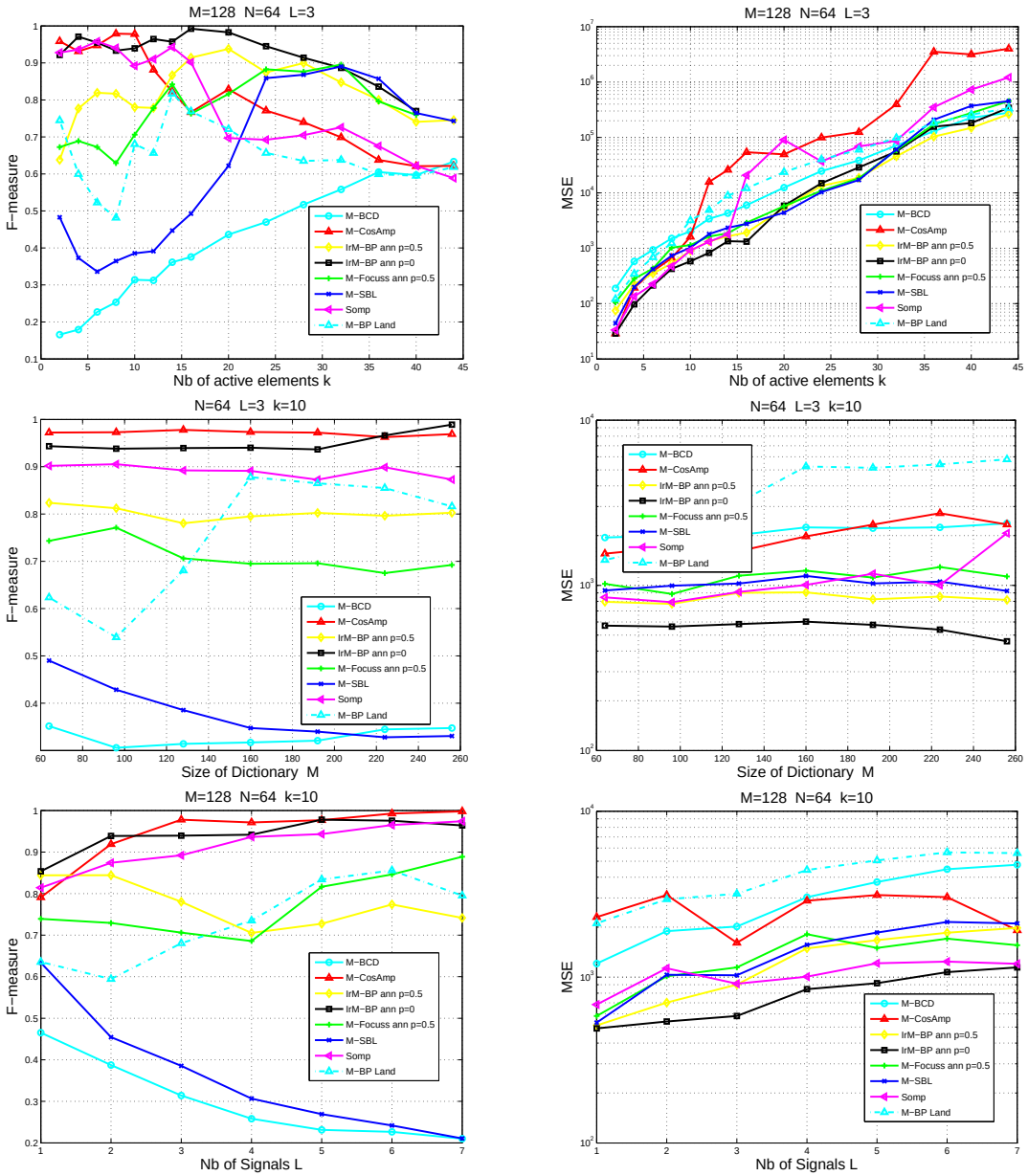

Fig. 6. Results comparing performances of different simultaneous sparse algorithms. We have varied (top) the number k of active elements in the dictionary. (middle) the dictionary size M and (bottom) the number of signal to approximate L. On the left columns are given the F-measure of all methods while the average mean-square errors are on the right column.

from 64 to 256 and when $L = 2, \cdots, 7$. When varying k, we can note that across the range of variation, our IrM-BP method with $p = 0$ is competitive compared to all other approaches both with respects to the F-measure and the mean-square error criterion. When k increases, IrM-BP and M-FOCUSS with $p = 0.5$

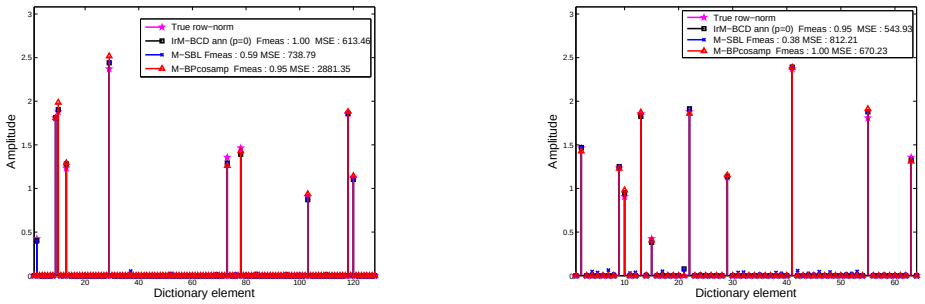

Fig. 7. Examples of estimated row-norm using 3 different algorithms. left) $M = 128$, $N = 64$, $k = 10$ and $L = 3$. right) $M = 64$, $N = 64$, $k = 10$ and $L = 3$. Here, we want to illustrate cases where a "good" sparsity recovery does not necessary lead to low mean-square error.

perform also very good. This may be explained by the fact that as k increases, the optimal solution becomes less and less sparse thus the need for a less aggressive penalty. CosAmp and S-OMP are very competitive for small k but as soon as the latter increases these two methods are not able anymore to recover a "reasonable" sparsity pattern. Interestingly, we remark that M-SBL yields to a poor sparsity recovery measure while the resulting model achieves good mean-square error. A reason for this is that the model selection procedure tends to under-estimate the noise level and thus it leads to a model which keeps many spurious dictionary elements as illustrated in Figure 7 and detailed in the sequel. From Figure 6, we can also notice that the two M-BP solvers, our M-BCD and the Landweber iteration approach perform poorly compared to other methods. However, the Fornasier's method seems to be less sensitive to noise and model selection since it provides a better sparsity pattern recovery. It is worth noting that M-SBL and these two latter methods always correctly select all the true dictionary elements but they also have the tendency to include other spurious ones.

In the middle and bottom plots, similar behavior as above can be highlighted. M-CosAmp yields to very sparsity recovery while the resulting mean-square error is rather poor. Again our IrM-BP with $p = 0$ yields the best mean-square error while providing a good sparsity pattern recovery. M-SBL and M-BCD keeps too many spurious dictionary elements. All other methods provide in-between performances both in term of F-measure and mean-square error.

Figure 7 illustrates the behaviour of M-CosAmp, M-SBL and our IrM-BP with $p = 0$ for two different experimental situations. On the left plot, we have a case where on one hand, M-CosAmp misses to recover the first active dictionary element yielding thus to high mean-square error. On the other hand, M-SBL achieves lower mean-square error while keeping few spurious dictionary elements in the model.

In the meantime, IrM-BP recovers perfectly the sparsity pattern and yields to low mean-square error. In the right plot, we have another case where M-CosAmp achieves perfect sparsity recovery but provides a model with higher mean-square error than IrM-BP.

In most of the experimental situations presented here, M-CosAmp and our IrM-BP seems to be the two algorithms that perform the best with however, a clear advantage for our IrM-BP. Nonetheless, these two methods are actually related since both approaches solve a simultaneous sparse approximation with a $J_{0,2}(\mathbf{C})$ penalty. The main difference lies in the algorithms since our IrM-BP owing to the ε term provides a smooth approximation of the ℓ_0 quasi-norm whereas M-CosAmp directly solves the approximation problem with the $J_{0,2}(\mathbf{C})$ penalty.

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-BCD. The algorithm we introduced is simple and efficient. It is based on a block-coordinate descent algorithm which only needs matrix multiplications. Then, we have considered the more general non-convex approximation problem with penalty $J_{p\geq 0,q\leq 2}(\mathbf{C})$ for which M-BP is a special case. We have shown that such a problem can also be understood as an ARD problem. Afterward, for addressing this ARD optimization problem, we derived an algorithm similar to M-FOCUSS which can handle any $q \in [1,2]$.

Finally, we have introduced an iterative reweighted M-BP algorithm for addressing the non-convex optimization problem with penalty $J_{p<1,1\leq q\leq 2}(\mathbf{C})$. We also made clear the relationship between M-SBL and such a reweighted algorithm. We 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, the experimental results show that our algorithms are provided with interesting features such as a better ability to recover the joint signal sparsity profile and a better estimation of the regression coefficients.

Owing to this formulation of the simultaneous sparse approximation 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 problem as well as the statistical properties of IrM-BP solutions. We believe that the recent theoretical advance on the Lasso and related methods can be extended in order to make clear in which situations M-BP and IrM-BP achieve consistency or better consistency compared to a single signal approximation. Recent works have investigated theoretical properties of related problems [29], [33],

[34] and we plan to contribute to such efforts in the context of simultaneous sparse approximation.

Further improvements on algorithm speed-up would also be interesting so that tackling very large-scale approximation problem may become 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 1

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

$$J_{1,2}(\mathbf{Z}) \geq J_{1,2}(\mathbf{B}) + \langle \mathbf{Z} - \mathbf{B}, \mathbf{G} \rangle_F \tag{27}$$

If we expand this equation we have the following equivalent expression

$$\sum_i \|z_{i,\cdot}\|_2 \geq \sum_i \|b_{i,\cdot}\|_2 + \sum_i \langle z_{i,\cdot} - b_{i,\cdot}, g_{i,\cdot} \rangle \tag{28}$$

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

$$\forall \mathbf{z}, \quad \|\mathbf{z}\|_2 \geq \|b_{i,\cdot}\|_2 + \langle z - b_{i,\cdot}, g_{i,\cdot} \rangle \tag{29}$$

A summation over all the rows then proves that $\mathbf{G}$ satisfies equation (28) 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 (28), 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 [2] that

$$\partial\|\mathbf{b}\|_2 = \begin{cases} \{\mathbf{g} \in \mathbb{R}^L : \|\mathbf{g}\|_2 \leq 1\} & \text{if } \mathbf{b} = \mathbf{0} \\ \frac{\mathbf{b}}{\|\mathbf{b}\|_2} & \text{otherwise} \end{cases} \tag{30}$$

B. Proof of Lemma 2

We aim at proving that

$$\min_{\mathbf{d}} \left\{ \sum_{t,k} \frac{|a_{t,k}|^2}{d_{t,k}} : d_{t,k} \geq 0, \sum_k \left(\sum_t d_{t,k}^{1/s} \right)^{\frac{s}{r+s}} \leq 1 \right\} = \left(\sum_k \left(\sum_t |a_{t,k}|^q \right)^{\frac{p}{q}} \right)^{\frac{2}{p}}$$

where $q = \frac{2}{s+1}$ and $p = \frac{2}{s+r+1}$. The proof proceeds by writing the Lagrangian of the optimization problem :

$$\mathcal{L} = \sum_{t,k} \frac{|a_{t,k}|^2}{d_{t,k}} + \lambda \left(\sum_k \left(\sum_t d_{t,k}^{1/s} \right)^{\frac{s}{r+s}} - 1 \right) - \sum_{t,k} \nu_{t,k} d_{t,k}$$

where λ and $\{\nu_{t,k}\}$ are the Lagrangian multipliers associated to the inequality constraint and the positivity constraints on $d_{t,k}$. By deriving the first-order optimality conditions, we get :

$$\frac{\partial \mathcal{L}}{\partial d_{m,n}} = -\frac{|a_{m,n}|^2}{d_{m,n}^2} - \nu_{m,n} + \frac{\lambda s}{r+s} \left(\sum_t d_{t,n}^{1/s} \right)^{\frac{-r}{r+s}} \cdot \frac{1}{s} \cdot d_{m,n}^{\frac{1-s}{s}}$$

According to these optimality conditions, at a stationary point, we have either $d_{m,n} = 0$ or

$$d_{m,n} = \left(\frac{\lambda}{r+s} \right)^{-s/(s+1)} |a_{m,n}|^{2s/(s+1)} \left(\sum_t d_{t,n}^{1/s} \right)^{rs/[(r+s)(s+1)]} \tag{31}$$

Then, we can derive

$$\left(\sum_m d_{m,n}^{1/s}) \right)^{(s+1)} = \left(\frac{r+s}{\lambda} \right) \left(\sum_m |a_{m,n}|^{2/(s+1)} \right)^{s+1} \left(\sum_m d_{m,n}^{1/s} \right)^{r/(r+s)} \tag{32}$$

and thus

$$\left(\sum_m d_{m,n}^{1/s}) \right)^{s} = \left(\frac{r+s}{\lambda} \left(\sum_m |a_{m,n}|^{2/(s+1)} \right)^{s+1} \right)^{(r+s)/(r+s+1)} \tag{33}$$

As $\lambda \neq 0$, the inequality on the mixed-norm on $d_{t,k}$ becomes an equality. Hence, after powering each side of Equation (33) to $1/(r+s)$ and summing each side over n, we have :

$$\frac{\lambda}{r+s} = \left(\sum_n g_n^{(s+1)/(r+s+1)} \right)^{r+s+1} \tag{34}$$

where $g_n = \sum_m |a_{m,n}|^{2/(s+1)}$. Then, plugging equations (34) and (33) into (31) gives the desired result :

$$d_{m,n} = \frac{|a_{m,n}|^{\frac{2s}{s+1}} g_n^{\frac{r}{s+r+1}}}{\left(\sum_n g_n^{\frac{s+1}{s+r+1}}\right)^{r+s}} \tag{35}$$

C. Proof of equation (24)

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] R. G. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory, Submitted*, 2009.

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

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

[4] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Analysis and Applications*, vol. 14, pp. 877–905, 2008.

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

[6] 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.

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

[8] 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.

[9] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communication Pure Applied Mathematics*, vol. 57, pp. 1413–1541, 2004.

[10] I. Daubechies, R. DeVore, M. Fornasier, and S. Gunturk, "Iteratively reweighted least squares minimization for sparse recovery," *Commun. Pure Appl. Math*, vol. to appear, 2009.

[11] 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.

[12] 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.

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

[14] Y. Eldar, P. Kuppinger, and H. Blocskei, "Compressed sensing of block-sparse signals : Uncertainty relations and efficient recovery," *IEEE Trans. Signal Processing, Submitted*, 2009.

[15] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Information Theory*, To appear.

[16] 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.

[17] M. Figueiredo, J. Bioucas-Dias, and N. R., "Majorization-minimization algorithms for wavelet-based image for restoration," *IEEE Trans. on Image Processing*, vol. 16, no. 12, pp. 2980–2991, 2007.

[18] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Convex Optimization Methods for Signal Processing*, vol. 1, no. 4, pp. 586–598, 2007.

[19] M. Fornasier and H. Rauhut, "Iterative thresholding algorithms," *Applied and Computational Harmonic Analysis*, vol. 25, no. 2, pp. 187–208, 2008.

[20] ——, "Recovery algorithms for vecto valued data with joint sparsity constraints," *SIAM Journal of Numerical Analysis*, vol. 46, no. 2, pp. 577–613, 2008.

[21] S. Foucart and M.-J. Lai, "Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$." *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 395–407, 2009.

[22] 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.

[23] G. Gasso, A. Rakotomamonjy, and S. Canu, "Recovering sparse signals with a certain family of non-convex penalties and dc programming," *IEEE Trans. Signal Processing, To appear*, 2009.

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

[25] 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.

[26] 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.

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

[28] 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," *Journal of Fourier Analysis and Applications*, vol. 14, pp. 655–687, 2008.

[29] J. Huang, S. Ma, H. Xie, and C. Zhang, "A group bridge approach for variable selection," *Biometrika*, vol. 96, no. 2, pp. 339–355, 2009.

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

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

[32] K. Knight and W. Fu, "Asymptotics for lasso-type estimators," *The Annals of statistics*, vol. 28, pp. 1356–1378, 2000.

[33] H. Liu and J. Zhang, "On the estimation and variable selection consistency of the sum of q-norm regularized regression," Department of Statistics, Carnegie Mellon University, Tech. Rep., 2009.

[34] K. Lounici, A. Tsybakov, M. Pontil, and S. V. de Geer, "Taking advantage of sparsity in multi-task learning," in *Proceedings of Computational Learning Theory*, 2009.

[35] 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.

[36] 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.

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

[38] M. Mishali and Y. Eldar, "Reduce and boost : Recovering arbitrary sets of jointly sparse vectors," *IEEE Trans. On Signal Processing*, vol. 56, no. 10, pp. 4692–4702, 2008.

[39] D. Needell and J. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[40] 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.

[41] 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.

[42] A. Rakotomamonjy, F. Bach, Y. Grandvalet, and S. Canu, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[43] R. Saab, R. Chartrand, and Özgür Yilmaz, "Stable sparse approximations via nonconvex optimization," in *33rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[44] S. Sardy, A. Bruce, and P. Tseng, "Block coordinate relaxation methods for non-parametric wavelet denoising," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 361–379, 2000.

[45] T. Simila, "Majorize-minimize algorithm for multiresponse sparse regression," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP*, 2007, pp. 553–556.

[46] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 46, pp. 267–288, 1996.

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

[48] J. Tropp, "Algorithms for simultaneous sparse approximation. part II: Convex relaxation," *Journal of Signal Processing*, vol. 86, pp. 589–602, 2006.

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

[50] 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.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

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

[57] 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.