Algorithms for Multiple Basis Pursuit Denoising

Alain Rakotomamonjy

Algorithms for Multiple Basis Pursuit Denoising

A. Rakotomamonjy
LITIS EA 4108, UniversitÃl' de Rouen Avenue de l'université
76800 Saint Etienne du Rouvray France
alain.rakotomamonjy@insa-rouen.fr

***Abstract*—We address the problem of learning a joint sparse approximation of several signals over a dictionary. We pose the problem as a matrix approximation problem with a row-sparsity inducing penalization on the coefficient matrix. We propose a simple algorithm based on iterative shrinking for solving the problem. At the present time, such a problem is solved either by using a Second-Order Cone programming or by means of a M-Focuss algorithm. While the former algorithm is computationally expensive, the latter is efficient but present some pitfalls like presences of fixed points which are undesiderable when solving a convex problem. By analyzing the optimality conditions of the problem, we derive a simple algorithm.**

The algorithm we propose is efficient and is guaranteed to converge to the optimal solution, up to a given tolerance. Furthermore, by means of a reweighted scheme, we are able to improve the sparsity of the solution.

I. Introduction

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

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

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

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

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

For the sparse simultaneous approximation problem, the goal is then to recover the matrix $\mathbf{C}$ given the signal matrix $\mathbf{S}$ and the dictionary $\mathbf{\Phi}$ under the hypothesis that all signals $\mathbf{s}_i$ share the same sparsity profile. Such a problem can be formulated as the following optimization problem

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

where λ is a user-defined parameter that balances the approximation error and the sparsity-inducing penalty $J_{p,q}(\mathbf{C})$ is

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

with typically $p \leq 1$ and $q \geq 1$. $J_{p,q}(\mathbf{C}$ can be interpreted as a relaxed version of a ℓ_0 quasi-norm on the row-sparsity measure of $\mathbf{C}$.

Several authors have proposed methods for solving problem (2). For instance, Cotter el al. [5] developed an algorithm for solving problem (2) when $p \leq 1$ and $q = 2$, known as M-FOCUSS. Such an algorithm based on factored gradient descent have been proven to converge towards a local or avglobal (when $p = 1$) minimum of problem (2) 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 the most natural extension of the so-called Lasso problem [10] or Basis Pursuit Denoising [4], since for $L = 1$, problem (2) 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. [8] have proposed an algorithm based on a second-order cone programming formulation, which at the contrary to M-FOCUSS, always converges to the problem global solution.

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

Afterwards, instead of directly deriving a proper algorithm for solving the non-convex optimization problem when $p < 1$ and $q = 2$, we introduce an iterative reweighted M-Basis pursuit (IrM-BP) algorithm. We then show that depending on the chosen weights, such an iterative scheme can actually solve problem (2). Our main contribution at this point is then to have translated the non-convex problem (2) into a series of convex problems which are easy to solve with our iterative method for M-BP.

II. Algorithms

This section describes two algorithms for solving the M-BP problem with convex and non-convex penalties $J_{p,q}$.

A. Iterative shrinking

The M-BP optimization problem is the following

$$\min_{\mathbf{C}} W(\mathbf{C}) = \frac{1}{2}\|\mathbf{S} - \mathbf{\Phi}\mathbf{C}\|_F^2 + \lambda \sum_i \|c_{i,\cdot}\|_2 \tag{3}$$

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

Algorithm 1 Solving M-BP through iterative shrinking

$\mathbf{C} = 0$, Loop = 1
while Loop **do**
  for $i = 1, 2, \cdots, M$ **do**
    if $c_{i,\cdot}$ KKT condition is not satisfied **then**
      $c_{i,\cdot} = \left(1 - \frac{\lambda}{\|T_i\|}\right)_+ T_i$
    end if
  end for
  if all KKT Conditions are satisfied **then**
    Loop = 0
  end if
end while

of (3) is that $\mathbf{0} \in \partial W(\mathbf{C}^\star)$ where $\partial W(\mathbf{C})$ denotes the subdifferential of our objective value $W(\mathbf{C})$ [1]. By computing the subdifferential of $W(\mathbf{C})$ with respect to each row $c_{i,\cdot}$ of $\mathbf{C}$, the optimality condition of problem (3) 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 $J_{1,2}(\mathbf{C}) = \sum_i \|c_{i,\cdot}\|_2$. According to this definition of $J_{1,2}$'s subdifferential, the optimality condition can be rewritten as

$$\begin{aligned} -\mathbf{r}_i + \lambda \frac{c_{i,\cdot}}{\|c_{i,\cdot}\|_2} &= \mathbf{0} \quad \forall i, \quad c_{i,\cdot} \neq \mathbf{0} \qquad (4)\\ \|\mathbf{r}_i\|_2 &\leq \lambda \quad \forall i, \quad c_{i,\cdot} = \mathbf{0} \end{aligned}$$

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

$$\begin{aligned} \mathbf{r}_i &= \phi_i^t(\mathbf{S} - \mathbf{\Phi}\mathbf{C}_{-i}) - \phi_i^t \phi_i c_{i,\cdot} \\ &= T_i - c_{i,\cdot} \end{aligned} \qquad (5)$$

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 (4) 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 (4) yields to an optimal solution that can be obtained as :

$$c_{i,\cdot} = \left(1 - \frac{\lambda}{\|T_i\|}\right)_+ T_i \qquad \forall i \qquad (6)$$

From this update equation, we can derive a simple algorithm which consists in iteratively applying the update (6) to each row of $\mathbf{C}$. Such an iterative scheme actually performs a block-coordinate optimization. Although, block-coordinate optimization does not converge in general for non-smooth optimization problem, Tseng [12] has shown that for an optimization problem which objective value is the sum of a smooth and convex function and a non-smooth but block-separable convex function, block-coordinate optimization converges towards the global minimum of the problem. Since for M-BP we are considering a quadratic function and a row-separable penalty function, Tseng's results can be directly applied in order to prove convergence of our algorithm.

Our approach, detailed in Algorithm (1), is a simple and efficient algorithm for solving M-BP especially when the dictionary size is large. A similar approach has also been proposed for solving the lasso [6], the group lasso [14] and the elastic net [15]. 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 (6), the smaller the resulting $c_{i,\cdot}$ is. Insight into this iterative shrinking algorithm can be further obtained by supposing that $M \leq N$ and that $\mathbf{\Phi}$ is composed of orthonormal elements of $\mathbb{R}^N$, hence $\mathbf{\Phi}^t\mathbf{\Phi} = \mathbf{I}$. In such situation, we have

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

and thus

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

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

B. Reweighted iterative shrinking algorithm

This subsection introduces an iterative reweighted M-Basis Pursuit (IrM-BP) algorithm which solves problem (2) when $p < 1$ and $q = 2$.

Recently, several works have advocated that sparse approximations can be recovered through iterative algorithms based on a reweighted ℓ_1 minimization [16], [2], [3]. Typically, for a single signal case, the idea consists in iteratively solving the following problem

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

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

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

$$\min_{\mathbf{C}} \frac{1}{2}\|\mathbf{S} - \mathbf{\Phi}\mathbf{C}\|_F^2 + \lambda \sum_i z_i \|c_{i,\cdot}\|_2 \qquad (7)$$

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

$$z_i = \frac{1}{(\|c_{i,\cdot}^{(n-1)}\|_2 + \varepsilon)^r} \quad \forall i \tag{8}$$

where $\{c_{i,\cdot}^{(n-1)}\}$ is the i-th row of $\mathbf{C}^{(n-1)}$, r a user-defined positive constant and ε a small regularization term that avoids numerical instabilities and prevents from having an infinite regularization term for $c_{i,\cdot}$as soon as $c_{i,\cdot}^{(n-1)}$ vanishes. This is a classical trick that has been used for instance by Candès et al. [2] . Note that for any positive weight vector $\mathbf{z}$, problem (7) is a convex problem that does not present local minima. Furthermore, it can be solved using our iterative shrinking algorithm by simply replacing λ with $\lambda_i = \lambda \cdot z_i$. Such a scheme is similar to the *adaptive lasso* algorithm of Zou et al. [16] but uses several iterations and addresses the simultaneous approximation problem.

The IrM-BP algorithm we proposed above can also be interpreted as an algorithm for solving problem (2) when $0 < p < 1$. Indeed, similarly to the reweighted ℓ_1 scheme of Candès et al. [2] or the one-step reweighted lasso of Zou et al. [17], our algorithm falls in the class of majorize-minimize (MM) algorithms [7]. MM algorithms consists in replacing a difficult optimization problem with a more easier one, for instance by linearizing the objective function, by solving the resulting optimization problem and by iterating such a procedure.

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

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

then for the minimization step, replacing in problem (2) $J_{p,2}$ with the above inequality and dropping constant terms lead to our optimization problem (7) with appropriately chosen z_i and r. Note that for the weights given in equation (8), $r = 1$ corresponds to the linearization of a log penalty $\sum_i \log(\|c_{i,\cdot}\|)$ whereas setting $r = 1 - p$ corresponds to a ℓ_p penalty ($0 < p < 1$). According to the convergence properties for MM algorithms towards a local minimum of their objective function [7], we can state that our IrM-BP algorithm converges towards a local minimum of problem (2) with p and r being appropriately related.

III. Results

In order to quantify the performance of our algorithms and compare them to other approaches, we have used simulated datasets with different redundancies $\frac{M}{N}$, number k of active elements and number L of signals to approximate. The dictionary is based on M vectors sampled from the unit hypersphere of $\mathbb{R}^N$. The true coefficient matrix $\mathbf{C}$ has been obtained as follows. The positions of the k non-zero rows in the matrix are randomly drawn. The non-zero coefficients of $\mathbf{C}$ are then drawn from a zero-mean unit variance Gaussian distribution. The signal matrix $\mathbf{S}$ is obtained as in equation (1) with the noise matrix being drawn i.i.d from a zero-mean Gaussian distribution and variance so that the signal-to-noise ratio of each single signal is 10 dB.

We compare performances of M-BP, IrM-BP (with $r = 0.5$ and $r = 1$), M-SBL [13] and M-OMP [11] for different experimental situations. The baseline context is $M = 50$, $N = 25$, $k = 10$ and $L = 3$. Note that for the M-OMP, we stop the algorithm after exactly k iterations. For this experiment, we did not performed model selection but instead tried several values of λ and σ and chosen the ones that maximize performances.

Figure 1 shows, from left to right, the performance averaged over 50 trials, on sparsity recovery when k increases from 2 to 20, when M goes from 25 to 150 and when $L = 2, \cdots, 8$. We can note that, M-BP performs worse that IrM-BP. This is a result that we could expected in views of the literature [16], [2] which compare Lasso and reweighted Lasso, the single signal approximation counterpart of M-BP and IrM-BP.

For all experimental situations, we remark that IrM-BP and M-SBL perform equally well. Again, this similar performances can easily be understood because of the strong relation between reweighted M-BP and M-SBL [9]. When considering M-OMP, although we suppose that k is known, we can see that the M-OMP performance is not as good as those of M-SBL and IrM-BP.

We have also empirically assessed the computational complexity of our algorithms (we used $r = 1$ for IrM-BP). We varied one of the different parameters (dictionary size M, signal dimensionality N, number of signals L) while keeping the others fixed. All matrices $\mathbf{\Phi}$, $\mathbf{C}$ and $\mathbf{S}$ are created as described above. Experiments have been run on a Pentium D-3 GHz with 4 GB of RAM using Matlab code. The results, averaged over 50 trials, in Figure 2 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. [13]. All algorithms need one hyperparameter to be set, for M-SBL and M-FOCUSS, we were able to choose the optimal one since the hyperparameter is dependent on a known noise level. For our algorithms, the choice of λ is more critical and has been manually set so as to achieve optimal performances. Note that our aim here is not give an exact comparison of computational complexity of the algorithms but just to give an order of magnitude of these complexities. Indeed, careful comparisons are difficult since the different algorithms do not solve the same problem and do not use the same stopping criterion.

We can remark in Figure 2 that with respects to the dictionary size, all algorithms present an empirical exponent between 1.4 and 2.4. Interestingly, we have theoretically evaluate the complexity of M-BP as quadratic whereas we measure a sub-quadratic complexity [9]. We suppose that this happens because at each iteration, only the non-optimal $c_{i,\cdot}$'s are updated and thus the number of updates drastically reduces

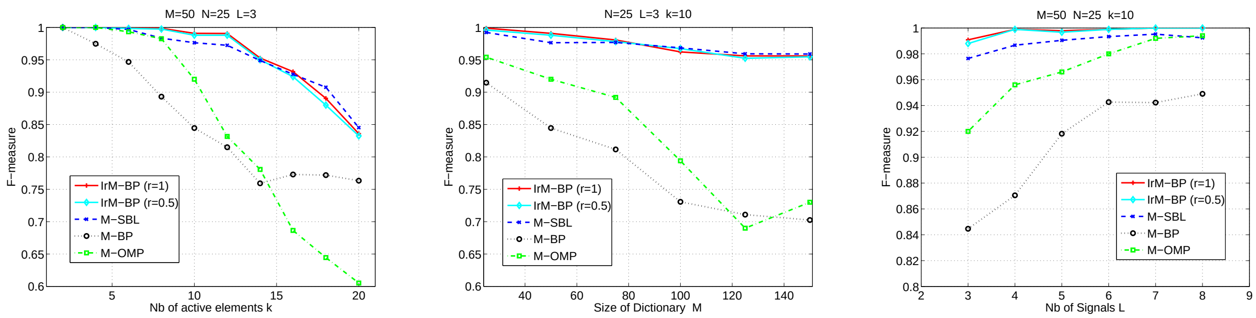

Fig. 1. Results comparing performances of different simultaneous sparse algorithms. from left to right, we have varied the number of generating functions k, the dictionary size M and the number of signal to approximate L. The default setting is $M = 50$, $N = 25$, $k = 10$ and $L = 3$.

along iterations. We can note that among all approaches, M-BP is the less demanding algorithm while IrM-BP is the less efficient one. This is clearly the cost to be paid for trading the resolution of a non-convex problem against several convex ones. Note however, that this complexity can be controlled by reducing the number of iterations while preserving good sparsity recovery. This is the case of many weighted Lasso algorithms which use only two iterations [17], [16].

The difference between the two top plots in Figure 2 shows that algorithm complexities not only depend on the dictionary size but also on the redundancy of the dictionary. Indeed, on the right plot, signal dimensionality is related to the dictionary size (redundancy is kept fixed) while on the left plot, the signal size is fixed. This results in a non-uniform variation of the complexities which is difficult to understand. It is not clear if it is related to the problem difficulty or is intrinsic to algorithms. Further researches are still needed to clarify this point.

Bottom left plot of Figure 2 depicts the complexity dependency of all algorithms with respects to the number of signal to approximate. The results we obtain is in agreement with theoretical exponents since for M-BP and IrM-BP we have exponents of approximately 1 while the other algorithm complexities do not depend on L. On the bottom right, we have evaluated these exponents with respects to signal dimension. Here again, we have results in accordance to theoretical expectations : M-BP and IrM-BP have lower complexities than M-SBL and M-FOCUSS. Furthermore, we note that IrM-BP has unexpectedly a very low exponent complexity. We assume that this is due to the fact that as dimension increases, the approximation problem becomes easier and thus needs less M-BP iterations.

IV. Conclusions

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

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

References

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

[2] 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.

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

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

[5] 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.

[6] 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.

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

[8] 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.

[9] A. Rakotomamonjy, "Simultaneous sparse approximations : insights and algorithms," Université de Rouen, hal-00328185, Tech. Rep., 2008.

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

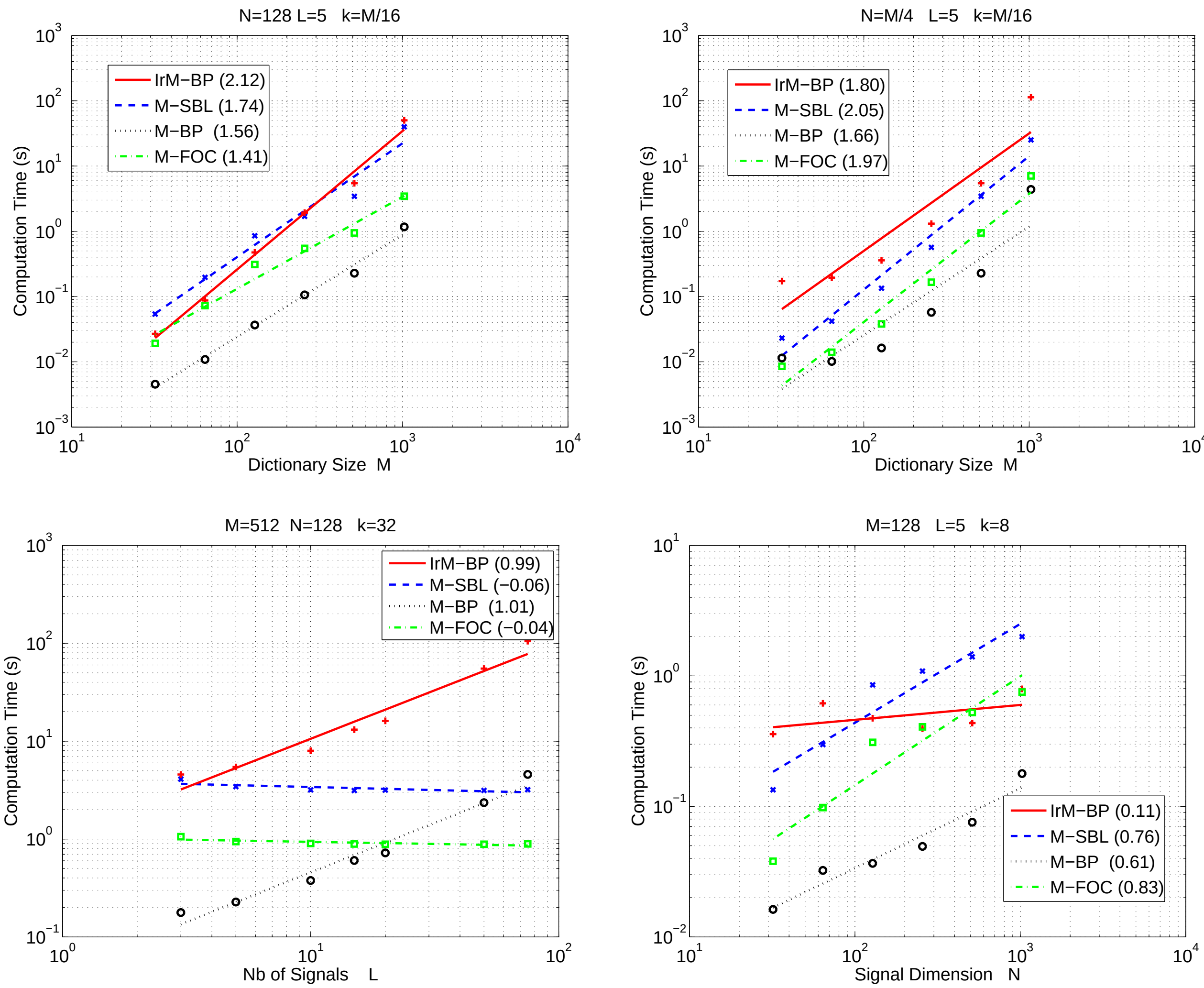

Fig. 2. Estimation of the empirical exponent of the computational complexity of different algorithms (M-BP, IrM-BP, M-SBL and M-FOCUSS). For the two last algorithms, we have used the code available from Wipf's website. The top plots give the computation time of the algorithms with respects to the dictionary size. On the top left, the signal dimension N has been kept fixed and equal to 128 whereas on the right one, the signal dimension is related to the dictionary size. The bottom plots respectively depict the computational complexity with respects to the number of signals to approximate and the dimensionality of these signals.

[11] 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.

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

[13] 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.

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

[15] 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.

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

[17] 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.