# Bayesian multiple measurement vector problem with spatial structured sparsity patterns 

Ningning Han ${ }^{\text {a,b,c }, ~ Z h a n j i e ~ S o n g ~}{ }^{\text {a,b,c,* }}$<br>${ }^{\text {a }}$ School of Mathematics, Tianjin University, Tianjin 300350, China<br>${ }^{\text {b }}$ Visual Pattern Analysis Research Lab, Tianjin University, Tianjin 300072, China<br>${ }^{\text {c }}$ Institute of TV and Image Information, Tianjin University, Tianjin 300072, China

## A R T I C L E I N F O

## Article history:

Available online 31 January 2018

## Keywords:

Multiple measurement vector (MMV)
Spatial correlation structure
Temporal correlation structure
Variational Bayes (VB)


#### Abstract

A promising research that has drawn considerable attentions is exploiting the inherent structures in the sparse signal. In this work, we apply the property to the multiple measurement vector (MMV) problem, in which a group of collected sparse signals that share an identical sparsity support are recovered from undersampled measurements. The main objective of this paper is to introduce a Bayesian model with taking both spatial and temporal dependencies into account and show how this model can be used for MMV with spatial structured sparsity patterns. Due to the property of the beta process that the sparse representation can be decomposed to values and sparsity indicators, the proposed algorithm ingeniously captures the temporal correlation structure by the learning of amplitudes and the spatial correlation structure by the estimation of clustered sparsity patterns. Detailed numerical experiments including synthetic and real data demonstrate the effectiveness of the proposed algorithm.


© 2018 Elsevier Inc. All rights reserved.

## 1. Introduction

There has been a recently emerged technique of signal sampling and reconstruction, known as compressive sensing (CS) [1-3] that recovers high dimensional signals from far less linear measurements with minimum loss of information. The measurement system is expressed as follows
$\boldsymbol{y}=\boldsymbol{\Phi} \boldsymbol{x}+\boldsymbol{e}$,
where $\boldsymbol{y} \in \mathbb{R}^{M \times 1}$ is the measurement vector, $\boldsymbol{x} \in \mathbb{R}^{N \times 1}$ denotes the underlying signal vector of interest, $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ represents the random projection matrix ( $M \ll N$ ), and $\boldsymbol{e}$ is the additive noise. Since most of natural signals are sparse or highly compressible under a basis, CS has a wide range of applications including image processing [4], sensor network [5], biological application [6], signal reconstruction [7], etc. CS algorithms can be broadly classified into three strategies. The large majority of algorithms are based on convex relation. A well-known instance based on such an approach is the Basis Pursuit (BP) [8], which replaces the $\ell_{0}$-norm with the $\ell_{1}$-norm. Another common used category for CS is greedy iterative

[^0]algorithms, such as matching pursuit (MP) [9] and its derivative orthogonal matching pursuits (OMP) [10]. One advantage of greedy algorithms is the low computational cost and the high recovered speed. The sparse signal recovery can also be formulated in a Bayesian framework, where the sparsity is described by a priori distribution such as Laplace.

The system in (1) is a typical CS model with a signal measurement vector (SMV). When the measurement procedure occurs at $L$ time instances, the basic model (1) can be extended to the multiple measurement vector (MMV) model, given by
$\boldsymbol{Y}=\boldsymbol{\Phi} \boldsymbol{X}+\boldsymbol{E}$,
where $\boldsymbol{Y} \in \mathbb{R}^{M \times L}$ is the measurement matrix containing $L$ measurement vectors, $\boldsymbol{X} \in \mathbb{R}^{N \times L}$ represents the underlying source matrix, and $\boldsymbol{E} \in \mathbb{R}^{M \times L}$ is the additive noise matrix. In general, the $L$ measurements in $\boldsymbol{X}$ are statistically related, particularly when repeated measurements are the same type of data that taken from similar scenes (e.g., repeated magnetic resonance imaging in a diagnostic task). In this case, different signals share an identical support, i.e., the locations of the nonzero elements of every column in $\boldsymbol{X}$ are identical. Specifically, if the $L$ measurements with $M$ samples are recovered independently, with desired accuracy, then ideally less than $M$ samples would be required by exploiting statistical correlations among $L$ measurements. In [11-14], authors have shown that compared to the SMV model, the accuracy of sparse signal recovery can be improved dramatically by performing MMV


Fig. 1. (a) A record of the channel impulse response of underwater acoustic channels measured off the coast of Martha's Vinyard, MA, USA. (b) The image of matrix $\boldsymbol{Q}$.
process. MMV is closely related to other fields, such as cognitive radio networks [15], dynamic compressive sensing [16], multi-task sparse learning [17], and source localization in electroencephalography (EEG) and magnetoencephalography (MEG) [18], etc.

The goal of MMV is to recover the underlying signal matrix $\boldsymbol{X}$ by solving the linear equation (2), in which $\boldsymbol{X}$ is row-wise sparse (i.e., only a few rows of $\boldsymbol{X}$ are nonzero). As an extension of SMV algorithms, greedy algorithms such as [19-21] have been developed to solve the model (2). Inspired by the fact that $\boldsymbol{X}$ is row-wise sparse, mixed norm regularizations [11,20,22] have been induced to obtain good performances. Among the MMV algorithms, Bayesian algorithms express the MMV as the solution of a Bayesian inference problem and apply statistical tools to solve it. In [23], the authors used the sparse Bayesian learning (SBL) to MMV and derived the MSBL algorithm. The multi-task compressive sensing algorithm [24] provided solutions to MMV where a shared prior was placed across all of the $L$ signals.

While most of the existing works concern on the row-wise sparsity of $\boldsymbol{X}$, temporal structures within the nonzero rows of $\boldsymbol{X}$ have been largely ignored. It contradicts the real scenarios, since the coefficients in a source may be strongly correlated. For example, in the electroencephalography (EEG) magnetoencephalography (MEG) source localization problem, each row of $\boldsymbol{X}$ is a temporally smooth signal reflecting the activity of a set of neurons, as it responses to the time course of stimulus caused by an equivalent electrical current. Recent contributions have suggested that the statistical relationship can potentially lead to the improved performance. Zhang et al. [25] presented a block sparse Bayesian learning framework where the correlation structure of each row of $\boldsymbol{X}$ was captured by a positive definite matrix. The positive definite matrix was adaptively learned from the measurements, and the resulted matrix was then used to estimate the underlying signals. The authors in [26] established a hierarchical Bayesian framework to model the temporally smooth signals. The matrix used to capture the temporal correlation was not restricted to be full rank by using a singular multinomial distribution instead of a multivariate Gaussian distribution. In [27], a Bayesian approximate message passing (AMP) was proposed for solving the MMV. The temporal correlation of signal amplitudes was modeled by the stationary first-order Gauss-Markov random process and the expectation-maximization (EM) algorithm that coupled with the message passing procedure was employed to learn the model parameters. Noted that outliers might be presented in the measurements, an unconstrained optimization problem [28] sought to research both temporally smooth and row-wise sparse in $\boldsymbol{X}$, which fitted the measurements well.

In practice, besides the temporal correlation structure, the distribution of elements in each sparse measurement is not truly pixel-wised sparsity but structured sparsity. For instance, most of the wavelet coefficients of a natural image are small, and however the large coefficients usually have group sparsity structures that can be utilized to enhance the image recovery [29]. The sparse outliers in background subtraction are also typically spatially continuous. Numerous works [30-33] prove that imposing structured sparsity on the support of the signal (the sparsity pattern) that goes beyond simple sparsity can boost the performance of sparse signal recovery. In order to take aim at the trajectory, we address the problem where each measurement of MMV is structured sparsity in this paper. Such problem can be seen as the combination of the traditional MMV and the canonical block-sparse pattern. The real signals often exhibit the two sparsity patterns, i.e., nonzero elements in each column of $\boldsymbol{X}$ are existed in clusters (spatial correlation) and entries in each nonzero row of $\boldsymbol{X}$ are correlated (temporal correlation). Fig. 1(a) shows a record of the channel impulse responses (CIR) of underwater acoustic channels (represented over the propagation delay and time domain) measured from the experiments conducted in Atlantic Ocean in USA [34]. We can observe that the sparsity structure of the CIR is varying slowly and the nonzero coefficients of each channel impulse response occur in clusters.

Two key points herein are to capture the temporal correlation within the nonzero rows of $\boldsymbol{X}$ and to exploit clustered sparsity structures within each column of $\boldsymbol{X}$. In this context, we employ a hierarchical sparse Bayesian framework to solve the problem. Taking advantage of the beta process factor analysis [35], the estimation of MMV can be decomposed to the magnitudes and the latent variables indicating sparsity patterns. According to the clustered pattern of structured sparsity signal, the current element and its neighbors are strongly correlated. To model the clustered prior of each measurement, we assume that the sparsity patterns satisfy a Markov property. Specifically, if an element is nonzero, it is very likely that its neighbors are nonzero. As done in [25], we employ a block-sparse Bayesian learning framework where the temporal correlation of $\boldsymbol{X}$ is captured by a positive definite matrix. Moreover, the parametric computation is cast within a Bayesian inference procedure and based on variational Bayes (VB) [36] approximation.

This paper makes several contributions. First and foremost, the proposed work not only exploits temporal correlation within each source of the signal (i.e., the temporal correlation existing in each nonzero row of $\boldsymbol{X}$ ), but also exploits spatial correlation among different sources of the signal (i.e., clustered sparsity structures existing in each column of $\boldsymbol{X}$ ). Almost all the existing MMV al-
gorithms only consider the temporal correlation, while the spatial correlation is rarely utilized. In fact, taking advantage of the spatial correlation for the signal recovery has been a promising research. Based on the property of the beta process that the sparse representation can be decomposed to values and sparsity indicators, our work ingeniously captures the temporal correlation structure by the learning of amplitudes and the spatial correlation structure by the estimation of clustered sparsity patterns. Second, we propose a variational Bayesian algorithm to automatically estimate the parameters of the proposed model. For the proposed Bayesian model, as done in [25], a positive definite matrix $\boldsymbol{B}$ is used to capture the temporal correlation structure. We would like to point out, however, the significant difference in estimating $\boldsymbol{B}$ of our method from [25]. Our approach incorporate the conjugate Wishart hyperprior on the positive definite matrix $\boldsymbol{B}$
$\boldsymbol{B} \sim \mathcal{W}\left(\boldsymbol{B} \mid \boldsymbol{V}_{0}, v_{0}\right)=\frac{1}{C}|\boldsymbol{B}|^{\left(v_{0}-L-1\right) / 2} \exp \left(-\frac{1}{2} \operatorname{Tr}\left(\boldsymbol{V}_{0}^{-1} \boldsymbol{B}\right)\right)$,
as defined in (6). The posterior estimation of $\boldsymbol{B}$ is inferred by VB algorithm, while [25] makes no prior about $\boldsymbol{B}$ and approximates it by expectation-maximization (EM) method. Besides, an indicator matrix is introduced to capture the spatial correlation and enhance the recovery. Specifically, the spatial correlation is described by the Markov dependency between the element and its neighbors, as shown in model (10).

Next, we introduce some notations used in this paper. Boldfaced upper-case letters, e.g., $\boldsymbol{A}$, denote matrices, while boldfaced lowercase letters, e.g., a, denote column vectors. $\operatorname{diag}(\boldsymbol{a})$ denotes a diagonal matrix with principal diagonal elements being $\boldsymbol{a}$. If $\boldsymbol{A}_{1}, \boldsymbol{A}_{2}$, $\ldots, \boldsymbol{A}_{n}$ is a series of square matrices, then $\operatorname{diag}\left(\boldsymbol{A}_{1}, \boldsymbol{A}_{2}, \ldots, \boldsymbol{A}_{n}\right)$ is a block diagonal matrix with principal diagonal blocks being $\boldsymbol{A}_{1}$, $\boldsymbol{A}_{2}, \ldots, \boldsymbol{A}_{n}$ in turn. For a vector $\boldsymbol{a}, \boldsymbol{a}_{i}$ is the $i$ th element of $\boldsymbol{a}$ and $\boldsymbol{a}_{[i]}$ is the $i$ th block of $\boldsymbol{a}$. For a matrix $\boldsymbol{A}, \boldsymbol{A}_{i}$. denotes the $i$ th row, $\boldsymbol{A}_{. i}$ denotes the $i$ th column, and $\boldsymbol{A}_{i, j}$ denotes the element that lies in the $i$ th row and the $j$ th column. An $m$ by $m$ identity matrix is denoted by $\boldsymbol{I}_{m} . \boldsymbol{J}_{m}$ denotes the $m$ by $m$ matrix in which all elements are equal to 1 and $\boldsymbol{I}_{m}$ is a $m$-dimension column vector of ones. $\boldsymbol{A}^{T}$ and $\operatorname{Tr}(\boldsymbol{A})$ denote the transpose and trace of $\boldsymbol{A}$, respectively.

## 2. Bayesian modeling

Structured sparsity is an expansion of standard sparsity pattern, in which variables in the same group tend to be zero or nonzero simultaneously. Thus the spatial correlation structure mainly depends on the clustered sparsity pattern. Under this structured setup, we use latent variables indicating sparsity patterns to capture the spatial correlations of MMV. In fact, MMV is a special structured sparse signal. If we stack $\boldsymbol{X}$ into a column vector, the vector is a structured sparse signal with the known clustered pattern. However, different from general structured sparse signals, the correlations in MMV embody not just in the fixed sparsity profile of all measurements but also in the temporal correlations among different measurements. Further exploring the temporal correlations is of high importance towards more effective recovery. To address the issue, the temporal correlations about magnitudes are modeled by a positive definite matrix.

As a result, it is believable that a Bayesian model can capture the temporal correlation within each source, as well as the spatial correlation among different sources. In this paper, we integrate the temporal correlation and the spatial correlation into a unified Bayesian framework. Concretely, we separate the learning of amplitude and support, in which the amplitude prior and the latent variable are employed to capture the temporal correlation and the spatial correlation, respectively.

In this section, we elaborate on the Bayesian model. As discussed above, we separate the learning of sparsity from the learning of amplitude. The proposed Bayesian model is

$$
\begin{equation*}
\boldsymbol{Y}=\boldsymbol{\Phi}(\boldsymbol{W} \circ \boldsymbol{Z})+\boldsymbol{E}, \tag{3}
\end{equation*}
$$

where $\boldsymbol{W} \in \mathbb{R}^{N \times L}$ is the weight matrix representing magnitudes, $Z \in \mathbb{R}^{N \times L}$ is the indicator matrix and o denotes the Hadamard product. Since each row of $\boldsymbol{W}$ shares a fixed sparsity profile, each column of $\boldsymbol{Z}$ is identical, i.e., $\boldsymbol{Z}_{. i}=\boldsymbol{s}(\forall i=1, \ldots, L)$, where $\boldsymbol{s} \in\{0,1\}^{N}$ is a binary vector. $\boldsymbol{E}$ is additive noise, which is drawn from a Gaussian distribution.

Bayesian algorithms express the sparse signal recovery problem as the solution of a Bayesian inference procedure and apply statistical tools to solve it. Regarding the choice of the prior, one can model the sparse signal as a continuous random variable whose distribution have a sharp peak at zero and heavy tails. For the signal measurement vector model (1), a widely used sparseness prior is the Laplace density function [37],
$\boldsymbol{x} \sim\left(\frac{\lambda}{2}\right)^{N} \exp \left(-\lambda \sum_{i=1}^{N}\left|\boldsymbol{x}_{i}\right|\right)$.
The Laplace sparseness prior is not conjugate to the Gaussian likelihood and hence the connected Bayesian inference may not be performed exactly. This issue has been addressed in sparse Bayesian learning with automatic relevance determination (ARD) [38]. Rather than imposing a Laplace prior on $\boldsymbol{x}$, a hierarchical prior has similar properties as the Laplace prior but allows convenient conjugate-exponential analysis. To this end, the first choice is a zero-mean Gaussian prior on each element of $\boldsymbol{x}$,
$\boldsymbol{x} \sim \prod_{i=1}^{N} \mathcal{N}\left(\boldsymbol{x}_{i} \mid 0, \alpha_{i}^{-1}\right)$,
where $\alpha_{i}$ is the inverse-variance of a Gaussian density function. Motivated by applications of sparse Bayesian learning in the signal measurement model (SMV), Wife et al. [23] extended it to the MMV model, i.e., $\boldsymbol{X}_{i} \sim \mathcal{N}\left(0, \gamma_{i}^{-1} \boldsymbol{I}\right)$. Many of the precisions $\gamma_{i}$ will assume very large values during inference, which encourage the sparsity of $\boldsymbol{X}$. To model the temporal correlation, the Bayesian MMV model was extended to $\boldsymbol{X}_{i .} \sim \mathcal{N}\left(\boldsymbol{X}_{i .} \mid 0, \gamma_{i}^{-1} \boldsymbol{B}\right)$ [25]. As done in [25], the distribution of $\boldsymbol{W}_{i .}(i=1, \ldots, N)$ is given by
$\boldsymbol{W}_{i .} \sim \mathcal{N}\left(\boldsymbol{W}_{i .} \mid \mathbf{0},\left(\gamma_{i} \boldsymbol{B}\right)^{-1}\right)$,
where $\gamma_{i}$ is a nonnegative hyperparameter controlling the sparsity. Following the conventional sparse Bayesian learning principle, we use Gamma distributions as hyperpriors over the precisions $\gamma_{i}$
$\gamma_{i} \sim \operatorname{Gamma}\left(\gamma_{i} \mid a, b\right)=\Gamma(a)^{-1} b^{a} \gamma_{i}^{a-1} \exp \left(-b \gamma_{i}\right)$,
where $\Gamma(a)=\int_{0}^{\infty} t^{a-1} e^{-t} d t$. In sparse Bayesian framework, very small values (e.g., $10^{-6}$ ) are assigned to the parameters $a$ and $b$. Such small values can lead to broad hyperpriors.

To avoid the overfitting and improve the adaptivity, we assign a positive definite $\boldsymbol{B} \in \mathbb{R}^{L \times L}$ to model all the source covariance matrices. In addition to (5), we incorporate the conjugate Wishart hyperprior on the positive definite $\boldsymbol{B}$, that is

$$
\begin{equation*}
\boldsymbol{B} \sim \mathcal{W}\left(\boldsymbol{B} \mid \boldsymbol{V}_{0}, v_{0}\right)=\frac{1}{C}|\boldsymbol{B}|^{\left(v_{0}-L-1\right) / 2} \exp \left(-\frac{1}{2} \operatorname{Tr}\left(\boldsymbol{V}_{0}^{-1} \boldsymbol{B}\right)\right) \tag{6}
\end{equation*}
$$

where $C$ is the normalizing constant, $v_{0}$ is the number of degrees of freedom and $\boldsymbol{V}_{0}$ is the scale matrix.

Here, we are interested in the Bayesian model that emphasizes not only the value but also its sparse pattern. Thus a binary vector
$\boldsymbol{s}$ is introduced to enhance the recovery and capture the spatial correlation. A standard choice for modeling the binary vector $\boldsymbol{s}$ is Bernoulli distribution,
$\boldsymbol{s} \sim \prod_{i=1}^{N} \operatorname{Bernoulli}\left(\boldsymbol{s}_{i} \mid \pi_{i}\right)$.
With this model, we know that the standard sparsity only captures simple primary data structure. With more data structure priors, we focus on the Markov dependency between the element and its neighbors. For an element, if its neighbors are nonzero (zero), then with a high probability the element is also nonzero (zero).

To illustrate the Markov property and the structured sparsity pattern, we carry out experiments on real audio signals, which have clustered sparsity patterns under a certain basis, such as discrete cosine transform (DCT) basis. The audio signals including piano, bird calls, jazz, jazz-trio, strings, glockenspiel, voicefemale and voicemale are available. ${ }^{1}$ We choose a short-time segment with consisting of $m(m=\{100,200, \ldots, 1000\})$ samples for each signal. By varying $m, 10$ different dimensional samples can be obtained for each audio signal. The sparse representations can be computed under the DCT basis $\boldsymbol{\Psi} \in \mathbb{R}^{m \times m}$. For a sparse representation $\boldsymbol{t}$, we find all elements of $\boldsymbol{t}$ whose neighbors are nonzero. The dependency implies that these elements tend to be nonzero. To examine the validity of the above mentioned Markov property, we compute the ratio denoted as $q / p$, where $p$ is the total number of these elements and $q$ is the number of nonzero ones. Turning to the Markov property, the ratio should be near 1 . Consequently, we compute the variable matrix $\boldsymbol{Q} \in \mathbb{R}^{8 \times 10}$, where $\boldsymbol{Q}_{i j}$ denotes the ratio of $i$ th audio signal with $j$ th $(j \in\{100,200, \ldots, 1000\})$ dimensional sample. The matrix $\boldsymbol{Q}$ is displayed in Fig. 1(b). The white color, which corresponds to near-one, is dominant. This illustrates that the Markov dependent property between the individual element and its neighbors characterizing the structured sparsity is desirable.

Beta distribution is a distribution over a continuous variable $\rho \in[0,1]$, which is often used to represent the probabilities for some binary events. The prior on parameter $\pi_{i}(i=1, \ldots, N)$ is expressed as
$\pi_{i} \sim \operatorname{Beta}\left(\pi_{i} \mid p, q\right)$.
The expectation of $\pi_{i}$ (the probability of $\boldsymbol{s}_{i}=1$ ) is
$E\left[\pi_{i}\right]=p /(p+q)$.
As discussed, if the neighbors of an element are non-zero, it is very likely that the element is non-zero. Assume that the neighbors of an element are zero, with a high probability the element is zero. The Markov dependency is then described as the Beta process:

$$
\begin{align*}
\pi_{i} & \sim\left\{\begin{array}{l}
\operatorname{Beta}\left(\pi_{i} \mid p_{h}, q_{h}\right) \text { if } \boldsymbol{s}_{i-1}=1 \text { and } \boldsymbol{s}_{i+1}=1, \\
\operatorname{Beta}\left(\pi_{i} \mid p_{l}, q_{l}\right) \text { if } \boldsymbol{s}_{i-1}=0 \text { and } \boldsymbol{s}_{i+1}=0, \\
\operatorname{Beta}\left(\pi_{i} \mid p_{u}, q_{u}\right) \text { if } \boldsymbol{s}_{i-1}=0, \boldsymbol{s}_{i+1}=1 \text { or } \boldsymbol{s}_{i-1}=1, \boldsymbol{s}_{i+1}=0,
\end{array}\right. \\
i & =1, \ldots, N, \tag{10}
\end{align*}
$$

where $p_{h}, q_{h}, p_{l}$, and $q_{l}$ are the fixed hyperparameters and $p_{h} /\left(p_{h}+q_{h}\right) \rightarrow 1$ and $p_{l} /\left(p_{l}+q_{l}\right) \rightarrow 0$, e.g., $p_{h}=\frac{N-1}{N}, q_{h}=\frac{1}{N}$ and $q_{l}=\frac{N-1}{N}, p_{l}=\frac{1}{N}$. Let $p_{u}=1 / 2$ and $q_{u}=1 / 2$, i.e., $p_{u} /\left(p_{u}+q_{u}\right)=$ $1 / 2$.

[^1]
## 3. Bayesian inference

For ease of inference, we transform (3) to the single measurement model. Let $\boldsymbol{y}=\operatorname{vec}\left(\boldsymbol{Y}^{T}\right) \in \mathbb{R}^{M L \times 1}, \boldsymbol{D}=\boldsymbol{\Phi} \otimes \boldsymbol{I}_{L} \in \mathbb{R}^{M L \times N L}$, $\boldsymbol{w}=\operatorname{vec}\left(\boldsymbol{W}^{T}\right) \in \mathbb{R}^{N L \times 1}, \boldsymbol{z}=\operatorname{vec}\left(\boldsymbol{Z}^{T}\right) \in \mathbb{R}^{N L \times 1}$ and $\boldsymbol{e}=\operatorname{vec}\left(\boldsymbol{E}^{T}\right)$, where $\otimes$ denotes the Kronecker product. The model is developed to
$\boldsymbol{y}=\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z})+\boldsymbol{e}$,
where $\circ$ denotes the Hadamard product, $\boldsymbol{w}_{[i]}=\boldsymbol{W}_{i}^{T}$ and $\boldsymbol{z}_{[i]}=$ $\boldsymbol{s}_{\boldsymbol{i}} \boldsymbol{l}_{L}(i=1, \ldots, N)$. Correspondingly, $\boldsymbol{z}$ can be constructed as $\boldsymbol{z}=\boldsymbol{\Omega} \boldsymbol{s}$, where $\boldsymbol{\Omega} \in \mathbb{R}^{N L \times N}$ with
$\boldsymbol{\Omega}_{i j}=\left\{\begin{array}{l}1, \text { if } i \in[(j-1) L+1, j L] . \\ 0, \text { else } .\end{array}\right.$
We assume that noise obey Gaussian distribution with zero mean and unknown precision $\beta$. Formally, the noise is modeled as
$\boldsymbol{e} \sim \mathcal{N}\left(\boldsymbol{e} \mid 0, \beta^{-1} \boldsymbol{J}_{M L}\right)$,
$\beta \sim \operatorname{Gamma}(\beta \mid c, d)$.
As in (5), we set hyperparameters $c=d=10^{-6}$. If necessary, one can learn different noise precisions for different parts of $\boldsymbol{e}$. But it will result in overfitting because of limited observations and too many parameters. To avoid the overfitting, we use one precision $\beta$ to model all noise as the way of choosing $\boldsymbol{B}$. The conditional distribution of the observation is expressed as

$$
\begin{equation*}
\boldsymbol{y} \mid \boldsymbol{D}, \boldsymbol{w}, \boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}), \beta^{-1} \boldsymbol{J}_{M L}\right) \tag{13}
\end{equation*}
$$

To facilitate the illustration of the proposed Bayesian model, Fig. 2 shows the simplified graphical model. In the figure, the circles represent the variables with prior distributions, while the squares denote the constants with fixed values.

We first let $\Theta=\left\{\boldsymbol{w}, \gamma_{i}, \boldsymbol{B}, \boldsymbol{s}, \pi_{i}, \beta\right\}$ denote all hidden variables. Bayesian inference is evaluating the posterior distributions of unknowns. However, the posterior distributions are computationally intractable since the marginal distribution $p(\boldsymbol{y})$ is not calculated analytically. In this paper, we use the VB [36] to deal with the tractable joint posterior distribution problem. The approximate posterior distribution is denoted by $q(\Theta)$. The underlying idea is to posit a parameterized family of distributions over the hidden variables and then optimize the parameters to minimize the Kullback-Leibler (KL) divergence between $q(\Theta)$ and true posterior distribution $p(\Theta \mid \boldsymbol{y})$, given by
$\min _{q(\Theta)} K L(q(\Theta) \| p(\Theta))=\int q(\Theta) \ln \frac{q(\Theta)}{p(\Theta \mid \boldsymbol{y})} d \Theta$.
Equivalently, it corresponds to the following problem
$\max _{q(\Theta)}-K L(q(\Theta) \| p(\Theta))=\int q(\Theta) \ln \frac{p(\Theta, \boldsymbol{y})}{q(\Theta) p(\boldsymbol{y})} d \Theta$.
Since $K L(q(\Theta) \| p(\Theta)) \geq 0$ and $\int q(\Theta) d \Theta=1$, it refers to the estimation of the marginal likelihood $p(\boldsymbol{y})$ with maximal lower bound, i.e.,
$\ln p(\boldsymbol{y}) \geq \mathcal{L}(\theta)=\int q(\Theta) \ln \frac{p(\Theta, \boldsymbol{y})}{q(\Theta)} d \Theta$.
It is assumed that $q(\Theta)$ factorizes with respect to these partitions as
$q(\Theta)=\prod_{k} q_{k}\left(\Theta_{k}\right)$.


Fig. 2. Bayesian model for MMV with structured sparse signals.

Let $q_{k}\left(\Theta_{k}\right)=q_{k}$ for simplicity. We aim to maximize the lower bound $\mathcal{L}(\theta)$.

$$
\begin{align*}
\mathcal{L}(\theta)= & \int \prod_{k} q_{k}\left[\ln p(\Theta, \boldsymbol{y})-\sum_{k} \ln q_{k}\right] d \Theta \\
= & \int \prod_{k} q_{k} \ln p(\Theta, \boldsymbol{y}) \prod_{k} d \Theta_{k}-\sum_{k} \int_{j} q_{j} \ln q_{k} \prod_{j} d \Theta_{j} \\
= & \int q_{k}\left[\ln p(\Theta, \boldsymbol{y}) \prod_{j \neq k}\left(q_{j} d \Theta_{j}\right)\right] d \Theta_{k}-\sum_{k} \int_{k} q_{k} \ln q_{k} d \Theta_{k} \\
= & \int q_{k}\left[\ln p(\Theta, \boldsymbol{y}) \prod_{j \neq k}\left(q_{j} d \Theta_{j}\right)\right] d \Theta_{k}-\int q_{k} \ln q_{k} d \Theta_{k} \\
& -\sum_{j \neq k} \int q_{j} \ln q_{j} d \Theta_{j} \\
= & \int q_{k} \ln \bar{p}\left(\Theta_{k}, \boldsymbol{y}\right) d \Theta_{k}-\int q_{k} \ln q_{k} d \Theta_{k}-\sum_{j \neq k} \int q_{j} \ln q_{j} d \Theta_{j} \\
= & -K L\left(q_{k} \| \bar{p}\right)-\sum_{j \neq k} \int q_{j} \ln q_{j} d \Theta_{j} \tag{14}
\end{align*}
$$

where $\ln \bar{p}\left(\Theta_{k}, \boldsymbol{y}\right)=E_{\Theta \backslash \Theta_{k}}[\ln p(\Theta, \boldsymbol{y})]=\int \ln p(\Theta, \boldsymbol{y}) \prod_{j \neq k}\left(q_{j} d \Theta_{j}\right)$ and $\int q_{j} d \Theta_{j}=1(j=1, \cdots)$. The expectation $E_{\Theta \backslash \Theta_{k}}$ is taken about the set $\Theta$ with $\Theta_{k}$ removed. Clearly the bound in (14) is maximized when the KL distance is zero, which is the case for $q_{k}\left(\Theta_{k}\right)=\bar{p}\left(\Theta_{k}, \boldsymbol{y}\right)$. Consequently, the expression of the optimal posterior approximation $q_{k}\left(\Theta_{k}\right)$ with other variables fixed is
$\ln q_{k}\left(\Theta_{k}\right)=E_{\Theta \backslash \Theta_{k}}[\ln p(\boldsymbol{y}, \Theta)]+C$,
where $C$ denotes a constant that does not depend on the current variable and can be obtained through normalization. We present
the update rules involved in the inference scheme (15) for all unknown variables.

In the next subsections, we update each parameter in its turn holding others fixed. For notational simplicity, the expectation over the approximate posterior $q(\cdot)$ is denoted by $\langle\cdot\rangle$. For more details, please refer to the Appendix.

### 3.1. Estimation of magnitudes

The estimation of magnitudes can be naturally implemented by inferring $\boldsymbol{w}, \boldsymbol{\gamma}_{i}$, and $\boldsymbol{B}$. Invoking the prior model in (4), the observation model in (13), and update rule in (15), one can obtain the posterior distribution of $\boldsymbol{w}_{[i]}\left(\boldsymbol{W}_{i .}\right)$,
$q\left(\boldsymbol{w}_{[i]}\right)=\mathcal{N}\left(\boldsymbol{w}_{[i]} \mid\left\langle\boldsymbol{w}_{[i]}\right\rangle, \Sigma_{\boldsymbol{w}_{[i]}}\right)$,
the posterior of $\boldsymbol{w}_{[i]}$ can be shown to be normal with mean, $\left\langle\boldsymbol{w}_{[i]}\right\rangle$, and covariance, $\Sigma_{\boldsymbol{w}_{[i]}}$, equal to

$$
\begin{align*}
& \left\langle\boldsymbol{w}_{[i]}\right\rangle=\langle\beta\rangle\left\langle\boldsymbol{s}_{i}\right\rangle \Sigma_{\boldsymbol{w}_{[i]}} \boldsymbol{D}_{[\cdot \cdot i]}^{T} \boldsymbol{y}_{-i}  \tag{17}\\
& \Sigma_{\boldsymbol{w}_{[i]}}=\left(\left\langle\gamma_{i}\right\rangle\langle\boldsymbol{B}\rangle+\left\langle\boldsymbol{s}_{i}^{2}\right\rangle\langle\beta\rangle \boldsymbol{D}_{[\cdot i]}^{T} \boldsymbol{D}_{[\cdot i]}\right)^{-1}
\end{align*}
$$

where $\boldsymbol{D}_{[\cdot k]}=\boldsymbol{D}(1: M L,(k-1) L+1: k L), k=1, \ldots, N$ and $\boldsymbol{y}_{-i}=$ $\boldsymbol{y}-\sum_{j \neq i}^{N}\left\langle\boldsymbol{s}_{j}\right\rangle \boldsymbol{D}_{[\cdot j]}\left\langle\boldsymbol{w}_{[j]}\right\rangle$.

By combining (4) and (5), we find that $\mathrm{q}\left(\gamma_{i}\right)$ follows a Gamma distribution, where
$q\left(\gamma_{i}\right)=\operatorname{Gamma}\left(\gamma_{i} \left\lvert\, a+\frac{L}{2}\right., \frac{2 b+\operatorname{Tr}\left(\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right)\langle\boldsymbol{B}\rangle\right)}{2}\right)$.

It is easy to get the mean of $q\left(\gamma_{i}\right)$,
$\left\langle\gamma_{i}\right\rangle=\frac{2 a+L}{2 b+\operatorname{Tr}\left(\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right)\langle\boldsymbol{B}\rangle\right)}$.

Similarly, the posterior approximation of $\boldsymbol{B}$ is Wishart distribution
$q(\boldsymbol{B})=\mathcal{W}\left(\boldsymbol{B} \mid\left(\boldsymbol{V}_{0}^{-1}+\sum_{i=1}^{N}\left\langle\gamma_{i}\right\rangle\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\right)\right)^{-1}, v_{0}+N\right)$.
The posterior expectation of $\boldsymbol{B}$ can be calculated as

$$
\begin{equation*}
\langle\boldsymbol{B}\rangle=\left(v_{0}+N\right)\left(\boldsymbol{V}_{0}^{-1}+\sum_{i=1}^{N}\left\langle\gamma_{i}\right\rangle\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right)^{-1}\right. \tag{21}
\end{equation*}
$$

### 3.2. Estimation of sparsity patterns

The parameters involved in the sparsity pattern are $\boldsymbol{s}$ and $\pi_{i}$.
With the Bernoulli prior (7) and the Gaussian observation likelihood (13), $q\left(\boldsymbol{s}_{i}\right)$ follows a Bernoulli distribution,
$q\left(\boldsymbol{s}_{i}\right)=\operatorname{Bernoulli}\left(\boldsymbol{s}_{i} \mid \xi^{\boldsymbol{s}_{i}}, \zeta^{1-\boldsymbol{s}_{i}}\right)$,
where

$$
\begin{aligned}
\xi= & \exp \left(\left\langle\ln \pi_{i}\right\rangle-\frac{\langle\beta\rangle}{2} \operatorname{Tr}\left(\left(\Sigma_{\boldsymbol{w}_{[i]}}\right.\right.\right. \\
& \left.\left.\left.+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right) \boldsymbol{D}_{[\cdot i]}^{T} \boldsymbol{D}_{[i]}\right)+\langle\beta\rangle\left(\boldsymbol{y}_{-i}\right)^{T} \boldsymbol{D}_{[\cdot i]}\left\langle\boldsymbol{w}_{[i]}\right\rangle\right), \\
\zeta= & \exp \left(\left\langle\ln \left(1-\pi_{i}\right)\right\rangle\right)
\end{aligned}
$$

Similarly, the approximate posterior of $\pi_{i}$ can be shown to be

$$
q\left(\pi_{i}\right)=\left\{\begin{array}{l}
\operatorname{Beta}\left(\pi_{i} \mid p_{h}+\left\langle\boldsymbol{s}_{i}\right\rangle, q_{h}+1-\left\langle\boldsymbol{s}_{i}\right\rangle\right)  \tag{24}\\
\text { if } \boldsymbol{s}_{i-1}=1 \text { and } \boldsymbol{s}_{i+1}=1, \\
\operatorname{Beta}\left(\pi_{i} \mid p_{l}+\left\langle\boldsymbol{s}_{i}\right\rangle, q_{l}+1-\left\langle\boldsymbol{s}_{i}\right\rangle\right) \\
\text { if } \boldsymbol{s}_{i-1}=0 \text { and } \boldsymbol{s}_{i+1}=0, \\
\operatorname{Beta}\left(\pi_{i} \mid p_{u}+\left\langle\boldsymbol{s}_{i}\right\rangle, q_{u}+1-\left\langle\boldsymbol{s}_{i}\right\rangle\right) \\
\text { if } \boldsymbol{s}_{i-1}=0, \boldsymbol{s}_{i+1}=1 \text { or } \boldsymbol{s}_{i-1}=1, \boldsymbol{s}_{i+1}=0,
\end{array}\right.
$$

therefore,

$$
\begin{align*}
& \left\langle\ln \pi_{i}\right\rangle=\left\{\begin{array}{l}
\psi\left(p_{h}+\left\langle\boldsymbol{s}_{i}\right\rangle\right)-\psi\left(p_{h}+q_{h}+1\right) \\
\text { if } \boldsymbol{s}_{i-1}=1 \text { and } \boldsymbol{s}_{i+1}=1, \\
\psi\left(p_{l}+\left\langle\boldsymbol{s}_{i}\right\rangle\right)-\psi\left(p_{l}+q_{l}+1\right) \\
\text { if } \boldsymbol{s}_{i-1}=0 \text { and } \boldsymbol{s}_{i+1}=0, \\
\psi\left(p_{u}+\left\langle\boldsymbol{s}_{i}\right\rangle\right)-\psi\left(p_{u}+q_{u}+1\right) \\
\text { if } \boldsymbol{s}_{i-1}=0, \boldsymbol{s}_{i+1}=1 \text { or } \boldsymbol{s}_{i-1}=1, \boldsymbol{s}_{i+1}=0,
\end{array}\right. \\
& \left\langle\ln \left(1-\pi_{i}\right)\right\rangle=\left\{\begin{array}{c}
\psi\left(q_{h}+1-\left\langle\boldsymbol{s}_{i}\right\rangle\right)-\psi\left(p_{h}+q_{h}+1\right) \\
\text { if } \boldsymbol{s}_{i-1}=1 \text { and } \boldsymbol{s}_{i+1}=1, \\
\psi\left(q_{l}+1-\left\langle\boldsymbol{s}_{i}\right\rangle\right)-\psi\left(p_{l}+q_{l}+1\right) \\
\text { if } \boldsymbol{s}_{i-1}=0 \text { and } \boldsymbol{s}_{i+1}=0, \\
\psi\left(q_{u}+1-\left\langle\boldsymbol{s}_{i}\right\rangle\right)-\psi\left(p_{u}+q_{u}+1\right) \\
\text { if } \boldsymbol{s}_{i-1}=0, \boldsymbol{s}_{i+1}=1 \text { or } \boldsymbol{s}_{i-1}=1, \boldsymbol{s}_{i+1}=0,
\end{array}\right. \tag{25}
\end{align*}
$$

where $\psi(x)=\frac{d \ln \Gamma(x)}{d x}$ is a digamma function.

### 3.3. Estimation of noise precision

Finally, the variational distribution of $\beta$ is a Gamma distribution,
$q(\beta)=\operatorname{Gamma}\left(\beta \left\lvert\, c+\frac{M L}{2}\right., \frac{\left\langle\|\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z})\|_{2}^{2}\right\rangle}{2}+d\right)$,
with expectation

```
Algorithm 1 VB for MMV with structured sparsity patterns.
    Input: multiple measurement vectors \(\boldsymbol{Y}\), the dictionary matrix \(\boldsymbol{\Phi}\)
    Output: source signals \(\boldsymbol{X}\)
    Initialize: \(a=b=c=d=10^{-6}, p_{h}=q_{l}=\frac{N L-1}{N L}, p_{l}=q_{h}=\frac{1}{N L}, v_{0}=L\),
    \(\boldsymbol{V}_{0}=\boldsymbol{I}_{L}, \boldsymbol{W}=\mathbf{0}, \gamma_{i}=1(i=1, \ldots N), \boldsymbol{Z}=\mathbf{0}\),.
    While not converged do.
    Update B using (21).
    Update \(\gamma\) using (19)
    Update \(\boldsymbol{W}\) using (17).
    Update \(\left\langle\ln \pi_{i}\right\rangle\) and \(\left\langle\ln \left(1-\pi_{i}\right)\right\rangle\) using (25).
    Update \(\boldsymbol{s}(\boldsymbol{Z})\) using (23).
    Update \(\beta\) using (27).
    end while.
    Set \(\boldsymbol{X}=\boldsymbol{W} \circ \boldsymbol{Z}\)
```

$$
\begin{equation*}
\langle\beta\rangle=\frac{2 c+M L}{2 d+\left\langle\|\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z})\|_{2}^{2}\right\rangle}, \tag{27}
\end{equation*}
$$

where

$$
\begin{align*}
& \left\langle\|\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z})\|_{2}^{2}\right\rangle \\
& =\boldsymbol{y}^{T} \boldsymbol{y}-2 \boldsymbol{y}^{T} \boldsymbol{D}(\langle\boldsymbol{w}\rangle \circ(\boldsymbol{\Omega}\langle\boldsymbol{s}\rangle))+\operatorname{Tr}\left(\left[\left(\Sigma_{\boldsymbol{w}}+\langle\boldsymbol{w}\rangle\langle\boldsymbol{w}\rangle^{T}\right)\right.\right.  \tag{28}\\
& \left.\left.\circ\left(\boldsymbol{\Omega}\left(\mathcal{Z}+\langle\boldsymbol{s}\rangle\langle\boldsymbol{s}\rangle^{T}\right) \boldsymbol{\Omega}^{T}\right)\right] \boldsymbol{D}^{T} \boldsymbol{D}\right),
\end{align*}
$$

$\mathcal{Z}=\operatorname{diag}(\langle\boldsymbol{s}\rangle \circ(1-\langle\boldsymbol{s}\rangle))$ and $\Sigma_{\boldsymbol{w}}=\operatorname{diag}\left(\Sigma_{\boldsymbol{w}_{1}}, \ldots, \Sigma_{\boldsymbol{w}_{N}}\right)$.
In summary, the variational Bayesian procedure infers the posterior distributions of the unknowns iteratively, where in each iteration the method first computes $\boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{B}$ using (17), (19) and (21), followed by the estimation of sparsity patterns $\boldsymbol{s}, \pi$ according to (23), (25), and finally the estimation of noise $\beta$ obtained from (27). The whole algorithm is outlined in Algorithm 1.

## 4. Discussions

### 4.1. Analysis of variational Bayesian inference

The ways to approximate the posterior density functions include maximum a posterior (MAP) estimation, Markov chain Monte Carlo (MCMC) analysis using a Gibbs sampler and variational Bayesian (VB) approximation. Although all of these methods may provide local minima, the full Bayesian inference including MCMC and VB is generally more effective in avoiding undesired local minima compared to deterministic methods such as MAP, since the full Bayesian inference approximates the full posterior distributions instead of point estimations. In addition, the MCMC method requires a large number of burn-in iterations, followed by a sufficient number of iterations to collect samples. As a result, the computational complexity of MCMC is significantly higher than ones of other algorithms. To simplify the computation and the inference, we propose using VB approximation. As a family of probability distribution approximation procedures, VB offers appealing advantages. First, VB algorithm directly updates the hyperparameters of $q(\Theta)$, and the latest update of $q(\Theta)$ approximate the distribution of $\Theta$. After several iterations, VB usually converges, making the procedure computationally efficient. On the other hand, VB approximation allows us to make the assumption on the factorization of posterior distributions so that we can determine the function form of the posterior distribution of each unknown. Since all distributions in the Bayesian model are in the conjugate exponential family, it is easy to calculate all posterior distributions. Finally, VB approximates the unknown posterior distributions with simple, analytically tractable distributions, which can compute the needed expectations, and therefore extend the applicability and the adaptability of Bayesian inference to a wide range of modeling options. More complicated priors that are needed in some typical problems can be utilized, resulting in the improved estimation accuracy.

In the above iterations, the dominant computational process is (17) and (21). For convenience, we attempt to simplify the $\boldsymbol{B}$ learning rule (21) using the Woodbury identity,
$\langle\boldsymbol{B}\rangle=\left(v_{0}+N\right)\left(\boldsymbol{V}_{0}-\boldsymbol{V}_{0} \Delta\left(\boldsymbol{I}_{L}+\boldsymbol{V}_{0} \boldsymbol{\Delta}\right)^{-1} \boldsymbol{V}_{0}\right)$,
where $\boldsymbol{\Delta}=\sum_{i=1}^{N}\left\langle\gamma_{i}\right\rangle\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right)$. Consequently, the dominant complexity of the algorithm is $\mathcal{O}\left((N+1) L^{3}\right)$. Most of precisions $\gamma_{i}$ assume very large values and most of sparsity labels $\boldsymbol{s}_{i}$ become numerically equal to zero in the very first iterations, so that the actual complexity of (17) reduces rapidly with the iterations.

We now draw connections of the proposed Bayesian model to optimization-based methods. To see the connection more clearly, the negative logarithm of the full posterior density function of our Bayesian model is
$-\log p(\Theta \mid \boldsymbol{y}, \mathcal{H})$
$=\frac{\beta}{2}\|\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z})\|_{2}^{2}+\sum_{i=1}^{N} \frac{\gamma_{i}}{2} \boldsymbol{w}_{[i]}^{T} \boldsymbol{B} \boldsymbol{w}_{[i]}-\log \left[\prod_{i=1}^{N} f\left(\boldsymbol{s}_{i}, \pi_{i} ; \mathcal{H}\right)\right]$

$$
\begin{equation*}
-\log \left[\prod_{i=1}^{N} \operatorname{Gamma}\left(\gamma_{i} \mid \mathcal{H}\right) \operatorname{Gamma}(\beta \mid \mathcal{H})\right]+\text { const }, \tag{30}
\end{equation*}
$$

where $\Theta$ represents all model parameters, $\mathcal{H}=\left\{a, b, c, d, p_{h}, q_{h}\right.$, $\left.p_{l}, q_{l}\right\}$ is hyperparameters and $f(;)$ represents the Bernoulli-beta prior in (7) and (10). The error term $\|\boldsymbol{Y}-\boldsymbol{\Phi} \boldsymbol{X}\|_{F}^{2}$ in deterministic optimization-based approaches such as [23], [28] corresponds to the Gaussian prior associated with the measurement noise in (12). In the deterministic optimization-based approaches, there is a perennial challenge about tuning the parameter to balance the error term and the regularization term. However, in the proposed Bayesian model, the prior about noise variance is not required. The model can learn the variance when performing inference. Zhang et al. $[39,40]$ demonstrated that the operation $\boldsymbol{w}_{[i]}^{T} \boldsymbol{B} \boldsymbol{w}_{[i]}$ functioned as a row regularization, which can replace the row-norms (such as the $\ell_{2}$ norm and the $\ell_{\infty}$ norm) in iterative reweighted algorithms for the MMV model. In [28], a pseudo norm $\|\boldsymbol{W}\|_{1,2}=$ $\sum_{i=1}^{N}\left(\boldsymbol{W}_{i} \cdot \boldsymbol{P} \boldsymbol{W}_{i .}^{T}\right)^{1 / 2}$ is related to this operation. The Bernoulli-beta distribution (like an $\ell_{0}$ prior) is employed to impose the sparsity and capture the spatial temporal correlation, which is connected to the row support $\mathcal{R}(\boldsymbol{W})=\left\{1 \leq i \leq N \mid \boldsymbol{W}_{i} \neq \mathbf{0}\right\}$ defined in [19,21]. Thus, we can note that there exist strong connections between the Bayesian formulation and the deterministic optimization-based approaches. Apart from the differences in the form of expressions between the Bayesian model and the related deterministic optimization-based approaches, the main difference is that Bayesian inference estimates the distribution of the unknown parameters, while one effectively seeks a single solution that minimizes the objective function analogous to $-\log p(\Theta \mid \boldsymbol{y}, \mathcal{H})$ in the deterministic optimization-based approaches.

### 4.2. Convergence analysis

Since the proposed algorithm is derived based on the variational Bayes algorithm, we can study the convergence of the variational Bayes algorithm to understand the convergence of the proposed algorithm. The expression (15) for the optimal solution $q\left(\Theta_{k}\right)$ depends on calculating the expectations with respect to the other factors $q\left(\Theta_{j}\right)$ for $j \neq k$. One needs to cycle through all the factors for obtaining the maximum of the lower bound (14). To
conduct the variational inference, all the factors $q\left(\Theta_{j}\right)$ need to be initialized. Then, each factor is estimated in turn with a updated value obtained by (15) using the current solutions for all of the other factors. Convergence is guaranteed since the bound is convex with respect to each of the factors $q\left(\Theta_{j}\right)$ [36,41]. The detailed inference process of updating factors is provided in the Appendix.

## 5. Simulation experiments

In this section, we illustrate the performance of the proposed model using an evaluation with simulated data. The simulations are conducted with two types of signals, which are generated by a first-order autoregressive (AR) process and a Hanning-window tapered sinusoid, respectively. The performance is examined and compared with those obtained from state-of-the-art algorithms.

For the synthetic data, we consider two scenarios: MMV with spatial unstructured and structured cases. The unstructured case without the spatial correlation structure corresponds to the data with the temporal correlation structure, which ignores the possible structures existing among different elements in each measurement so that the nonzero rows of MMV are distributed randomly. MMV with structured case assumes that the nonzero entries in each measurement are clustered. The structured case corresponds to the data with both the temporal correlation structure and the spatial correlation structure. For block-sparse signal, the $N$ dimensional sparse signal contains $K$ nonzero elements ( $K$ is also called sparsity level), which are partitioned into $g$ blocks with random sizes and random locations. Since all measurements of MMV share the same sparse support, the $K$ nonzero rows are actually divided into $g$ blocks with random sizes and random locations.

All the experiments consist of 1000 independent runs. The random measurement matrix $\boldsymbol{\Phi}$ corresponds to a uniform spherical ensemble, so each column of $\boldsymbol{\Phi}$ is drawn from a uniform distribution on the $M$-sphere with radius 1 . Independent and identically distributed Gaussian noise is used with a desired value of SNR, which is defined as $\operatorname{SNR}(\mathrm{db})=10 \log _{10}\left(\|\boldsymbol{\Phi} \boldsymbol{X}\|_{F}^{2} /\|\boldsymbol{E}\|_{F}^{2}\right)$. Two performance metrics are used throughout the experiments. The first metric, which refers to the mean square error (MSE), is quantified as MSE $=\|\widehat{\boldsymbol{X}}-\boldsymbol{X}\|_{F}^{2} /\|\boldsymbol{X}\|_{F}^{2}$, where $\boldsymbol{X}$ and $\widehat{\boldsymbol{X}}$ denote the ground truth and the reconstructed signal, respectively. Normalized mean squared error (NMSE) is calculated by averaging MSE of 1000 independent trials. The second metric, used to gauge the accuracy of the sparse support, is the support failure rate (SFR), which is the number of indices in which the estimated and true support differ, normalized by the cardinality of the original support $K$. Normalized support failure rate (NSFR) is defined as averaged SFR over 1000 independent trials.

The following methods are included in the experiment: MFOCUSS [11], M-SBL [23], TMSBL [25], and AMP-MMV [27]. For M-SBL, TMSBL and AMP-MMV, MATLAB implementations are obtained from the authors' websites while MATLAB implementation from the Multiple-Spars Toolbox ${ }^{2}$ is used for M-FOCUSS. Among these algorithms, TMSBL, and AMP-MMV are the ones that take temporal correlation into account, while M-FOCUSS and M-SBL are effective techniques that do not consider the temporal correlation. The proposed algorithm, termed BMMV, exploits both temporal and spatial correlations by (4) and (10), respectively. In the experiments, the proposed iteration algorithm is repeated until a convergence criterion, e.g., $\left\|\widehat{\boldsymbol{X}}_{k}-\widehat{\boldsymbol{X}}_{k-1}\right\| \leq 10^{-6}$, where $\widehat{\boldsymbol{X}}_{k}$ and $\widehat{\boldsymbol{X}}_{k-1}$ are estimates of $\boldsymbol{X}$ in the $k$ th and $(k-1)$ th iterations, respectively or when the number of iterations $k$ attains a specified maximum number of iterations $k_{\text {max }}$.

[^2]

Fig. 3. Performance versus $N / M$. (a) NMSE of unstructured case, (b) NMSE of structured case, (c) NFSR of unstructured case, (d) NFSR of structured case, (e) Running time of unstructured case, (f) Running time of structured case.

### 5.1. Signals generated by an $A R(1)$ process

In this experiment, nonzero rows (sources) are generated by the $\operatorname{AR}(1)$ model. Higher order process can be readily manipulated, and however it increases the complexity of the experiment. In fact, $\operatorname{AR}(1)$ process can assure the intra-row correlation structure and make a good compromise between performance and complexity. The $i$-th source that satisfies an $\operatorname{AR}(1)$ process is generated according to
$\boldsymbol{X}_{i, j}=a \boldsymbol{X}_{i, j-1}+\sqrt{1-a^{2}} \varepsilon_{i, j}, i=1, \ldots, N ; j=1, \ldots, L$,
where $a$ is the AR model parameter controlling the temporal correlation and $\varepsilon_{i, j} \sim \mathcal{N}\left(0, \sigma_{i}^{2}\right)$. To generate $K$-sparsity signals, most of $\sigma_{i}^{2}$ are set to zero, except for $K$ rows where the $\sigma_{i}^{2}$ are drawn uniform in the range [1, 2]. As discussed, the $K$ nonzero rows of MMV with spatial unstructured measurements are distributed randomly, while the $K$ nonzero rows are divided into $g$ blocks for MMV with spatial structured measurements.
(1) Performance versus underdetermined level $N / M$ : The highlighted principle of CS theory is to derive better recovery performance, in terms of the minimum number of measurements necessary to obtain perfect or nearly perfect reconstruction. Consequently, experiments are first manipulated by varying the sampling rate. To this end, we first study the performance of the considered algorithms across the undersampling level $N / M$. In the experiment, $N / M$ varies from 5 to 25 together with $N=500, L=10$, $\operatorname{SNR}=25, a=0.9$ and $K=M / 2$. For the structured case, the number of nonzero blocks $g$ is $M / 5$. It is seen that the methods with temporal correlation i.e. TMSBL, AMP-MMV, BMMV, are uniformly better than the methods without temporal priors both in terms of NMSE and NSFR. As shown in Fig. 3, the probability of false recovery increases as the underdetermined level $N / M$ becomes larger. Although the performance of M-FOCUSS is stable, it offers worse result than other methods. Fig. 3(a) and (c) manifest that the performance of TMSBL, AMP-MMV, and BMMV are in the same level. In the structured case, BMMV achieves lowest NMSE and NSFR among all algorithms. It demonstrates the ability of our method in capturing both temporal and spatial correlation structures.


Fig. 4. Performance versus L. (a) NMSE of unstructured case, (b) NMSE of structured case, (c) NFSR of unstructured case, (d) NFSR of structured case, (e) Running time of unstructured case, (f) Running time of structured case.
(2) Performance versus the number of measurement vectors $L$ : In this experiment, we study the effects of the number of measurements on the recovery accuracy of MMV. The parametric setting is $N=500, M=100, \mathrm{SNR}=25, a=0.9, K=50$ and $g=10 . L$ varies from 2 to 10. Fig. 4 shows the NMSE and NSFR as functions of the number of measurements $L$. As shown, all algorithms have better performance with increasing $L$. It verifies the theory that the recovery accuracy can be improved using multiple measurement vectors compared to SMV case. In fact, these approaches achieve the performance of the support-aware oracle estimator around $L=8$ in terms of NSFR. The unstructured results in Fig. 4(a) and (c) show that TMSBL, AMP-MMV BMMV and M-SBL are significantly better than ones of M-FOCUSS. Beyond this observation, TMSBL achieves the lowest NMSE, while AMP-MMV outperforms M-SBL, TMSBL, and BMMV in NSFR. Fig. 4(b) and (d) depict the performance of the structured data. Although TMSBL and AMP-MMV present similar performance as BMMV in unstructured case, they make no improvement in structured case. In contrast, BMMV provides a modest improvement compared to the unstructured case,
reflecting the fact that it is necessary to exploit the spatial correlations of the structured signal.
(3) Performance versus sparsity $K$ : This experiment tests how the performance changes as a function of the sparsity. The dimension is also fixed at $N=500$ with $M=100, \operatorname{SNR}=25, a=0.9$, $L=10$ and $g=10$. The sparsity $K$ changes from 10 to 80 , implying that the ratio of measurements-to-active-elements, $M / K$, ranges from 1.25 to 10 . We can see in Fig. 5 that the performances of all algorithms degrade with increasing the number of nonzero elements. For unstructured case, all algorithms except MFOCUSS appear to stall at around the same sparsity and BMMV remains competitive with other algorithms, in particular for low sparsity. However, BMMV provides the most accurate recovery of the structured data, including the reconstructed magnitude and the recovered support.
(4) Performance versus signal scale $N$ : The MMV technique relies on the fact that the acquired high-dimensional signals reside in a low-dimensional structure and intends to recover highdimensional signals from a small number of samples. Therefore,


Fig. 5. Performance versus sparsity K. (a) NMSE of unstructured case, (b) NMSE of structured case, (c) NFSR of unstructured case, (d) NFSR of structured case, (e) Running time of unstructured case, (f) Running time of structured case.
we execute the experiment, in which the signal dimension $N$ is in the range $[100,1000]$. For each scale, we set $M=N / 2, K=M / 2$, $a=0.9, L=10$, and $g=10$. Fig. 6 sums up the performance obtained by different algorithms under the two scenarios. As the signal scale increases the reconstructed errors of algorithms become more and more smooth and eventually give rise to fixed precisions. In the unstructured case, all algorithms enjoy significant performance in identifying the true support, while algorithms except M-FOCUSS deliver superior NMSE. It can be observed that BMMV succeeds in keeping both the lowest NMSE and NSFR on a large range of signal scales when embarking on structured data. BMMV nicely takes benefit from the additional constraint on the support and thus exhibits appealing results. Experiments confirm the advantages of taking the spatial structures of MMV into account.
(5) Performance versus SNR: The SNR can be used to quantify the robustness of MMV recovery. To sufficiently analyze the robustness of considered approaches, the range of SNR values varies from 5 db to 30 db . We choose the parameters: $N=500, M=100$,
$a=0.9, K=50, L=10$ and $g=10$. Fig. 7 plots the reconstructed results as a function of the noise level. As shown, all algorithms present acceptable robustness. When $S N R \geq 15$, the performances of TMSBL, AMP-MMV, and BMMV are significantly better than one of M-FOCUSS. However, one should roughly know or estimate the noise level when using some methods such as the TMSBL. It is practical prohibitive for real data. BMMV regards the noise as a variable, which can be endowed with a prior and updated adaptively at each iteration. Referring to the structured case, BMMV perform consistently better compared with the other methods. Therefore, BMMV gives relatively effective recovery at different level of noise, which verifies its robustness.
(6) The running time: It is instructive to examine the computational complexity of different approaches at each iteration. Due to M-FOCUSS and M-SBL ignore all correlation structures in MMV, we mainly focus on TMSBL, AMP-MMV and BMMV. AMP-MMV exploits a Bayesian approximate message passing algorithm to solve MMV problem, it follows that the overall complexity of AMPMMV is $\mathcal{O}(L M N)$. The dominant computational process of TMSBL

 unstructured case, (f) Running time of structured case.
and BMMV is calculating the matrix inversion. The complexity of TMSBL is $\mathcal{O}\left(\left(M^{3}+1\right) L^{3}\right)$, while the complexity of BMMV is $\mathcal{O}\left((N+1) L^{3}\right)$. In addition, most of precisions $\gamma_{i}$ assume very large values and most of sparsity labels $\boldsymbol{s}_{i}$ become numerically equal to zero in the very first iterations, so that the actual complexity of BMMV reduces rapidly with the iterations. Referring to the computational complexity, AMP-MMV is computationally considerably more efficient compared to other methods and our algorithm has a similar time performance with TMSBL. We show the running times versus different variables in above experiments. In summary, M-FOCUSS and AMP-MMV deliver good time performances, which is slightly better than ones of M-SBL, TMSBL, BMMV. However, as discussed, the effectiveness of M-FOCUSS is limited because of ignoring the correlation structure in MMV problem, while AMPMMV fails to capture the spatial correlation structure. The correlation structure including the temporal correlation structure and the spatial correlation structure is fully considered in the proposed algorithm. It can be noticed that our method, although not as good as AMP-MMV in terms of time performance, still delivers acceptable performance.

### 5.2. Signals generated by a Hanning window tapered sinusoid

In this part, each active signal with $L$ measurements is generated by a Hanning window tapered sinusoid, where the number of periods is uniformly distributed between 1 and 3 and the phase of the sine wave is uniformly drawn between 0 and $\pi$. As in the previous section, we carry out five experiments with setting the same parameters and assess the performance of the proposed algorithm in both unstructured and structured cases. For each experiment, only a primary variable is varied, while the others are fixed. For each experiment, the results including NMSE and NSFR averaged over all scales are shown in Table 1 and 2.

In the unstructured case, AMP-MMV achieves the better performance than other methods in most cases, and TMSBL presents a lower NMSE than other methods for signals generated by an $\operatorname{AR}(1)$ process as indicated in the previous experiment. The merit of incorporating smoothness into MMV is even compelling when comparing M-FOCUSS, M-SBL with TMSBL, AMP-MMV and BMMV. Although BMMV is slightly inferior to AMP-MMV in the unstruc-


Fig. 7. Performance versus noise SNR. (a) NMSE of unstructured case, (b) NMSE of structured case, (c) NFSR of unstructured case, (d) NFSR of structured case, (e) Running time of unstructured case, (f) Running time of structured case.
tured case, it obtains the best results in both NMSE and NSFR for structured case.

## 6. Application to EEG data

Electroencephalogram (EEG) can record brain waves of electricity, which is a commonly used brain imaging way and play an important part in psychology and neurology. In a practical clinical setting, the same stimulus may be repeated many times to obtain high SNR estimation of the evoked response so that the EEG signals are recorded over multiple channels with multiple trials. The major challenge for the collection and analysis of EEG data is the storage and the processing of the huge amount of data. EEG activity of a single subject can correspond to a lot of the data. Therefore, there is a need for compression or data reduction processes that can reduce the number of samples and also recover the most important features. An effective compression of the EEG data should not only reduce the number of sampled data and allow for a fast wireless transmission in a clinical diagnosis.

A potential application of MMV is source location in EEG data. In the formulation of the source location problem, $\boldsymbol{Y}$ represents the recordings from $L$ channels and $\boldsymbol{X}$ denotes the unknown sources of $N$ current dipoles distributed over the cortical surface. The measurements correspond to a small number of current dipoles that represent the active brain regions. The algorithm with exploiting the temporal correlation can obtain continuous estimation resulting in accurate estimation. Turning to the sparse recovery problem, we represent the signals in the transformed domain under the DCT dictionary. The model (2) can be written as
$\boldsymbol{Y}=\boldsymbol{\Phi} \Psi \widetilde{\boldsymbol{X}}+\boldsymbol{E}$,
$\boldsymbol{\Psi}$ is a DCT dictionary under which $\boldsymbol{X}$ has a sparse representation. In this experiment, the model (32) will be adapted. The nonzero elements of the DCT coefficients $\widetilde{\boldsymbol{X}}_{. i}$ are distributed as concatenation of a number of nonzero blocks. The block-sparse structures in which the nonzero elements occur in clusters result in the spatial

Table 1
The reconstructed results (NMSE(NSFR)) for the unstructured case.

| Method/variable | $N / M$ | $L$ | K | $N$ | SNR |
| :---: | :---: | :---: | :---: | :---: | :---: |
| M-FOCUSS | $\begin{aligned} & 1.2 \times 10^{-1} \\ & \left(6.6 \times 10^{-2}\right) \end{aligned}$ | $\begin{aligned} & 6.0 \times 10^{-4} \\ & \left(3.9 \times 10^{-2}\right) \end{aligned}$ | $\begin{aligned} & 2.7 \times 10^{-3} \\ & \left(1.1 \times 10^{-1}\right) \end{aligned}$ | $\begin{aligned} & 2.3 \times 10^{-3} \\ & \left(1.4 \times 10^{-1}\right) \end{aligned}$ | $\begin{aligned} & 1.2 \times 10^{-2} \\ & \left(6.2 \times 10^{-1}\right) \end{aligned}$ |
| M-SBL | $\begin{aligned} & 5.4 \times 10^{-2} \\ & \left(1.2 \times 10^{-2}\right) \end{aligned}$ | $\begin{aligned} & 2.6 \times 10^{-4} \\ & \left(2.1 \times 10^{-2}\right) \end{aligned}$ | $\begin{aligned} & 7.0 \times 10^{-4} \\ & \left(3.8 \times 10^{-2}\right) \end{aligned}$ | $\begin{aligned} & 8.4 \times 10^{-4} \\ & \left(9.1 \times 10^{-2}\right) \end{aligned}$ | $\begin{aligned} & 8.3 \times 10^{-3} \\ & \left(5.4 \times 10^{-1}\right) \end{aligned}$ |
| TMSBL | $\begin{aligned} & 9.8 \times 10^{-4} \\ & \left(1.2 \times 10^{-2}\right) \end{aligned}$ | $\begin{aligned} & 2.1 \times 10^{-4} \\ & \left(6.7 \times 10^{-3}\right) \end{aligned}$ | $\begin{aligned} & \mathbf{4 . 8 \times 1 0 ^ { - 4 }} \\ & \left(1.6 \times 10^{-2}\right) \end{aligned}$ | $\begin{aligned} & 6.6 \times 10^{-4} \\ & \left(2.9 \times 10^{-2}\right) \end{aligned}$ | $\begin{aligned} & 3.7 \times 10^{-3} \\ & \left(1.1 \times 10^{-1}\right) \end{aligned}$ |
| AMP-MMV | $\begin{aligned} & 2.2 \times 10^{-4} \\ & \left(8.2 \times 10^{-3}\right) \end{aligned}$ | $\begin{aligned} & 1.6 \times 10^{-4} \\ & \left(4.7 \times 10^{-3}\right) \end{aligned}$ | $\begin{aligned} & 1.3 \times 10^{-3} \\ & \left(\mathbf{7 . 0} \times \mathbf{1 0}^{-\mathbf{3}}\right) \end{aligned}$ | $\begin{aligned} & 3.8 \times 10^{-4} \\ & \left(1.4 \times 10^{-2}\right) \end{aligned}$ | $\begin{aligned} & 2.1 \times 10^{-3} \\ & \left(5.7 \times 10^{-\mathbf{2}}\right) \end{aligned}$ |
| BMMV | $\begin{aligned} & 7.2 \times 10^{-3} \\ & \left(1.1 \times 10^{-2}\right) \end{aligned}$ | $\begin{aligned} & 1.8 \times 10^{-4} \\ & \left(5.6 \times 10^{-3}\right) \end{aligned}$ | $\begin{aligned} & 9.4 \times 10^{-4} \\ & \left(9.0 \times 10^{-3}\right) \end{aligned}$ | $\begin{aligned} & 6.1 \times 10^{-4} \\ & \left(1.9 \times 10^{-2}\right) \end{aligned}$ | $\begin{aligned} & \mathbf{1 . 9 \times 1 0 ^ { - \mathbf { 3 } }} \\ & \left(8.1 \times 10^{-2}\right) \end{aligned}$ |

Table 2
The reconstructed results (NMSE(NSFR)) for the structured case.

| Method/variable | $N / M$ | $L$ | $K$ | $N$ | SNR |
| :--- | :--- | :--- | :--- | :--- | :--- |
| M-FOCUSS | $5.3 \times 10^{-1}$ | $1.4 \times 10^{-3}$ | $6.5 \times 10^{-3}$ | $7.6 \times 10^{-3}$ | $5.6 \times 10^{-2}$ |
|  | $\left(1.0 \times 10^{-1}\right)$ | $\left(8.6 \times 10^{-2}\right)$ | $\left(5.1 \times 10^{-1}\right)$ | $\left(4.4 \times 10^{-1}\right)$ | $\left(8.7 \times 10^{-1}\right)$ |
| M-SBL | $8.4 \times 10^{-2}$ | $7.6 \times 10^{-4}$ | $4.0 \times 10^{-3}$ | $3.6 \times 10^{-3}$ | $2.0 \times 10^{-2}$ |
|  | $\left(6.2 \times 10^{-2}\right)$ | $\left(5.9 \times 10^{-2}\right)$ | $\left(6.2 \times 10^{-2}\right)$ | $\left(2.5 \times 10^{-1}\right)$ | $\left(6.9 \times 10^{-1}\right)$ |
| TMSBL | $1.9 \times 10^{-3}$ | $4.2 \times 10^{-4}$ | $1.2 \times 10^{-3}$ | $1.3 \times 10^{-3}$ | $6.4 \times 10^{-3}$ |
|  | $\left(3.7 \times 10^{-2}\right)$ | $\left(8.2 \times 10^{-3}\right)$ | $\left(4.5 \times 10^{-2}\right)$ | $\left(5.4 \times 10^{-2}\right)$ | $\left(4.8 \times 10^{-1}\right)$ |
| AMP-MMV | $2.6 \times 10^{-3}$ | $4.6 \times 10^{-4}$ | $4.1 \times 10^{-3}$ | $8.7 \times 10^{-4}$ | $4.5 \times 10^{-3}$ |
|  | $\left(1.2 \times 10^{-2}\right)$ | $\left(6.3 \times 10^{-\mathbf{3}}\right)$ | $\left(1.3 \times 10^{-2}\right)$ | $\left(5.2 \times 10^{-2}\right)$ | $\left(9.7 \times 10^{-2}\right)$ |
| BMMV | $\mathbf{3 . 0 \times \mathbf { 1 0 } ^ { - \mathbf { 4 } }}$ | $\mathbf{1 . 2 \times \mathbf { 1 0 } ^ { - \mathbf { 4 } }}$ | $\mathbf{1 . 2 \times \mathbf { 1 0 } ^ { \mathbf { - 3 } }}$ | $\mathbf{3 . 4 \times \mathbf { 1 0 } ^ { - \mathbf { 4 } }}$ | $\mathbf{1 . 3 \times \mathbf { 1 0 } ^ { - \mathbf { 3 } }}$ |
|  | $\left(\mathbf{8 . 4 \times \mathbf { 1 0 } ^ { \mathbf { - 3 } } )}\right.$ | $\left(\mathbf{3 . 5 \times \mathbf { 1 0 } ^ { - \mathbf { 4 } } )}\right.$ | $\left(\mathbf{8 . 3 \times \mathbf { 1 0 } ^ { - \mathbf { 3 } } )}\right.$ | $\left(\mathbf{7 . 3 \times \mathbf { 1 0 } ^ { \mathbf { - 3 } } )}\right.$ | $\left(\mathbf{3 . 4 \times \mathbf { 1 0 } ^ { \mathbf { - 2 } } )}\right.$ |


 references to color in this figure, the reader is referred to the web version of this article.)

 color in this figure, the reader is referred to the web version of this article.)

Table 3
The reconstructed results for the EEG data.

| Method | NMSE | NSFR |
| :--- | :--- | :--- |
| M-FOCUSS | $3.1 \times 10^{-1}$ | $5.9 \times 10^{-1}$ |
| M-SBL | $1.4 \times 10^{-1}$ | $4.7 \times 10^{-1}$ |
| TMSBL | $4.9 \times 10^{-2}$ | $2.8 \times 10^{-1}$ |
| AMP-MMV | $5.7 \times 10^{-2}$ | $2.4 \times 10^{-1}$ |
| BMMV | $\mathbf{2 . 5} \times \mathbf{1 0}^{-\mathbf{2}}$ | $\mathbf{1 . 1} \times \mathbf{1 0}^{\mathbf{- 1}}$ |

correlation structure. According to the model (3), the model (32) can be developed to
$\boldsymbol{Y}=\boldsymbol{\Phi} \boldsymbol{\Psi}(\widetilde{\boldsymbol{W}} \circ \widetilde{\boldsymbol{Z}})+\boldsymbol{E}$.
A set of EEG data from [42] are used in the experiment. The data is divided into a number of short time segments, with each consisting of 250 data samples. Fig. 8 and Fig. 9 show the reconstructed results of arbitrarily selected six channels (represented by different colors). It is clear that AMP-MMV and BMMV deliver the better vision effect and results of other methods contain false variations. Although the gap between AMP-MMV and BMMV are small, the reconstructed signals by BMMV are much closer and smoother to the original signals. One can focus on the channel represented by black color (especially the tail of the channel). It can be observed that the reconstructed signals by BMMV (Fig. 9(c)) are much closer and smoother to the original signals. The quantitative compared results are shown in Table 3. It can be seen that BMMV achieves smallest error among all algorithms and shows perfect performance for real data. The recovery of EEG data confirms that BMMV is highly effective in recovering MMV with spatial structured sparsity.

## 7. Conclusions

To solve the MMV with spatial structured sparsity patterns encountered in the signal processing, we point out a Bayesian model with considering both spatial and temporal dependencies. By resorting to the beta process that separates the learning of sparseness from the learning of magnitudes, the proposed algorithm can exploit such multi-task dependencies. The spatial contiguity prior that corresponds to the spatial correlation structure is assumed to satisfy a Markov property. Moreover, the row smoothness priors are incorporated into the sparse Bayesian model, which character the temporal correlation structure. The model assumptions in our approach are flexible and thus cover broad class of real signals to be reconstructed from undersampled measurements. For the fully Bayesian estimation of the model parameters, the efficient variational Bayesian inference is employed. We provide a comprehensive comparison between the suggested method and the state-of-the-art algorithms. Experimental evaluations with simulated data and real EEG data demonstrate the power of our method. Future work will concern on the use of the proposed algorithm in practical applications, in particular in compressed audio signal where the temporal correlation and the spatial structured sparsity are favorably developed for efficient reconstruction.

## Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant No. 91746107 and Grant No. 61379014) and the Natural Science Foundation of Tianjin (No. 16JCYBJC15900).

## Appendix

In this appendix, we give the detailed VB inference.
Estimation of magnitudes
Infer $\boldsymbol{w}_{[i]}\left(\boldsymbol{W}_{i .}\right), \quad(i=1, \cdots, N)$ :
By combining the related terms (4) and (13), we have
$p(\boldsymbol{y}, \Theta) \propto \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}), \beta^{-1} \boldsymbol{I}_{M L}\right) \mathcal{N}\left(w_{[i]} \mid 0,\left(\gamma_{i} \boldsymbol{B}\right)^{-1}\right)$

$$
\begin{aligned}
& \propto \exp \left(-\frac{\gamma_{i}}{2} \boldsymbol{w}_{[i]}^{T} \boldsymbol{B} \boldsymbol{w}_{[i]}\right) \exp \left[-\frac{\beta}{2}\left(\boldsymbol{y}-\boldsymbol{D}_{[\cdot i]} \boldsymbol{w}_{[i]} \boldsymbol{s}_{i}-\sum_{j \neq i}^{N} \boldsymbol{D}_{[\cdot j]} \boldsymbol{w}_{[j]} \boldsymbol{s}_{j}\right)^{T}\left(\boldsymbol{y}-\boldsymbol{D}_{[\cdot i]} \boldsymbol{w}_{[i]} \boldsymbol{s}_{i}-\sum_{j \neq i}^{N} \boldsymbol{D}_{[\cdot j]} \boldsymbol{w}_{[j]} \boldsymbol{s}_{j}\right)\right] \\
& \propto \exp \left(-\frac{1}{2} \boldsymbol{w}_{[i]}^{T}\left(\gamma_{i} \boldsymbol{B}+\beta \mathbf{s}_{i}^{2} \boldsymbol{D}_{[\cdot i]}^{T} \boldsymbol{D}_{[\cdot i]}\right) \boldsymbol{w}_{[i]}+\beta \mathbf{s}_{i} \boldsymbol{w}_{[i]}^{T} \boldsymbol{D}_{[\cdot i]}^{T}\left(\boldsymbol{y}-\sum_{j \neq i}^{N} \boldsymbol{s}_{j} \boldsymbol{D}_{[\cdot j]_{[j]}}\right)\right),
\end{aligned}
$$

$\ln p(\boldsymbol{y}, \Theta) \propto\left(-\frac{1}{2} \boldsymbol{w}_{[i]}^{T}\left(\gamma_{i} \boldsymbol{B}+\beta \boldsymbol{s}_{i}^{2} \boldsymbol{D}_{[\cdot i]}^{T} \boldsymbol{D}_{[\cdot i]}\right) \boldsymbol{w}_{[i]}+\beta \boldsymbol{s}_{i} \boldsymbol{w}_{[i]}^{T} \boldsymbol{D}_{[\cdot i]}^{T}\left(\boldsymbol{y}-\sum_{j \neq i}^{N} \boldsymbol{s}_{j} \boldsymbol{D}_{[\cdot j]} \boldsymbol{w}_{[j]}\right)\right)$,
$E_{\Theta \backslash \boldsymbol{w}_{[i]}}[\ln p(\boldsymbol{y}, \Theta)] \propto\left(-\frac{1}{2} \boldsymbol{w}_{[i]}^{T}\left(\left\langle\gamma_{i}\right\rangle\langle\boldsymbol{B}\rangle+\langle\beta\rangle\left\langle\mathbf{s}_{i}^{2}\right\rangle \boldsymbol{D}_{[\cdot i]}^{T} \boldsymbol{D}_{[\cdot i]}\right) \boldsymbol{w}_{[i]}+\langle\beta\rangle\left\langle\boldsymbol{s}_{i}\right\rangle \boldsymbol{w}_{[i]}^{T} \boldsymbol{D}_{[\cdot i]}^{T}\left(\boldsymbol{y}-\sum_{j \neq i}^{N}\left\langle\boldsymbol{s}_{j}\right\rangle \boldsymbol{D}_{[\cdot j]}\left\langle\boldsymbol{w}_{[j]}\right\rangle\right)\right)$,
$q\left(\boldsymbol{w}_{[i]}\right) \propto \exp \left(\left(-\frac{1}{2} \boldsymbol{w}_{[i]}^{T}\left(\left\langle\gamma_{i}\right\rangle\langle\boldsymbol{B}\rangle+\langle\beta\rangle\left\langle\boldsymbol{s}_{i}^{2}\right\rangle \boldsymbol{D}_{[\cdot i]}^{T} \boldsymbol{D}_{[\cdot i]}\right) \boldsymbol{w}_{[i]}+\langle\beta\rangle\left\langle\boldsymbol{s}_{i}\right\rangle \boldsymbol{w}_{[i]}^{T} \boldsymbol{D}_{[\cdot i]}^{T}\left(\boldsymbol{y}-\sum_{j \neq i}^{N}\left\langle\boldsymbol{s}_{j}\right\rangle \boldsymbol{D}_{[\cdot j]}\left\langle\boldsymbol{w}_{[j]}\right\rangle\right)\right)\right)$,
where $\boldsymbol{D}_{[\cdot i]}=\boldsymbol{D}(1: M L,(i-1) L+1: i L)$.
Therefore,
$q\left(\boldsymbol{w}_{[i]}\right)=\mathcal{N}\left(\boldsymbol{w}_{[i]} \mid\left\langle\boldsymbol{w}_{[i]}\right\rangle, \Sigma_{\boldsymbol{w}_{[i]}}\right)$,
where

$$
\begin{aligned}
\left\langle\boldsymbol{w}_{[i]}\right\rangle & =\Sigma_{\boldsymbol{w}_{[i]}}\langle\beta\rangle\left\langle\boldsymbol{s}_{i}\right\rangle \boldsymbol{D}_{[\cdot i]}^{T}\left(\boldsymbol{y}-\sum_{j \neq i}^{N}\left\langle\boldsymbol{s}_{j}\right\rangle \boldsymbol{D}_{[\cdot j]}\left\langle\boldsymbol{w}_{[j]}\right\rangle\right), \\
\Sigma_{\boldsymbol{w}_{[i]}} & =\left(\left\langle\gamma_{i}\right\rangle\langle B\rangle+\langle\beta\rangle\left\langle s_{i}^{2}\right\rangle \boldsymbol{D}_{[\cdot i]}^{T} \boldsymbol{D}_{[\cdot i]}\right)^{-1} .
\end{aligned}
$$

Infer $\gamma_{i},(i=1, \cdots, N)$ :
By combining the related terms (4) and (5), we have
$p(\boldsymbol{y}, \Theta) \propto \mathcal{N}\left(\boldsymbol{w}_{[i]} \mid 0,\left(\gamma_{i} \boldsymbol{B}\right)^{-1}\right) \operatorname{Gamma}\left(\gamma_{i} \mid a, b\right)$

$$
\propto\left|\gamma_{i} \boldsymbol{B}\right|^{\frac{1}{2}} \exp \left(-\frac{\gamma_{i}}{2} \boldsymbol{w}_{[i]}^{T} \boldsymbol{B} \boldsymbol{w}_{[i]}\right) \Gamma(a)^{-1} b^{a} \gamma_{i}^{a-1} \exp \left(-b \gamma_{i}\right),
$$

$\ln p(\boldsymbol{y}, \Theta) \propto \ln \left(\gamma_{i}^{\frac{L}{2}}\right)+\left(-\frac{\gamma_{i}}{2} \boldsymbol{w}_{[i]}^{T} \boldsymbol{B} \boldsymbol{w}_{[i]}\right)+\ln \left(\gamma_{i}^{a-1}\right)+\left(-b \gamma_{i}\right)$,
$E_{\Theta \backslash \gamma_{i}}[\ln p(\boldsymbol{y}, \Theta)] \propto \ln \left(\gamma_{i}^{\frac{L}{2}+a-1}\right)+\left(-\frac{\gamma_{i}}{2}\left\langle\boldsymbol{w}_{[i]}^{T} \boldsymbol{B} \boldsymbol{w}_{[i]}\right\rangle\right)+\left(-b \gamma_{i}\right)$,
where

$$
\begin{aligned}
\left\langle\boldsymbol{w}_{[i]}^{T} \boldsymbol{B} \boldsymbol{w}_{[i]}\right\rangle & =\left\langle\operatorname{Tr}\left(\boldsymbol{w}_{[i]}^{T} \boldsymbol{B} \boldsymbol{w}_{[i]}\right)\right\rangle \\
& =\left\langle\operatorname{Tr}\left(\boldsymbol{w}_{[i]} \boldsymbol{w}_{[i]}^{T} \boldsymbol{B}\right)\right\rangle \\
& =\operatorname{Tr}\left(\left\langle\boldsymbol{w}_{[i]} \boldsymbol{w}_{[i]}^{T}\right\rangle\langle\boldsymbol{B}\rangle\right) \\
& =\operatorname{Tr}\left(\left\langle\left(\boldsymbol{w}_{[i]}-\left\langle\boldsymbol{w}_{[i]}\right\rangle+\left\langle\boldsymbol{w}_{[i]}\right\rangle\right)\left(\boldsymbol{w}_{[i]}-\left\langle\boldsymbol{w}_{[i]}\right\rangle+\left\langle\boldsymbol{w}_{[i]}\right\rangle\right)^{T}\right\rangle\langle\boldsymbol{B}\rangle\right) \\
& =\operatorname{Tr}\left[\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}^{T}\right\rangle^{T}\right)\langle\boldsymbol{B}\rangle\right] .
\end{aligned}
$$

Therefore,
$q\left(\gamma_{i}\right) \propto \gamma_{i}^{\frac{L}{2}+a-1} \exp \left[-\left(\frac{1}{2} \operatorname{Tr}\left[\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right)\langle\boldsymbol{B}\rangle\right]+b\right) \gamma_{i}\right]$.
Furthermore,
$q\left(\gamma_{i}\right)=\operatorname{Gamma}\left(\gamma_{i} \left\lvert\, \frac{L}{2}+a\right.,\left(\frac{1}{2} \operatorname{Tr}\left[\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right)\langle B\rangle\right]+b\right)\right)$,
$\left\langle\gamma_{i}\right\rangle=\frac{L+2 a}{\operatorname{Tr}\left[\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right)\langle\boldsymbol{B}\rangle\right]+2 b}$.
Infer B:
By combining the related terms (4) and (6), we have
$p(\boldsymbol{y}, \Theta) \propto \prod_{i=1}^{N} \mathcal{N}\left(\boldsymbol{w}_{[i]} \mid 0,\left(\gamma_{i} \boldsymbol{B}\right)^{-1}\right) \mathcal{W}\left(\boldsymbol{B} \mid \boldsymbol{V}_{0}, v_{0}\right)$

$$
\propto \prod_{i=1}^{N}\left(|\boldsymbol{B}|^{\frac{1}{2}} \exp \left(-\frac{\gamma_{i}}{2} \boldsymbol{w}_{[i]}^{T} \boldsymbol{B} \boldsymbol{w}_{[i]}\right)\right) \frac{1}{C}|\boldsymbol{B}|^{\left(v_{0}-L-1\right) / 2} \exp \left(-\frac{1}{2} \operatorname{Tr}\left(\boldsymbol{V}_{0}^{-1} \boldsymbol{B}\right)\right),
$$

$\ln p(\boldsymbol{y}, \Theta) \propto \ln |\boldsymbol{B}|^{\frac{N}{2}}+\sum_{i=1}^{N}\left(-\frac{\gamma_{i}}{2} \boldsymbol{w}_{[i]}^{T} \boldsymbol{B} \boldsymbol{w}_{[i]}\right)+\ln |\boldsymbol{B}|^{\left(v_{0}-L-1\right) / 2}+\left(-\frac{1}{2} \operatorname{Tr}\left(\boldsymbol{V}_{0}^{-1} \boldsymbol{B}\right)\right)$,
$E_{\Theta \backslash \boldsymbol{B}}[\ln p(\boldsymbol{y}, \Theta)] \propto \ln |\boldsymbol{B}|^{\frac{N+v_{0}-L-1}{2}}-\frac{1}{2}\left(\Sigma_{i=1}^{N}\left\langle\gamma_{i}\right\rangle\left\langle\boldsymbol{w}_{[i]}^{T} \boldsymbol{B} \boldsymbol{w}_{[i]}\right\rangle+\operatorname{Tr}\left(\boldsymbol{V}_{0}^{-1} \boldsymbol{B}\right)\right)$,
$q(\boldsymbol{B}) \propto|\boldsymbol{B}| \frac{N+v_{0}-L-1}{2}+\exp \left(-\frac{1}{2} \operatorname{Tr}\left(\sum_{i=1}^{N}\left\langle\gamma_{i}\right\rangle\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right)+\boldsymbol{V}_{0}^{-1}\right) \boldsymbol{B}\right)$.
Therefore,
$q(\boldsymbol{B})=\mathcal{W}\left(\boldsymbol{B} \mid\left(\sum_{i=1}^{N}\left\langle\gamma_{i}\right\rangle\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right)+\boldsymbol{V}_{0}^{-1}\right)^{-1}, N+v_{0}\right)$.
Furthermore,
$\langle\boldsymbol{B}\rangle=\left(N+v_{0}\right)\left(\sum_{i=1}^{N}\left\langle\gamma_{i}\right\rangle\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right)+\boldsymbol{V}_{0}^{-1}\right)^{-1}$.
Estimation of sparsity patterns
Infer $\boldsymbol{s}_{i}(i=1, \cdots, N)$ :
By combining the related terms (7) and (13), we have
$p(\boldsymbol{y}, \Theta) \propto \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}), \beta^{-1} \boldsymbol{I}_{M L}\right)$ Bernoulli $\left(\boldsymbol{s}_{i} \mid \pi_{i}\right)$

$$
\begin{aligned}
& \propto \exp \left[-\frac{\beta}{2}\left(\boldsymbol{y}-\boldsymbol{s}_{i} \boldsymbol{D}_{[\cdot i]} \boldsymbol{w}_{[i]}-\sum_{j \neq i}^{N} \boldsymbol{D}_{[\cdot j]} \boldsymbol{w}_{[j]}\right)^{T}\left(\boldsymbol{y}-\boldsymbol{s}_{i} \boldsymbol{D}_{[\cdot i]} \boldsymbol{w}_{[i]}-\sum_{j \neq i}^{N} \boldsymbol{D}_{[\cdot j]} \boldsymbol{w}_{[j]}\right)\right] \\
& \pi_{i}^{\boldsymbol{s}_{i}}\left(1-\pi_{i}\right)^{1-\boldsymbol{s}_{i}}
\end{aligned}
$$

$\ln p(\boldsymbol{y}, \Theta) \propto-\frac{\beta}{2}\left(\boldsymbol{y}-\boldsymbol{s}_{i} \boldsymbol{D}_{[\cdot i]} \boldsymbol{w}_{[i]}-\sum_{j \neq i}^{N} \boldsymbol{D}_{[\cdot j]} \boldsymbol{w}_{[j]}\right)^{T}\left(\boldsymbol{y}-\boldsymbol{s}_{i} \boldsymbol{D}_{[\cdot i]} \boldsymbol{w}_{[i]}-\sum_{j \neq i}^{N} \boldsymbol{D}_{[\cdot j]} \boldsymbol{w}_{[j]}\right)+\boldsymbol{s}_{i} \ln \left(\pi_{i}\right)$

$$
+\left(1-\boldsymbol{s}_{i}\right) \ln \left(1-\pi_{i}\right)
$$

$E_{\Theta \backslash \boldsymbol{s}_{i}}[\ln p(\boldsymbol{y}, \Theta)] \propto-\frac{\langle\beta\rangle}{2} \boldsymbol{s}_{i}\left\langle\boldsymbol{w}_{[i]}^{T} \boldsymbol{D}_{[\cdot i]}^{T} \boldsymbol{D}_{[\cdot i]} \boldsymbol{w}_{[i]}\right\rangle+\langle\beta\rangle \boldsymbol{s}_{i}\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T} \boldsymbol{D}_{[\cdot i]}^{T}\left(\boldsymbol{y}-\sum_{j \neq i}^{N} \boldsymbol{D}_{[\cdot j]}\left\langle\boldsymbol{w}_{[j]}\right\rangle\right)+\boldsymbol{s}_{i}\left\langle\ln \left(\pi_{i}\right)\right\rangle$

$$
\begin{aligned}
& +\left(1-\boldsymbol{s}_{i}\right)\left\langle\ln \left(1-\pi_{i}\right)\right\rangle \\
& \propto \boldsymbol{s}_{i}\left[-\frac{\langle\beta\rangle}{2} \operatorname{Tr}\left(\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right) \boldsymbol{D}_{[\cdot i]}^{T} \boldsymbol{D}_{[\cdot i]}\right)+\langle\beta\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T} \boldsymbol{D}_{[\cdot i]}^{T}\left(\boldsymbol{y}-\sum_{j \neq i}^{N} \boldsymbol{D}_{[\cdot j]}\left\langle w_{[j]}\right\rangle\right)\right. \\
& \left.+\left\langle\ln \left(\pi_{i}\right)\right\rangle\right]+\left(1-\boldsymbol{s}_{i}\right)\left\langle\ln \left(1-\pi_{i}\right)\right\rangle,
\end{aligned}
$$

$q\left(\boldsymbol{s}_{i}\right) \propto \exp \left(\boldsymbol{s}_{i}\left[-\frac{\langle\beta\rangle}{2} \operatorname{Tr}\left(\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right) \boldsymbol{D}_{[\cdot i]}^{T} \boldsymbol{D}_{[\cdot i]}\right)+\langle\beta\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T} \boldsymbol{D}_{[\cdot i]}^{T}\left(\boldsymbol{y}-\sum_{j \neq i}^{N} \boldsymbol{D}_{[\cdot j]}\left\langle\boldsymbol{w}_{[j]}\right\rangle\right)+\left\langle\ln \left(\pi_{i}\right)\right\rangle\right]\right)$

$$
\exp \left(\left(1-\boldsymbol{s}_{i}\right)\left\langle\ln \left(1-\pi_{i}\right)\right\rangle\right)
$$

Therefore,
$q\left(\boldsymbol{s}_{i}\right)=\operatorname{Bernoulli}\left(\boldsymbol{s}_{i} \mid \xi^{\boldsymbol{s}_{i}}, \zeta^{1-\boldsymbol{s}_{i}}\right)$.
where,
$\xi=\exp \left(\left\langle\ln \pi_{i}\right\rangle-\frac{\langle\beta\rangle}{2} \operatorname{Tr}\left(\left(\Sigma_{\boldsymbol{w}_{[i]}}+\left\langle\boldsymbol{w}_{[i]}\right\rangle\left\langle\boldsymbol{w}_{[i]}\right\rangle^{T}\right) \boldsymbol{D}_{[\cdot i]}^{T} \boldsymbol{D}_{[\cdot i]}\right)+\langle\beta\rangle\left(\boldsymbol{y}^{-i}\right)^{T} \boldsymbol{D}_{[\cdot i]}\left\langle\boldsymbol{w}_{[i]}\right\rangle\right)$,
$\zeta=\exp \left(\left\langle\ln \left(1-\pi_{i}\right)\right\rangle\right)$.
Furthermore,
$\left\langle\boldsymbol{s}_{i}\right\rangle=\frac{\xi}{\xi+\zeta}$.
Infer $\pi_{i}(i=1, \cdots, N)$ :
By combining the related terms (7) and (10), we have
$p(\boldsymbol{y}, \Theta) \propto \operatorname{Bernoulli}\left(\boldsymbol{s}_{i} \mid \pi_{i}\right) \operatorname{Beta}\left(\pi_{i} \mid p, q\right)$

$$
\propto \pi_{i}^{s_{i}}\left(1-\pi_{i}\right)^{1-s_{i}} \pi_{i}^{p-1}\left(1-\pi_{i}\right)^{q-1},
$$

$\ln p(\boldsymbol{y}, \Theta) \propto\left(\boldsymbol{s}_{i}+p-1\right) \ln \left(\pi_{i}\right)+\left(q-\boldsymbol{s}_{i}\right) \ln \left(1-\pi_{i}\right)$,
$E_{\Theta \backslash \pi_{i}}[\ln p(\boldsymbol{y}, \Theta)] \propto\left(\left\langle\boldsymbol{s}_{i}\right\rangle+p-1\right) \ln \left(\pi_{i}\right)+\left(q-\left\langle\boldsymbol{s}_{i}\right\rangle\right) \ln \left(1-\pi_{i}\right)$,
$q\left(\pi_{i}\right) \propto \pi_{i}^{\left\langle\boldsymbol{s}_{i}\right\rangle+p-1}\left(1-\pi_{i}\right)^{q-\left\langle\boldsymbol{s}_{i}\right\rangle}$.
Therefore,
$q\left(\pi_{i}\right)=\operatorname{Beta}\left(\pi_{i} \mid\left\langle\mathbf{s}_{i}\right\rangle+p, q-\left\langle\mathbf{s}_{i}\right\rangle+1\right)$.
Furthermore,
$q\left(\pi_{i}\right)=\left\{\begin{array}{l}\operatorname{Beta}\left(\pi_{i} \mid p_{h}+\left\langle\boldsymbol{s}_{i}\right\rangle, q_{h}+1-\left\langle\boldsymbol{s}_{i}\right\rangle\right) \text { if } \boldsymbol{s}_{i-1}=1 \text { and } \boldsymbol{s}_{i+1}=1, \\ \operatorname{Beta}\left(\pi_{i} \mid p_{l}+\left\langle\boldsymbol{s}_{i}\right\rangle, q_{l}+1-\left\langle\boldsymbol{s}_{i}\right\rangle\right) \text { if } \boldsymbol{s}_{i-1}=0 \text { and } \boldsymbol{s}_{i+1}=0, \\ \operatorname{Beta}\left(\pi_{i} \mid p_{u}+\left\langle\boldsymbol{s}_{i}\right\rangle, q_{u}+1-\left\langle\boldsymbol{s}_{i}\right\rangle\right) \text { if } \boldsymbol{s}_{i-1}=0, \boldsymbol{s}_{i+1}=1 \text { or } \boldsymbol{s}_{i-1}=1, \boldsymbol{s}_{i+1}=0 .\end{array}\right.$
If $p(x)=\operatorname{Beta}(x \mid \alpha, \beta)$, then $\langle\ln x\rangle=\psi(\alpha)-\psi(\alpha+\beta)$ and $\langle\ln (1-x)\rangle=\psi(\beta)-\psi(\alpha+\beta)$, where $\psi(x)=\frac{d \ln \Gamma(x)}{d x}$ is a digamma function. It then follows by
$\left\langle\ln \left(\pi_{i}\right)\right\rangle=\left\{\begin{array}{l}\psi\left(p_{h}+\left\langle\boldsymbol{s}_{i}\right\rangle\right)-\psi\left(p_{h}+q_{h}+1\right) \text { if } \boldsymbol{s}_{i-1}=1 \text { and } \boldsymbol{s}_{i+1}=1, \\ \psi\left(p_{l}+\left\langle\boldsymbol{s}_{i}\right\rangle\right)-\psi\left(p_{l}+q_{l}+1\right) \text { if } \boldsymbol{s}_{i-1}=0 \text { and } \boldsymbol{s}_{i+1}=0, \\ \psi\left(p_{u}+\left\langle\boldsymbol{s}_{i}\right\rangle\right)-\psi\left(p_{u}+q_{u}+1\right) \text { if } \boldsymbol{s}_{i-1}=0, \boldsymbol{s}_{i+1}=1 \text { or } \boldsymbol{s}_{i-1}=1, \boldsymbol{s}_{i+1}=0,\end{array}\right.$
$\left\langle\ln \left(1-\pi_{i}\right)\right\rangle=\left\{\begin{array}{l}\psi\left(q_{h}+1-\left\langle\boldsymbol{s}_{i}\right\rangle\right)-\psi\left(p_{h}+q_{h}+1\right) \text { if } \boldsymbol{s}_{i-1}=1 \text { and } \boldsymbol{s}_{i+1}=1, \\ \psi\left(q_{l}+1-\left\langle\boldsymbol{s}_{i}\right\rangle\right)-\psi\left(p_{l}+q_{l}+1\right) \text { if } \boldsymbol{s}_{i-1}=0 \text { and } \boldsymbol{s}_{i+1}=0, \\ \psi\left(q_{u}+1-\left\langle\boldsymbol{s}_{i}\right\rangle\right)-\psi\left(p_{u}+q_{u}+1\right) \text { if } \boldsymbol{s}_{i-1}=0, \boldsymbol{s}_{i+1}=1 \text { or } \boldsymbol{s}_{i-1}=1, \boldsymbol{s}_{i+1}=0 .\end{array}\right.$
Estimation of noise precision
Infer $\beta$ :
By combining the related terms (12) and (13), we have
$p(\boldsymbol{y}, \Theta) \propto \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}), \beta^{-1} \boldsymbol{I}_{M L}\right) \operatorname{Gamma}(\beta \mid c, d)$

$$
\begin{aligned}
& \propto \beta^{\frac{M L}{2}} \exp \left[-\frac{\beta}{2}(\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}))^{T}(\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}))\right] \beta^{c-1} \exp (-d \beta) \\
& \propto \beta^{\frac{M L}{2}+c-1} \exp \left[-\left(\frac{1}{2}(\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}))^{T}(\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}))+d\right) \beta\right]
\end{aligned}
$$

$\ln p(\boldsymbol{y}, \Theta) \propto\left(\frac{M L}{2}+c-1\right) \ln (\beta)-\left[\frac{1}{2}(\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}))^{T}(\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}))+d\right] \beta$,
$E_{\Theta \backslash \beta}[\ln p(\boldsymbol{y}, \Theta)] \propto\left(\frac{M L}{2}+c-1\right) \ln (\beta)-\left[\frac{1}{2}\left\langle\|(\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}))\|^{2}\right\rangle+d\right] \beta$,
$q(\beta) \propto \beta^{\frac{M L}{2}+c-1} \exp \left(-\left[\frac{1}{2}\left\langle\|(\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}))\|^{2}\right\rangle+d\right] \beta\right)$.
Therefore,
$q(\beta)=\operatorname{Gamma}\left(\beta \left\lvert\, \frac{M L}{2}+c\right.,\left(\frac{\left\langle\|(\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}))\|^{2}\right\rangle}{2}+d\right)\right.$.
Furthermore,
$\langle\beta\rangle=\frac{M L+2 c}{\left\langle\|(\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z}))\|^{2}\right\rangle+2 d}$,
where
$\left\langle\|\boldsymbol{y}-\boldsymbol{D}(\boldsymbol{w} \circ \boldsymbol{z})\|_{2}^{2}\right\rangle$
$=\boldsymbol{y}^{T} \boldsymbol{y}-2 \boldsymbol{y}^{T} \boldsymbol{D}(\langle\boldsymbol{w}\rangle \circ(\boldsymbol{\Omega}\langle\boldsymbol{s}\rangle))+\operatorname{Tr}\left(\left[\left(\Sigma_{\boldsymbol{w}}+\langle\boldsymbol{w}\rangle\langle\boldsymbol{w}\rangle^{T}\right)\right.\right.$
$\left.\left.\circ\left(\Omega\left(\mathcal{Z}+\langle\boldsymbol{s}\rangle\langle\boldsymbol{s}\rangle^{T}\right) \boldsymbol{\Omega}^{T}\right)\right] \boldsymbol{D}^{T} \boldsymbol{D}\right)$,
$\mathcal{Z}=\operatorname{diag}(\langle\boldsymbol{s}\rangle \circ(1-\langle\boldsymbol{s}\rangle))$ and $\Sigma_{\boldsymbol{w}}=\operatorname{diag}\left(\Sigma_{\boldsymbol{w}_{1}}, \ldots, \Sigma_{\boldsymbol{w}_{N}}\right)$.

## References

[1] E.J. Candes, J.K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Commun. Pure Appl. Math. 59 (8) (2006) 1207-1223.
[2] D.L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory 52 (4) (2006) 1289-1306.
[3] K. Kose, O. Gunay, A.E. Cetin, Compressive sensing using the modified entropy functional, Digit. Signal Process. 24 (2014) 63-70.
[4] J. Hahn, C. Debes, M. Leigsnering, A.M. Zoubir, Compressive sensing and adaptive direct sampling in hyperspectral imaging, Digit. Signal Process. 26 (2014) 113-126.
[5] J. Haupt, W.U. Bajwa, M. Rabbat, R. Nowak, Compressed sensing for networked data, IEEE Signal Process. Mag. 25 (2) (2008) 92-101.
[6] F. Parvaresh, H. Vikalo, S. Misra, B. Hassibi, Recovering sparse signals using sparse measurement matrices in compressed DNA microarrays, IEEE J. Sel. Top. Signal Process. 2 (3) (2008) 275-285.
[7] Z. Sadeghigol, M.H. Kahaei, F. Haddadi, Generalized beta Bayesian compressive sensing model for signal reconstruction, Digit. Signal Process. 60 (2017) 163-171.
[8] S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM Rev. 43 (1) (2001) 129-159.
[9] S.G. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Trans. Signal Process. 41 (12) (1993) 3397-3415.
[10] J.A. Tropp, A.C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, IEEE Trans. Inf. Theory 53 (12) (2007) 4655-4666.
[11] S.F. Cotter, B.D. Rao, K. Engan, K. Kreutz-Delgado, Sparse solutions to linear inverse problems with multiple measurement vectors, IEEE Trans. Signal Process. 53 (7) (2005) 2477-2488.
[12] Y.C. Eldar, M. Mishali, Robust recovery of signals from a structured union of subspaces, IEEE Trans. Inf. Theory 55 (11) (2009) 5302-5316
[13] Y.C. Eldar, H. Rauhut, Average case analysis of multichannel sparse recovery using convex relaxation, IEEE Trans. Inf. Theory 56 (1) (2010) 505-519.
[14] Y. Jin, B.D. Rao, Insights into the stable recovery of sparse solutions in overcomplete representations using network information theory, in: Acoustics, Speech and Signal Processing, 2008, ICASSP 2008, IEEE International Conference on, IEEE, 2008, pp. 3921-3924.
[15] J. Wang, Y. Huang, J. Zhou, L. Sun, On imperfect pricing in globally constrained noncooperative games for cognitive radio networks, Signal Process. 117 (2015) 96-101.
[16] N. Vaswani, Kalman filtered compressed sensing, in: Image Processing, 2008, ICIP 2008, 15th IEEE International Conference on, IEEE, 2008, pp. 893-896.
[17] T. Zhang, B. Ghanem, S. Liu, N. Ahuja, Robust visual tracking via multi-task sparse learning, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2042-2049.
[18] M. Hämäläinen, R. Hari, R.J. Ilmoniemi, J. Knuutila, O.V. Lounasmaa, Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain, Rev. Mod. Phys. 65 (2) (1993) 413.
[19] J. Chen, X. Huo, Theoretical results on sparse representations of multiplemeasurement vectors, IEEE Trans. Signal Process. 54 (12) (2006) 4634-4643.
[20] J.A. Tropp, A.C. Gilbert, M.J. Strauss, Algorithms for simultaneous sparse approximation. Part I: greedy pursuit, Signal Process. 86 (3) (2006) 572-588.
[21] X. Du, L. Cheng, G. Cheng, A heuristic search algorithm for the multiple measurement vectors problem, Signal Process. 100 (2014) 1-8.
[22] F.R. Bach, Consistency of the group lasso and multiple kernel learning, J. Mach. Learn. Res. 9 (2008) 1179-1225.
[23] D.P. Wipf, B.D. Rao, An empirical Bayesian strategy for solving the simultaneous sparse approximation problem, IEEE Trans. Signal Process. 55 (7) (2007) 3704-3716.
[24] S. Ji, D. Dunson, L. Carin, Multitask compressive sensing, IEEE Trans. Signal Process. 57 (1) (2009) 92-106.
[25] Z. Zhang, B.D. Rao, Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning, IEEE J. Sel. Top. Signal Process. 5 (5) (2011) 912-926.
[26] M. Luessi, S. Derin Babacan, R. Molina, A.K. Katsaggelos, Bayesian simultaneous sparse approximation with smooth signals, IEEE Trans. Signal Process. 61 (22) (2013) 5716-5729.
[27] J. Ziniel, P. Schniter, Efficient high-dimensional inference in the multiple measurement vector problem, IEEE Trans. Signal Process. 61 (2) (2013) 340-354.
[28] Z. Chen, R. Molina, A.K. Katsaggelos, Robust recovery of temporally smooth signals from under-determined multiple measurements, IEEE Trans. Signal Process. 63 (7) (2015) 1779-1791.
[29] G. Zhang, N. Kingsbury, Variational Bayesian image restoration with groupsparse modeling of wavelet coefficients, Digit. Signal Process. 47 (2015) 157-168.
[30] R.G. Baraniuk, V. Cevher, M.F. Duarte, C. Hegde, Model-based compressive sensing, IEEE Trans. Inf. Theory 56 (4) (2010) 1982-2001.
[31] J. Chen, X. Zhang, H. Meng, Self-adaptive sampling rate assignment and image reconstruction via combination of structured sparsity and non-local total variation priors, Digit. Signal Process. 29 (2014) 54-66.
[32] S.D. Babacan, S. Nakajima, M.N. Do, Bayesian group-sparse modeling and variational inference, IEEE Trans. Signal Process. 62 (11) (2014) 2906-2921.
[33] K.E. Themelis, A.A. Rontogiannis, K.D. Koutroumbas, Variational Bayes group sparse time-adaptive parameter estimation with either known or unknown sparsity pattern, IEEE Trans. Signal Process. 64 (12) (2016) 3194-3206.
[34] J.W. Choi, B. Shim, Statistical recovery of simultaneously sparse time-varying signals from multiple measurement vectors, IEEE Trans. Signal Process. 63 (22) (2015) 6136-6148.
[35] J. Paisley, L. Carin, Nonparametric factor analysis with beta process priors, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 777-784.
[36] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[37] M. Figueiredo, Adaptive sparseness using Jeffreys prior, in: Advances in Neural Information Processing Systems, 2002, pp. 697-704.
[38] R.M. Neal, Bayesian Learning for Neural Networks, vol. 118, Springer Science \& Business Media, 2012.
[39] Z. Zhang, B.D. Rao, Iterative reweighted algorithms for sparse signal recovery with temporally correlated source vectors, in: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011, pp. 3932-3935.
[40] Z. Zhang, B.D. Rao, Exploiting correlation in sparse signal recovery problems: Multiple measurement vectors, block sparsity, and time-varying sparsity, arXiv preprint arXiv:1105.0725.
[41] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[42] B.C.-T. Lin, L.-W. Ko, J.-C. Chiou, J.-R. Duann, R.-S. Huang, S.-F. Liang, T.-W. Chiu, T.-P. Jung, Noninvasive neural prostheses using mobile and wireless EEG, Proc. IEEE 96 (7) (2008) 1167-1183.

Ningning Han received the Ph.D. degree in probability theory and mathematical statistics from the School of Mathematics, Tianjin University, Tianjin, China. His work focused on the research of recovering the sparse or low-rank signal. Most recently, he has concentrated his efforts on direction of arrival (DOA) estimations in radar signal processing application platforms, extremely parallel with the classical harmonic retrieval problems.

Zhanjie Song received the Ph.D. degree in probability theory and mathematical statistics from the School of Mathematical Science, Nankai University, Tianjin, China. He is currently a Full Professor with the School of mathematics, and a Vice-Director with the Institute of TV and Image Information, all at Tianjin University (TJU), Tianjin, China. He was a PostDoctoral Fellow in signal and information processing at TJU. His current research interests include approximation of deterministic signals, reconstruction of random signals, compressed sampling of multidimensional signals, and statistical analysis of random processes.


[^0]:    * Corresponding author at: School of Mathematics, Tianjin University, Tianjin 300350, China.

    E-mail addresses: ning_ninghan@tju.edu.cn (N. Han), zhanjiesong@tju.edu.cn (Z. Song).

[^1]:    ${ }^{1}$ homepage.univie.ac.at/monika.doerfler/StrucAudio.html.

[^2]:    ${ }^{2}$ http://asi.insa-rouen.fr/enseignants/arakotom/code/SSAindex.html.

