Sparse Signal Recovery with Temporally Correlated Source Vectors Using Sparse Bayesian Learning

Zhilin Zhang, *Student Member, IEEE* and Bhaskar D. Rao, *Fellow, IEEE*

***Abstract*— We address the sparse signal recovery problem in the context of multiple measurement vectors (MMV) when elements in each nonzero row of the solution matrix are temporally correlated. Existing algorithms do not consider such temporal correlation and thus their performance degrades significantly with the correlation. In this work, we propose a block sparse Bayesian learning framework which models the temporal correlation. We derive two sparse Bayesian learning (SBL) algorithms, which have superior recovery performance compared to existing algorithms, especially in the presence of high temporal correlation. Furthermore, our algorithms are better at handling highly underdetermined problems and require less row-sparsity on the solution matrix. We also provide analysis of the global and local minima of their cost function, and show that the SBL cost function has the very desirable property that the global minimum is at the sparsest solution to the MMV problem. Extensive experiments also provide some interesting results that motivate future theoretical research on the MMV model.**

***Index Terms*— Sparse Signal Recovery, Compressed Sensing, Sparse Bayesian Learning (SBL), Multiple Measurement Vectors (MMV), Temporal Correlation**

I. Introduction

Sparse signal recovery, or compressed sensing, is an emerging field in signal processing [1]–[4]. The basic mathematical model is

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{v}, \tag{1}$$

where $\mathbf{\Phi} \in \mathbb{R}^{N\times M}(N \ll M)$ is a known dictionary matrix, and any N columns of $\mathbf{\Phi}$ are linearly independent (i.e. satisfies the Unique Representation Property (URP) condition [5]), $\mathbf{y} \in \mathbb{R}^{N\times 1}$ is an available measurement vector, and $\mathbf{v}$ is an unknown noise vector. The task is to estimate the source vector $\mathbf{x}$. To ensure a unique global solution, the number of nonzero entries in $\mathbf{x}$ has to be less than a threshold [5], [6]. This single measurement vector (SMV) model (1) has a wide range of applications, such as electroencephalography (EEG)/Magnetoencephalography (MEG) source localization [7], direction-of-arrival (DOA) estimation [8], radar detection [9], and magnetic resonance imaging (MRI) [10].

Motivated by many applications such as EEG/MEG source localization and DOA estimation, where a sequence of measurement vectors are available, the basic model (1) has been extended to the multiple measurement vector (MMV) model in [11], [12], given by

$$\mathbf{Y} = \mathbf{\Phi}\mathbf{X} + \mathbf{V}, \tag{2}$$

where $\mathbf{Y} \triangleq [\mathbf{Y}_{\cdot 1}, \cdots, \mathbf{Y}_{\cdot L}] \in \mathbb{R}^{N\times L}$ is an available measurement matrix consisting of L measurement vectors, $\mathbf{X} \triangleq [\mathbf{X}_{\cdot 1}, \cdots, \mathbf{X}_{\cdot L}] \in \mathbb{R}^{M\times L}$ is an unknown source matrix (or called a solution matrix) with each row representing a possible source [1], and $\mathbf{V}$ is an unknown noise matrix. A key assumption in the MMV model is that the support (i.e. indexes of nonzero entries) of every column in $\mathbf{X}$ is identical (referred as *the common sparsity assumption* in literature [12]). In addition, similar to the constraint in the SMV model, the number of nonzero rows in $\mathbf{X}$ has to be below a threshold to ensure a unique and global solution [12]. This leads to the fact that $\mathbf{X}$ has a small number of nonzero rows.

It has been shown that compared to the SMV case, the successful recovery rate can be greatly improved using multiple measurement vectors [12]–[15]. For example, Cotter and Rao [12] showed that by taking advantage of the MMV formulation, one can relax the upper bound in the uniqueness condition for the solution. Tang, Eldar and their colleagues [14], [16] showed that under certain mild assumptions the recovery rate increases exponentially with the number of measurement vectors L. Jin and Rao [15], [17] analyzed the benefits of increasing L by relating the MMV model to the capacity regions of MIMO communication channels. All these theoretical results reveal the advantages of the MMV model and support increasing L for better recovery performance.

However, under the common sparsity assumption we cannot obtain many measurement vectors in practical applications. The main reason is that the sparsity profile of practical signals is (slowly) time-varying, so the common sparsity assumption is valid for only a small L in the MMV model. For example, in EEG/MEG source localization there is considerable evidence [18] that a given pattern of dipole-source distributions [2] may only exist for 10-20 ms. Since the EEG/MEG sampling frequency is generally 250 Hz, a dipole-source pattern may only exist through 5 snapshots (i.e. in the MMV model $L = 5$). In DOA estimation [19], directions of targets [3] are continuously changing, and thus the source vectors that satisfy the common sparsity assumption are few. Of course, one can increase the measurement vector number at the cost of increasing the source number, but a larger source number can result in degraded recovery performance.

Thanks to numerous algorithms for the basic SMV model,

Z.Zhang and B.D.Rao are with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407, USA. Email:{z4zhang,brao}@ucsd.edu. The work was supported by NSF grant CCF-0830612.

[1]Here for convenience we call each row in $\mathbf{X}$ a source. The term is often used in application-oriented literature. Throughout the work, the i-th source is denoted by $\mathbf{X}_{i\cdot}$.

[2]In this application the set of indexes of nonzero rows in $\mathbf{X}$ is called a pattern of dipole-source distribution.

[3]In this application the index of a nonzero row in $\mathbf{X}$ indicates a direction.

most MMV algorithms [4] are obtained by straightforward extension of the SMV algorithms; for example, calculating the ℓ_2 norm of each row of $\mathbf{X}$, forming a vector, and then imposing the sparsity constraint on the vector. These algorithms can be roughly divided into greedy algorithms [20], [21], algorithms based on mixed norm optimization [22]–[24], iterative reweighted algorithms [12], [25], and Bayesian algorithms [26], [27].

Among the MMV algorithms, Bayesian algorithms have received much attention recently since they generally achieve the best recovery performance. Sparse Bayesian learning (SBL) is one important family of Bayesian algorithms. It was first proposed by Tipping [28], [29], and then was greatly enriched and extended by many researchers [25]–[27], [30]–[36]. For example, Wipf and Rao first introduced SBL to sparse signal recovery [30] for the SMV model, and later extended it to the MMV model, deriving the MSBL algorithm [26]. One attraction of SBL/MSBL is that, different from the popular ℓ_1 minimization based algorithms [37], [38], whose global minimum is generally not the sparsest solution [30], [39], the global minima of SBL/MSBL are always the sparsest one. In addition, SBL/MSBL have much fewer local minima than some classic algorithms, such as the FOCUSS family [5], [12].

Motivated by applications where signals and other types of data often contain some kind of structures, many algorithms have been proposed [13], [40]–[42], which exploit special structures in the source matrix $\mathbf{X}$. However, most of these works focus on exploiting spatial structures (i.e. the dependency relationship among different sources) and completely ignore temporal structures. Besides, for tractability purposes, almost all the existing MMV algorithms (and theoretical analysis) assume that the sources are independent and identically distributed (i.i.d.) processes. This contradicts the real-world scenarios, since a practical source often has rich temporal structures. For example, the waveform smoothness of biomedical signals has been exploited in signal processing for several decades. Besides, due to high sampling frequency, amplitudes of successive samplings of a source are strongly correlated. Recently, Zdunek and Cichocki [43] proposed the SOB-MFOCUSS algorithm, which exploits the waveform smoothness via a pre-defined smoothness matrix. However, the design of the smoothness matrix is completely subjective and not data-adaptive. In fact, in the task of sparse signal recovery, learning temporal structures of a source is a difficult problem. Generally, such structures are learned via a training dataset (which often contains sufficient data without noise for robust statistical inference) [44], [45]. Although effective for some specific signals, this method is limited. Having noticed that the temporal structures strongly affect the performance of existing algorithms, in [31] we derived the AR-SBL algorithm, which models each source as a first-order autoregressive (AR) process and learns AR coefficients from the data per se. Although the algorithm has superior performance compared to MMV algorithms in the presence of temporal correlation, it is slow, which limits its applications. As such, there is a need for efficient algorithms that can deal more effectively with temporal correlation.

In this work, we present a block sparse Bayesian learning (bSBL) framework, which transforms the MMV model (2) to a SMV model. This framework allows us to easily model the temporal correlation of sources. Based on it, we derive an algorithm, called T-SBL, which is very effective but is slow due to its operation in a higher dimensional parameter space resulting from the MMV-to-SMV transformation. Thus, we make some approximations and derive a fast version, called T-MSBL, which operates in the original parameter space. Similar to T-SBL, T-MSBL is also effective but has much lower computational complexity. Interestingly, when compared to MSBL, the only change of T-MSBL is the replacement of $\|\mathbf{X}_{i\cdot}\|_2^2$ with the Mahalanobis distance measure, i.e. $\mathbf{X}_{i\cdot}\mathbf{B}^{-1}\mathbf{X}_{i\cdot}^T$, where $\mathbf{B}$ is a positive definite matrix estimated from data and can be partially interpreted as a covariance matrix. We analyze the global minimum and the local minima of the two algorithms' cost function. One of the key results is that in the noiseless case the global minimum is at the sparsest solution. Extensive experiments not only show the superiority of the proposed algorithms, but also provide some interesting (even counter-intuitive) phenomena that may motivate future theoretical study.

The rest of the work is organized as follows. In Section II we present the bSBL framework. In Section III we derive the T-SBL algorithm. Its fast version, the T-MSBL algorithm, is derived in Section IV. Section V provides theoretical analysis on the algorithms. Experimental results are presented in Section VI. Finally, discussions and conclusions are drawn in the last two sections.

We introduce the notations used in this paper:

- $\|\mathbf{x}\|_1, \|\mathbf{x}\|_2, \|\mathbf{A}\|_{\mathcal{F}}$ denote the ℓ_1 norm of the vector $\mathbf{x}$, the ℓ_2 norm of $\mathbf{x}$, and the Frobenius norm of the matrix $\mathbf{A}$, respectively. $\|\mathbf{A}\|_0$ and $\|\mathbf{x}\|_0$ denote the number of nonzero rows in the matrix $\mathbf{A}$ and the number of nonzero elements in the vector $\mathbf{x}$, respectively;
- Bold symbols are reserved for vectors and matrices. Particularly, $\mathbf{I}_L$ denotes the identity matrix with size $L \times L$. When the dimension is evident from the context, for simplicity, we just use $\mathbf{I}$;
- $\mathrm{diag}\{a_1, \cdots, a_M\}$ denotes a diagonal matrix with principal diagonal elements being $a_1, \cdots, a_M$ in turn; if $\mathbf{A}_1, \cdots, \mathbf{A}_M$ are square matrices, then $\mathrm{diag}\{\mathbf{A}_1, \cdots, \mathbf{A}_M\}$ denotes a block diagonal matrix with principal diagonal blocks being $\mathbf{A}_1, \cdots, \mathbf{A}_M$ in turn;
- For a matrix $\mathbf{A}$, $\mathbf{A}_{i\cdot}$ denotes the i-th row, $\mathbf{A}_{\cdot i}$ denotes the i-th column, and $\mathbf{A}_{i,j}$ denotes the element that lies in the i-th row and the j-th column;
- $\mathbf{A} \otimes \mathbf{B}$ represents the Kronecker product of the two matrices $\mathbf{A}$ and $\mathbf{B}$. $\mathrm{vec}(\mathbf{A})$ denotes the vectorization of the matrix $\mathbf{A}$ formed by stacking its columns into a single column vector. $\mathrm{Tr}(\mathbf{A})$ denotes the trace of $\mathbf{A}$. $\mathbf{A}^T$ denotes the transpose of $\mathbf{A}$.

[4]For convenience, algorithms for the MMV model are called MMV algorithms; algorithms for the SMV model are called SMV algorithms.

II. Block Sparse Bayesian Learning Framework

Most existing works do not deal with the temporal correlation of sources. For many non-Bayesian algorithms, incorporating temporal correlation is not easy due to the lack of a well defined methodology to modify the diversity measures employed in the optimization procedure. For example, it is not clear how to best incorporate correlation in ℓ_1 norm based methods. For this reason, we adopt a probabilistic approach to incorporate correlation structure. Particularly, we have found it convenient to incorporate correlation into the sparse Bayesian learning (SBL) methodology.

Initially, SBL was proposed for regression and classification in machine learning [28]. Then Wipf and Rao [30] applied it to the SMV model (1) for sparse signal recovery. The idea is to find the posterior probability $p(\mathbf{x}|\mathbf{y};\Theta)$ via the Bayesian rule, where Θ indicates the set of all the hyperparameters. Given the hyperparameters, the solution $\widehat{\mathbf{x}}$ is given by the Maximum-A-Posterior (MAP) estimate. The hyperparameters are estimated from data by marginalizing over $\mathbf{x}$ and then performing evidence maximization or Type-II Maximum Likelihood [28]. To solve the MMV problem (2), Wipf and Rao [26] proposed the MSBL algorithm, which implicitly applies the ℓ_2 norm on each source $\mathbf{X}_{i\cdot}$. One drawback of this algorithm is that the temporal correlation of sources is not exploited to improve performance.

To exploit the temporal correlation, we propose another SBL framework, called the block sparse Bayesian learning (bSBL) framework. In this framework, the MMV model is transformed to a block SMV model. In this way, we can easily model the temporal correlation of sources and derive new algorithms.

First, we assume all the sources $\mathbf{X}_{i\cdot}$ $(\forall i)$ are mutually independent, and the density of each $\mathbf{X}_{i\cdot}$ is Gaussian, given by

$$p(\mathbf{X}_{i\cdot};\gamma_i,\mathbf{B}_i) \sim \mathcal{N}(\mathbf{0},\gamma_i\mathbf{B}_i), \quad i=1,\cdots,M$$

where γ_i is a nonnegative hyperparameter controlling the row sparsity of $\mathbf{X}$ as in the basic SBL [26], [28], [30]. When $\gamma_i = 0$, the associated $\mathbf{X}_{i\cdot}$ becomes zeros. $\mathbf{B}_i$ is a positive definite matrix that captures the correlation structure of $\mathbf{X}_{i\cdot}$ and needs to be estimated.

By letting $\mathbf{y} = \mathrm{vec}(\mathbf{Y}^T) \in \mathbb{R}^{NL\times 1}$, $\mathbf{D} = \mathbf{\Phi}\otimes\mathbf{I}_L$, $\mathbf{x} = \mathrm{vec}(\mathbf{X}^T) \in \mathbb{R}^{ML\times 1}$, $\mathbf{v} = \mathrm{vec}(\mathbf{V}^T)$, we can transform the MMV model to the block SMV model

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{v}. \tag{3}$$

To elaborate the block sparsity model (3), we rewrite it as $\mathbf{y} = [\boldsymbol{\phi}_1\otimes\mathbf{I}_L,\cdots,\boldsymbol{\phi}_M\otimes\mathbf{I}_L][\mathbf{x}_1^T,\cdots,\mathbf{x}_M^T]^T+\mathbf{v} = \sum_{i=1}^M(\boldsymbol{\phi}_i\otimes\mathbf{I}_L)\mathbf{x}_i+\mathbf{v}$, where $\boldsymbol{\phi}_i$ is the i-th column in $\mathbf{\Phi}$, and $\mathbf{x}_i \in \mathbb{R}^{L\times 1}$ is the i-th block in $\mathbf{x}$ and $\mathbf{x}_i = \mathbf{X}_{i\cdot}^T$. K nonzero rows in $\mathbf{X}$ means K nonzero blocks in $\mathbf{x}$. Thus $\mathbf{x}$ is block-sparse.

Assume elements in the noise vector $\mathbf{v}$ are independent and each has a Gaussian distribution, i.e. $p(v_i) \sim \mathcal{N}(0,\lambda)$, where v_i is the i-th element in $\mathbf{v}$ and λ is the variance. For the block model (3), the Gaussian likelihood is

$$p(\mathbf{y}|\mathbf{x};\lambda) \sim \mathcal{N}_{y|x}(\mathbf{D}\mathbf{x},\lambda\mathbf{I}).$$

The prior for $\mathbf{x}$ is given by

$$p(\mathbf{x};\gamma_i,\mathbf{B}_i,\forall i) \sim \mathcal{N}_x(\mathbf{0},\mathbf{\Sigma}_0),$$

where $\mathbf{\Sigma}_0$ is

$$\mathbf{\Sigma}_0 = \begin{bmatrix} \gamma_1\mathbf{B}_1 & & \\ & \ddots & \\ & & \gamma_M\mathbf{B}_M \end{bmatrix}. \tag{4}$$

Using the Bayes rule we obtain the posterior density of $\mathbf{x}$, which is also Gaussian,

$$p(\mathbf{x}|\mathbf{y};\lambda,\gamma_i,\mathbf{B}_i,\forall i) = \mathcal{N}_x(\boldsymbol{\mu}_x,\mathbf{\Sigma}_x)$$

with the mean

$$\boldsymbol{\mu}_x = \frac{1}{\lambda}\mathbf{\Sigma}_x\mathbf{D}^T\mathbf{y} \tag{5}$$

and the covariance matrix

$$\begin{aligned}\mathbf{\Sigma}_x &= (\mathbf{\Sigma}_0^{-1}+\frac{1}{\lambda}\mathbf{D}^T\mathbf{D})^{-1} \\ &= \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0\mathbf{D}^T\big(\lambda\mathbf{I}+\mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T\big)^{-1}\mathbf{D}\mathbf{\Sigma}_0. \end{aligned}\tag{6}$$

So given all the hyperparameters $\lambda,\gamma_i,\mathbf{B}_i,\forall i$, the MAP estimate of $\mathbf{x}$ is given by:

$$\begin{aligned}\mathbf{x}^* \triangleq \boldsymbol{\mu}_x &= (\lambda\mathbf{\Sigma}_0^{-1}+\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{y} \\ &= \mathbf{\Sigma}_0\mathbf{D}^T\big(\lambda\mathbf{I}+\mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T\big)^{-1}\mathbf{y} \end{aligned}\tag{7}$$

where the last equation follows the matrix identity $(\mathbf{I}+\mathbf{A}\mathbf{B})^{-1}\mathbf{A} \equiv \mathbf{A}(\mathbf{I}+\mathbf{B}\mathbf{A})^{-1}$, and $\mathbf{\Sigma}_0$ is the block diagonal matrix given by (4) with many diagonal block matrices being zeros. Clearly, the block sparsity of $\mathbf{x}^*$ is controlled by the γ_i's in $\mathbf{\Sigma}_0$: during the learning procedure, when $\gamma_k = 0$, the associated k-th block in $\mathbf{x}^*$ becomes zeros, and the associated dictionary vectors $\boldsymbol{\phi}_k\otimes\mathbf{I}_L$ are pruned out [5].

To estimate the hyperparameters we can use evidence maximization or Type-II maximum likelihood [28]. This involves marginalizing over the weights $\mathbf{x}$ and then performing maximum likelihood estimation. We refer to the whole framework including the solution (7) and the hyperparameter estimation as the block sparse Bayesian learning (bSBL) framework. Note that in contrast to the original SBL framework, the bSBL framework models the temporal structures of sources in the prior density via the matrices $\mathbf{B}_i$ $(i=1,\cdots,M)$. Different ways to learn the matrices result in different algorithms. We will discuss the learning of these matrices and other hyperparameters in the following sections.

III. Estimation of Hyperparameters

Before estimating the hyperparameters, we note that assigning a different matrix $\mathbf{B}_i$ to each source $\mathbf{X}_{i\cdot}$ will result in overfitting [46], [47] due to limited data and too many parameters. To avoid the overfitting, we consider using one positive definite matrix $\mathbf{B}$ to model all the source covariance matrices up to a scalar [6]. Thus Eq.(4) becomes $\mathbf{\Sigma}_0 = \mathbf{\Gamma}\otimes\mathbf{B}$

[5] In practice, we judge whether γ_k is less than a small threshold, e.g. 10^{-5}. If it is, then the associated dictionary vectors are pruned out from the learning procedure and the associated block in $\mathbf{x}$ is set to zeros.

[6] Note that the covariance matrix in the density of $\mathbf{X}_{i\cdot}$ is $\gamma_i\mathbf{B}_i$.

with $\mathbf{\Gamma} \triangleq \mathrm{diag}(\gamma_1, \cdots, \gamma_M)$. Although this strategy is equivalent to assuming all the sources have the same correlation structure, it leads to very good results even if all the sources have totally different correlation structures (see Section VI). More importantly, this constraint does not destroy the global minimum property (i.e. the global unique solution is the sparsest solution) of our algorithms, as confirmed by Theorem 1 in Section V.

To find the hyperparameters $\Theta = \{\gamma_1, \cdots, \gamma_M, \mathbf{B}, \lambda\}$, we employ the Expectation-Maximization (EM) method to maximize $p(\mathbf{y}; \Theta)$. This is equivalent to minimizing $-\log p(\mathbf{y}; \Theta)$, yielding the effective cost function:

$$\mathcal{L}(\Theta) = \mathbf{y}^T \mathbf{\Sigma}_y^{-1} \mathbf{y} + \log |\mathbf{\Sigma}_y|, \tag{8}$$

where $\mathbf{\Sigma}_y \triangleq \lambda \mathbf{I} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T$. The EM formulation proceeds by treating $\mathbf{x}$ as hidden variables and then maximizing:

$$\begin{aligned} Q(\Theta) &= E_{x|y;\Theta^{(\mathrm{old})}}\big[\log p(\mathbf{y}, \mathbf{x}; \Theta)\big] \\ &= E_{x|y;\Theta^{(\mathrm{old})}}\big[\log p(\mathbf{y}|\mathbf{x}; \lambda)\big] \\ &\quad + E_{x|y;\Theta^{(\mathrm{old})}}\big[\log p(\mathbf{x}; \gamma_1, \cdots, \gamma_M, \mathbf{B})\big] \end{aligned} \tag{9}$$

where $\Theta^{(\mathrm{old})}$ denotes the estimated hyperparameters in the previous iteration.

To estimate $\boldsymbol{\gamma} \triangleq [\gamma_1, \cdots, \gamma_M]$ and $\mathbf{B}$, we notice that the first term in (9) is unrelated to $\boldsymbol{\gamma}$ and $\mathbf{B}$. So, the Q function (9) can be simplified to:

$$Q(\boldsymbol{\gamma}, \mathbf{B}) = E_{x|y;\Theta^{(\mathrm{old})}}\big[\log p(\mathbf{x}; \boldsymbol{\gamma}, \mathbf{B})\big].$$

It can be shown that[7]

$$\log p(\mathbf{x}; \boldsymbol{\gamma}, \mathbf{B}) \propto -\frac{1}{2}\log\big(|\mathbf{\Gamma}|^L |\mathbf{B}|^M\big) - \frac{1}{2}\mathbf{x}^T(\mathbf{\Gamma}^{-1} \otimes \mathbf{B}^{-1})\mathbf{x},$$

which results in

$$\begin{aligned} Q(\boldsymbol{\gamma}, \mathbf{B}) \propto\ & -\frac{L}{2}\log\big(|\mathbf{\Gamma}|\big) - \frac{M}{2}\log\big(|\mathbf{B}|\big) \\ & -\frac{1}{2}\mathrm{Tr}\Big[\big(\mathbf{\Gamma}^{-1} \otimes \mathbf{B}^{-1}\big)\big(\mathbf{\Sigma}_x + \boldsymbol{\mu}_x\boldsymbol{\mu}_x^T\big)\Big], \end{aligned} \tag{10}$$

where $\boldsymbol{\mu}_x$ and $\mathbf{\Sigma}_x$ are evaluated according to (5) and (6), given the estimated hyperparameters $\Theta^{(\mathrm{old})}$.

The derivative of (10) with respect to γ_i $(i = 1, \cdots, M)$ is given by

$$\frac{\partial Q}{\partial \gamma_i} = -\frac{L}{2\gamma_i} + \frac{1}{2\gamma_i^2}\mathrm{Tr}\Big[\mathbf{B}^{-1}\big(\mathbf{\Sigma}_x^i + \boldsymbol{\mu}_x^i(\boldsymbol{\mu}_x^i)^T\big)\Big],$$

where we define (using the MATLAB notations)

$$\begin{cases} \boldsymbol{\mu}_x^i \triangleq \boldsymbol{\mu}_x((i-1)L+1\ :\ iL) \\ \mathbf{\Sigma}_x^i \triangleq \mathbf{\Sigma}_x((i-1)L+1\ :\ iL\ ,\ (i-1)L+1\ :\ iL) \end{cases} \tag{11}$$

So the learning rule for γ_i $(i = 1, \cdots, M)$ is given by

$$\gamma_i \leftarrow \frac{\mathrm{Tr}\big[\mathbf{B}^{-1}\big(\mathbf{\Sigma}_x^i + \boldsymbol{\mu}_x^i(\boldsymbol{\mu}_x^i)^T\big)\big]}{L}, \quad i = 1, \cdots, M \tag{12}$$

On the other hand, the gradient of (10) over $\mathbf{B}$ is given by

$$\frac{\partial Q}{\partial \mathbf{B}} = -\frac{M}{2}\mathbf{B}^{-1} + \frac{1}{2}\sum_{i=1}^{M}\frac{1}{\gamma_i}\mathbf{B}^{-1}\big(\mathbf{\Sigma}_x^i + \boldsymbol{\mu}_x^i(\boldsymbol{\mu}_x^i)^T\big)\mathbf{B}^{-1}.$$

[7]The $\propto$ notation is used to indicate that terms that do not contribute to the subsequent optimization of the parameters have been dropped. This convention will be followed through out the paper.

Thus we obtain the learning rule for $\mathbf{B}$:

$$\mathbf{B} \leftarrow \frac{1}{M}\sum_{i=1}^{M}\frac{\mathbf{\Sigma}_x^i + \boldsymbol{\mu}_x^i(\boldsymbol{\mu}_x^i)^T}{\gamma_i}. \tag{13}$$

To estimate λ, the Q function (9) can be simplified to

$$\begin{aligned} Q(\lambda) &= E_{x|y;\Theta^{(\mathrm{old})}}\big[\log p(\mathbf{y}|\mathbf{x}; \lambda)\big] \\ &\propto -\frac{NL}{2}\log\lambda - \frac{1}{2\lambda}E_{x|y;\Theta^{(\mathrm{old})}}\big[\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2\big] \\ &= -\frac{NL}{2}\log\lambda - \frac{1}{2\lambda}\Big[\|\mathbf{y} - \mathbf{D}\boldsymbol{\mu}_x\|_2^2 \\ &\quad + E_{x|y;\Theta^{(\mathrm{old})}}\big[\|\mathbf{D}(\mathbf{x} - \boldsymbol{\mu}_x)\|_2^2\big]\Big] \\ &= -\frac{NL}{2}\log\lambda - \frac{1}{2\lambda}\Big[\|\mathbf{y} - \mathbf{D}\boldsymbol{\mu}_x\|_2^2 + \mathrm{Tr}\big(\mathbf{\Sigma}_x\mathbf{D}^T\mathbf{D}\big)\Big] \\ &= -\frac{NL}{2}\log\lambda - \frac{1}{2\lambda}\Big[\|\mathbf{y} - \mathbf{D}\boldsymbol{\mu}_x\|_2^2 \\ &\quad + \widehat{\lambda}\mathrm{Tr}\big(\mathbf{\Sigma}_x(\mathbf{\Sigma}_x^{-1} - \mathbf{\Sigma}_0^{-1})\big)\Big] \qquad (14) \\ &= -\frac{NL}{2}\log\lambda - \frac{1}{2\lambda}\Big[\|\mathbf{y} - \mathbf{D}\boldsymbol{\mu}_x\|_2^2 \\ &\quad + \widehat{\lambda}\big[ML - \mathrm{Tr}(\mathbf{\Sigma}_x\mathbf{\Sigma}_0^{-1})\big]\Big], \qquad (15) \end{aligned}$$

where (14) follows from the first equation in (6), and $\widehat{\lambda}$ denotes the estimated λ in the previous iteration. The λ learning rule is obtained by setting the derivative of (15) over λ to zero, leading to

$$\lambda \leftarrow \frac{\|\mathbf{y} - \mathbf{D}\boldsymbol{\mu}_x\|_2^2 + \lambda\big[ML - \mathrm{Tr}(\mathbf{\Sigma}_x\mathbf{\Sigma}_0^{-1})\big]}{NL}, \tag{16}$$

where the λ on the right-hand side is the $\widehat{\lambda}$ in (15). There are some challenges to estimate λ in SMV models. This, however, is alleviated in MMV models when considering temporal correlation. We elaborate on this next.

In the SBL framework (either for the SMV model or for the MMV model), many learning rules for λ have been derived [26], [28], [30], [34]. However, in noisy environments some of the learning rules probably cannot provide an optimal λ, thus leading to degraded performance. For the basic SBL/MSBL algorithms, Wipf et al [26] pointed out that the reason is that λ and appropriate N nonzero hyperparameters γ_i make an identical contribution to the covariance $\mathbf{\Sigma}_y = \lambda\mathbf{I} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T$ in the cost functions of SBL/MSBL. To explain this, they gave an example: let a dictionary matrix $\mathbf{\Phi}' = [\mathbf{\Phi}_0, \mathbf{I}]$, where $\mathbf{\Phi}' \in \mathbb{R}^{N\times M}$ and $\mathbf{\Phi}_0 \in \mathbb{R}^{N\times(M-N)}$. Then the λ as well as the N hyperparameters $\{\gamma_{M-N+1}, \cdots, \gamma_M\}$ associated with the columns of the identity matrix in Φ' are not identifiable, because

$$\begin{aligned} \mathbf{\Sigma}_y &= \lambda\mathbf{I} + \mathbf{\Phi}'\mathbf{\Gamma}\mathbf{\Phi}'^T \\ &= \lambda\mathbf{I} + [\mathbf{\Phi}_0, \mathbf{I}]\mathrm{diag}\{\gamma_1, \cdots, \gamma_M\}[\mathbf{\Phi}_0, \mathbf{I}]^T \\ &= \lambda\mathbf{I} + \mathbf{\Phi}_0\mathrm{diag}\{\gamma_1, \cdots, \gamma_{M-N}\}\mathbf{\Phi}_0^T \\ &\quad + \mathrm{diag}\{\gamma_{M-N+1}, \cdots, \gamma_M\} \end{aligned}$$

indicating a nonzero value of λ and appropriate values of the N nonzero hyperparameters, i.e. $\gamma_{M-N+1}, \cdots, \gamma_M$, can make an identical contribution to the covariance matrix $\mathbf{\Sigma}_y$. This

problem can be worse when the noise covariance matrix is $\mathrm{diag}(\lambda_1, \cdots, \lambda_N)$ with arbitrary nonzero λ_i, instead of $\lambda\mathbf{I}$.

However, our learning rule (16) does not have such ambiguity problem. To see this, we now examine the covariance matrix $\mathbf{\Sigma}_y$ in our cost function (8). Noting that $\mathbf{D} = \mathbf{\Phi}' \otimes \mathbf{I}$, we have

$$\begin{aligned}\mathbf{\Sigma}_y &= \lambda\mathbf{I} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T \\ &= \lambda\mathbf{I} + (\mathbf{\Phi}' \otimes \mathbf{I})\big(\mathrm{diag}\{\gamma_1, \cdots, \gamma_M\} \otimes \mathbf{B}\big)(\mathbf{\Phi}' \otimes \mathbf{I})^T \\ &= \lambda\mathbf{I} + [\mathbf{\Phi}_0 \otimes \mathbf{I}, \mathbf{I} \otimes \mathbf{I}]\big(\mathrm{diag}\{\gamma_1, \cdots, \gamma_M\} \otimes \mathbf{B}\big) \\ &\quad \cdot[\mathbf{\Phi}_0 \otimes \mathbf{I}, \mathbf{I} \otimes \mathbf{I}]^T \\ &= \lambda\mathbf{I} + (\mathbf{\Phi}_0 \mathrm{diag}\{\gamma_1, \cdots, \gamma_{M-N}\}\mathbf{\Phi}_0^T) \otimes \mathbf{B} \\ &\quad + \mathrm{diag}\{\gamma_{M-N+1}, \cdots, \gamma_M\} \otimes \mathbf{B}.\end{aligned}$$

Obviously, since $\mathbf{B}$ is not an identity matrix [8], λ and $\{\gamma_{M-N+1}, \cdots, \gamma_M\}$ cannot identically contribute to $\mathbf{\Sigma}_y$.

The SBL algorithm using the learning rules (6), (7), (12), (13) and (16) is denoted by ***T-SBL***.

IV. An Efficient Algorithm Processing in the Original Problem Space

The proposed T-SBL algorithm has excellent performance in terms of recovery performance (see Section VI). But it is not fast because it learns the parameters in a higher dimensional space instead of the original problem space [9]. For example, the dictionary matrix is of the size $NL \times ML$ in the bSBL framework, while it is only of the size $N \times M$ in the original MMV model. Interestingly, the MSBL developed for i.i.d. sources has complexity $\mathcal{O}(N^2M)$ and does not exhibit this drawback [26]. Motivated by this, we make a reasonable approximation and back-map T-SBL to the original space [10].

For convenience, we first list the MSBL algorithm derived in [26]:

$$\mathbf{\Xi}_x = \big(\mathbf{\Gamma}^{-1} + \frac{1}{\lambda}\mathbf{\Phi}^T\mathbf{\Phi}\big)^{-1} \tag{17}$$

$$\mathbf{X} = \mathbf{\Gamma}\mathbf{\Phi}^T\big(\lambda\mathbf{I} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T\big)^{-1}\mathbf{Y} \tag{18}$$

$$\gamma_i = \frac{1}{L}\|\mathbf{X}_{i\cdot}\|_2^2 + (\mathbf{\Xi}_x)_{ii}, \quad \forall i \tag{19}$$

An important observation is the lower dimension of the matrix operations involved in this algorithm. We attempt to achieve similar complexity for the T-SBL algorithm by adopting the following approximation:

$$\begin{aligned}\big(\lambda\mathbf{I}_{NL} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T\big)^{-1} &= \big(\lambda\mathbf{I}_{NL} + (\mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T) \otimes \mathbf{B}\big)^{-1} \\ &\approx \big(\lambda\mathbf{I}_N + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T\big)^{-1} \otimes \mathbf{B}^{-1}\end{aligned} \tag{20}$$

which is exact when $\lambda = 0$ or $\mathbf{B} = \mathbf{I}_L$. For high signal-to-noise ratio (SNR) or low correlation the approximation is quite reasonable. But our experiments will show that our algorithm adopting this approximation performs quite well over a broader range of conditions (see Section VI).

Now we use the approximation to simplify the γ_i learning rule (12). First, we consider the following term in (12):

$$\begin{aligned}\frac{1}{L}\mathrm{Tr}\big(\mathbf{B}^{-1}\mathbf{\Sigma}_x^i\big) &= \frac{1}{L}\mathrm{Tr}\Big[\gamma_i\mathbf{I}_L - \gamma_i^2(\boldsymbol{\phi}_i^T \otimes \mathbf{I}_L)(\lambda\mathbf{I}_{NL} + \\ &\qquad \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T)^{-1}(\boldsymbol{\phi}_i \otimes \mathbf{I}_L) \cdot \mathbf{B}\Big] \qquad (21) \\ &\approx \gamma_i - \frac{\gamma_i^2}{L}\mathrm{Tr}\Big[\Big(\big[\boldsymbol{\phi}_i^T\big(\lambda\mathbf{I}_N + \\ &\qquad \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T\big)^{-1}\boldsymbol{\phi}_i\big] \otimes \mathbf{B}^{-1}\Big)\mathbf{B}\Big] \\ &= \gamma_i - \frac{\gamma_i^2}{L}\mathrm{Tr}\Big[\Big(\boldsymbol{\phi}_i^T\big(\lambda\mathbf{I}_N + \\ &\qquad \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T\big)^{-1}\boldsymbol{\phi}_i\Big)\mathbf{I}_L\Big] \\ &= \gamma_i - \gamma_i^2\boldsymbol{\phi}_i^T\big(\lambda\mathbf{I}_N + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T\big)^{-1}\boldsymbol{\phi}_i \\ &= (\mathbf{\Xi}_x)_{ii} \qquad (22)\end{aligned}$$

where (21) follows the second equation in (6), and $\mathbf{\Xi}_x$ is given in (17). Using the same approximation (20), the $\boldsymbol{\mu}_x$ in (12) can be expressed as

$$\begin{aligned}\boldsymbol{\mu}_x &\approx (\mathbf{\Gamma} \otimes \mathbf{B})(\mathbf{\Phi}^T \otimes \mathbf{I}) \\ &\quad \cdot\big[\big(\lambda\mathbf{I} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T\big)^{-1} \otimes \mathbf{B}^{-1}\big]\mathrm{vec}(\mathbf{Y}^T) \qquad (23) \\ &= \big[\mathbf{\Gamma}\mathbf{\Phi}^T\big(\lambda\mathbf{I} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T\big)^{-1}\big] \otimes \mathbf{I} \cdot \mathrm{vec}(\mathbf{Y}^T) \\ &= \mathrm{vec}\big(\mathbf{Y}^T\big(\lambda\mathbf{I} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T\big)^{-1}\mathbf{\Phi}\mathbf{\Gamma}\big) \\ &= \mathrm{vec}(\mathbf{X}^T) \qquad (24)\end{aligned}$$

where (23) follows (5) and the approximation (20), and $\mathbf{X}$ is given in (18). Therefore, based on (22) and (24), we can transform the γ_i learning rule (12) to the following form:

$$\gamma_i \leftarrow \frac{1}{L}\mathbf{X}_{i\cdot}\mathbf{B}^{-1}\mathbf{X}_{i\cdot}^T + (\mathbf{\Xi}_x)_{ii}, \quad \forall i \tag{25}$$

To simplify the $\mathbf{B}$ learning rule (13), we note that

$$\begin{aligned}\mathbf{\Sigma}_x &= \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0\mathbf{D}^T(\lambda\mathbf{I} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{\Sigma}_0 \\ &= \mathbf{\Gamma} \otimes \mathbf{B} - (\mathbf{\Gamma} \otimes \mathbf{B})(\mathbf{\Phi}^T \otimes \mathbf{I})(\lambda\mathbf{I} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T)^{-1} \\ &\quad \cdot(\mathbf{\Phi} \otimes \mathbf{I})(\mathbf{\Gamma} \otimes \mathbf{B}) \\ &\approx \mathbf{\Gamma} \otimes \mathbf{B} - \big[(\mathbf{\Gamma}\mathbf{\Phi}^T) \otimes \mathbf{B}\big]\big[(\lambda\mathbf{I} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T)^{-1} \otimes \mathbf{B}^{-1}\big] \\ &\quad \cdot\big[(\mathbf{\Phi}\mathbf{\Gamma}) \otimes \mathbf{B}\big] \qquad (26) \\ &= \big(\mathbf{\Gamma} - \mathbf{\Gamma}\mathbf{\Phi}^T(\lambda\mathbf{I} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T)^{-1}\mathbf{\Phi}\mathbf{\Gamma}\big) \otimes \mathbf{B} \\ &= \mathbf{\Xi}_x \otimes \mathbf{B},\end{aligned}$$

where (26) uses the approximation (20). Using the definition (11), we have $\mathbf{\Sigma}_x^i = (\mathbf{\Xi}_x)_{ii}\mathbf{B}$. Therefore, the learning rule (13) becomes:

$$\mathbf{B} \leftarrow \Big(\frac{1}{M}\sum_{i=1}^{M}\frac{(\mathbf{\Xi}_x)_{ii}}{\gamma_i}\Big)\mathbf{B} + \frac{1}{M}\sum_{i=1}^{M}\frac{\mathbf{X}_{i\cdot}^T\mathbf{X}_{i\cdot}}{\gamma_i}. \tag{27}$$

From the learning rule above, we can directly construct a fixed-point learning rule, given by

$$\mathbf{B} \leftarrow \frac{1}{M(1-\rho)}\sum_{i=1}^{M}\frac{\mathbf{X}_{i\cdot}^T\mathbf{X}_{i\cdot}}{\gamma_i}$$

[8] Note that even all the sources are i.i.d. processes, the estimated $\mathbf{B}$ in practice is not an exact identity matrix.

[9] T-SBL can be directly used to solve the block sparsity models [13], [22], [41]. In this case, the algorithm directly performs in the original parameter space and thus it is not slow (compared to the speed of some other algorithms for the block sparsity models).

[10] By back-mapping, we mean we use some approximation to simplify the algorithm such that the simplified version directly operates in the parameter space of the original MMV model.

where $\rho = \frac{1}{M}\sum_{i=1}^{M}\gamma_i^{-1}(\mathbf{\Xi}_x)_{ii}$. To increase the robustness, however, we suggest using the rule below:

$$\widetilde{\mathbf{B}} \leftarrow \sum_{i=1}^{M}\frac{\mathbf{X}_{i\cdot}^T\mathbf{X}_{i\cdot}}{\gamma_i} \tag{28}$$

$$\mathbf{B} \leftarrow \widetilde{\mathbf{B}}/\|\widetilde{\mathbf{B}}\|_{\mathcal{F}} \tag{29}$$

where (29) is to remove the ambiguity between $\mathbf{B}$ and γ_i ($\forall i$). This learning rule performs well in high SNR cases and noiseless cases [11]. However, in low or medium SNR cases (e.g. $\mathrm{SNR} \leq 20\mathrm{dB}$) it is not robust due to errors from the estimated γ_i and $\mathbf{X}_{i\cdot}$. For these cases, we suggest adding a regularization item in $\widetilde{\mathbf{B}}$, namely,

$$\widetilde{\mathbf{B}} \leftarrow \sum_{i=1}^{M}\frac{\mathbf{X}_{i\cdot}^T\mathbf{X}_{i\cdot}}{\gamma_i} + \eta\mathbf{I} \tag{30}$$

where η is a positive scalar. This regularized form (30) ensures that $\widetilde{\mathbf{B}}$ is positive definite.

Similarly, we simplify the λ learning rule (16) as follows:

$$\begin{aligned}
\lambda &\leftarrow \frac{\|\mathbf{y}-\mathbf{D}\boldsymbol{\mu}_x\|_2^2 + \lambda\big[ML - \mathrm{Tr}(\mathbf{\Sigma}_x\mathbf{\Sigma}_0^{-1})\big]}{NL} \\
&= \frac{\|\mathbf{y}-\mathbf{D}\boldsymbol{\mu}_x\|_2^2 + \lambda\mathrm{Tr}(\mathbf{\Sigma}_0\mathbf{D}^T\mathbf{\Sigma}_y^{-1}\mathbf{D})}{NL} \qquad (31)\\
&\approx \frac{1}{NL}\|\mathbf{Y}-\mathbf{\Phi}\mathbf{X}\|_{\mathcal{F}}^2 + \frac{\lambda}{NL}\mathrm{Tr}\Big[(\mathbf{\Gamma}\otimes\mathbf{B})(\mathbf{\Phi}^T\otimes\mathbf{I}) \\
&\quad \cdot\big((\lambda\mathbf{I}+\mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T)^{-1}\otimes\mathbf{B}^{-1}\big)(\mathbf{\Phi}\otimes\mathbf{I})\Big] \qquad (32)\\
&= \frac{1}{NL}\|\mathbf{Y}-\mathbf{\Phi}\mathbf{X}\|_{\mathcal{F}}^2 + \frac{\lambda}{N}\mathrm{Tr}\big[\mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T(\lambda\mathbf{I}+\mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T)^{-1}\big] \qquad (33)
\end{aligned}$$

where in (31) we use the first equation in (6), and in (32) we use the approximation (20). Empirically, we find that setting the off-diagonal elements of $\mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T$ to zeros further improves the robustness of the λ learning rule in strongly noisy cases. In our experiments we will use the modified version when $\mathrm{SNR} \leq 20\mathrm{dB}$.

We denote the algorithm using the learning rules (17), (18), (25), (28), (29) (or (30)), and (33) by ***T-MSBL*** (the name emphasizes the algorithm is a *temporal* extension of MSBL). Note that T-MSBL cannot be derived by modifying the cost function of MSBL.

Comparing the γ_i learning rule of T-MSBL (Eq.(25)) with the one of MSBL (Eq.(19)), we observe that the only change is the replacement of $\|\mathbf{X}_{i\cdot}\|_2^2$ with $\mathbf{X}_{i\cdot}\mathbf{B}^{-1}\mathbf{X}_{i\cdot}^T$, which incorporates the temporal correlation of the sources. Hence, T-MSBL has only extra computational load for calculating the matrix $\mathbf{B}$ and the item $\mathbf{X}_{i\cdot}\mathbf{B}^{-1}\mathbf{X}_{i\cdot}^T$ [12]. Since the matrix $\mathbf{B}$ has a small size and is positive definite and symmetric, the extra computational load is low.

Note that $\mathbf{X}_{i\cdot}\mathbf{B}^{-1}\mathbf{X}_{i\cdot}^T$ is the quadratic Mahalanobis distance between $\mathbf{X}_{i\cdot}$ and its mean (a vector of zeros). In the following section we will get more insight into this change.

V. Analysis of Global Minimum and Local Minima

Since our bSBL framework generalizes the basic SBL framework, many proofs below are rooted in the theoretic work on the basic SBL [30]. However, some essential modifications are necessary in order to adapt the results to the bSBL model. Due to the equivalence of the original MMV model (2) and the transformed block sparsity model (3), in the following discussions we use (2) or (3) interchangeably and per convenience.

Throughout our analysis, the true source matrix is denoted by $\mathbf{X}_{\mathrm{gen}}$, which is the sparsest solution among all the possible solutions. The number of nonzero rows in $\mathbf{X}_{\mathrm{gen}}$ is denoted by K_0. We assume that $\mathbf{X}_{\mathrm{gen}}$ is full column-rank, the dictionary matrix $\mathbf{\Phi}$ satisfies the URP condition [5], and the matrix $\mathbf{B}$ (or $\mathbf{B}_i, \forall i$) and its estimate are positive definite.

A. Analysis of the Global Minimum

We have the following result on the global minimum of the cost function (8) [13]:

Theorem 1: In the limit as $\lambda \to 0$, assuming $K_0 < (N+L)/2$, for the cost function (8) the unique global minimum $\widehat{\boldsymbol{\gamma}} \triangleq [\widehat{\gamma}_1, \cdots, \widehat{\gamma}_M]$ produces a source estimate $\widehat{\mathbf{X}}$ that equals to $\mathbf{X}_{\mathrm{gen}}$ irrespective of the estimated $\widehat{\mathbf{B}}_i$, $\forall i$, where $\widehat{\mathbf{X}}$ is obtained from $\mathrm{vec}(\widehat{\mathbf{X}}^T) = \widehat{\mathbf{x}}$ and $\widehat{\mathbf{x}}$ is computed using Eq.(7).

The proof is given in the Appendix.

If we change the condition $K_0 < (N+L)/2$ to $K_0 < N$, then we have the conclusion that the source estimate $\widehat{\mathbf{X}}$ equals to $\mathbf{X}_{\mathrm{gen}}$ with probability 1, irrespective of $\widehat{\mathbf{B}}_i$, $\forall i$. This is due to the result in [48] that if $K_0 < N$ the above conclusion still holds for all $\mathbf{X}$ except on a set with zero measure.

Note that $\widehat{\boldsymbol{\gamma}}$ is a function of the estimated $\widehat{\mathbf{B}}_i$ ($\forall i$). However, the theorem implies that even when the estimated $\widehat{\mathbf{B}}_i$ is different from the true $\mathbf{B}_i$, the estimated sources are the true sources at the global minimum of the cost function. As a reminder, in deriving our algorithms, we assumed $\mathbf{B}_i = \mathbf{B}$ ($\forall i$) to avoid overfitting. Theorem 1 ensures our algorithms using this strategy also have the global minimum property. Also, the theorem explains why MSBL has the ability to exactly recover true sources in noiseless cases even when sources are temporally correlated. But we hasten to add that this does not mean $\mathbf{B}$ is not important for the performance of the algorithms. For instance, MSBL is more frequently attracted to local minima than our proposed algorithms, as experiments show later.

B. Analysis of the Local Minima

In this subsection we discuss the local minimum property of the cost function $\mathcal{L}$ in (8) with respect to $\boldsymbol{\gamma} \triangleq [\gamma_1, \cdots, \gamma_M]$,

[11] Note that in (28) when the number of distinct nonzero rows in $\mathbf{X}$ is smaller than the number of measurement vectors, the matrix $\widetilde{\mathbf{B}}$ is not invertible. But this case is rarely encountered in practical problems, since in practice the number of measurement vectors is generally small, as we explained previously. The presence of noise in practical problems also requires the use of the regularized form (30), which is always invertible.

[12] Here we do not compare the λ learning rules of both algorithms, since in some cases one can feed the algorithms with suitable fixed values of λ, instead of using the λ learning rules. However, the computational load of the simplified λ learning rule of T-MSBL is also not high.

[13] For convenience, in this theorem we consider the cost function with $\mathbf{\Sigma}_0$ given by (4), i.e. the one before we use our strategy to avoid the overfitting.

in which $\mathbf{\Sigma}_0 = \mathbf{\Gamma} \otimes \mathbf{B}$ for fixed $\mathbf{B}$. Before presenting our results, we provide two lemmas needed to prove the results.

Lemma 1: $\log|\mathbf{\Sigma}_y| \triangleq \log|\lambda\mathbf{I} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T|$ is concave with respect to $\boldsymbol{\gamma}$.

This can be shown using the composition property of concave functions [49].

Lemma 2: $\mathbf{y}^T\mathbf{\Sigma}_y^{-1}\mathbf{y}$ equals a constant C when $\boldsymbol{\gamma}$ satisfies the linear constraints

$$\mathbf{A} \cdot (\boldsymbol{\gamma} \otimes \mathbf{1}_L) = \mathbf{b} \tag{34}$$

with

$$\mathbf{b} \triangleq \mathbf{y} - \lambda\mathbf{u} \tag{35}$$

$$\mathbf{A} \triangleq (\mathbf{\Phi} \otimes \mathbf{B})\mathrm{diag}(\mathbf{D}^T\mathbf{u}) \tag{36}$$

where $\mathbf{A}$ is full row rank, $\mathbf{1}_L$ is an $L \times 1$ vector of ones, and $\mathbf{u}$ is any fixed vector such that $\mathbf{y}^T\mathbf{u} = C$.

The proof is given in the Appendix. According to the definition of basic feasible solution (BFS) [50], we know that if $\boldsymbol{\gamma}$ satisfies Equation (34), then it is a BFS to (34) if $\|\boldsymbol{\gamma}\|_0 \leq NL$, or a degenerate BFS to (34) if $\|\boldsymbol{\gamma}\|_0 < NL$. Now we give the following result:

Theorem 2: Every local minimum of the cost function $\mathcal{L}$ with respect to $\boldsymbol{\gamma}$ is achieved at a solution with $\|\widehat{\boldsymbol{\gamma}}\|_0 \leq NL$, regardless of the values of λ and $\mathbf{B}$.

The proof is given in the Appendix.

Admittedly, the bound on the local minima $\|\widehat{\boldsymbol{\gamma}}\|_0$ is loose, and it is not meaningful when $NL > M$. However, it is empirically found that $\|\widehat{\boldsymbol{\gamma}}\|_0$ is very smaller than NL, typically smaller than N.

Now, we calculate the local minima of the cost function $\mathcal{L}$. The result can provide some insights to the role of $\mathbf{B}$. Particularly, we are more interested in the local minima satisfying $\|\widehat{\boldsymbol{\gamma}}\|_0 \leq N$, since the global minimum satisfies $\|\widehat{\boldsymbol{\gamma}}\|_0 < N$. For these local minima, we have the following result:

Lemma 3: In noiseless cases ($\lambda \to 0$), for every local minimum of $\mathcal{L}$ that satisfies $\|\widehat{\boldsymbol{\gamma}}\|_0 \triangleq K \leq N$, its i-th nonzero element is given by $\widehat{\gamma}_{(i)} = \frac{1}{L}\widetilde{\mathbf{X}}_{i\cdot}\mathbf{B}^{-1}\widetilde{\mathbf{X}}_{i\cdot}^T\,(i = 1, \cdots, K)$, where $\widetilde{\mathbf{X}}_{i\cdot}$ is the i-th nonzero row of $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{X}}$ is the basic feasible solution to $\mathbf{Y} = \mathbf{\Phi}\mathbf{X}$.

The proof is given in the Appendix.

From this lemma we immediately have the closed form of the global minimum.

$\mathbf{B}$ actually plays a role of temporally whitening the sources during the learning of $\boldsymbol{\gamma}$. To see this, assume all the sources have the same correlation structure, i.e. share the same matrix $\mathbf{B}$. Let $\mathbf{Z}_{i\cdot} \triangleq \widetilde{\mathbf{X}}_{i\cdot}\mathbf{B}^{-1/2}$. From Lemma 3, at the global minimum we have $\widehat{\gamma}_{(i)} = \frac{1}{L}\mathbf{Z}_{i\cdot}\mathbf{Z}_{i\cdot}^T\,(i = 1, \cdots, K_0)$. On the other hand, in the case of i.i.d. sources, at the global minimum we have $\widehat{\gamma}_{(i)} = \frac{1}{L}\widetilde{\mathbf{X}}_{i\cdot}\widetilde{\mathbf{X}}_{i\cdot}^T\,(i = 1, \cdots, K_0)$. So the results for the two cases have the same form. Since $E\{\mathbf{Z}_{i\cdot}^T\mathbf{Z}_{i\cdot}\} = \gamma_i\mathbf{I}$, we can see in the learning of $\boldsymbol{\gamma}$, $\mathbf{B}$ plays the role of whitening each source. This gives us a motivation to modify most state-of-the-art iterative reweighted algorithms by temporally whitening the estimated sources during iterations [32], [33].

VI. Computer Experiments

Extensive computer experiments have been conducted and a few representative and informative results are presented. All the experiments consisted of 1000 independent trials. In each trial a dictionary matrix $\mathbf{\Phi} \in \mathbb{R}^{N\times M}$ was created with columns uniformly drawn from the surface of a unit hypersphere (except the experiment in Section VI-G), as advocated by Donoho et al [51]. And the source matrix $\mathbf{X}_{\mathrm{gen}} \in \mathbb{R}^{M\times L}$ was randomly generated with K nonzero rows (i.e. sources). In each trial the indexes of the sources were randomly chosen. In most experiments (except to the experiment in Section VI-D) each source was generated as AR(1) process. Thus the AR coefficient of the i-th source, denoted by β_i, indicated its temporal correlation. As done in [20], [24], for noiseless cases, the ℓ_2 norm of each source was rescaled to be uniformly distributed between $1/3$ and 1; for noisy cases, rescaled to be unit norm. Finally, the measurement matrix $\mathbf{Y}$ was constructed by $\mathbf{Y} = \mathbf{\Phi}\mathbf{X}_{\mathrm{gen}} + \mathbf{V}$ where $\mathbf{V}$ was a zero-mean homoscedastic Gaussian noise matrix with variance adjusted to have a desired value of SNR, which is defined by $\mathrm{SNR(dB)} \triangleq 20\log_{10}(\|\mathbf{\Phi}\mathbf{X}_{\mathrm{gen}}\|_{\mathcal{F}}/\|\mathbf{V}\|_{\mathcal{F}})$.

We used two performance measures. One was the *Failure Rate* defined in [26], which indicated the percentage of failed trials in the total trials. In noiseless cases, a failed trial was recognized if the indexes of estimated sources were not the same as the true indexes. In noisy cases, since any algorithm cannot recover $\mathbf{X}_{\mathrm{gen}}$ exactly in these cases, a failed trial was recognized if the indexes of estimated sources with the K largest ℓ_2 norms were not the same as the true indexes. In noisy cases, the *mean square error* (MSE) was also used as a performance measure, defined by $\|\widehat{\mathbf{X}} - \mathbf{X}_{\mathrm{gen}}\|_{\mathcal{F}}^2/\|\mathbf{X}_{\mathrm{gen}}\|_{\mathcal{F}}^2$, where $\widehat{\mathbf{X}}$ was the estimated source matrix.

In our experiments we compared our T-SBL and T-MSBL with the following algorithms:

- MSBL, proposed in [26] [14];
- MFOCUSS, the regularized M-FOCUSS proposed in [12]. In all the experiments, we set its p-norm $p = 0.8$, as suggested by the authors [15];
- SOB-MFOCUSS, a smoothness constrained M-FOCUSS proposed in [43]. In all the experiments, we set its p-norm $p = 0.8$. For its smoothness matrix, we chose the identity matrix when $L \leq 2$, and a second-order smoothness matrix when $L \geq 3$, as suggested by the authors. Since in our experiments L is small, no overlap blocks were used [16];
- ISL0, an improved smooth ℓ_0 algorithm for the MMV model which was proposed in [52]. The regularization parameters were chosen according to the authors' sug-

[14]The MATLAB code was downloaded at `http://dsp.ucsd.edu/~zhilin/MSBL_code.zip`.

[15]The MATLAB code was downloaded at `http://dsp.ucsd.edu/~zhilin/MFOCUSS.m`.

[16]The MATLAB code was provided by the first author of [43] in personal communication. In the code the second-order smoothness matrix $\mathbf{S}$ was defined as (in MATLAB notations): $\mathbf{S} = \mathrm{eye}(L) - 0.25 * (\mathrm{diag}(\mathbf{e}(1:L-1),-1) + \mathrm{diag}(\mathbf{e}(1:L-1),1) + (\mathrm{diag}(\mathbf{e}(1:L-2),-2) + \mathrm{diag}(\mathbf{e}(1:L-2),2)))$, where $\mathbf{e}$ is an $L \times 1$ vector with ones.

gestions [17];

- Reweighted ℓ_1/ℓ_2, an iterative reweighted ℓ_1/ℓ_2 algorithm suggested in [25]. It is an MMV extension of the iterative reweighted ℓ_1 algorithm [39] via the mixed ℓ_1/ℓ_2 norm. The algorithm is given by
  1) Set the iteration count k to zero and $w_i^{(0)} = 1, i = 1, \cdots, M$
  2) Solve the weighted MMV ℓ_1 minimization problem
  $$\mathbf{X}^{(k)} = \arg\min \sum_{i=1}^{M} w_i^{(k)} \|\mathbf{X}_{i\cdot}\|_2 \quad \text{s.t. } \mathbf{Y} = \mathbf{\Phi X}$$
  3) Update the weights for each $i = 1, \cdots, M$
  $$w_i^{(k+1)} = \frac{1}{\|\mathbf{X}_{i\cdot}^{(k)}\|_2 + \epsilon^{(k)}}$$
  where $\epsilon^{(k)}$ is adaptively adjusted as in [39];
  4) Terminate on convergence or when k attains a specified maximum number of iterations $k_{\max}$. Otherwise, increment k and go to Step 2).

  For noisy cases, Step 2) is modified to
  $$\mathbf{X}^{(k)} = \arg\min \sum_{i=1}^{M} w_i^{(k)} \|\mathbf{X}_{i\cdot}\|_2 \quad \text{s.t. } \|\mathbf{Y} - \mathbf{\Phi X}\|_{\mathcal{F}} \leq \delta$$
  Throughout our experiments, $k_{\max} = 5$. We implemented it using the CVX optimization toolbox [18].

In noisy cases, we chose the optimal values for the regularization parameter λ in MFOCUSS and the parameter δ in Reweighted ℓ_1/ℓ_2 by exhaustive search. Practically, we used a set of candidate parameter values and for each value we ran an algorithm for 50 trials, and then picked up the one which gave the smallest averaged failure rate. By comparing enough number of candidate values we could ensure a nearly optimal value of the regularization parameter for this algorithm. For T-MSBL, T-SBL and MSBL, we fixed $\lambda = 10^{-9}$ for noiseless cases, and used their λ learning rules for noisy cases. Besides, for T-MSBL we chose the learning rule (30) with $\eta = 2$ to estimate $\mathbf{B}$ when $\text{SNR} \leq 15\text{dB}$.

For reproducibility, the experiment codes can be downloaded at `http://dsp.ucsd.edu/~zhilin/TSBL_code.zip`.

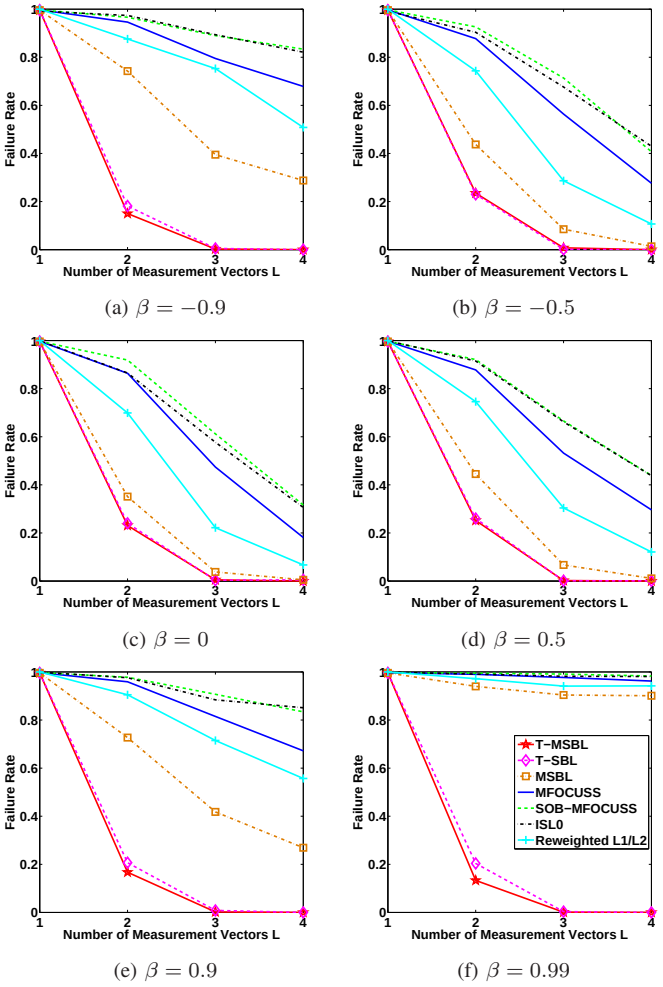
A. Benefit from Multiple Measurement Vectors at Different Temporal Correlation Levels

In this experiment we study how algorithms benefit from multiple measurement vectors and how the benefit is discounted by the temporal correlation of sources. The dictionary matrix $\mathbf{\Phi}$ was of the size 25×125 and the number of sources $K = 12$. The number of measurement vectors L varied from 1 to 4. No noise was added. All the sources were AR(1) processes with the common AR coefficient β, such that we could easily observe the relationship between temporal correlation and algorithm performance. Note that for small L, modeling sources as AR(1) processes, instead of AR(p) processes with $p > 1$, is sufficient to cover wide ranges of temporal structure. We compared algorithms at six different temporal correlation levels, i.e. $\beta = -0.9, -0.5, 0, 0.5, 0.9, 0.99$.

Figure 1 shows that with L increasing, all the algorithms had better performance. But as $|\beta| \to 1$, for all the compared algorithms the benefit from multiple measurement vectors diminished. One surprising observation is that our T-MSBL and T-SBL had excellent performance in all cases, no matter what the temporal correlation was. Notice that even sources had no temporal correlation ($\beta = 0$), T-MSBL and T-SBL still had better performance than MSBL.

Next we compare all the algorithms in noisy environments. We set $\text{SNR} = 25\text{dB}$ while kept other experimental settings unchanged. The behaviors of all the algorithms were similar to the noiseless case. To save space, we only present the cases of $\beta = 0.7$ and $\beta = 0.9$ in Fig.2.

Since the performance of all the algorithms at a given correlation level β is the same as their performance at the correlation level $-\beta$, in the following we mainly show their performance at positive correlation levels.

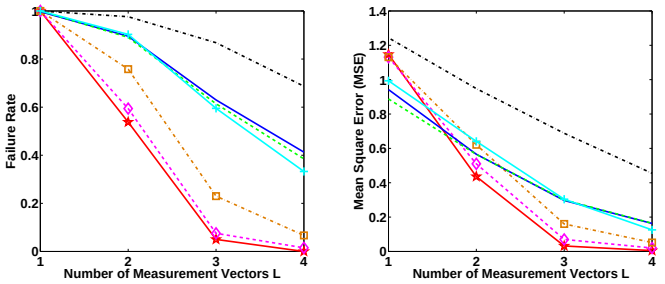


(a) $\beta = -0.9$ (b) $\beta = -0.5$ (c) $\beta = 0$ (d) $\beta = 0.5$ (e) $\beta = 0.9$ (f) $\beta = 0.99$

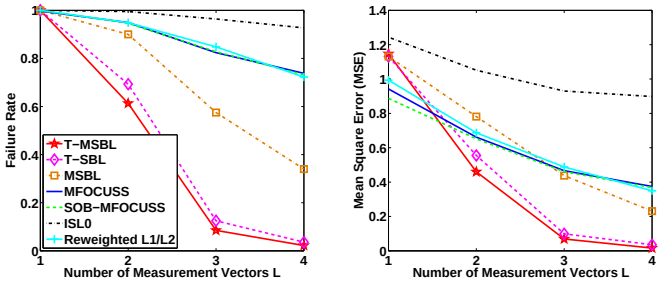
Fig. 1. Performance of all the algorithms at different temporal correlation levels when L varied from 1 to 4.

[17]The MATLAB code was provided by the first author of [52] in personal communication.

[18]The toolbox was downloaded at: `http://cvxr.com/cvx/`

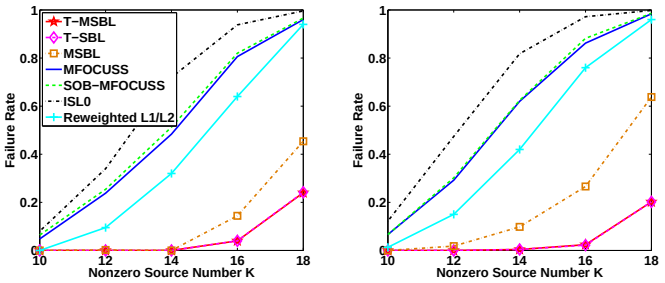


(a) $\beta = 0.7$ (b) $\beta = 0.7$

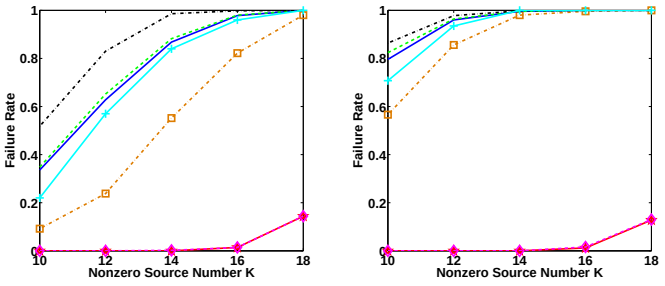


(c) $\beta = 0.9$ (d) $\beta = 0.9$

Fig. 2. Performance of all the algorithms at different temporal correlation levels when L varied from 1 to 4 and SNR was 25 dB.



(a) $\beta = 0$ (b) $\beta = 0.5$



(c) $\beta = 0.9$ (d) $\beta = 0.99$

Fig. 3. Failure rates of all the algorithms when K varied from 10 to 18 at different temporal correlation levels.

B. Recovered Source Number at Different Temporal Correlation Levels

In this experiment we study the effects of temporal correlation on the number of accurately recovered sources in a noiseless case. The dictionary matrix $\mathbf{\Phi}$ was of the size 25×125. L was 4. K varied from 10 to 18. The sources were generated in the same manner as before. Algorithms were compared at four different temporal correlation levels, i.e. $\beta = 0$, 0.5, 0.9, and 0.99. Results (Fig.3) show that T-MSBL and T-SBL accurately recovered much more sources than other algorithms, especially at high temporal correlation levels. This indicates that our proposed algorithms are very advantageous in the cases when the source number is large.

C. Ability to Handle Highly Underdetermined Problem

Most published works only compared algorithms in mildly underdetermined cases, namely, the ratio of M/N was about $2 \sim 5$. However, in some applications such as neuroimaging, one can easily have $N \approx 100$ and $M \approx 100000$. So, in this experiment we compare the algorithms in the highly underdetermined cases when N was fixed at 25 and M/N varied from 1 to 25. The source number K was 12, and the measurement vector number L was 4. SNR was 25 dB. Different to previous experiments, all the sources were AR(1) processes but with different AR coefficients. Their AR coefficients were uniformly chosen from $(0.5, 1)$ at random. Results are presented in Fig.4, from which we can see that when $M/N \geq 10$, all the compared algorithms had large errors. In contrast, our proposed algorithms had much lower errors. Note that due to the performance trade-off between N and M, if one increases N, algorithms can keep the same recovery performance for larger M/N.

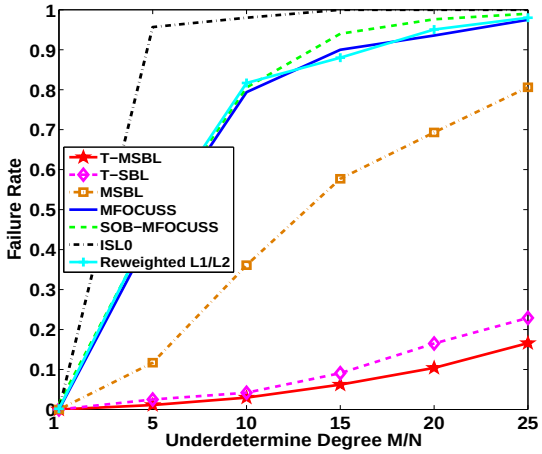


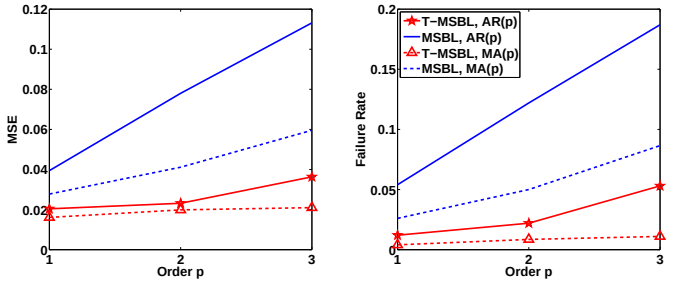
Fig. 4. Performance comparison in highly underdetermined cases.

D. Recovery Performance for Different Kinds of Sources

In previous experiments all the sources were AR(1) processes. Although we have pointed out that for small L modeling sources by AR(1) processes is sufficient, here we carry out an experiment to show our algorithms maintaining the same superiority for various time series. Since from previous experiments we have seen that T-SBL has similar performance to T-MSBL, and that MSBL has the best performance among the compared algorithms, in this experiment we only compare T-MSBL with MSBL.

The dictionary matrix was of the size 25×125. L was 4. K was 14. SNR was 25dB. First we generated sources as three kinds of AR processes, i.e. AR(p) $(p = 1, 2, 3)$. All the AR coefficients were randomly uniformly chosen from the feasible regions such that the processes were stable. We

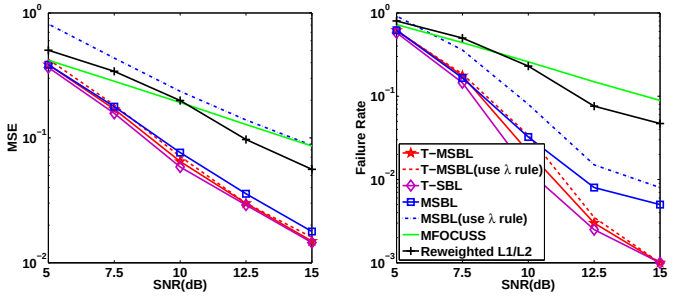
examined the algorithms' performance as a function of the AR order p. Results are given in Fig.5, showing that T-MSBL again outperformed MSBL. With large p, the performance gap between the two algorithms increased. We repeated the previous experiment with the same experiment settings except that we replaced the $\mathrm{AR}(p)$ sources by moving-averaging sources $\mathrm{MA}(p)$ ($p = 1, 2, 3$). The MA coefficients were uniformly chosen from $(0, 1]$ at random. Again, we obtained the same results. These results imply that our algorithms maintain their superiority for various temporally structured sources, not only AR processes.



(a) (b)

Fig. 5. Performance of T-MSBL and MSBL for different AR(p) sources and different MA(p) sources measured in terms of MSE and failure rates.

E. Recovery Ability at Different Noise Levels



(a) (b)

Fig. 6. Performance of various algorithms at different noise levels.

From previous experiments we have seen that the proposed algorithms significantly outperformed all the compared algorithms in noiseless scenarios and mildly noisy cases, even though to derive T-MSBL we used the approximation (20) which takes the equal sign only when $\mathbf{B} = \mathbf{I}$ (no temporal correlation) or $\lambda = 0$ (no noise). Some natural questions may be raised: What is the performance of T-SBL and T-MSBL in strongly noisy cases? Is it still beneficial to exploit temporal correlation in these cases? To answer these questions, we carry out the following experiment.

The dictionary matrix was of the size 25×125. The number of measurement vectors L was 4. The source number K was 7. All the sources were AR(1) processes and the temporal correlation of each source was 0.8. SNR varied from 5 dB to 15 dB. The experiment was repeated 2000 trials. We compared the proposed T-SBL, T-MSBL with three representative algorithms, i.e. MSBL, MFOCUSS, and Reweighted ℓ_1/ℓ_2.

Note that in low SNR cases, the estimated $\mathbf{B}$ of T-SBL and T-MSBL can include large errors, and thus the estimated amplitudes of sources are distorted. To reduce the distortion, we set $\mathbf{B} = \mathbf{I}$ once the number of nonzero γ_i was less than N during the learning procedure. The reason is that the role of $\mathbf{B}$ is to prevent T-SBL/T-MSBL from arriving at local minima; once the algorithms approach global minima very closely, $\mathbf{B}$ is no longer useful.

Also note that the λ learning rules of T-SBL, T-MSBL and MSBL may not lead to optimal performance in low SNR cases. To avoid the potential disturbance of these λ learning rules, we provided the three SBL algorithms with the optimal λ^*'s, which were obtained by the exhaustive search method stated previously.

Figure 6 shows that T-SBL and T-MSBL outperformed other algorithms in all the noise levels. This implies that even in low SNR cases exploiting temporal correlation of sources is beneficial.

But we want to emphasize that although the λ learning rules of the three SBL algorithms may not be optimal in low SNR cases, our proposed λ learning rules can lead to near-optimal performance, compared to the one of MSBL. To see this, we ran T-MSBL and MSBL again, but this time both algorithms used their λ learning rules. T-MSBL used the modified version of the λ learning rule (33), i.e. setting the off-diagonal elements of $\mathbf{\Phi\Gamma\Phi}^T$ to zeros. The results (Fig. 6) show that MSBL had very poor performance when using its λ learning rule. In contrast, T-MSBL's performance was very close to its performance when using its optimal λ^* [19]. The results indicate our proposed algorithms are advantageous in practical applications, since in practice the optimal λ^*'s are difficult to obtain, if not impossible.

F. Temporal Correlation: Beneficial or Detrimental?

From previous experiments one may think that temporal correlation is always harmful to algorithms' performance, at least not helpful. However, in this experiment we will show that when SNR is high, the performance of our proposed algorithms increases with increasing temporal correlation.

We set $N = 25$, $L = 4$, $K = 14$, and $\mathrm{SNR} = 50\mathrm{dB}$. The underdeterminacy ratio M/N varied from 5 to 20. Sources were generated as AR(1) processes with the common AR coefficient β. We considered the performance of T-MSBL and MSBL in three cases, i.e. the temporal correlation β was 0, 0.5, and 0.9, respectively. Results are shown in Fig.7. As expected, the performance of MSBL deteriorated with increasing temporal correlation. But the behavior of T-MSBL was rather counterintuitive. It is surprising that the best performance of T-MSBL was not achieved at $\beta = 0$, but at $\beta = 0.9$. Clearly, high temporal correlation enabled T-MSBL to handle more highly underdetermined problems. For example, its performance at $M/N = 20$ with $\beta = 0.9$ was better than that at $M/N = 15$

[19]T-SBL had the same behavior. But for clarity we do not present its performance curve.

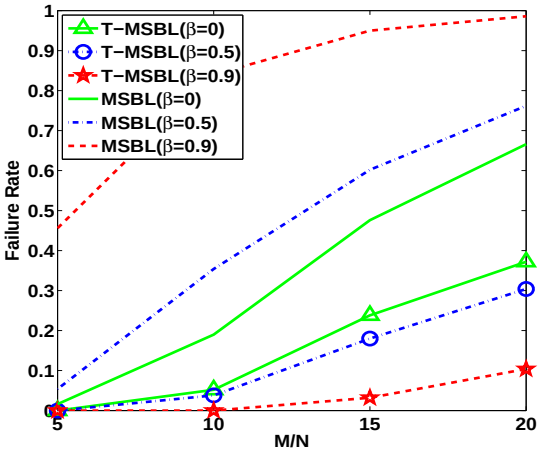


Fig. 7. Behaviors of MSBL and T-MSBL at different temporal correlation levels when SNR $= 50$dB.

with $\beta = 0.5$ or $\beta = 0$. The same phenomenon was observed in noiseless cases as well, and was observed for T-SBL.

The results indicating that temporal correlation is helpful may appear counterintuitive at first glance. A closer examination of the sparse recovery problems indicates a plausible explanation. There are two elements to the sparse recovery task; one is the location of the nonzero entries and the other is the value for the nonzero entries. Both tasks interact and combine to determine the overall performance. Correlation helps the estimation of the values for the nonzero entries and this may be important for the problem when dealing with finite matrices and may be lost when dealing with limiting results as the matrix dimension go to infinity. A more rigorous study of the interplay between estimation of the values and estimation of the locations is an interesting topic.

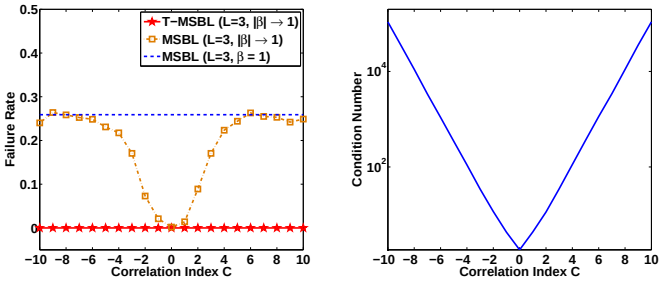
G. An Extreme Experiment on the Importance of Exploiting Temporal Correlation

It may be natural to take for granted that in noiseless cases, when source vectors are almost identical, algorithms have almost the same performance as in the case when only one measurement vector is available. In the following we show that it is not the case.

We designed a noiseless experiment. First, we generated a Hadamard matrix of the size 128×128. From the matrix, 40 rows were randomly selected in each trial and formed a dictionary matrix of the size 40×128. The source number K was 12, and the measurement vector number L was 3. Sources were generated as AR(1) processes with the common AR coefficient β, where $\beta = \text{sign}(C)(1 - 10^{-|C|})$. We varied C from -10 to 10 in order to see how algorithms behaved when the absolute temporal correlation, $|\beta|$, approximated to 1.

Figure 8 (a) shows the performance curves of T-MSBL and MSBL when $|\beta| \to 1$, and also shows the performance curve of MSBL when $\beta = 1$. We observe an interesting phenomenon. First, as $|\beta| \to 1$, MSBL's performance closely approximated to its performance in the case of $\beta = 1$. It seems to make sense, because when $|\beta| \to 1$, every source vector provides almost the same information on locations and amplitudes of nonzero elements. Counter-intuitively, no matter how close $|\beta|$ was to 1, the performance of T-MSBL did not change. Figure 8 (b) shows the averaged condition numbers of the submatrix formed by the sources (i.e. nonzero rows in $\mathbf{X}_{\text{gen}}$) at different correlation levels. We can see that the condition numbers increased with the increasing temporal correlation. This suggests that T-MSBL was not sensitive to the ill-condition issue in the source matrix, while MSBL is very sensitive. Although not shown here, we found that T-SBL had the same behavior as T-MSBL, while other MMV algorithms had the same behaviors as MSBL. The phenomenon was also observed when using other dictionary matrices, such as random Gaussian matrices.

These results emphasize the importance of exploiting the temporal correlation, and also motivate future theoretical studies on the temporal correlation and the ill-condition issue of source matrices.



(a) (b)

Fig. 8. (a) The performance and (b) the condition numbers of the submatrix formed by sources when the temporal correlation approximated to 1. The temporal correlation $\beta = \text{sign(C)}(1 - 10^{-|\text{C}|})$, where C was the correlation index varying from -10 to 10.

VII. Discussions

Although there are a few works trying to exploit temporal correlation in the MMV model, based on our knowledge no works have explicitly studied the effects of temporal correlation, and no existing algorithms are effective in the presence of such correlation. Our work is a starting point in the direction of considering temporal correlation in the MMV model. However, there are many issues that are unclear so far. In this section we discuss some of them.

A. The Matrix $\mathbf{B}$*: Trade-off Between Accurately Modeling and Preventing Overfitting*

In our algorithm development we used one single matrix $\mathbf{B}$ as the covariance matrix (up to a scalar) for each source model in order to avoid overfitting. Mathematically, it is straightforward to extend our algorithms to use multiple matrices to capture the covariance structures of sources. For example, one can classify sources into several groups, say G groups, and the sources in a group are all assigned by a common matrix $\mathbf{B}_i$ $(i = 1, \cdots, G,\ G \ll M)$ as the covariance matrix (up to a scalar). It seems that this extension can better capture the

covariance structures of sources while still avoiding overfitting. However, we find that this extension (even for $G = 2$) has much poorer performance than our proposed algorithms and MSBL. One possible reason is that during the early stage of the learning procedure of our algorithms, the estimated sources from each iteration are far from the true sources, and thus grouping them based on their covariance structures is difficult, if not impossible. The grouping error may cause avalanche effect, leading to the noted poor performance. Reducing the grouping error and more accurately capturing the temporal correlation structures is an area for future work.

However, as we have stated, $\mathbf{B}$ plays a role of whitening each source. In our recent work [32], [33] we found that the operation $\mathbf{X}_{i\cdot}\mathbf{B}^{-1}\mathbf{X}_{i\cdot}^T$ $(\forall i)$ can replace the row-norms (such as the ℓ_2 norm and the ℓ_∞ norm) in iterative reweighted ℓ_2 and ℓ_1 algorithms for the MMV model, functioning as a row regularization. This indicates that using one single matrix $\mathbf{B}$ may be a better method than using multiple matrices $\mathbf{B}_1, \cdots, \mathbf{B}_G$.

On the other hand, there may be many ways to parameterize and estimate $\mathbf{B}$. In this work we provide a general method to estimate $\mathbf{B}$. In [31] we proposed a method to parameterize $\mathbf{B}$ by a hyperparameter β, i.e.,

$$\mathbf{B} = \begin{bmatrix} 1 & \beta & \cdots & \beta^{L-1} \\ \beta & 1 & \cdots & \beta^{L-2} \\ \vdots & \vdots & \ddots & \vdots \\ \beta^{L-1} & \beta^{L-2} & \cdots & 1 \end{bmatrix}$$

which equivalently assumes the sources are AR(1) processes with the common AR coefficient β. The resulting algorithms have good performance as well. Also, for low SNR cases in our experiments, we added an identity matrix (with a scalar) to the estimated $\mathbf{B}$ in T-MSBL, and achieved satisfying performance. All these imply that $\mathbf{B}$ could have many forms. Finding the forms that are advantageous in strongly noisy environments is an important issue and needs further study.

B. The Parameter λ: Noise Variance or Regularization Parameter?

In our algorithms the covariance matrix of the multi-channel noise $\mathbf{V}_{\cdot i}$ $(i = 1, \cdots, L)$ is $\lambda\mathbf{I}_N$ with the implicit assumption that each channel noise has the same variance λ. It is straightforward to extend our algorithms to consider the general noise covariance matrix $\mathrm{diag}([\lambda_1, \cdots, \lambda_N])$, i.e. assuming different channel noise have different variance. However, this largely increases parameters to estimate, and thus we may once again encounter an overfitting problem (similar to the overfitting problem in learning the matrix $\mathbf{B}_i$).

Some works [34], [53] considered alternative noise covariance models. In [34] the authors assumed that the covariance matrix of multi-channel noise is $\lambda\mathbf{C}$, instead of $\lambda\mathbf{I}_N$, where $\mathbf{C}$ is a known positive definite and symmetric matrix and λ is an unknown noise-variance parameter. This model may better capture the noise covariance structures, but generally one does not know the exact value of $\mathbf{C}$. Thus there is no clear benefit from this covariance model. In [53], instead of deriving a learning rule for the noise covariance inside the SBL framework, the authors estimated the noise covariance by a method independent of the SBL framework. But this method is based on a large number of measurement vectors, and has a high computational load.

On the other hand, due to the works in [25], [27], [54], which connected SBL algorithms to traditional convex relaxation methods such as Lasso [37] and Basis Pursuit Denoising [38], it was found that λ is functionally the same as the regularization parameters of those convex relaxation algorithms. This suggests the use of methods such as the modified L-curve procedure [55] or the cross-validation [37], [38] to choose λ especially in strongly noisy environments. It is also interesting to see that SBL algorithms could adopt the continuation strategies [56], [57], used in Lasso-type algorithms, to adjust the value of λ for better recovery performance or faster speed.

However, if some channels contain very large noise (e.g. outliers) and the number of such channels is very small, then as suggested in [58], we can extend the dictionary matrix $\mathbf{\Phi}$ to $[\mathbf{\Phi}, \mathbf{I}]$ and perform any sparse signal recovery algorithms without modification. The estimated 'sources' associated with the identity dictionary matrix are these large noise components.

C. Connections to Other Models

In fact, our bSBL framework is a block sparsity model [13], [22], [41], and thus the derived T-SBL algorithm can be directly used for this model. Compared to most existing algorithms derived in this model [22], [41], [59], an important difference is that T-SBL considers the correlation within each block.

The time-varying sparsity model [60], [61] is another related model. Different to our MMV model that assumes the support of each source vector is the same, the time-varying sparsity model assumes the support is slowly time-varying. It is interesting to note that this model can be approximated by concatenation of several MMV models, where in each MMV model the support does not change. Thus our proposed T-SBL and T-MSBL can be used for this model. The results are appealing, as shown in our recent work [33].

It should be noted that the proposed algorithms can be directly used for the SMV model. In this case the matrix $\mathbf{B}$ reduces to a scalar, and the γ_i learning rules are the same as the one in the basic SBL algorithm [30]. But due to the effective λ learning rules, our algorithms are superior to the basic SBL algorithm, especially in noisy cases.

VIII. Conclusions

We addressed a multiple measurement vector (MMV) model in practical scenarios, where the source vectors are temporally correlated and the number of measurement vectors is small due to the common sparsity constraint. We showed that existing algorithms have poor performance when temporal correlation is present, and thus they have limited ability in practice. To solve this problem, we proposed a block sparse Bayesian learning framework, which allows for easily modeling the temporal correlation and incorporating this information into derived algorithms. Based on this framework, we derived two algorithms, namely, T-SBL and T-MSBL. The latter can be

seen as an extension of MSBL by replacing the ℓ_2 norm imposed on each source with a Mahalanobis distance measure. Extensive experiments have shown that the proposed algorithms have superior performance to many state-of-the-art algorithms. Theoretical analysis also has shown that the proposed algorithms have desirable global and local minimum properties.

ACKNOWLEDGEMENT

Z.Z would like to thank Dr. David Wipf for his considerable help with the study of SBL, Ms. Jing Wan for kind help in performing some experiments, Mr. Tim Mullen for kind help in the paper writing, Dr. Rafal Zdunek for providing the code of SOB-MFOCUSS, and Mr. Md Mashud Hyder for providing the code of ISL0. The authors thank the reviewers for their helpful comments and especially thank a reviewer for the idea of using multiple covariance matrices, which is discussed in Section VII.A.

APPENDIX

A. Outline of the Proof of Theorem 1

Since the proof is a generalization of the Theorem 1 in [53], we only give an outline.

For convenience we consider the equivalent model (3). Let $\widehat{\mathbf{x}}$ be computed using $\widehat{\mathbf{x}} = (\lambda\widehat{\boldsymbol{\Sigma}}_0^{-1}+\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{y}$ with $\widehat{\boldsymbol{\Sigma}}_0 = \mathrm{diag}\{\widehat{\gamma}_1\widehat{\mathbf{B}}_1,\cdots,\widehat{\gamma}_M\widehat{\mathbf{B}}_M\}$, and $\widehat{\boldsymbol{\gamma}} \triangleq [\widehat{\gamma}_1,\cdots,\widehat{\gamma}_M]$ is obtained by globally minimizing the cost function for given $\widehat{\mathbf{B}}_i\,(\forall i)$ [20]:

$$\mathcal{L}(\gamma_i) = \mathbf{y}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{y} + \log|\boldsymbol{\Sigma}_y|.$$

It can be shown [53] that when $\lambda \to 0$ (noiseless case), the above problem is equivalent to

$$\min: \quad g(\mathbf{x}) \triangleq \min_{\boldsymbol{\gamma}} \left[\mathbf{x}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{x} + \log|\boldsymbol{\Sigma}_y|\right] \tag{37}$$

$$\text{s.t.}: \quad \mathbf{y} = \mathbf{D}\mathbf{x} \tag{38}$$

So we only need to show the global minimizer of (37) satisfies the property stated in the theorem.

Assume in the noiseless problem $\mathbf{Y} = \boldsymbol{\Phi}\mathbf{X}$, $\boldsymbol{\Phi}$ satisfies the URP condition [5]. For its any solution $\widehat{\mathbf{X}}$, denote the number of nonzero rows by K. Thus following the method in [53], we can show the above $g(\mathbf{x})$ satisfies

$$g(\mathbf{x}) = \mathcal{O}(1) + \big(NL - \min[NL, KL]\big)\log\lambda, \tag{39}$$

providing $\widehat{\mathbf{B}}_i$ is full rank. Here we adopt the notation $f(s) = \mathcal{O}(1)$ to indicate that $|f(s)| < C_1$ for all $s < C_2$, with C_1 and C_2 constants independent of s. Therefore, by globally minimizing (39), i.e. globally minimizing (37), K will achieve its minimum value, which will be shown to be K_0, the number of nonzero rows in $\mathbf{X}_{\mathrm{gen}}$.

According to the result in [6], [12], if $\mathbf{X}_{\mathrm{gen}}$ satisfies

$$K_0 < \frac{N+L}{2}$$

then there is no other solution (with K nonzero rows) such that $\mathbf{Y} = \boldsymbol{\Phi}\mathbf{X}$ with $K < \frac{N+L}{2}$. So, $K \geq K_0$, i.e. the minimum value of K is K_0. Once K achieves its minimum, we have $\widehat{\mathbf{X}} = \mathbf{X}_{\mathrm{gen}}$.

In summary, the global minimum solution $\widehat{\boldsymbol{\gamma}}$ leads to the solution that equals to the unique sparsest solution $\mathbf{X}_{\mathrm{gen}}$. And we can see, providing $\widehat{\mathbf{B}}_i$ is full rank, it does not affect the conclusion.

[20] In the proof we fix $\widehat{\mathbf{B}}_i$ because we will see $\widehat{\mathbf{B}}_i$ has no effect on the global minimum property.

B. Proof of Lemma 2

Re-write the equation $\mathbf{y}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{y} = C$ by $\mathbf{y}^T\mathbf{u} = C$, where $\mathbf{u} \triangleq \boldsymbol{\Sigma}_y^{-1}\mathbf{y} = \left(\lambda\mathbf{I} + \mathbf{D}\boldsymbol{\Sigma}_0\mathbf{D}^T\right)^{-1}\mathbf{y}$, from which we have $\mathbf{y} - \lambda\mathbf{u} = \mathbf{D}\boldsymbol{\Sigma}_0\mathbf{D}^T\mathbf{u} = \mathbf{D}(\boldsymbol{\Gamma}\otimes\mathbf{B})\mathbf{D}^T\mathbf{u} = \mathbf{D}(\mathbf{I}_M\otimes\mathbf{B})(\boldsymbol{\Gamma}\otimes\mathbf{I}_L)\mathbf{D}^T\mathbf{u} = \mathbf{D}(\mathbf{I}_M\otimes\mathbf{B})\mathrm{diag}(\mathbf{D}^T\mathbf{u})\mathrm{diag}(\boldsymbol{\Gamma}\otimes\mathbf{I}_L) = (\boldsymbol{\Phi}\otimes\mathbf{B})\mathrm{diag}(\mathbf{D}^T\mathbf{u})(\boldsymbol{\gamma}\otimes\mathbf{1}_L)$. It can be seen that the matrix $\mathbf{A} \triangleq (\boldsymbol{\Phi}\otimes\mathbf{B})\mathrm{diag}(\mathbf{D}^T\mathbf{u})$ is full row rank.

C. Proof of Theorem 2

The proof follows along the lines of Theorem 2 in [30] using our Lemma 1 and Lemma 2. Consider the optimization problem:

$$\begin{cases} \min: & f(\boldsymbol{\gamma}) \triangleq \log|\lambda\mathbf{I} + \mathbf{D}\boldsymbol{\Sigma}_0\mathbf{D}^T| \\ \text{s.t.}: & \mathbf{A}\cdot(\boldsymbol{\gamma}\otimes\mathbf{1}_L) = \mathbf{b} \\ & \boldsymbol{\gamma} \succeq \mathbf{0} \end{cases} \tag{40}$$

where $\mathbf{A}$ and $\mathbf{b}$ are defined in Lemma 2. From Lemma 1 and Lemma 2 we can see the optimization problem (40) is optimizing a concave function over a closed, bounded convex polytope. Obviously, any local minimum of $\mathcal{L}$, e.g. $\boldsymbol{\gamma}^*$, must also be a local minimum of the above optimization problem with $C = \mathbf{y}^T\left(\lambda\mathbf{I} + \mathbf{D}(\boldsymbol{\Gamma}^*\otimes\mathbf{B})\mathbf{D}^T\right)^{-1}\mathbf{y}$, where $\boldsymbol{\Gamma}^* \triangleq \mathrm{diag}(\boldsymbol{\gamma}^*)$. Based on the Theorem 6.5.3 in [50] the minimum of (40) is achieved at an extreme point. Further, based on the Theorem in Chapter 2.5 of [50] the extreme point is a BFS to

$$\begin{cases} \mathbf{A}\cdot(\boldsymbol{\gamma}\otimes\mathbf{1}_L) = \mathbf{b} \\ \boldsymbol{\gamma} \succeq \mathbf{0} \end{cases}$$

which indicates $\|\boldsymbol{\gamma}\|_0 \leq NL$.

D. Proof of Lemma 3

For convenience we first consider the case of $K = N$. Let $\widetilde{\boldsymbol{\gamma}}$ be the vector consisting of nonzero elements in $\widehat{\boldsymbol{\gamma}}$, and $\widetilde{\boldsymbol{\Phi}}$ be a matrix consisting of the columns of $\boldsymbol{\Phi}$ whose indexes are the same as those of nonzero elements in $\widehat{\boldsymbol{\gamma}}$. Thus, the equation $\mathbf{Y} = \boldsymbol{\Phi}\widehat{\mathbf{X}}$ can be rewritten as $\mathbf{Y} = \widetilde{\boldsymbol{\Phi}}\widetilde{\mathbf{X}}$. By transferring it to its equivalent block sparse Bayesian learning model, we have $\mathbf{y} = \widetilde{\mathbf{D}}\widetilde{\mathbf{x}}$, where $\mathbf{y} \triangleq \mathrm{vec}(\mathbf{Y}^T)$, $\widetilde{\mathbf{D}} \triangleq \widetilde{\boldsymbol{\Phi}}\otimes\mathbf{I}_L$, and $\widetilde{\mathbf{x}} \triangleq \mathrm{vec}(\widetilde{\mathbf{X}}^T)$. Since $\widetilde{\mathbf{D}}$ is a square matrix with full rank, we have $\widetilde{\mathbf{x}} = \widetilde{\mathbf{D}}^{-1}\mathbf{y}$. For convenience, let $\widetilde{\mathbf{x}}_i \triangleq \widetilde{\mathbf{x}}_{[(i-1)L+1:iL]}$, i.e. $\widetilde{\mathbf{x}}_i$ consists of elements of $\widetilde{\mathbf{x}}$ with indexes from $(i-1)L+1$ to iL. Now consider the cost function $\mathcal{L}$, which becomes

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}) \;=\; & \sum_{i=1}^{N}\Big(\frac{\widetilde{\mathbf{x}}_i^T\mathbf{B}^{-1}\widetilde{\mathbf{x}}_i}{\widetilde{\gamma}_i} + L\log\widetilde{\gamma}_i\Big) + M\log|\mathbf{B}| \\ & +2\log|\widetilde{\mathbf{D}}|.\end{aligned}$$

Letting $\frac{\partial \mathcal{L}(\boldsymbol{\gamma})}{\partial \widetilde{\gamma}_i} = 0$ gives

$$\widetilde{\gamma}_i = \frac{1}{L}\widetilde{\mathbf{x}}_i^T \mathbf{B}^{-1}\widetilde{\mathbf{x}}_i, \quad i = 1, \cdots, K$$

The second derivative of $\mathcal{L}$ at $\widetilde{\gamma}_i = \frac{1}{L}\widetilde{\mathbf{x}}_i^T \mathbf{B}^{-1}\widetilde{\mathbf{x}}_i$ is given by

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\gamma})}{\partial \widetilde{\gamma}_i^2}\Big|_{\widetilde{\gamma}_i = \widetilde{\mathbf{x}}_i^T \mathbf{B}^{-1}\widetilde{\mathbf{x}}_i} = \frac{\widetilde{\mathbf{x}}_i^T \mathbf{B}^{-1}\widetilde{\mathbf{x}}_i}{\widetilde{\gamma}_i^3}.$$

Since $\mathbf{B}$ is positive definite and $\widetilde{\mathbf{x}}_i \neq \mathbf{0}$, $\frac{\widetilde{\mathbf{x}}_i^T \mathbf{B}^{-1}\widetilde{\mathbf{x}}_i}{\widetilde{\gamma}_i^3} > 0$. So $\widetilde{\gamma}_i = \frac{1}{L}\widetilde{\mathbf{x}}_i^T \widehat{\mathbf{B}}^{-1}\widetilde{\mathbf{x}}_i$ $(i = 1, \cdots, K)$ is a local minimum.

If $\|\widetilde{\boldsymbol{\gamma}}\|_0 \triangleq K < N$, which implies there exists $\widetilde{\mathbf{x}} \in \mathbb{R}^{KL \times 1}$ such that $\mathbf{y} = \widetilde{\mathbf{D}}\widetilde{\mathbf{x}}$, then we can expand the matrix $\widetilde{\mathbf{D}}$ to a full-rank square matrix $[\widetilde{\mathbf{D}}, \mathbf{D}_e]$ by adding an arbitrary full column-rank matrix $\mathbf{D}_e$. And we expand $\widetilde{\mathbf{x}}$ to $[\widetilde{\mathbf{x}}^T, \boldsymbol{\varepsilon}^T]^T$, where $\boldsymbol{\varepsilon} \in \mathbb{R}^{(N-K)L \times 1}$ and $\boldsymbol{\varepsilon} \to \mathbf{0}$. Therefore, $[\widetilde{\mathbf{D}}, \mathbf{D}_e][\widetilde{\mathbf{x}}^T, \boldsymbol{\varepsilon}^T]^T \to \widetilde{\mathbf{D}}\widetilde{\mathbf{x}} = \mathbf{y}$. Similarly, we also expand $\widetilde{\boldsymbol{\gamma}}$ to $[\widetilde{\boldsymbol{\gamma}}^T, \boldsymbol{\zeta}^T]^T$ with $\boldsymbol{\zeta} \to \mathbf{0}$. Then, following the above steps, we can obtain the same result. Therefore, we finish the proof.

REFERENCES

[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. on Information Theory*, vol. 52, no. 4, pp. 1289 – 1306, 2006.

[2] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[3] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–124, 2007.

[4] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[5] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm," *IEEE Trans. on Signal Processing*, vol. 45, no. 3, pp. 600–616, 1997.

[6] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization," *PNAS*, vol. 100, no. 5, pp. 2197–2202, 2003.

[7] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm," *Electroencephalography and Clinical Neurophysiology*, vol. 95, pp. 231–251, 1995.

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

[9] J. H. G. Ender, "On compressive sensing applied to radar," *Signal Processing*, vol. 90, pp. 1402–1414, 2010.

[10] U. Gamper, P. Boesiger, and S. Kozerke, "Compressed sensing in dynamic MRI," *Magnetic Resonance in Medicine*, vol. 59, pp. 365–373, 2008.

[11] B. D. Rao and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," in *Proc. IEEE Digital Signal Processing Workshop*, Bryce Canyon, UT, 1998.

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

[13] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[14] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. on Information Theory*, vol. 56, no. 1, pp. 505–519, 2010.

[15] Y. Jin and B. D. Rao, "Insights into the stable recovery of sparse solutions in overcomplete representations using network information theory," in *Proc. of the 33th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, Las Vegas, USA, pp. 3921–3924.

[16] G. Tang and A. Nehorai, "Performance analysis for sparse support recovery," *IEEE Trans. on Information Theory*, vol. 56, no. 3, pp. 1383–1399, 2010.

[17] Y. Jin and B. D. Rao, "On the role of the properties of the nonzero entries on sparse signal recovery," in *Proc. of the 44th Asilomar Conference on Signals, Systems, and Computers*, USA, 2010, pp. 753–757.

[18] C. M. Michel, T. Koenig, D. Brandeis, and et al, *Electrical Neuroimaging*, 1st ed. Cambridge University Press, 2009.

[19] S. F. Cotter, "Multiple snapshot matching pursuit for direction of arrival (DOA) estimation," in *Proc. of the 15th European Signal Processing Conference (EUSIPCO 2007)*, Poznan, Poland, 2007.

[20] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing*, vol. 86, pp. 572–588, 2006.

[21] K. Lee and Y. Bresler, "Subspace-augmented MUSIC for joint sparse recovery," 2011. [Online]. Available: http://arxiv.org/abs/1004.3071v3

[22] S. Negahban and M. J. Wainwright, "Simultaneous support recovery in high dimensions: benefits and perils of block ℓ_1/ℓ_∞-regularization," *IEEE Trans. on Information Theory*, vol. 57, no. 6, pp. 3841–3863, 2011.

[23] J. A. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Processing*, vol. 86, pp. 589–602, 2006.

[24] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *Journal of Machine Learning Research*, vol. 9, pp. 1179–1225, 2008.

[25] D. Wipf and S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.

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

[27] D. Wipf, B. D. Rao, and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity," *accepted by IEEE Trans. on Information Theory*, 2010.

[28] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[29] A. C. Faul and M. E. Tipping, "Analysis of sparse bayesian learning," in *Advances in Neural Information Processing Systems 14*, 2002, pp. 383–389.

[30] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

[31] Z. Zhang and B. D. Rao, "Sparse signal recovery in the presence of correlated multiple measurement vectors," in *Proc. of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Texas, USA, 2010, pp. 3986–3989.

[32] ——, "Iterative reweighted algorithms for sparse signal recovery with temporally correlated source vectors," in *Proc. of the 36th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, the Czech Republic, 2011.

[33] ——, "Exploiting correlation in sparse signal recovery problems: Multiple measurement vectors, block sparsity, and time-varying sparsity," in *ICML 2011 Workshop on Structured Sparsity: Learning and Inference*, Washington, USA, 2011. [Online]. Available: http://arxiv.org/pdf/1105.0725v1

[34] K. Qiu and A. Dogandzic, "Variance-component based sparse signal reconstruction and model selection," *IEEE Trans. on Signal Processing*, vol. 58, no. 6, pp. 2935–2952, 2010.

[35] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. on Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[36] G. Tzagkarakis, D. Milioris, and P. Tsakalides, "Multiple-measurement Bayesian compressed sensing using GSM priors for DOA estimation," in *Proc. of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Texas, USA, 2010, pp. 2610–2613.

[37] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. R. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[38] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[39] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J Fourier Anal Appl*, vol. 14, pp. 877–905, 2008.

[40] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[41] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Statist. Soc. B*, vol. 68, pp. 49–67, 2006.

[42] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *The Annals of Statistics*, vol. 37, no. 6A, pp. 3468–3497, 2009.

[43] R. Zdunek and A. Cichocki, "Improved M-FOCUSS algorithm with overlapping blocks for locally smooth sparse signals," *IEEE Trans. on Signal Processing*, vol. 56, no. 10, pp. 4752–4761, 2008.
[44] Y. Cho and L. K. Saul, "Sparse decomposition of mixed audio signals by basis pursuit with autoregressive models," in *Proc. of the 34th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, Taipei, pp. 1705–1708.
[45] A. Hyvärinen, "Optimal approximation of signal priors," *Neural Computation*, vol. 20, no. 12, pp. 3087–3110, 2008.
[46] G. C. Cawley and N. L. C. Talbot, "Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters," *Journal of Machine Learning Research*, vol. 8, pp. 841–861, 2007.
[47] I. Guyon, A. Saffari, G. Dror, and G. Cawley, "Model selection: beyond the Bayesian/frequentist divide," *Journal of Machine Learning Research*, vol. 11, pp. 61–87, 2010.
[48] M. Elad, "Sparse representations are most likely to be the sparsest possible," *EUROSIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–12, 2006.
[49] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[50] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed. Springer, 2005.
[51] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Stanford University Technical Report*, 2004.
[52] M. M. Hyder and K. Mahata, "A robust algorithm for joint-sparse recovery," *IEEE Signal Processing Letters*, vol. 16, no. 12, pp. 1091–1094, 2009.
[53] D. Wipf, J. P. Owen, H. T. Attias, and et al, "Robust Bayesian estimation of the location, orientation, and time course of multiple correlated neural sources using meg," *NeuroImage*, vol. 49, pp. 641–655, 2010.
[54] D. Wipf and S. Nagarajan, "A new view of automatic relevance determination," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 1625–1632.
[55] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. on Signal Processing*, vol. 51, no. 3, pp. 760–770, 2003.
[56] S. Becker, J. Bobin, and E. J. Candes, "NESTA: A fast and accurate first-order method for sparse recovery," *SIAM Journal on Imaging Sciences*, vol. 4, no. 1, pp. 1–39, 2011.
[57] E. T. Hale, W. Yin, and Y. Zhang, "A fixed-point continuation method for ℓ_1-regularized minimization with applications to compressed sensing," *CAAM Technical Report TR07-07, Rice University*, 2007.
[58] J. Wright, A. Y. Yang, A. Ganesh, and et al, "Robust face recognition via sparse representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.
[59] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: uncertainty relations and efficient recovery," *IEEE Trans. on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.
[60] N. Vaswani, "Kalman filtered compressed sensing," in *Proc. of the15th IEEE International Conference on Image Processing (ICIP 2008)*, San Diego, USA, 2008, pp. 893–896.
[61] J. Ziniel, L. C. Potter, and P. Schniter, "Tracking and smoothing of time-varying sparse signals via approximate belief propagation," in *Proc. of the 44th Asilomar Conference on Signals, Systems and Computers*, 2010, pp. 808–812.

Zhilin Zhang (S'08) received the B.S. degree in automatics and the M.S. degree in electrical engineering from the University of Electronic Science and Technology of China. Since 2007 he has been working toward the Ph.D. degree in the Department of Electrical and Computer Engineering at University of California, San Diego.

His research interests include sparse signal recovery/compressed sensing, blind source separation, neuroimaging, computational and cognitive neuroscience.

Bhaskar D. Rao (F'00) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, in 1979 and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1981 and 1983, respectively. Since 1983, he has been with the University of California at San Diego, La Jolla, where he is currently a Professor with the Electrical and Computer Engineering Department and holder of the Ericsson endowed chair in wireless access networks. His interests are in the areas of digital signal processing, estimation theory, and optimization theory, with applications to digital communications, speech signal processing, and human-computer interactions.

He is the holder of the Ericsson endowed chair in Wireless Access Networks and is the Director of the Center for Wireless Communications. His research group has received several paper awards. His paper received the best paper award at the 2000 speech coding workshop and his students have received student paper awards at both the 2005 and 2006 International conference on Acoustics, Speech and Signal Processing conference as well as the best student paper award at NIPS 2006. A paper he co-authored with B. Song and R. Cruz received the 2008 Stephen O. Rice Prize Paper Award in the Field of Communications Systems. He was elected to the fellow grade in 2000 for his contributions in high resolution spectral estimation. Dr. Rao has been a member of the Statistical Signal and Array Processing technical committee, the Signal Processing Theory and Methods technical committee, the Communications technical committee of the IEEE Signal Processing Society. He has also served on the editorial board of the EURASIP Signal Processing Journal.