# SUBSPACE PENALIZED SPARSE LEARNING FOR JOINT SPARSE RECOVERY

*Jong Chul Ye[a], Jong Min Kim[a] and Yoram Bresler[b]*

[a]Dept. of Bio/Brain Engineering, KAIST, Daejon 305-701, Republic of Korea
[b]Coordinated Science Lab, Univ. of Illinois at Urbana-Champaign., Urbana, USA;

## ABSTRACT

The multiple measurement vector problem (MMV) is a generalization of the compressed sensing problem that addresses the recovery of a set of jointly sparse signal vectors. One of the important contributions of this paper is to reveal that the seemingly least related state-of-art MMV joint sparse recovery algorithms - M-SBL (multiple sparse Bayesian learning) and subspace-based hybrid greedy algorithms - have a very important link. More specifically, we show that replacing the $\log\det(\cdot)$ term in M-SBL by a $\log\det(\cdot)$ rank proxy that exploits the spark reduction property discovered in subspace-based joint sparse recovery algorithms, provides significant improvements. Theoretical analysis demonstrates that even though M-SBL is often unable to remove all local minimizers, the proposed method can do so under fairly mild conditions, without affecting the global minimizer.

***Index Terms***— Compresse sensing, joint sparse recovery, multiple measurement vector problem

## 1. INTRODUCTION

The multiple measurement vector problem (MMV) is a generalization of the compressed sensing problem, which addresses the recovery of a set of sparse signal vectors that share common non-zero support [1, 2]. Then, for a given noisy observation matrix $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_N] \in \mathbb{C}^{m\times N}$ and a sensing matrix $A \in \mathbb{C}^{m\times n}$, the multiple measurement vector (MMV) problem can be formulated as:

$$\begin{aligned} &\text{minimize} \quad \|X\|_0 \\ &\text{subject to} \quad \|Y - AX\|_F < \epsilon, \end{aligned} \tag{1}$$

where $X = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{n\times N}$ and $\|X\|_0 = |\text{supp}X|$, where $\text{supp}X = \{1 \le i \le n : \mathbf{x}^i \neq 0\}$ and $\mathbf{x}^i$ is the $i$-th row of $X$. In a noisy environment, Obozinski *et al* showed that a near optimal sampling rate reduction up to $\text{rank}(Y)$ can be achieved using $l_1/l_2$ mixed norm penalty [3]. Similar gain was observed in computationally inexpensive greedy approaches such as compressive MUSIC (CS-MUSIC) [1] and subspace augmented MUSIC (SA-MUSIC) [2].

This work was supported by the Korea Science and Engineering Foundation, in Grant number 20120000173.

While the aforementioned mixed norm approach and subspace based greed approaches provide theoretical performance guarantees, there also exist a very different class of powerful MMV algorithms that are based on empirical Bayesian and Automatic Relevance Determination (ARD) principle from machine learning. Among these, so-called multiple sparse Bayesian learning (M-SBL) is best known [4]. Even though M-SBL is more computationally expensive than greedy algorithms such as CS-MUSIC or SA-MUSIC, empirical results show that M-SBL is quite robust to noise and unfavorable restricted isometry property constant (RIC) of the sensing matrix [5]. Since Bayesian approaches are very different from classical compressed sensing, such high performance appears mysterious at first glance. However, a recent breakthrough by Wipf *et al* unveiled that M-SBL can be converted to a standard compressed sensing framework with an additional $\log|\cdot|$ (log determinant) penalty - a non-separable sparsity inducing prior [6].

One of the important contributions of this paper is that we show that the seemingly least related algorithms - M-SBL and subspace-based hybrid greedy algorithms - have a very important link. More specifically, we show that the $\log|\cdot|$ term in M-SBL is a proxy for the rank of a partial sensing matrix corresponding to the true support. Furthermore, we show that replacing this proxy by a $\log|\cdot|$ proxy for the rank of restricted partial sensing matrix that was discovered in subspace-based hybrid greedy algorithm to exploit the spark reduction property of MMV, provides significant performance improvements. We show that the global minimizer of the cost function of the proposed subspace penalized sparse learning (SPL) algorithm is identical to the original $l_0$ minimization problem under some regularity conditions. Furthermore, even though M-SBL is often impossible to remove local minimizers, our theoretical analysis demonstrates that SPL can eliminate all local minimizers without affecting the global minimizers under fairly mild conditions.

## 2. SUBSPACE-PENALIZED SPARSE LEARNING

### 2.1. M-SBL: A Review

Under appropriate assumptions of noise and signal Gaussian statistics, one can show that M-SBL minimizes the following

cost function in a so-called $\boldsymbol{\gamma}$ space [4]:

$$\mathcal{L}^{\gamma}(\boldsymbol{\gamma}) = \mathrm{Tr}\left(\Sigma_y^{-1} Y Y^*\right) + N \log|\Sigma_y| \tag{2}$$

where $\Sigma_y = \lambda I + A\Gamma A^*, \Gamma = \mathrm{diag}(\boldsymbol{\gamma})$. With an estimate of $\Gamma$, the solution of M-SBL is given by

$$X = \Gamma A^*(\lambda I + A\Gamma A^*)^{-1} Y \,. \tag{3}$$

One of the most important contributions by Wipf is that the minimization problem of the cost function (2) can be equivalently represented as the following standard sparse recovery framework [6]:

$$\min_X \mathcal{L}^{\mathbf{x}}(X), \quad \mathcal{L}^{\mathbf{x}}(X) = \|Y - AX\|_F^2 + \lambda g_{msbl}(X) \tag{4}$$

where $g_{msbl}(X)$ is a penalty given by

$$g_{msbl}(X) \equiv \min_{\boldsymbol{\gamma}\geq \mathbf{0}} \mathrm{Tr}\left(X^*\Gamma^{-1}X\right) + N\log|\lambda I + A\Gamma A^*| \,. \tag{5}$$

Note that due to the non-negativity constraint for $\boldsymbol{\gamma}$, a critical solution should satisfy the first order Karush-Kuhn-Tucker (KKT) necessary conditions [7]. This result in the following fixed point equation:

$$\gamma_i = \lim_{\lambda\to 0} \frac{\frac{1}{N}\|\mathbf{x}^i\|^2}{\gamma_i \mathbf{a}_i^H(\lambda I + A\Gamma A^*)^{-1}\mathbf{a}_i} = \frac{\frac{1}{N}\|\mathbf{x}^i\|^2}{\left(P_{R(\Gamma^{\frac{1}{2}}A^*)}\right)_{ii}} \tag{6}$$

where $P_{R(\Gamma^{1/2}A^*)}$ denote the projection matrix for the range space of $\Gamma^{\frac{1}{2}}A^*$. By plugging in (6) , we have

$$\begin{aligned} g_{msbl}(X) &= \min_{\boldsymbol{\gamma}\geq \mathbf{0}} \mathrm{Tr}\left(X^*\Gamma^{-1}X\right) + N\log|\lambda I + A\Gamma A^*| \\ &= N\|\boldsymbol{\gamma}_*\|_0 + N\log|\lambda I + A\Gamma_* A^*| \end{aligned} \tag{7}$$

where $\boldsymbol{\gamma}_*$ denotes a $\boldsymbol{\gamma}$ that satisfies (6). Note that the first term in (7) imposes the row sparsity since $\gamma_i = 0$ for $\|\mathbf{x}^i\| = 0$ due to (6). Then, what is the meaning of the $\log|\cdot|$ term ?

### 2.2. Key Observation

Note that $\log\det(\cdot)$ is often used for proxy for a matrix rank [8]. This leads us to an another interpretation that the penalty term in M-SBL is equivalent to

$$g_{msbl}(X) = N\|\boldsymbol{\gamma}\|_0 + N\ \mathrm{Rprox}(A\Gamma^{\frac{1}{2}}) \tag{8}$$

where $\mathrm{Rprox}(\cdot)$ dentoes a rank proxy; so the penalty simultaneously imposes the row sparsity of $X$ as well as the low rankness of the matrix $A\Gamma^{\frac{1}{2}}$. However, it is not clear why $\mathrm{Rank}(A\Gamma^{\frac{1}{2}})$ needs to be minimized. In fact, this paper shows that we may replace the second term, $\mathrm{Rprox}(A\Gamma^{\frac{1}{2}})$ by geometrically more intuitive rank proxys:

$$\begin{aligned} g_{SPL}(X) &= N\|\boldsymbol{\gamma}\|_0 + N\ \mathrm{Rprox}(Q^*A\Gamma^{\frac{1}{2}}) \\ &= \mathrm{Tr}\left(X^*\Gamma^{-1}X\right) + N\log|Q^*A\Gamma A^*Q + \epsilon I| \end{aligned}$$

where $Q$ denotes a basis for noise subspace such that $R(Q) = R^{\perp}(Y)$. This is due to the following theorem.

**Theorem 2.1** *Assume that $A \in \mathbb{R}^{m\times n}$, $X_* \in \mathbb{R}^{n\times r}$, $Y \in \mathbb{R}^{m\times r}$ satisfy $AX_* = Y$ where $\|X_*\|_0 = k$ and the columns of $Y$ are linearly independent and $r = \mathrm{rank}(Y)$. If $A$ satisfies a RIP condition $0 \leq \delta^L_{2k-r+1}(A) < 1$, then for noiseless measurement we have*

$$k - r = \min_{|I|\geq k} \mathrm{rank}\left(Q^*A_I\right),$$

*and*

$$\mathrm{supp}X_* = \arg\min_{|I|\geq k} \mathrm{rank}\left(Q^*A_I\right).$$

### 2.3. Alternating Minimization Algorithm

Therefore, following the derivation that leads to (7), we propose the following SPL penalty:

$$g_{SPL,\epsilon}(X) \equiv \min_{\boldsymbol{\gamma}\geq \mathbf{0}} \mathcal{G}_{SPL,\epsilon}(\boldsymbol{\gamma}, X) \tag{9}$$

where

$$\mathcal{G}_{SPL,\epsilon}(\boldsymbol{\gamma}, X) = \mathrm{Tr}\left(X^*\Gamma^{-1}X\right) + N\log|Q^*A\Gamma A^*Q + \epsilon I| \tag{10}$$

Using the proposed SPL penalty, we formulate the following SPL minimization problem:

$$\min_X \|Y - AX\|_F^2 + \lambda g_{SPL,\epsilon}(X) \,. \tag{11}$$

This can be solved using the following alternating minimization.

#### *2.3.1. Minimization with respect to $X$*

For a given estimate $\boldsymbol{\gamma}^{(t)}$, we have close form solution for $X^{(t+1)}$:

$$X^{(t+1)} = \Gamma^{(t)}A^*(\lambda I + A\Gamma^{(t)}A^*)^{-1}Y, \quad \Gamma^{(t)} = \mathrm{diag}(\boldsymbol{\gamma}^{(t)}).$$

#### *2.3.2. Estimation of $\boldsymbol{\gamma}$*

For a given$X^{(t)}$, we have the following update equation:

$$\gamma_i^{(t)} = \left(\frac{\frac{1}{N}\sum_j |x_{ij}^{(t)}|^2}{\mathbf{a}_i^* Q(Q^*A\Gamma^{(t)}A^*Q + \epsilon I)^{-1}Q^*\mathbf{a}_i}\right)^{\frac{1}{2}}, \tag{12}$$

which encompasses the case of $\sum_j |x_{ij}^{(t)}|^2 = 0$.

## 3. THEORETICAL ANALYSIS

While the overall cost function is convex for each variable $X, \boldsymbol{\gamma}$ separately, it is not convex for all these variable simultaneously due to the existence of bi-convex term $\mathrm{Tr}\left(X^*\Gamma^{-1}X\right)$. Indeed, this is a typical example of the d.c. algorithm (DCA) for the difference of convex functions programming [9, 10], and the alternating minimization algorithm

converges to a *local* minimizer or a critical point. Similar to [6], we investigate the global minimizer and local minimizers. The global minimizer proof is essentially the same as [6], so here we provide the conditions to remove local minimizers.

**Theorem 3.1** *Let $X_*$ denote a maximally sparse solution to be $Y = AX_*$ with $\|X_*\|_0 = k$ and the elements of $A$ are drawn from a random distribution. Let $A$ satisfy the RIP condition with $0 \leq \delta^L_{2k-r+1}(A) < 1$, where $r = \mathrm{rank}(Y)$. Suppose $X$ represent a coefficient such that $S = \mathrm{supp}X$ and $k < |S| = p \leq m$, and $X^S = A_S^\dagger Y$. Then, the following statements are true:*

1. *If $X$ is not a basic feasible solution (BFS), it is not a local minimizer.*
2. *Suppose $X$ be a BFS. For some $j \notin S$, if we have*

$$\mathbf{v}^*_{S,j}\left(\bar{W}_{S,j}\bar{R}_S\bar{W}_{S,j}\right)\mathbf{v}_{S,j} > 1, \tag{13}$$

*where $\mathbf{v}_{S,j} = A_S^\dagger \mathbf{a}_j$, $\bar{R}_S = [\bar{r}_{ii'}]_{i,i'\in S}$ and $\bar{W}_{S,j} = \mathrm{diag}([\bar{w}_i]_{i\in S})$ such that*

$$\bar{r}_{ii'} = \frac{\mathbf{x}^i(\mathbf{x}^{i'})^*}{\|\mathbf{x}^i\|\|\mathbf{x}^{i'}\|}, \quad \bar{w}_i = \sqrt{\frac{\mathbf{a}_i^* Q\Psi_{\setminus i}Q^*\mathbf{a}_i}{\mathbf{a}_j^* Q\Psi Q^*\mathbf{a}_j}}, \tag{14}$$

*for $i, i' \in S$ and $j \notin S$ and $\Psi = (Q^*A\Gamma AQ + \epsilon)^{-1}$, then $X$ is* not *a local minimizer.*

3. *In particular, if $\mathrm{rank}(Q^*A_S) \leq m - r$ and $|S \cap \mathrm{supp}X_*| \geq k - r$ and the rows of $X^{S\setminus \mathrm{supp}X_*}$ are in general position, then there always exists $\epsilon_1 > 0$ such that for any $0 < \epsilon < \epsilon_1$, $X$ is* not *a local minimizer almost surely.*

Our local minimizer analysis clearly show why SPL is better than M-SBL in finding the global minimizer. In particular, unlike M-SBL, SPL has very unique way to eliminate the local minimizer. More specifically, as long as $k - r$ correct supports are included in a local minimizer, SPL can escape from the local minimizer almost surely. In addition, the proposed SPL is more robust to the condition number of the unknown signal $X$ compared to M-SBL. This can be shown from the following observation. Aside from $\bar{W}_{S,j}$ and $\mathbf{v}_{S,j}$, another important components in (13) is $\bar{R}_S$. As $\mathrm{rank}(\bar{R}_S) \leq k$, there always exist null spaces when the rank of $\bar{R}_S$ is not full. In particular, in noisy cases, the corresponding numerical rank can be reduced when the condition number of $\bar{R}_S$ is bad, which can be happened when the rows of $X^S$ are highly correlated. In this case, there are more chances that $\bar{W}_{S,j}\mathbf{v}_{S,j}$ can fall into the null space so that the condition (13) cannot be met. However, in SPL, we can still prevent such situation as long as $k - r$ correct support are included in $X^S$. This implies that SPL can be more robust to the condition number of the unknown signal compared to M-SBL.

## 4. NUMERICAL RESULTS

The elements of a sensing matrix $A$ were generated either from a Gaussian distribution having zero mean and variance of $1/m$, and then each column of $A$ was normalized to have an unit norm. An unknown signal $X$ with $\mathrm{rank}(X) = r \leq k$ was generated using the same procedure as in [2]. Specifically, we randomly generated a support $I$, and then the corresponding nonzero signal components were obtained by

$$X^I = \Psi\Lambda\Phi\,, \tag{15}$$

where $\Psi \in \mathbb{R}^{k\times r}$ was set to random orthonormal columns, and $\Lambda = \mathrm{diag}([\lambda_i]_{i=1}^r)$ is a diagonal matrix whose $i$-th element is given by

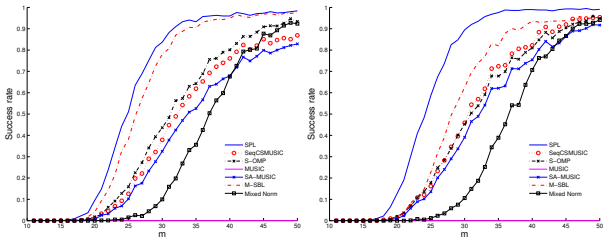
$$\lambda_i = \tau^i, \quad 0 < \tau < 1, \tag{16}$$

and $\Phi \in \mathbb{R}^{r\times N}$ were made using Gaussian random distribution with zero mean and variance of $1/N$. After generating noiseless data, we added zero mean white Gaussian noise. We declared success if an estimated support from a certain algorithm was the same as a true $\mathrm{supp}X$.

As the proposed algorithm does not require a prior knowledge of sparsity level, we need to define a stoping criterion. Here, the stopping criterion is defined by monitoring relative ratio of the variable $\boldsymbol{\gamma}$:

$$\frac{\|\boldsymbol{\gamma}^{(t)} - \boldsymbol{\gamma}^{(t-1)}\|_2}{\|\boldsymbol{\gamma}^{(t)}\|_2} < 10^{-3}\,.$$

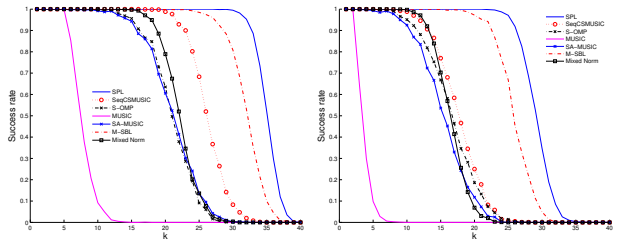
From our experiments, usually 20-30 iterations are required for SPL to converge.




(a) (b)

**Fig. 1**. Performance of various joint sparse recovery algorithms at $n = 128, k = 10, r = 6$ when (a) $SNR = 30dB, N = 16, \tau = 0.1$, (b) $SNR = 30dB, N = 256, \tau = 0.1$, respectively.

To compare the proposed algorithm with various state-of-art joint sparse recovery methods, the recovery rates of various state-of-art joint sparse recovery algorithms such as MUSIC, S-OMP, SA-MUSIC, sequential CS-MUSIC, and M-SBL, and $l_1/l_2$ mixed norm approach are plotted in Fig. 1 along with those of SPL. Since M-SBL, mixed norm approach as well as SPL do not provide a exact $k$-sparse solution, we




(a) (b)

**Fig. 2**. Various joint sparse recovery algorithm for varying sparsity level at $N = 256$. The simulation parameters are (a) $m = 40, r = 5, \tau = 1$ and SNR=30dB, and (b) $m = 40, r = 15, \tau = 0.5$ and SNR=30dB, respectively.

used the support for the largest $k$ coefficients as a support estimate in calculating the perfect recovery ratio. For MUSIC, S-OMP, SA-MUSIC, sequential CS-MUSIC, we assume that $k$ is known. For subspace based algorithms such as MUSIC, SA-MUSIC, sequential CS-MUSIC as well as SPL, we determine the signal subspace using the following criterion

$$\max_{i\in\{1,\cdots,m\}} \frac{\sigma_i - \sigma_{i+1}}{\sigma_i - \sigma_m} > 0.1,$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq_m$ denotes the singular values of $YY^*$. Here, the success rates were averaged over 1000 experiments. The simulation parameters were as follows: $m \in \{1, 2, \ldots, 50\}$, $n = 128, k = 8, r = 5$, $SNR = 30dB, 10dB$, respectively. Figs. 1(a)-(b) illustrates the comparison results under various snapshot number conditions and SNR conditions. Note that SPL consistently outperforms all other algorithms at various snapshots numbers. In particular, the gain increases with increasing number of snapshots, since it provides better subspace estimation. Also, note that SPL consistently outperforms M-SBL at all SNR ranges. Figs. 1(a)(b) illustrates that SPL significantly outperforms M-SBL when the condition number of $X$ is very bad. Moreover, as the subspace estimation becomes accurate with increasing $N$, the gain becomes more significant.

Figs. 2(a)(b) shows the performance comparison of various MMV algorithm by varying sparsity level. Here, $m$ and $\mathrm{rank}(Y)$ are fixed and the sparsity levels changes, and we calculated the perfect reconstruction ratio. Again, SPL outperforms all existing methods for various SNR and conditions numbers.

## 5. CONCLUSION

Our joint sparse recovery algorithm was inspired from the observation that the $\log|\cdot|$ term in M-SBL is a rank proxy for partial sensing matrix, and similar rank criteria exist in subspace-based greedy MMV algorithms like CS-MUSIC and SA-MUSIC. Furthermore, we proved that instead of $\mathrm{rank}(A\Gamma^{1/2})$, minimizing $\mathrm{rank}(Q^*A\Gamma^{1/2})$ is more direct way of imposing joint sparsity since its global minimizer can provide a true joint support. Theoretical analysis demonstrated that even though M-SBL is often impossible to remove all local minimizers, the proposed method can do that if $k - r$ partial support are included in a intermediate solution.

## 6. REFERENCES


[1] J. Kim, O. Lee, and J. Ye, "Compressive MUSIC: revisiting the link between compressive sensing and array signal processing," *IEEE Trans. on Information Theory*, vol. 58, no. 1, pp. 278–301, 2012.

[2] K. Lee, Y. Bresler, and M. Junge, "Subspace methods for joint sparse recovery," *IEEE Trans. on Information Theory*, vol. 58, no. 6, pp. 3613–3641, 2012.

[3] G. Obozinski, M. Wainwright, and M. Jordan, "Support union recovery in high-dimensional multivariate regression," *The Annals of Statistics*, vol. 39, no. 1, pp. 1–47, 2011.

[4] 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 Part 2, pp. 3704–3716, 2007.

[5] J. M. Kim, O. K. Lee, and J. C. Ye, "Improving noise robustness in subspace-based joint sparse recovery," *IEEE Trans. on Signal Processing*, vol. 60, no. 11, pp. 5799–5809, 2012.

[6] D. Wipf, B. Rao, and S. Nagarajan, "Latent variable bayesian models for promoting sparsity," *IEEE Trans. on Information Theory*, vol. 57, no. 9, p. 6236, 2011.

[7] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*. New York: Wiley-Interscience, 1996.

[8] K. Mohan and M. Fazel, "Iterative reweighted least squares for matrix rank minimization," in *48th IEEE Annual Allerton Conference on Communication, Control, and Computing*, 2010, pp. 653–661.

[9] P. Tao and L. An, "Convex analysis approach to dc programming: Theory, algorithms and applications," *Acta Mathematica Vietnamica*, vol. 22, no. 1, pp. 289–355, 1997.

[10] P. Tao and L. T. H. An, "A DC optimization algorithm for solving the trust-region subproblem," *SIAM Journal on Optimization*, vol. 8, no. 2, pp. 476–505, 1998.