Structured Bayesian Compressive Sensing Exploiting Dirichlet Process Priors

Qisong Wu[1,2], *Member, IEEE*, Yin Fu[1], *Student Member, IEEE*, Yimin D. Zhang[3], *Fellow, IEEE*
Moeness G. Amin[4], *Fellow, EURASIP and Life Fellow, IEEE*

[1] Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing, 210096, China

[2] Purple Mountain Laboratories, Nanjing, 211111, China

[3] Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA, 19122, USA

[4] Center for Advanced Communications, Villanova University, Villanova, PA 19085, USA

Abstract

Conventional multi-task Bayesian compressive sensing methods, which compute the sparse representations of signals with a group sparse pattern, generally ignore the inner sparse structures of signals and/or their statistical correlations. These structures are naturally exhibited among clustered tasks with different sparse patterns. In this paper, a novel structured and clustered multi-task compressive sensing framework based on a hierarchical Bayesian model is proposed to exploit the inner sparse pattern and the statistical dependence between tasks. This is achieved by adopting a signal model that exploits the spike-and-slab priors and the Dirichlet Process priors. The former encode sparse patterns of the signals and are further generalized by imposing the Gaussian process for modeling inner structures. The Dirichlet Process priors, on the other hand, imposed on the support reveal the clustering mechanisms among tasks. In so doing, these priors provide a new means to simultaneously infer the clusters and perform compressive sensing inversion, yielding enhanced sparse reconstruction performance. A new inference scheme based on expectation propagation is derived to approximate the posterior distribution for simplifying the computation and deriving analytical expression. Experimental results verify the performance superiorities of the proposed algorithm over existing state-of-the-art methods.

Index Terms

Compressive sensing, structured priors, Dirichlet process, cluster, expectation propagation

The work of Q. Wu and Y. Fu was supported in part by the National Natural Science Foundation under Grants No. 61701109 and 91938203, and by the National Natural Science Foundation of Jiangsu Province under Grant No. BK20160701.

Q. Wu and Y. Fu are with Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing, 210096, China, and Q. Wu is also with Purple Mountain Laboratories, Nanjing, 211111, China. (Email: qisong.wu@seu.edu.cn)

Y. D. Zhang is Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA, 19122, USA (Email: ydzhang@temple.edu)

M. G. Amin is with the Center for Advanced Communications, Villanova University, Villanova, PA 19085, USA.

I. INTRODUCTION

Compressive sensing (CS) has become a powerful technique for high precision sparse signal recovery using a small number of measurements [1]. CS and sparse signal reconstruction have been widely used in many applications, such as radar imaging [2]–[4], direction-of-arrival (DOA) estimation [5]–[7], radio astronomy [8]–[10], and time-frequency analysis [11]–[13].

A typical single-task CS model addresses the problem of finding the sparse solution of $\boldsymbol{x} \in \mathbb{R}^{K}$ in the following linear inversion problem

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^{P}$ denotes the measurement vector, $\boldsymbol{D} \in \mathbb{R}^{P \times K}$ is a known sensing matrix, and $\boldsymbol{\varepsilon}$ is an unknown additive zero-mean Gaussian noise vector. We are mainly interested in the sparse recovery problem that deals with the ill-posed regime with $P \ll K$. The sparse recovery problem can be formulated in an l_0-regularized form

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \|\boldsymbol{x}\|_0 \\ \text{s.t.} \quad & \|\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y}\|_2^2 \leq \sigma, \end{aligned} \tag{2}$$

where $\|\cdot\|_0$ denotes the canonical l_0 sparsity metric, i.e., the number of nonzero elements in a vector, and σ is a scalar to be determined by the input signal-to-noise ratio (SNR). Inference is in general intractable for this NP-hard problem. A feasible way is using l_1 norm in lieu of the l_0 norm. The l_1 regularization minimizes the residual sum of squares subject to an l_1 penalty on the solution expressed as

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 + \alpha \|\boldsymbol{x}\|_1 \right\}, \tag{3}$$

where α is an elastic scalar which is used to balance the least square term and the l_1-norm term in Eq. (3). This classic framework is also referred to as LASSO and has been the driver for several CS inversion algorithms, including linear programming [14] and greedy (constructive) algorithm [15], [16].

A Bayesian compressive sensing (BCS) methodology is proposed in [17] by imposing an independent and identically distributed (i.i.d.) Laplace prior distribution on the desired solution $\boldsymbol{x}$ which is proved to be equivalent to the sparsity regularization term in Eq. (3). However, unlike the regularization technique that aims to find a point estimate of the underlying solution, the Bayesian formulation strives to obtain the full posterior distribution of the desired solution $\boldsymbol{x}$. The latter provides probabilistic prediction solutions and leads to determination of model complexity using the observed data alone. There have been numerous algorithms developed to perform posterior inference of the BCS methods, such as the variational Bayesian (VB) analysis which performs inference of the posterior distribution [18]. Reference [19] proposed an inference scheme based on Markov chain Monte Carlo (MCMC) to approximate the posterior for clustered structured sparse signals.

The above techniques, however, typically perform separate inversions independent of other tasks, thus ignoring the statistical structures that naturally exist in real-world signals. Such dependence is especially exhibited when measurements are acquired from the same physical phenomena. Examples include magnetic resonance imaging (MRI) when repeated images are taken from the same diagnostic object [20] and face recognition where a person's

face is captured from different look directions and illumination conditions [21]. Multi-task CS provides a framework that utilizes the statistical structures present in the different measurements with the aim to achieve significant reduction in the number of measurements required for sparse reconstruction [20].

In this paper, we consider multi-task CS problems that generalizes Eq. (1) as

$$\boldsymbol{y}_i = \mathbf{D}_i \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i, \quad i \in \{1, \cdots, M\}, \tag{4}$$

where $\boldsymbol{y}_i \in \mathbb{R}^P$ denotes the measurement vector in the ith task, $\boldsymbol{x}_i \in \mathbb{R}^K$ is the corresponding sparse solution, $\boldsymbol{\varepsilon}_i$ is an unknown zero-mean Gaussian noise vector, and $\mathbf{D}_i \in \mathbb{R}^{P \times K}$ is the ith sensing dictionary matrix with $P \ll K$. In Eq. (4), $\boldsymbol{x}_i, i = 1, ..., M$ can possess different forms of joint sparsity or statistical sparsity structures that enable recover $\{\boldsymbol{x}_i\}_{i=1}^M$ of the sparse signals together with fewer observations compared to solving each of the M tasks in Eq. (1) separately. Various algorithms have been developed to perform joint signal reconstructions by exploiting the statistical correlations among tasks [20], [22]. These conventional multi-task models generally assume the sparse vectors $\{\boldsymbol{x}_i\}_{i=1}^M$ to have identical or similar support, i.e., the respective positions of the non-zero entries are identical or similar across tasks. However, tasks are typically grouped into several clusters with different statistical correlations within each cluster [23]. As such, jointly reconstructing the signals from different clusters would lead to severe performance degradations. On the other hand, structured sparsity is a generalization of simple sparsity and seeks to exploit the fact that the sparsity pattern of each signal contains a richer structure than a simple pattern, e.g., the block sparsity [24] and the tree structure [25].

Towards enhanced signal reconstruction performance, we seek to combine the above two approaches to specifically deal with tasks that exhibit both structured and clustered patterns. In particular, the sparsity structures of the tasks are augmented with latent multivariate variables and the clustered mechanism is implemented with nonparametric techniques. Application of such approach includes image denoising [26], DOA estimation [7], and electroencephalogram (EEG) source localization [27].

A. Related Work

A large body of research has been dedicated to enhance the sparse signal reconstruction performance by exploiting the underlying statistical relationships within and between signals. Reference [28] generalizes the model for the selection of group variables to propose group LASSO by selecting or dropping an entire group of predictors depending on the trade-off parameter. From a probabilistic perspective, the hierarchical Bayesian framework provides effective representations to model both the individuality and the statistical dependence of different signals. The multi-task compressive sensing (MT-CS) algorithm [20] exploits a hierarchical model with a shared gamma distribution prior to characterize the statistic information of different tasks. It incorporates an empirical Bayesian procedure for fast point estimation of hyper-parameters and full posterior density function inference for each task. This approach is further generalized in [22] to recover complex signals, defining the complex multi-task Bayesian compressive sensing (CMT-BCS) approach. The structured spike-and-slab prior [29] imposes a spatial Gaussian distribution with a standard normal cumulative distribution function (CDF) on the support vector to encode the structure of a sparse pattern using generic covariance functions. It generalizes the model to multi-measurement vector (MMV) problems

by imposing a transformed Gaussian process on the spike-and-slab probabilities to incorporate both spatial and temporal structure information of different tasks [30]. A MT-CS algorithm proposed in [31] exploits the intra-group correlation and the continuous structure using two Toeplitz matrices.

Although the above strategies have successfully utilized the structures within or between signals to improve the reconstruction performance, they stop short in accounting for the prior knowledge associated with the different group structures. To avoid the complex hierarchical model and guarantee the sparsity of the results, a 'semi-Bayesian' strategy based on a simple empirical prior is proposed in [32], where the central mechanisms of clustered sparse estimation are revealed using rigorous properties of the cost function. However, this approach requires that the cluster structure, i.e., the upper bound on the number of clusters, L, to be set appropriately based on prior knowledge. In the parametric modeling, nevertheless, the true number of clusters is difficult to estimate in advance.

Fortunately, this problem can be tackled from a nonparametric Bayesian approach, which utilizes a model with an unbounded complexity, i.e., $L \to \infty$. A nonparametric model in the context of the Dirichlet process (DP) can be employed to automatically infer the actual number of clusters that best describes the data. A DP is a probability distribution whose realizations are a set of probability distributions. Samples drawn from a DP are usually discrete and have clustering property. A widely employed metaphor for the DP is based on the so-called Chinese restaurant process [33]. At each step of generating data points, the DP can either assign a data point to a previously-generated cluster or can start a new cluster. More importantly, the number of clusters grows automatically as new data points arrive. Unlike a finite parametric model, the number of clusters in the DP can be automatically inferred from the data set. A number of solutions have been developed to modeling the clustering mechanism with DP. Reference [34] uses a hierarchical DP to handle a clustering problem involving multiple groups of data, where each group of data is modeled with a mixture of components and an inference procedure based on Gibbs sampling is adopted. In a MT-CS framework, DP priors are employed on the variance of Gaussian priors to reveal the sharing mechanisms as well as perform CS inversion simultaneously [23]. A mean-field variational approximation procedure is then adopted to infer the clusters of signals and perform the sparse inversion. Although these MT-CS methods have set foot in the cluster learning, the hierarchical model is either too complex or the underlying mechanics are not clearly analyzed.

The choice of sparsity-promoting prior plays a crucial role in the BCS methods. These methods with appropriate priors would offer improved reconstruction results with noise robustness. A non-exhaustive list of sparsity-promoting priors includes the Laplace prior [17], the automatic relevance determination prior [35], and the spike-and-slab prior [36]. Particularly, the spike-and-slab prior (also called the Bernoulli-Gaussian prior) has recently become increasingly popular. It takes the following form

$$x_i \sim (1 - z_i)\delta(x_i) + z_i\mathcal{N}(x_i|0, \tau_0), \tag{5}$$

where x_i is the ith element of $\boldsymbol{x}$, $\delta(\cdot)$ denotes a Dirac delta distribution concentrated at zero (spike), $\mathcal{N}(\cdot)$ denotes a Gaussian distribution (slab), and τ_0 is the variance scalar. z_i is referred to as the support of x_i, which is a binary variable determining the sparse level or the sparse pattern of $\boldsymbol{x}$. When $z_i \neq 0$, the corresponding x_i is active, i.e.,

$x_i \neq 0$. On the other hand, when $z_i = 0$, the corresponding element x_i is inactive, i.e., $x_i = 0$. This prior is the starting point of our work.

Our work is closely related to the work [7] and [23]. Reference [23] proposed a method which employs the DP prior over the latent parametric space to model the clustering characteristic. Reference [7] applies this model to DOA estimation and extends to the off-grid problem in integrated and separated manners. However, the structured information within each task was not considered therein. In our work, the *a priori* knowledge of the structure is injected into the model using generic covariance functions rather than independent probability distributions, and an expectation propagation framework with structured spike-and-slab priors is proposed. The comparisons between the proposed approach and the above existing methods are provided in order to highlight the offering of the former and show its performance superiority.

B. Contributions

The main novelty of this paper lies in the exploitation of both the clustering mechanisms among tasks with various sparsity patterns and the structured patterns of inner signals. We propose a novel MT-CS technique based on the structured Gaussian process and DP priors. This technique enhances sparse signal constructions by automatically learning and inferring the structure and clustering of the tasks in a hierarchical Bayesian framework. The spike-and-slab priors are first generalized to encode the sparse pattern of each task and induce the relationships among the tasks. A Gaussian process is then introduced to facilitate the sparsity and smooth structure properties of the tasks. Motivated by the nonparametric clustering technique in the mixture learning model [37], the DP priors are then introduced to learn the clustering mechanisms among tasks. A stick-breaking construction is used to describe the DP. A novel inference algorithm based on expectation propagation (EP) is used to perform the approximate posterior inference induced by the extended spike-and-slab priors and DP priors. Furthermore, the Woodbury identity is employed to significantly accelerate the inversion of the involved high-dimensional matrices. Considering that the hierarchical Bayesian model allows the estimation of the posterior and the clustering parameters in an unsupervised manner, the proposed algorithm is capable of automatically inferring the sparsity pattern and learning the clustering structure across tasks without requiring the knowledge of the sparsity and the number of clusters in advance.

C. Structure of the Paper

The remainder of the paper is organized as follows. In Section II, the generalized spike-and-slab priors incorporating the Gaussian process are described. An approach to impose the DP to encode the clustering structure is then illustrated. After introducing the generative model, an algorithm based on the EP framework is proposed. The basics of EP are reviewed and the proposed algorithm, termed EP based Structured BCS (EP-SBCS), is described in Section III. Section IV demonstrates simulation and experimental results.

D. Notations

We use lower-case (upper-case) bold characters to denote vectors (matrices). $f(w|a,b)$ is the conditional probability distribution function (pdf) of variable w depending on a and b. $\mathcal{N}(w|a,b)$ denotes that random variable w

follows a Gaussian distribution with mean a and variance b. Bern$(z|\pi)$ denotes that variable z follows a Bernoulli distribution with probability of $p(z = 1) = \pi$. Beta$(x|a, b)$ means that variable x follows a beta distribution parameterized by a and b. Multi $(x|\boldsymbol{\beta})$ denotes a multinomial distribution with the probability vector of $\boldsymbol{\beta}$. $\delta(\cdot)$ is the Dirac delta function, and $(\cdot)^T$ denotes the transpose of a matrix or vector. $\boldsymbol{I}_k$ denotes the $K \times K$ identity matrix. $|\cdot|$ denotes the cardinality of a set. $Q^{\backslash i} = Q/q_i$ denotes the distribution function Q except the function q_i, and $^{\backslash(i)}$ represents the removal of the ith function from the joint distribution function.

II. The Proposed Model

A. Generative Model

This subsection describes the proposed generative model. In general, the measurement vectors follow a Gaussian distribution with the following likelihood function

$$\boldsymbol{Y}|\boldsymbol{X}, \sigma_0 \sim \prod_{i=1}^{M} \mathcal{N}(\boldsymbol{D}_i\boldsymbol{x}_i, \sigma_0^2\boldsymbol{I}_K), \tag{6}$$

where $\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_i, \cdots, \boldsymbol{y}_M] \in \mathbb{R}^{P\times M}$ is a measurement data matrix consisting of M tasks collected by each measurement vector described in Eq. (4), $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_M] \in \mathbb{R}^{K\times M}$ denotes the M-task sparse matrix to be reconstructed, and σ_0^2 represents the noise variance.

To encourage sparsity, spike-and-slab priors are imposed on each task and take the following form

$$\boldsymbol{x}_i|\boldsymbol{z}_i \sim \prod_{j=1}^{K} [(1 - z_{ji})\delta(x_{ji}) + z_{ji}\mathcal{N}(x_{ji}|0, \tau_0)], \tag{7}$$

where x_{ji} is the jth element of signal vector $\boldsymbol{x}_i$. As stated in Eq. (5), $\boldsymbol{z}_i$ is the support or sparse pattern of task $\boldsymbol{x}_i$ which follows Bernoulli distribution and z_{ji} is the jth element of $\boldsymbol{z}_i$. It is important to note that the factorization form of Eq. (7) implies that the variables x_{ji} and x_{mi} in the ith task are assumed to be independent for $m \neq j$. That is, the number of active variables follows a binomial distribution and hence the marginal probability for x_{ji} and x_{mi} to be jointly active is given by the product $p(x_{ji} \neq 0, x_{mi} \neq 0) = p(x_{ji} \neq 0)p(x_{mi} \neq 0)$. In practice, however, the signals may exhibit correlated statistical relationships or structures within each task. Herein, we extend the conventional spike-and-slab priors to model such structure information by introducing the Gaussian process.

The Bernoulli distribution is first imposed on the sparse pattern $\boldsymbol{z}_i$,

$$\boldsymbol{z}_i|\boldsymbol{\pi}_i \sim \prod_{j=1}^{K} \text{Bern}(z_{ji}|\phi(\pi_{ji})), \tag{8}$$

where π_{ji} is the probability weight parameter with $p(z_{ji} = 1) = \phi(\pi_{ji})$, $\phi(\cdot) : \mathbb{R} \to (0, 1)$ is the normal CDF, which squeezes π_{ji} into the unit interval and thereby $\phi(\pi_{ji})$ represents the probability of $z_{ji} = 1$. The Gaussian distribution is then placed on the parameter vector $\boldsymbol{\pi}_i = [\pi_{1i}, \cdots, \pi_{Ki}]^T$ and is expressed as

$$\boldsymbol{\pi}_i \sim G_0 = \mathcal{N}(\boldsymbol{\pi}_i|\boldsymbol{a}_0, \boldsymbol{\Sigma}_0). \tag{9}$$

The marginal prior distribution of $\boldsymbol{z}_i$ can be computed after integrating out $\boldsymbol{\pi}_i$ using the following formula

$$\begin{aligned} p(z_{ji}=1) &= \int p(z_{ji}|\pi_{ji})p(\pi_{ji})d\pi_{ji} \\ &= \int \text{Bern}(z_{ji}|\phi(\pi_{ji}))\mathcal{N}(\pi_{ji}|a_{j0},\Sigma_{0,jj})d\pi_{ji} \\ &= \phi\left(\frac{a_{j0}}{\sqrt{1+\Sigma_{0,jj}}}\right), \end{aligned} \tag{10}$$

where a_{j0} is the jth element of $\boldsymbol{a}_{j0}$, $\Sigma_{0,jj}$ is the jth diagonal element of $\boldsymbol{\Sigma}_0$. The latent variable $\boldsymbol{z}_i$ controls the structure of the sparsity pattern of each task. From Eq. (10), it can be seen that, when $a_{j0}=0$, the prior belief of x_{ji} being active is unbiased since $p(z_{ji}=1)=0.5$, whereas x_{ji} is biased towards being inactive when $a_{j0}<0$ and vice versa. If it is known a priori that x_{ji} is more likely to be active than x_{mi}, then we can encode this information by assigning the prior mean of $\boldsymbol{\pi}$ such that $\pi_{ji}>\pi_{mi}$. The prior probability of two variables being joint active is given by

$$\begin{aligned} p(z_{ji}=1,z_{mi}=1) &= \int p(z_{ji}|\pi_{ji})p(z_{mi}|\pi_{mi})p(\boldsymbol{\pi}_i)d\boldsymbol{\pi}_i \\ &= \int \phi(\pi_{ji})\phi(\pi_{mi})\mathcal{N}(\boldsymbol{\pi}_i|\boldsymbol{a}_0,\boldsymbol{\Sigma}_0)d\boldsymbol{\pi}_i. \end{aligned} \tag{11}$$

From Eq. (11), the marginal probability of x_{ji} and x_{mi} being jointly active are controlled by the covariance matrix $\boldsymbol{\Sigma}_0$ rather than being independent as in the conventional spike-and-slab priors. In practice, the *a priori* knowledge of sparsity correlation within the signal can be encoded by choosing different forms of generic kernel functions for $\boldsymbol{\Sigma}_0$, such as the squared exponential kernel [29] and the nearest neighbors-type kernels [30]. Using this model, the expected degree of sparsity of each task is modeled by $\boldsymbol{a}_0$ and $\boldsymbol{\Sigma}_0$ which control the prior correlation of the support. In conventional MT-CS, all binary variable vectors are i.i.d. drawn from the Bernoulli distribution with identical weight parameter vector $\boldsymbol{\pi}$, i.e., $\boldsymbol{\pi}_i=\boldsymbol{\pi}$ for $i=1,...,M$. It encourages consistent sparse patterns across tasks with the shared $\boldsymbol{\pi}$. The effectiveness of this hierarchical model is proved in [38] where the underlying signal $\boldsymbol{X}$ is extracted from nature images.

The conventional MT-CS assumes that all M tasks are clustered in a single class, i.e., all tasks hold the same probability to exhibit the same sparse pattern. However, in practice, these tasks may often be clustered into different sets of tasks and, as a result, data sharing is appropriate only within each cluster, rather than a single mechanism is shared across all tasks [23]. In this case, the most challenging issue is to determine the appropriate number of clusters that best describes the underlying signals. For this purpose, the DP provides a popular and effective nonparametric probabilistic structure that forms clusters by assuming an infinite number of components and automatically learn the actual number of clusters. In this paper, we exploit the fact that tasks belonging to the same cluster share identical sparse pattern, and a hierarchical probability framework with DP priors is used to carry out task clustering. By employing DP on sparse pattern parameters $\boldsymbol{\pi}_i$ described in the next subsection, the proposed model encourages sharing of information within each cluster, yielding an effective means to simultaneously cluster tasks and reconstruct the sparse signals.

B. Dirichlet Process for Multi-Task CS Formulation

A DP denoted as DP $(\lambda,\ G_0)$ is parameterized by a positive scaling parameter λ and a base non-atomic probability distribution G_0 [39]. An explicit and intuitive way of constructively forming a DP, called stick-breaking, is provided by Sethuraman [40], which is formulated as,

$$v_l \sim \text{Beta}\,(1,\lambda)\,, \tag{12}$$

$$\boldsymbol{\pi}_l \overset{i.i.d.}{\sim} G_0, \tag{13}$$

$$\beta_l = v_l \prod_{h=1}^{l-1}(1-v_h), \tag{14}$$

$$G = \sum_{l=1}^{\infty} \beta_l \delta(\boldsymbol{\pi}_l), \tag{15}$$

where β_l is the length of the lth fraction break from a 'stick' of original one, whereas the fraction of the rest of the stick broken off on break l is v_l. It can also be viewed that the random measure G is a discrete distribution with the probability β_l being equal to $\boldsymbol{\pi}_l$ drawn from G_0. It is found that $\sum_{l=1}^{\infty}\beta_l = 1$. In order to satisfy the properties of incorporating sparsity structure mentioned above, we set $G_0 = \mathcal{N}(\boldsymbol{\pi}_i|\boldsymbol{a}_0, \boldsymbol{\Sigma}_0)$. Therefore, the binary variable vectors can be generated using the following steps:

1. Draw $v_l|\lambda \sim \text{Beta}\,(1,\lambda)\,, l = \{1,2,...\}$;
2. Draw $\boldsymbol{\pi}_l|G_0 \overset{i.i.d.}{\sim} G_0, l = \{1,2,...\}$;
3. For the ith task:
   a. Draw $c_i \sim \text{Multi}\,(\beta_1,\beta_2,...)$, where $\beta_l = v_l\prod_{h=1}^{l-1}(1-v_h)$,
   b. Draw $\boldsymbol{z}_i|c_i, \{\boldsymbol{\pi}_l\}_{l=1,2...} \sim \prod_{j=1}^{K}\text{Bern}\,(\pi_{jc_i})$,

where c_i is an indicator variable which denotes the label of ith cluster. $c_i \sim \text{Multi}\,(\beta_1,\beta_2,...)$ implies that random variable c_i follows a discrete distribution with probability of $p(c_i = l) = \beta_l$. This Dirichlet process mixture (DPM) model on the sparse pattern $\{\boldsymbol{z}_i\}_{l=1}^{\infty}$ can be viewed as the limit of a normal parametric finite Bernoulli mixture model with L components with $L \to \infty$ [41],

$$p\left(\boldsymbol{z}_i|\{v_l\}_{l=1}^{L}, \{\boldsymbol{\pi}_l\}_{l=1}^{L}\right) = \sum_{l=1}^{L}\beta_l \prod_{j=1}^{K}\text{Bern}\,(z_{ji}|\pi_{jl}). \tag{16}$$

It was shown that the number of components typically used to model M signals is independent of L and is approximately $\mathcal{O}(\lambda\log M)$ [39]. Hence, considering the computational efficiency, the above infinite DPM is usually truncated into a finite number of components with sufficiently large L proportional to the logarithm of the number of data points, which can be initialized as follows:

$$v_L = 1;\ \beta_l = 0, l > L;\ \sum_{l=1}^{L}\beta_l = 1. \tag{17}$$

The posterior distribution automatically infers and learns the number of clusters based on the data, which is determined by how many elements of β_l are of a significant value [37]. Therefore, the generative model of the

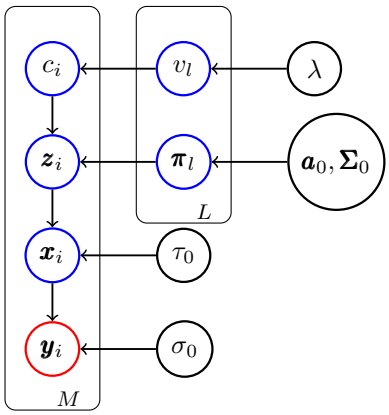


Figure 1. The probabilistic graphical model of the proposed sparse Bayesian framework. The red node represents the known observations and blue nodes represent unknown hidden variables. Black nodes are the hyper-parameters. The edges denote possible dependence, and plates denote replications.

structured MT-CS with DP priors can be summarized as following:

$$
\begin{aligned}
\boldsymbol{y}_i &= \boldsymbol{D}_i \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i, \\
\boldsymbol{x}_i | \boldsymbol{z}_i &\sim \textstyle\prod_{j=1}^{K} \left[(1 - z_{ji})\, \delta\left(x_{ji}\right) + z_{ji} \mathcal{N}\left(x_{ji} | 0, \tau_0\right) \right], \\
\boldsymbol{z}_i | c_i, \{\boldsymbol{\pi}_l\}_{l=1,2...L} &\sim \textstyle\prod_{j=1}^{K} \text{Bern}\left(\phi(\pi_{jc_i})\right), \\
c_i &\sim \text{Multi}\left(\beta_1, \beta_2, ..., \beta_L\right), \\
\boldsymbol{\pi}_i &\sim G = \textstyle\sum_{l=1}^{\infty} \beta_l \delta\left(\boldsymbol{\pi}_l\right), \\
\beta_l &= v_l \textstyle\prod_{h=1}^{l-1} \left(1 - v_h\right), \\
v_h &\sim \text{Beta}\left(1, \lambda\right), \\
\boldsymbol{\pi}_l &\sim G_0 = \mathcal{N}(\boldsymbol{\pi}_l | \boldsymbol{a}_0, \boldsymbol{\Sigma}_0).
\end{aligned}
\tag{18}
$$

The corresponding graphical model of the above equations is shown in Fig.1.

III. Posterior Distribution Inference Based on Expectation Propagation method

In this section, the posterior of the proposed hierarchical model is derived firstly and the framework of the application of the expectation propagation method is demonstrated in Section III-A. Detailed iteration criteria and formulations are computed in Section III-B, Section III-C and Section III-D. Finally, the reduction of computational complexity and update techniques are described in Section III-E. The entire procedure is summarized in Algorithm 1. Fundamental to the development of the analysis in this section is using key existing algorithms which include Gibbs sampler, variational Baysian (VB), and expectation propagation (EP). Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution [38]. Unlike Monte Carlo technique, which provides a numerical approximation to the exact posterior using a set of samples, variational Bayes provides a locally-optimal, exact analytical solution

to an approximation of the posterior [23]. EP is shown to be an effective method for approximate inference in a linear model with spike-and-slab priors and provides a better approximation of the posterior for the spike-and-slab model [42]. Therefore, we propose a novel algorithm based on expectation propagation to carry out the approximate inference of the posterior distribution in the proposed hierarchical Bayesian framework.

A. The Expectation Propagation Approximation

A brief summary of the expectation propagation is presented in [42]. According to Bayes' rule, the full posterior pdf of the proposed generative model can be formulated as

$$\begin{aligned} f\left(\boldsymbol{H}|\boldsymbol{Y},\boldsymbol{r}\right) &= \frac{f\left(\boldsymbol{Y}|\boldsymbol{H}\right) f\left(\boldsymbol{H}|\boldsymbol{r}\right)}{\int f\left(\boldsymbol{Y}|\boldsymbol{H}\right) f\left(\boldsymbol{H}|\boldsymbol{r}\right) d\boldsymbol{H}} \\ &\propto \prod_{i=1}^{M} f(\boldsymbol{y}_i|\boldsymbol{x}_i) \prod_{i=1}^{M} f(\boldsymbol{x}_i|\boldsymbol{z}_i) \prod_{i=1}^{M} f(\boldsymbol{z}_i|\boldsymbol{\Pi},\boldsymbol{c}) \prod_{l=1}^{M} f(c_i|\boldsymbol{v}) \prod_{l=1}^{L} f(v_l) \prod_{l=1}^{L} f(\boldsymbol{\pi}_l), \end{aligned}$$

where $\boldsymbol{r} = \{\sigma_0,\ \tau_0, \boldsymbol{a}_0, \boldsymbol{\Sigma}_0,\ \lambda,\ L\}$ is the set of hyper-parameters, $\boldsymbol{H} = \{\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Pi}, \boldsymbol{c}, \boldsymbol{v}\}$ is the set of all latent variables with $\boldsymbol{Z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_i, \cdots, \boldsymbol{z}_M] \in \mathbb{R}^{K\times M}$, $\boldsymbol{\Pi} = [\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_i, \cdots, \boldsymbol{\pi}_L] \in \mathbb{R}^{K\times L}$, and $\boldsymbol{v} = [v_1, \cdots, v_l, \cdots, v_L]^T$. We omit all the dependence of hyper-parameters in Eq. (19) for simplicity. It is observed that the posterior density can be decomposed into three terms, i.e., f_a for $a = 1, 2, 3$, and each of the three terms can be further decomposed as follows,

$$f_1(\boldsymbol{X}) = \prod_{i=1}^{M} f_{1i}(\boldsymbol{x}_i) = \prod_{i=1}^{M} \mathcal{N}\left(\boldsymbol{y}_i|\boldsymbol{D}_i\boldsymbol{x}_i, \sigma_0^2\boldsymbol{I}\right), \tag{19}$$

$$f_2(\boldsymbol{X},\boldsymbol{Z}) = \prod_{i=1}^{M} f_{2i}\left(\boldsymbol{x}_i,\boldsymbol{z}_i\right) = \prod_{i=1}^{M}\prod_{j=1}^{K} f_{2i,j}\left(x_{ji}, z_{ji}\right) = \prod_{i=1}^{M}\prod_{j=1}^{K} \left(\left(1-z_{ji}\right)\delta\left(x_{ji}\right) + z_{ji}\mathcal{N}\left(x_{ji}|0,\tau_0\right)\right), \tag{20}$$

$$f_3(\boldsymbol{Z},\boldsymbol{c},\boldsymbol{v},\boldsymbol{\Pi}) = \prod_{i=1}^{M} f_{3i}\left(\boldsymbol{z}_i, c_i, \boldsymbol{v}, \boldsymbol{\Pi}\right) = \prod_{i=1}^{M} \text{Bern}\left(\boldsymbol{z}_i|\boldsymbol{\pi}_{c_i}\right) \prod_{l=1}^{L} (1-v_l)^{\text{I}[c_i>l]} v_l^{c_i^l} \prod_{l=1}^{L} \text{Beta}(v_l|1,\lambda) \prod_{l=1}^{L} \mathcal{N}(\boldsymbol{\pi}_l|\boldsymbol{a}_0,\boldsymbol{\Sigma}_0), \tag{21}$$

where $\text{I}[\cdot]$ is an indicator function such that if the condition is true, $\text{I} = 1$, else $\text{I} = 0$ and $c_i^l = 1$ if $c_i = l$ else equals to 0. The idea is then to approximate each term in the true posterior density, i.e., f_a, by simpler terms, i.e., q_a, for $a = 1, 2, 3$. The EP framework provides flexibility in the choice of the approximating factors. This choice is a trade-off between analytical tractability and sufficient flexibility for capturing the important characteristics of the true density. It is observed that each f_{1i} term only depends on $\boldsymbol{x}_i$, f_{2i} only depends on $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, whereas f_{3i}

depends on $\boldsymbol{z}_i$, c_i, $\boldsymbol{v}$ and $\boldsymbol{\Pi}$. We choose q_{1i}, q_{2i} and q_{3i} of the following form,

$$q_{1i}(\boldsymbol{x}_i) = \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{m}_{i1}, \boldsymbol{\Sigma}_{i1}), \tag{22}$$

$$\begin{aligned} q_{2i}\left(\boldsymbol{x}_i, \boldsymbol{z}_i\right) &= \prod_{j=1}^{K} q_{2i,j}\left(x_{ji}, z_{ji}\right) \\ &= \prod_{j=1}^{K} \mathcal{N}\left(x_{ji}|m_{ji2}, \Sigma_{i2,jj}\right) \text{Bern}\left(z_{ji}|\mu_{ji2}\right) = \mathcal{N}\left(\boldsymbol{x}_i|\boldsymbol{m_{i2}}, \boldsymbol{\Sigma_{i2}}\right) \prod_{j=1}^{K} \text{Bern}\left(z_{ji}|\mu_{ji2}\right), \end{aligned} \tag{23}$$

$$\begin{aligned} q_{3i}\left(\boldsymbol{z}_i, c_i, \boldsymbol{v}, \boldsymbol{\Pi}\right) &= \prod_{j=1}^{K} \text{Bern}\left(z_{ji}|\mu_{ji3}\right) \prod_{l=1}^{L} \text{Beta}\left(v_l|g_{il}, h_{il}\right) \prod_{l=1}^{L}\prod_{j=1}^{K} \mathcal{N}\left(\pi_{jl}|a_{jil}, B_{il,jj}\right) \text{Multi}\left(c_i\,|w_{1i}, , .., w_{Li}\right) \\ &= \prod_{j=1}^{K} \text{Bern}\left(z_{ji}|\mu_{ji3}\right) \prod_{l=1}^{L} \text{Beta}\left(v_l|g_{il}, h_{il}\right) \prod_{l=1}^{L} \mathcal{N}\left(\boldsymbol{\pi}_l|\boldsymbol{a}_{il}, \boldsymbol{B}_{il}\right) \text{Multi}\left(c_i\,|w_{1i}, , .., w_{Li}\right). \end{aligned} \tag{24}$$

where $\boldsymbol{m}_{i2} = [m_{1i2}, m_{2i2}, ..., m_{Ki2}]^T$, $\boldsymbol{\Sigma}_{i2} = \text{diag}\left(\Sigma_{i2,11}, \Sigma_{i2,22}, ..., \Sigma_{i2,KK}\right), \boldsymbol{a}_{il} = [a_{1il}, a_{2il}, ..., a_{Kil}]^T$, and $\boldsymbol{B}_{il} = \text{diag}\left(B_{il,11}, B_{il,22}, ..., B_{il,KK}\right)$. Then the resulting global approximation $Q(\boldsymbol{H})$ becomes

$$Q(\boldsymbol{H}) = \frac{1}{Z_{EP}} \prod_{i=1}^{M} q_{1i} q_{2i} q_{3i} \propto \prod_{i=1}^{M} (\mathcal{N}\left(\boldsymbol{x}_i|\boldsymbol{m}_i, \boldsymbol{\Sigma}_i\right) \prod_{j=1}^{K} \text{Bern}\left(z_{ji}|\mu_{ji}\right) \text{Multi}(c_i)) \prod_{l=1}^{L} \text{Beta}(v_l|g_l, h_l) \prod_{l=1}^{L} \mathcal{N}\left(\pi_l|\boldsymbol{a}_l, \boldsymbol{B}_l\right), \tag{25}$$

where

$$\boldsymbol{\Sigma}_i^{-1} = \left(\boldsymbol{\Sigma}_{i1}^{-1} + \boldsymbol{\Sigma}_{i2}^{-1}\right)^{-1}, \tag{26}$$

$$\boldsymbol{m}_i = \boldsymbol{\Sigma}_i \left(\boldsymbol{\Sigma}_{i1}^{-1}\boldsymbol{m}_{i1} + \boldsymbol{\Sigma}_{i2}^{-1}\boldsymbol{m}_{i2}\right), \tag{27}$$

$$\mu_{ji} = \left(\frac{(1-\mu_{ji2})(1-\mu_{ji3})}{\mu_{ji2}\mu_{ji3}} + 1\right)^{-1}, \tag{28}$$

$$g_l = \textstyle\sum_{i=1}^{M} g_{il} - M, \tag{29}$$

$$h_l = \textstyle\sum_{i=1}^{M} h_{il} - M, \tag{30}$$

$$\boldsymbol{B}_l = \left(\textstyle\sum_{i=1}^{M} \boldsymbol{B}_{il}^{-1}\right)^{-1}, \tag{31}$$

$$\boldsymbol{a}_l = \boldsymbol{B}_l \left(\textstyle\sum_{i=1}^{M} \boldsymbol{B}_{il}^{-1}\boldsymbol{a}_{il}\right), \tag{32}$$

and Z_{EP} is the normalization constant which can be computed by $Z_{EP} = \int \prod_{i=1}^{M} q_{1i} q_{2i} q_{3i} d\boldsymbol{H}$.

B. Estimating Parameters for q_{1i}

The estimation procedure for $q_{1i}(\boldsymbol{x}_i)$ depends on the choice of observation model in Eq. (6), where it is the Gaussian noise model. Thus, f_{1i} is already in the exponential family for all M tasks and therefore needs not to be approximated by EP. Consider the quadratic form in the exponent of the Gaussian distribution $q_{1i}(\boldsymbol{x}_i)$ and $f_{1i}(\boldsymbol{x}_i)$,

$$f_{1i}(\boldsymbol{x}_i) : -\frac{1}{2\sigma_0^2}\boldsymbol{x}_i^T\boldsymbol{D}_i^T\boldsymbol{D}_i\boldsymbol{x}_i + \frac{1}{\sigma_0^2}\boldsymbol{x}_i^T\boldsymbol{D}_i^T\boldsymbol{y}_i + \text{const}, \tag{33}$$

$$q_{1i}(\boldsymbol{x}_i) : -\frac{1}{2}\boldsymbol{x}_i^T\boldsymbol{\Sigma}_{i1}^{-1}\boldsymbol{x}_i + \boldsymbol{x}_i^T\boldsymbol{\Sigma}_{i1}^{-1}\boldsymbol{m}_{i1} + \text{const}. \tag{34}$$

The parameters for q_{1i} are determined by completing the square between $f_{1i}(\boldsymbol{x}_i)$ and $q_{1i}(\boldsymbol{x}_i)$ using the following relations

$$\boldsymbol{\Sigma}_{i1}^{-1} = \frac{1}{\sigma_0^2}\boldsymbol{D}_i^T\boldsymbol{D}_i, \tag{35}$$

$$\boldsymbol{m}_{i1} = \frac{1}{\sigma_0^2}\boldsymbol{\Sigma}_{i1}\boldsymbol{D}_i^T\boldsymbol{y}_i. \tag{36}$$

In this paper, the noise variance is assumed to be constant for all tasks for simplicity.

C. Estimating Parameters for q_{2i}

According to Eq. (23), the term q_{2i} is factorized over j, which implies that we only need to update the parameters underlying each pair of x_{ji} and z_{ji} sequentially and iteratively. Consider the update of the jth term of the ith task $q_{2i,j}(x_{ji}, z_{ji})$, the first step is to compute the marginal cavity distribution by removing the contribution of $q_{2i,j}(x_{ji}, z_{ji})$ from the global distribution $Q(\boldsymbol{H})$

$$Q^{\backslash 2i,j}(x_{ji}, z_{ji}) = \frac{Q(x_{ji}, z_{ji})}{q_{2i,j}(x_{ji}, z_{ji})} = \mathcal{N}(x_{ji}|m_{ji}^{\backslash 2i,j}, \Sigma_{i,jj}^{\backslash 2i,j})\text{Bern}(z_{ji}|\mu_{ji}^{\backslash 2i,j}). \tag{37}$$

Note that $Q(\boldsymbol{H})$ is the product of distributions from exponential family. The parameters of the cavity distribution can be obtained by computing the differences of natural parameters, which can be expressed as

$$\Sigma_{i,jj}^{\backslash 2i,j} = (\Sigma_{i,jj}^{-1} - \Sigma_{i2,jj}^{-1})^{-1}, \tag{38}$$

$$m_{ji}^{\backslash 2i,j} = \Sigma_{i,jj}^{\backslash 2i,j}(m_{ji}\Sigma_{i,jj}^{-1} - m_{ji2}\Sigma_{i2,jj}^{-1}), \tag{39}$$

$$\mu_{ji}^{\backslash 2i,j} = \left(\frac{(1-\mu_{ji})\mu_{ji2}}{(1-\mu_{ji2})\mu_{ji}} + 1\right)^{-1} = \mu_{ji3}. \tag{40}$$

Since μ_{ji2} and μ_{ji3} are the only two parameters contributing to the distribution over z_{ji}, the cavity parameter for z_{ji} in $q_{2i,j}$ simply equals to μ_{ji3} and equals to μ_{ji2} in $q_{3i,j}$. The next step is to form the tilted distribution

$$\widetilde{Q}(x_{ji}, z_{ji}) = \frac{1}{Z_{2i,j}}Q^{\backslash 2i,j}(x_{ji}, z_{ji})f_{2i,j}(x_{ji}, z_{ji}), \tag{41}$$

where $Z_{2i,j}$ is the normalization constant given by

$$\begin{aligned} Z_{2i,j} &= \sum_{z_{ji}} \int Q^{\backslash 2i,j}(x_{ji}, z_{ji})f_{2i,j}(x_{ji}, z_{ji})dx_{ji} \\ &= \mu_{ji}^{\backslash 2i,j}\mathcal{N}\left(0|m_{ji}^{\backslash 2i,j}, \Sigma_{i,jj}^{\backslash 2i,j} + \tau_0\right) + (1-\mu_{ji}^{\backslash 2i,j})\mathcal{N}\left(0|m_{ji}^{\backslash 2i,j}, \Sigma_{i,jj}^{\backslash 2i,j}\right). \end{aligned} \tag{42}$$

Then the KL divergence between the titled distribution Eq. (41) and $Q^{\text{new}}(\boldsymbol{H})$, w.r.t. $Q^{\text{new}}(\boldsymbol{H})$, is minimized to obtain the revised parameters for $q(x_{ji}, z_{ji})$. For distributions from the exponential family, minimizing this form

of KL divergence amounts to matching moment between the tilted distribution and $Q^{\text{new}}(\boldsymbol{H})$. The first and second moments of x_{ji} w.r.t. the titled distribution are given by

$$\begin{aligned} X_1 &= \sum_{z_{ji}} \int x_{ji} \frac{1}{Z_{2i,j}} Q^{\backslash 2i,j}(x_{ji}, z_{ji}) f_{2i,j}(x_{ji}, z_{ji}) dx_{ji} \\ &= \frac{1}{Z_{2i,j}} \mu_{ji}^{\backslash 2i,j} \mathcal{N}\left(0 | m_{ji}^{\backslash 2i,j}, \Sigma_{i,jj}^{\backslash 2i,j} + \tau_0\right) \frac{\tau_0 m_{ji}^{\backslash 2i,j}}{\Sigma_{i,jj}^{\backslash 2i,j} + \tau_0}, \end{aligned} \tag{43}$$

$$\begin{aligned} X_2 &= \sum_{z_{ji}} \int x_{ji}^2 \frac{1}{Z_{2i,j}} Q^{\backslash 2i,j}(x_{ji}, z_{ji}) f_{2i,j}(x_{ji}, z_{ji}) dx_{ji} \\ &= \frac{1}{Z_{2i,j}} \mu_{ji}^{\backslash 2i,j} \mathcal{N}\left(0 | m_{ji}^{\backslash 2i,j}, \Sigma_{i,jj}^{\backslash 2i,j} + \tau_0\right) \left[\frac{\tau_0 \Sigma_{i,jj}^{\backslash 2i,j}}{\Sigma_{i,jj}^{\backslash 2i,j} + \tau_0} + \left(\frac{\tau_0 m_{ji}^{\backslash 2i,j}}{\Sigma_{i,jj}^{\backslash 2i,j} + \tau_0} \right)^2 \right], \end{aligned} \tag{44}$$

and the first moment of z_{ji} is given by

$$\begin{aligned} Z_1 &= \sum_{z_{ji}} \int z_{ji} \frac{1}{Z_{2i,j}} Q^{\backslash 2i,j}(x_{ji}, z_{ji}) f_{2i,j}(x_{ji}, z_{ji}) dx_{ji} \\ &= \frac{1}{Z_{2i,j}} \mu_{ji}^{\backslash 2i,j} \mathcal{N}\left(0 | m_{ji}^{\backslash 2i,j}, \Sigma_{i,jj}^{\backslash 2i,j} + \tau_0\right). \end{aligned} \tag{45}$$

Alternatively, the moments can be obtained by computing the partial derivatives of the log normalizer of the tilted distribution [37]. Hence, the revised parameters for $Q^{\text{new}}(x_{ji}, z_{ji})$ can be computed using the following formula

$$m_{ji}^{\text{new}} = X_1, \tag{46}$$

$$\Sigma_{i,jj}^{\text{new}} = X_2 - m_{ji}^{\text{new}2}, \tag{47}$$

$$\mu_{ji}^{\text{new}} = Z_1. \tag{48}$$

Once $Q^{\text{new}}(x_{ji}, z_{ji})$ is obtained, the new update of site distribution for $q_{2i,j}(x_{ji}, z_{ji})$ can be computed using the following

$$q_{2i,j}^{\text{new}}(x_{ji}, z_{ji}) = \frac{Q^{\text{new}}(x_{ji}, z_{ji})}{Q^{\backslash 2i,j}(x_{ji}, z_{ji})} = \mathcal{N}(x_{ji} | m_{ji2}^{\text{new}}, \Sigma_{i2,jj}^{\text{new}}) \text{Bern}(z_{ji} | \mu_{ji2}^{\text{new}}). \tag{49}$$

The parameters m_{ji2}^{new}, $\Sigma_{i2,jj}^{\text{new}}$, and μ_{ji2}^{new} of the new site distribution $q_{2i,j}^{\text{new}}(x_{ji}, z_{ji})$ can be obtained by computing the differences of natural parameters in the same manner as for the cavity distribution in Eq. (38)-Eq. (40)

$$\Sigma_{i2,jj}^{\text{new}} = (\Sigma_{i,jj}^{\text{new}-1} - \Sigma_{i,jj}^{\backslash 2i,j-1})^{-1}, \tag{50}$$

$$m_{ji2}^{\text{new}} = \Sigma_{i2,jj}^{\text{new}} (m_{ji}^{\text{new}} \Sigma_{i,jj}^{\text{new}-1} - m_{ji}^{\backslash 2i,j} \Sigma_{i,jj}^{\backslash 2i,j}), \tag{51}$$

$$\mu_{ji2}^{\text{new}} = \left(\frac{(1 - \mu_{ji}^{\text{new}}) \mu_{ji}^{\backslash 2i,j}}{\left(1 - \mu_{ji}^{\backslash 2i,j}\right) \mu_{ji}^{\text{new}}} + 1 \right)^{-1}. \tag{52}$$

D. Estimating Parameters for q_{3i}

The procedure for updating $q_{3i}(\boldsymbol{z}_i, c_i, \boldsymbol{v}, \boldsymbol{\pi})$ is completely analogously to that for $q_{2i}(\boldsymbol{x}_i, \boldsymbol{z}_i)$. The first step is to remove it from the global distribution $Q(\boldsymbol{H})$ to compute the cavity distribution $Q^{\backslash 3i}(\boldsymbol{v}, \boldsymbol{\Pi}, \boldsymbol{c}, \boldsymbol{z}_i)$, where the

corresponding parameters can be obtained by computing the differences of natural parameters using the following formula

$$\boldsymbol{B}_l^{\backslash 3i} = (\boldsymbol{B}_l^{-1} - \boldsymbol{B}_{il}^{-1})^{-1}, \tag{53}$$

$$\boldsymbol{a}_l^{\backslash 3i} = \boldsymbol{B}_l^{\backslash 3i}(\boldsymbol{B}_l^{-1}\boldsymbol{a}_l - \boldsymbol{B}_{il}^{-1}\boldsymbol{a}_{il}), \tag{54}$$

$$g_l^{\backslash 3i} = g_l - g_{il} + 1, \tag{55}$$

$$h_l^{\backslash 3i} = h_l - h_{il} + 1, \tag{56}$$

$$\mu_{ji}^{\backslash 3i} = \left(\tfrac{(1-\mu_{ji})\mu_{ji3}}{(1-\mu_{ji3})\mu_{ji}} + 1\right)^{-1} = \mu_{ji2}, \tag{57}$$

and the tilted distribution is

$$\widetilde{Q}(\boldsymbol{H}) = \frac{1}{Z_{3i}} Q^{\backslash 3i}(\boldsymbol{v}, \boldsymbol{\Pi}, \boldsymbol{c}, \boldsymbol{z}_i) f_{3i}(\boldsymbol{z}_i, c_i, \boldsymbol{v}, \boldsymbol{\Pi}), \tag{58}$$

where the normalization constant is given by

$$\begin{aligned} Z_{3i} &= \sum_{\boldsymbol{z}_i} \int Q^{\backslash 3i}(\boldsymbol{v}, \boldsymbol{\Pi}, \boldsymbol{c}, \boldsymbol{z}_i) f_{3i}(\boldsymbol{z}_i, c_i, \boldsymbol{v}, \boldsymbol{\Pi}) d\boldsymbol{c} d\boldsymbol{v} d\boldsymbol{\Pi} \\ &= \sum_{l=1}^{L} \frac{g_l^{\backslash 3i}}{g_l^{\backslash 3i} + h_l^{\backslash 3i}} \prod_{s=1}^{l-1} \left(1 - \frac{g_s^{\backslash 3i}}{g_s^{\backslash 3i} + h_s^{\backslash 3i}}\right) \\ &\cdot \prod_{j=1}^{K} \left(\phi(\frac{a_{jl}^{\backslash 3i}}{\sqrt{1 + B_{l,jj}^{\backslash 3i}}}) \mu_{ji}^{\backslash 3i} + (1 - \phi(\frac{a_{jl}^{\backslash 3i}}{\sqrt{1 + B_{l,jj}^{\backslash 3i}}}))(1 - \mu_{ji}^{\backslash 3i}) \right). \end{aligned} \tag{59}$$

Analogously, the first and second moments of π_{jl} can be computed as

$$\begin{aligned} X_1 &= \sum_{\boldsymbol{z}_i} \int \pi_{jl} \frac{1}{Z_{3i}} Q^{\backslash 3i}(\boldsymbol{v}, \boldsymbol{\Pi}, \boldsymbol{c}, \boldsymbol{z}_i) f_{3i}(\boldsymbol{z}_i, c_i, \boldsymbol{v}, \boldsymbol{\Pi}) d\boldsymbol{c} d\boldsymbol{v} d\boldsymbol{\Pi} \\ &= \frac{1}{Z_{3i}} \frac{g_l^{\backslash 3i}}{g_l^{\backslash 3i} + h_l^{\backslash 3i}} \prod_{s=1}^{l-1} \left(1 - \frac{g_s^{\backslash 3i}}{g_s^{\backslash 3i} + h_s^{\backslash 3i}}\right) \\ &\cdot \left(\mu_{ji}^{\backslash 3i} e_{jl} + (1 - \mu_{ji}^{\backslash 3i})(a_{jl}^{\backslash 3i} - e_{jl})\right) \prod_{j'=1, j' \neq j}^{K} \left(\mu_{j'i}^{\backslash 3i} \rho_{j'l} + (1 - \mu_{j'i}^{\backslash 3i})(1 - \rho_{j'l})\right) \\ &+ \frac{1}{Z_{3i}} \sum_{l'=1, l' \neq l}^{L} \frac{g_{l'}^{\backslash 3i}}{g_{l'}^{\backslash 3i} + h_{l'}^{\backslash 3i}} \prod_{s=1}^{l'-1} \left(1 - \frac{g_s^{\backslash 3i}}{g_s^{\backslash 3i} + h_s^{\backslash 3i}}\right) a_{jl}^{\backslash 3i} \prod_{j=1}^{K} \left(\mu_{ji}^{\backslash 3i} \rho_{jl'} + (1 - \mu_{ji}^{\backslash 3i})(1 - \rho_{jl'})\right), \end{aligned} \tag{60}$$

and

$$\begin{aligned} X_2 &= \sum_{\boldsymbol{z}_i} \int \pi_{jl}^2 \frac{1}{Z_{3i}} Q^{\backslash 3i} f_{3i}(\boldsymbol{z}_i, c_i, \boldsymbol{v}, \boldsymbol{\Pi}) d\boldsymbol{c} d\boldsymbol{v} d\boldsymbol{\Pi} \\ &= \frac{1}{Z_{3i}} \frac{g_l^{\backslash 3i}}{g_l^{\backslash 3i} + h_l^{\backslash 3i}} \prod_{s=1}^{l-1} \left(1 - \frac{g_s^{\backslash 3i}}{g_s^{\backslash 3i} + h_s^{\backslash 3i}}\right) \cdot \left(\mu_{ji}^{\backslash 3i} u_{jl} + (1 - \mu_{ji}^{\backslash 3i})(a_{j'l}^{\backslash 3i\ 2} + B_{l,jj}^{\backslash 3i} - u_{jl})\right) \\ &\cdot \prod_{j'=1, j' \neq j}^{K} \left(\mu_{j'i}^{\backslash 3i} \rho_{j'l} + (1 - \mu_{j'i}^{\backslash 3i})(1 - \rho_{j'l})\right) + \frac{1}{Z_{3i}} \sum_{l'=1, l' \neq l}^{L} \frac{g_{l'}^{\backslash 3i}}{g_{l'}^{\backslash 3i} + h_{l'}^{\backslash 3i}} \prod_{s=1}^{l'-1} \left(1 - \frac{g_s^{\backslash 3i}}{g_s^{\backslash 3i} + h_s^{\backslash 3i}}\right) \\ &\cdot (a_{j'l}^{\backslash 3i\ 2} + B_{l,jj}^{\backslash 3i}) \prod_{j=1}^{K} \left(\mu_{ji}^{\backslash 3i} \rho_{jl'} + (1 - \mu_{ji}^{\backslash 3i})(1 - \rho_{jl'})\right), \end{aligned} \tag{61}$$

where

$$e_{jl} = \int \pi_{jl} \mathcal{N}\left(\pi_{jl} | a_{jl}^{\backslash 3i}, B_{l,jj}^{\backslash 3i}\right) \phi(\pi_{jl}) d\pi_{jl} = \phi\left(\frac{a_{jl}^{\backslash 3i}}{\sqrt{1+B_{l,jj}^{\backslash 3i}}}\right) a_{jl}^{\backslash 3i} + \frac{B_{l,jj}^{\backslash 3i} N\left(r_{jl}|0,1\right)}{\sqrt{1+B_{l,jj}^{\backslash 3i}}} \tag{62}$$

$$u_{jl} = \int \pi_{jl}^2 \mathcal{N}(\pi_{jl} | a_{jl}^{\backslash 3i}, B_{l,jj}^{\backslash 3i}) \phi(\pi_{jl}) d\pi_{jl} = 2a_{jl}^{\backslash 3i} e_{jl} + \phi(r_{jl})[B_{l,jj}^{\backslash 3i} - (a_{jl}^{\backslash 3i})^2] - \frac{(B_{l,jj}^{\backslash 3i})^2 r_{jl} N(r_{jl}|0,1)}{1+B_{l,jj}^{\backslash 3i}} \tag{63}$$

$$\rho_{jl} = \phi(r_{jl}) \tag{64}$$

$$r_{jl} = \frac{a_{jl}^{\backslash 3i}}{\sqrt{1+B_{l,jj}^{\backslash 3i}}}. \tag{65}$$

The first moment of z_{ji} and c_{li} w.r.t. the titled distribution Eq. (58) is computed using

$$\begin{aligned} Z_1 &= \sum_{\boldsymbol{z}_i} \int Z_{jl} \frac{1}{Z_{3i}} Q^{\backslash 3i} f_{3i}(\boldsymbol{z}_i, c_i, \boldsymbol{v}, \boldsymbol{\Pi}) d\boldsymbol{c} d\boldsymbol{v} d\boldsymbol{\Pi} \\ &= \sum_{l=1}^{L} \frac{g_l^{\backslash 3i}}{g_l^{\backslash 3i} + h_l^{\backslash 3i}} \prod_{s=1}^{l-1} \left(1 - \frac{g_s^{\backslash 3i}}{g_s^{\backslash 3i} + h_s^{\backslash 3i}}\right) \prod_{j=1}^{K} \phi(\frac{a_{jl}^{\backslash 3i}}{\sqrt{1+B_{l,jj}^{\backslash 3i}}}) \mu_{ji}^{\backslash 3i}, \end{aligned} \tag{66}$$

$$\begin{aligned} C_{1l} &= \sum_{\boldsymbol{z}_i} \int c_{il} \frac{1}{Z_{3i}} Q^{\backslash 3i} f_{3i}(\boldsymbol{z}_i, c_i, \boldsymbol{v}, \boldsymbol{\Pi}) d\boldsymbol{c} d\boldsymbol{v} d\boldsymbol{\Pi} \\ &= \frac{g_l^{\backslash 3i}}{g_l^{\backslash 3i} + h_l^{\backslash 3i}} \prod_{s=1}^{l-1} \left(1 - \frac{g_s^{\backslash 3i}}{g_s^{\backslash 3i} + h_s^{\backslash 3i}}\right) \\ &\times \prod_{j=1}^{K} \left(\phi(\frac{a_{jl}^{\backslash 3i}}{\sqrt{1+B_{l,jj}^{\backslash 3i}}}) \mu_{ji}^{\backslash 3i} + (1 - \phi(\frac{a_{jl}^{\backslash 3i}}{\sqrt{1+B_{l,jj}^{\backslash 3i}}}))(1 - \mu_{ji}^{\backslash 3i})\right). \end{aligned} \tag{67}$$

Consider that $\boldsymbol{v}$ follows the Beta distribution, computing the moments of natural parameters w.r.t. $\widetilde{Q}(\boldsymbol{H})$ directly is extremely complex. Alternatively, we compute the partial derivatives of $\ln Z_{3i}$ to obtain the natural moments using the following formula

$$\begin{aligned} \nabla_{g_l^{\backslash 3i}} \ln Z_{3i} &= \frac{1}{Z_{3i}} \int \frac{\partial Q^{\backslash 3i}(\boldsymbol{H})}{\partial g_l^{\backslash 3i}} f_{3i}(\boldsymbol{z}_i, c_i, \boldsymbol{v}, \boldsymbol{\Pi}) d\boldsymbol{H} \\ &= \int \frac{\widetilde{Q}(\boldsymbol{H})}{\text{Beta}(v_l | g_l^{\backslash 3i}, h_l^{\backslash 3i})} \frac{\partial \text{Beta}(v_l | g_l^{\backslash 3i}, h_l^{\backslash 3i})}{\partial g_l^{\backslash 3i}} d\boldsymbol{H} \\ &= \int \widetilde{Q}(\boldsymbol{H}) \left(\ln(v_l) + \Psi(g_l^{\backslash 3i} + h_l^{\backslash 3i}) - \Psi(g_l^{\backslash 3i})\right) d\boldsymbol{H} \\ &= E_{\widetilde{Q}(\boldsymbol{H})}(\ln(v_l)) + \Psi(g_l^{\backslash 3i} + h_l^{\backslash 3i}) - \Psi(g_l^{\backslash 3i}), \end{aligned} \tag{68}$$

and

$$\begin{aligned} \nabla_{h_l^{\backslash 3i}} \ln Z_{3i} &= \frac{1}{Z_{3i}} \int \frac{\partial Q^{\backslash 3i}(\boldsymbol{H})}{\partial h_l^{\backslash 3i}} f_{3i}(\boldsymbol{z}_i, c_i, \boldsymbol{v}, \boldsymbol{\Pi}) d\boldsymbol{H} \\ &= E_{\widetilde{Q}(\boldsymbol{H})}(\ln(1 - v_l)) + \Psi(g_l^{\backslash 3i} + h_l^{\backslash 3i}) - \Psi(g_l^{\backslash 3i}), \end{aligned} \tag{69}$$

where $\Psi(x) = \frac{d \ln \Gamma(x)}{dx}$ denotes the digamma function. The moment matching for v_l can be expressed as

$$E_{f(v_l)}\left(\ln\left(v_l\right)\right) = E_{q^{\text{new}}(v_l)}\left(\ln\left(v_l\right)\right), \tag{70}$$

$$E_{f(v_l)}\left(\ln\left(1 - v_l\right)\right) = E_{q^{\text{new}}(v_l)}\left(\ln\left(1 - v_l\right)\right). \tag{71}$$

Note that the left hand side terms of Eqs. (68)-(69) can be computed analytically from Eq. (59). Combined with Eqs. (70)-(71), the expectations of the approximate posterior distribution $Q^{\text{new}}(\boldsymbol{H})$ are matched to that of the titled distribution $\widetilde{Q}(\boldsymbol{H})$ as

$$\mu_{ji}^{\text{new}} = Z_1, \tag{72}$$

$$w_{li}^{\text{new}} = \frac{C_{1l}}{\sum_{l=1}^{L}(C_{1l})}, \tag{73}$$

$$a_{jl}^{\text{new}} = X_1, \tag{74}$$

$$B_{l,jj}^{\text{new}} = X_2 - a_{jl}^{\text{new}2}, \tag{75}$$

$$\Psi(g_l^{\text{new}}) - \Psi(g_l^{\text{new}} + h_l^{\text{new}}) + \Psi(g_l^{\backslash 3i} + h_l^{\backslash 3i}) - \Psi(g_l^{\backslash 3i}) = \nabla_{g_l^{\backslash 3i}} \ln Z_{3i}, \tag{76}$$

$$\Psi(h_l^{\text{new}}) - \Psi(g_l^{\text{new}} + h_l^{\text{new}}) + \Psi(h_l^{\backslash 3i} + g_l^{\backslash 3i}) - \Psi(h_l^{\backslash 3i}) = \nabla_{h_l^{\backslash 3i}} \ln Z_{3i}. \tag{77}$$

Note that resolving Eqs. (76)-(77) requires inverting of Ψ function and thus, there exists no general closed form solutions. Motivated by [43], a numerical fixed-pointed iteration is adopted to find the solutions for $a_{jl}^{\text{new}}, b_{jl}^{\text{new}}, g_l^{\text{new}}$ and h_l^{new} in our work, which takes the following form

$$g_l^{\text{new new}} = \Psi^{-1}(\Psi(g_l^{\text{new old}} + h_l^{\text{new old}}) + \nabla_{g_l^{\backslash 3i}} \ln Z_{3i} + \Psi(g_l^{\backslash 3i}) - \Psi(g_l^{\backslash 3i} + h_l^{\backslash 3i})), \tag{78}$$

$$h_l^{\text{new new}} = \Psi^{-1}(\Psi(g_l^{\text{new old}} + h_l^{\text{new old}}) + \nabla_{h_l^{\backslash 3i}} \ln Z_{3i} + \Psi(h_l^{\backslash 3i}) - \Psi(g_l^{\backslash 3i} + h_l^{\backslash 3i})), \tag{79}$$

where Ψ^{-1} denotes the inversion of function Ψ. [43] also provides a generalized Newton-Raphson iteration method, which does not require inverting Ψ function or the Hessian matrix explicitly. Once $Q^{\text{new}}(\boldsymbol{H})$ is obtained, similar to Eqs. (53) - (57), the parameters for the new site distribution take the following form

$$\mu_{ji3}^{\text{new}} = \left(\frac{\left(1-\mu_{ji}^{\text{new}}\right)\mu_{ji}^{\backslash 3i}}{\left(1-\mu_{ji}^{\backslash 3i}\right)\mu_{ji}^{\text{new}}} + 1 \right)^{-1}, \tag{80}$$

$$\boldsymbol{a}_{il}^{\text{new}} = \boldsymbol{B}_{il}^{\text{new}}(\boldsymbol{B}_l^{\text{new }-1}\boldsymbol{a}_l^{\text{new}} - \boldsymbol{B}_l^{\backslash 3i\ -1}\boldsymbol{a}_l^{\backslash 3i}), \tag{81}$$

$$\boldsymbol{B}_{il}^{\text{new}} = (\boldsymbol{B}_l^{\text{new }-1} - \boldsymbol{B}_l^{\backslash 3i\ -1})^{-1}, \tag{82}$$

$$g_{il}^{\text{new}} = g_l^{\text{new}} - g_l^{\backslash 3i} + 1, \tag{83}$$

$$h_{il}^{\text{new}} = h_l^{\text{new}} - h_l^{\backslash 3i} + 1. \tag{84}$$

E. Reduction of Computational Complexity

The inversion of the $K \times K$ dimensional covariance matrix in Eq. (26) renders computational complexity of $O\left(K^3\right)$, which is extremely high when its dimension K is large. Observing that $\boldsymbol{\Sigma}_{i1}$ is low rank and $\boldsymbol{\Sigma}_{i2}$ is diagonal, the Woodbury matrix identity can be applied to Eq. (26) to get:

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_{i2} - \boldsymbol{\Sigma}_{i2} D_i^T \left(\sigma_0^2 I + D_i \boldsymbol{\Sigma}_{i2} D_i^T\right)^{-1} D_i. \tag{85}$$

For $P \ll K$, the complexity scales as $O(PK^2)$ which reduces the complexity by K/P compare to that when performing inversion directly.

Algorithm 1 EP-based Structured BCS

Set truncation level L

Initialize all the approximation terms q_{ai}, $a = 1, 2, 3$, $i = 1, 2, ..., M$

Pre-compute $\boldsymbol{m}_{i1}$, $\boldsymbol{\Sigma}_{i1}$ by Eqs. (35)-(36), $i = 1, 2, ..., M$

repeat

  for the ith task **do**

    for each $q_{2i,j}$ **do**

      Compute cavity $Q^{\backslash 2i,j} \propto \frac{Q}{q_{2i,j}}$ by Eqs. (38) -(40)

      Minimize: $KL\left(f_{2i,j}Q^{\backslash 2i,j}||Q^{\text{new}}\right)$ w.r.t. Q^{new} by Eq. (46)-(48)

      Compute: $q_{2i,j} \propto \frac{Q^{\text{new}}}{Q^{\backslash 2i,j}}$ to update m_{ji2}, $\Sigma_{i2,jj}, \mu_{ji2}$ by Eqs. (50)-(52)

    end for

    Compute cavity $Q^{\backslash 3i} \propto \frac{Q}{q_{3i}}$ by Eqs. (53)-(57)

    Minimize: $KL\left(f_{3i}Q^{\backslash 3i}||Q^{\text{new}}\right)$ w.r.t. Q^{new} by Eqs. (72)-(77)

    Compute: $q_{3i} \propto \frac{Q^{\text{new}}}{Q^{\backslash 3i}}$ to update $\boldsymbol{a}_{il}, \boldsymbol{B}_{il}, g_{il}$, $h_{il}, \mu_{ji3}, w_{li}$ by Eqs. (80)-(84)

  end for

  Update joint approximation parameters $\boldsymbol{m}_i, \boldsymbol{\Sigma}_i$,$\mu_{ji}$, $\boldsymbol{a}_l, \boldsymbol{B}_l, g_l, h_l, w_{li}$ by Eqs. (26)-(32)

until Convergence criterion is reached

The site approximation is updated in a sequential manner in the conventional EP framework, i.e., refining a single $f_{2i,j}$ or f_{3i} at a time. This means that we need to update global approximation every time a site approximation is refined, which lacks computational efficiency and is unnecessary. Parallel update scheme is adopted in this work to reduce the computational complexity. In essence, all the site approximations of the form $f_{2i,j}$ or f_{3i}, for $i = 1, 2, ..., M$, $j = 1, 2, ..., K$, are first updated and then the global joint approximation is updated. This can be interpreted as a particular scheduling of message from a message passing perspective [44]. The entire approximation procedure, referred to as the EP-based structured Bayesian compressive sensing (EP-SBCS) method, is summarized in Algorithm 1.

IV. Simulation and Experiment Results

In this section, experiments using both synthetic data and real data sets are conducted to investigate and verify the effectiveness of the proposed method. Three competing state-of-the-art methods, namely, MT-CS [20], multi-task adaptive matching pursuit (MT-AMP) [45], and WBDOA [7], are considered for performance comparisons.

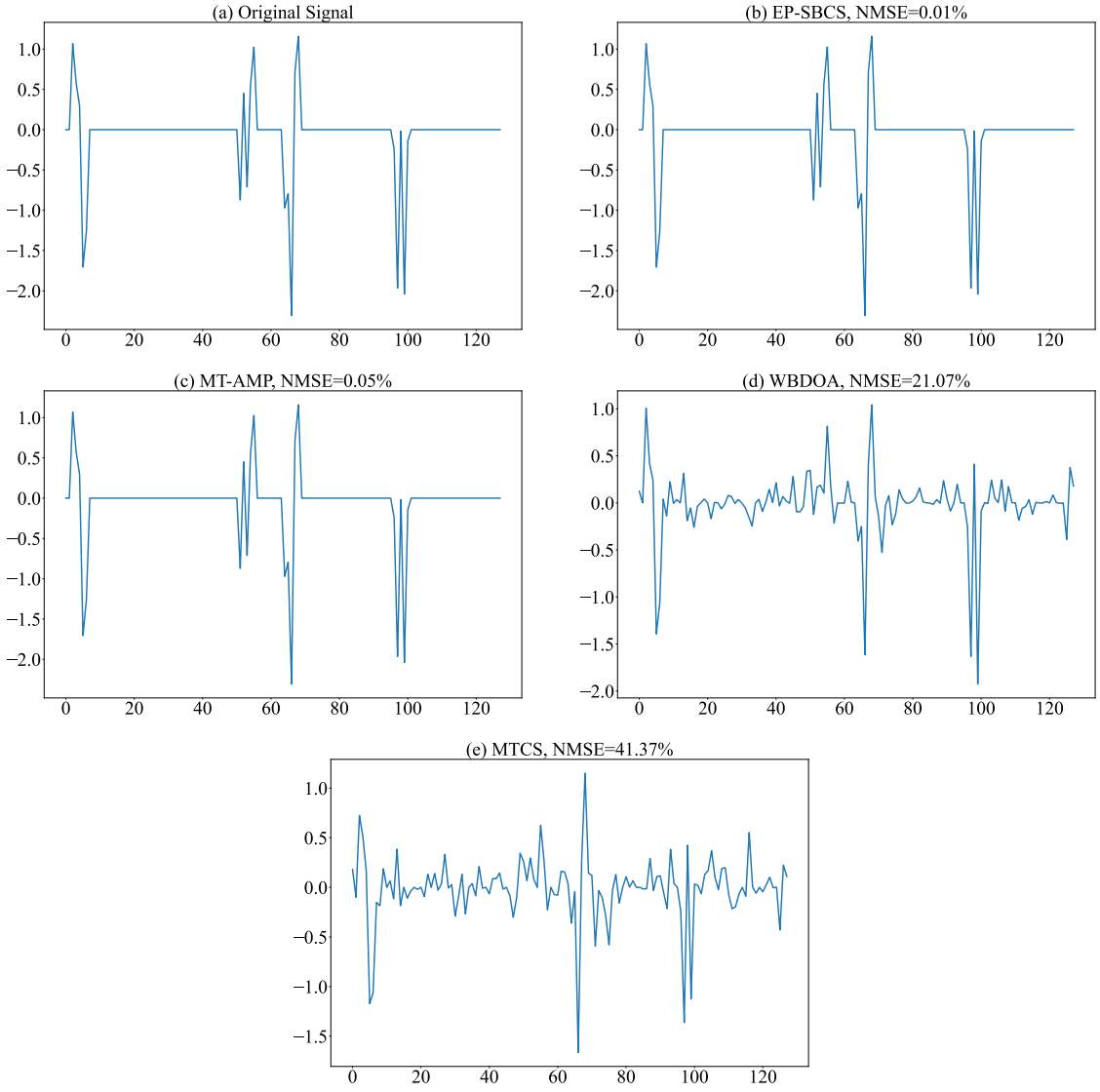


Figure 2. Reconstruction result examples of one task profiles.

The normalized root mean square error (NMSE) $||\hat{\boldsymbol{x}} - \boldsymbol{x}||_2/||\boldsymbol{x}||_2$ is used as the performance index. The number of components is typically independent of L and is approximately $\mathcal{O}(\lambda \log M)$ [39], and thus the stick-breaking truncation level is assumed to be $L = 10$ for all the experiments due to the consideration of the computational efficiency, and the scalar $\lambda = K/(K-1)$ is used according to the choice guidance [38]. All experimental results were obtained by averaging 100 trials conducted on a 2.80 GHz PC using Python3.

A. One-Dimensional Synthetic Data

In this simulation, the sensing dictionary matrices are i.i.d. and drawn from the Gaussian distribution with zero mean and unit variance, and the length of the sparse vectors is $K = 128$. The number of non-zero elements of each task is 20, and their coefficients are randomly drawn from $\mathcal{N}(0, 1)$. We generate sparse vector templates from

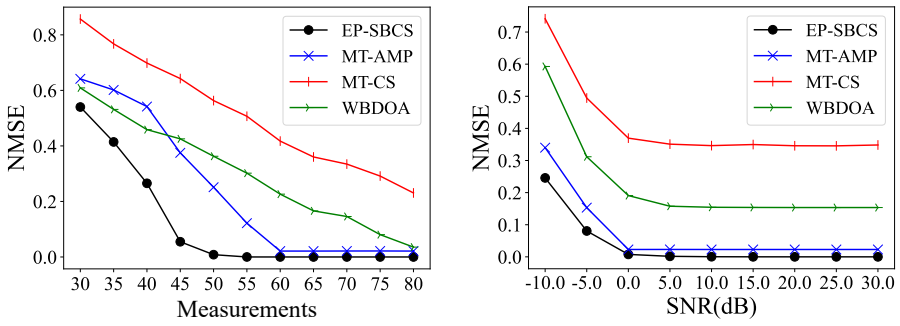


Figure 3. Performance comparison. (a) NMSE versus the number of measurements. (b) NMSE versus SNR.

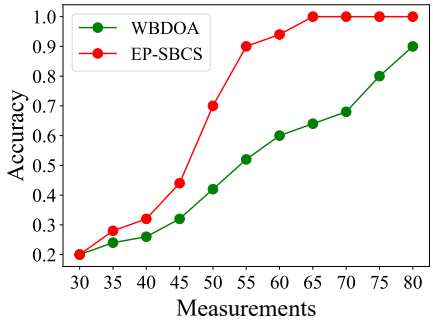


Figure 4. Cluster accuracies versus the number of measurements.

five clusters with different sparse patterns, and 10 observed measurement data in each cluster corresponding to 10 tasks. Accordingly, the total number of tasks is 50. We take a specific data generation mechanism to ensure that tasks within the same cluster are highly correlated and any two tasks from different clusters exhibit entirely different sparseness patterns. That is, locations of 20 non-zero elements are chosen using squared exponential kernel $\mathbf{\Sigma}_{0\ i,j} = 25\exp(-||i-j||_2^2/(2\times 100^2))$ and entirely different for each template. For each cluster, the 10 observed measurements are generated by randomly selecting two non-zero elements from the associated template and setting the coefficients to zero, and randomly selecting two zero-amplitude points in the template and setting them to be non-zero. The ith sensing matrix $\boldsymbol{D}_i$ is generated by following Gaussian distribution $\mathcal{N}(0,1)$. Without loss of generality, additive noise is considered with an input SNR of 20 dB. Both MT-CS and MT-AMP methods are able to reconstruct the sparse signals by exploiting the global sharing mechanism for all tasks. However, it is unfair to compare their performance when directly applied over all tasks data. This would lead to rapid performance degradation due to inappropriate sparse pattern sharing across clusters. Therefore, we in advance manually group these 50 tasks into 5 correctly formed clusters. The above two methods are then carried out over tasks within the same cluster one at a time. Five clusters of tasks are reconstructed respectively and the average NMSE of 5 clusters is used for performance comparison.

Fig. 2 demonstrates examples of the reconstruction results of one task profile when $P = 60$ and the reconstruction

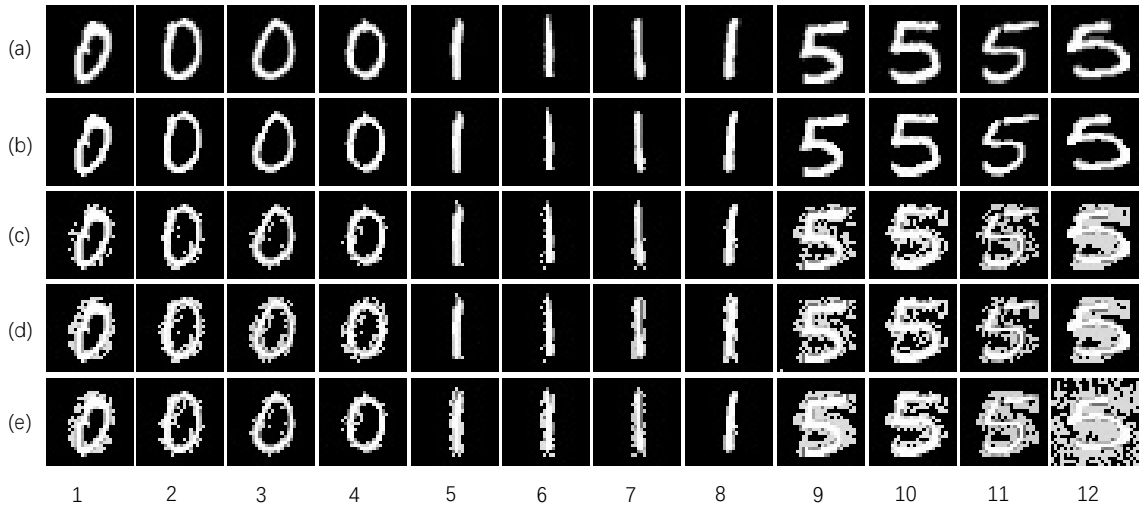


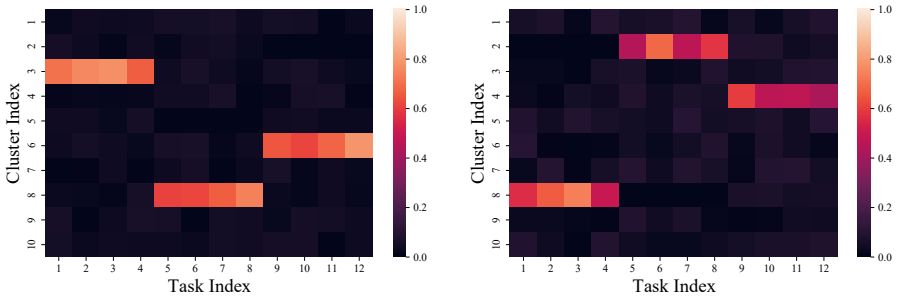
Figure 5. Reconstructed images from MNIST data set. (a) Origin, (b) EP-SBCS, (c) MT-AMP, (d) WBDOA, (e) MT-CS. Digits below the figure denotes the task index.

NMSE performances are indicated above each subplot. The full performance versus the number of measurements and the input SNR is shown in Figs. 3(a)–(b), respectively. It is observed that the reconstructed NMSE generally decreases with the increased measurements, and the proposed method has the lowest NMSE across all the level of measurements in Fig. 3(a), compared to those in WBDOA, MT-CS, and MT-AMP. This is despite the fact that in the latter two methods, the tasks are grouped manually. Fig. 3(b) shows the performance comparisons versus the input SNR with $P = 70$. It is clear that the NMSE decreases with the increased SNR, and the NMSE obtained by the proposed method is lower than those obtained from the other three methods, and thus is more robust to noise. These results verify the benefit of the generalized spike-and-slab priors with exploiting statistical relations within and between tasks.

Fig. 4 shows the clustering accuracy of both the proposed EP-SBCS and the WBDOA method versus the number of measurements. It is evident that the clustering accuracy improves with the increased number of measurements with the proposed method offering higher accuracy than WBDOA. The lower NMSE offered by our algorithm implies a positive relationship between clustering accuracy and sparse reconstruction precision. That is a high classification accuracy leads to a high reconstruction precision. The clustering accuracy of the proposed method increases up to 100% when $P \geq 65$ measurements are used. It can be concluded that the proposed method can infer the clusters and perform CS inversion simultaneously, and shows clear superiorities over the MT-CS, MT-AMP, and WBDOA methods.

B. Real Image Recovery

In the following set of experiments, the performance of EP-SBCS is compared to the other three methods on two example problems that involve 2D images.



(a) (b)

Figure 6. Probability heatmap results of clustering for MNIST data set. (a) EP-SBCS. (b) WBDOA.

Table I

PERFORMANCE OF RECONSTRUCTED MNIST IMAGES.

	NMSE												Time(s)
	Task 1	Task 2	Task 3	Taks 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11	Task 12	Average
EP-SBCS	**0.0015**	**0.0012**	**0.0013**	**0.0011**	**0.0012**	**0.0014**	**0.0010**	**0.0011**	**0.0012**	**0.0019**	**0.0020**	**0.0020**	1.1080
MT-AMP	0.0457	0.0426	0.0389	0.0267	0.0176	0.0460	0.0334	0.0217	0.0795	0.0738	0.0928	0.0947	1.2401
WBDOA	0.0624	0.0698	0.0729	0.0474	0.0176	0.0436	0.0585	0.0649	0.0845	0.0709	0.0894	0.0935	1.7031
MT-CS	0.0688	0.0446	0.0414	0.0411	0.0611	0.0871	0.0612	0.0179	0.0976	0.0770	0.1121	0.1149	**0.7031**

1) Handwritten Digit Images: In the following experiment, the real data of handwritten digit images from the MNIST data set [46] are used for performance comparison. The reconstructed results using 12 tasks from 3 different digit clusters are shown in Fig. 5. All of the samples are of size 28×28. Since gray values of most pixels in each image are zero (black), each image is naturally sparse. We simply reshape each image into a spare vector $\boldsymbol{x}_i$ of size 784×1. It is observed that sparse patterns of the same digits are highly similar, whereas different digits have different sparse patterns. The experiment hyper-parameters settings, the sensing matrices, and observation vectors are constructed in the same manner as in the first set of examples. Following a similar processing, both MT-CS and MT-AMP methods are used to recover sparse signals over the same digits, which are manually classified as one cluster in advance. Herein, the number of measurements is set as $P = 400$ and the input SNR is 20 dB. The reconstructed results are shown in Fig. 5 and the full performance comparison between them is summarized in Table I. Comparing to those images reconstructed by the WBDOA, MT-CS, and MT-AMP methods, the images recovered by the proposed EP-SBCS are closest to the original images and have the lowest NMSE. This was enabled by exploiting the structured and clustered statistical correlations within clusters.

It is observed that the reconstruction results vary from images of different digits. Generally, images of digit 1 have the best reconstruction performance, whereas digit 5 is the poorest, which is more obvious for WBDOA, MT-CS, and MT-AMP. This is caused by different sparsity levels of the images corresponding to the different digits. The average sparsity level of the images of digit 1, digit 0, and digit 5 is 0.06, 0.13, and 0.20, respectively, and the reconstructed results naturally degrade as the sparsity level increases. Furthermore, considering the negligible change of performance of the EP-SBCS for the images corresponding to the different digits, we can conclude that

Table II

PERFORMANCE OF RECONSTRUCTED VIDEO IMAGES.

	NMSE									Time(s)
	Task 1	Task 2	Task 3	Taks 4	Task 5	Task 6	Task 7	Task 8	Task 9	Average
Linear	0.0511	0.0517	0.0512	0.0611	0.0699	0.0650	0.1540	0.1450	0.1519	**0.0090**
EP-SBCS	**0.0570**	**0.0580**	**0.0574**	**0.0752**	**0.0815**	**0.0799**	**0.1551**	**0.1499**	**0.1598**	64.4526
MT-AMP	0.0611	0.0616	0.0617	0.0768	0.0866	0.0910	0.1623	0.1554	0.1627	70.1541
WBDOA	0.0650	0.0644	0.0637	0.0775	0.0867	0.0819	0.1629	0.1562	0.1629	81.5942
MT-CS	0.0624	0.0638	0.0617	0.0818	0.0990	0.0923	0.1619	0.1560	0.1627	31.1561

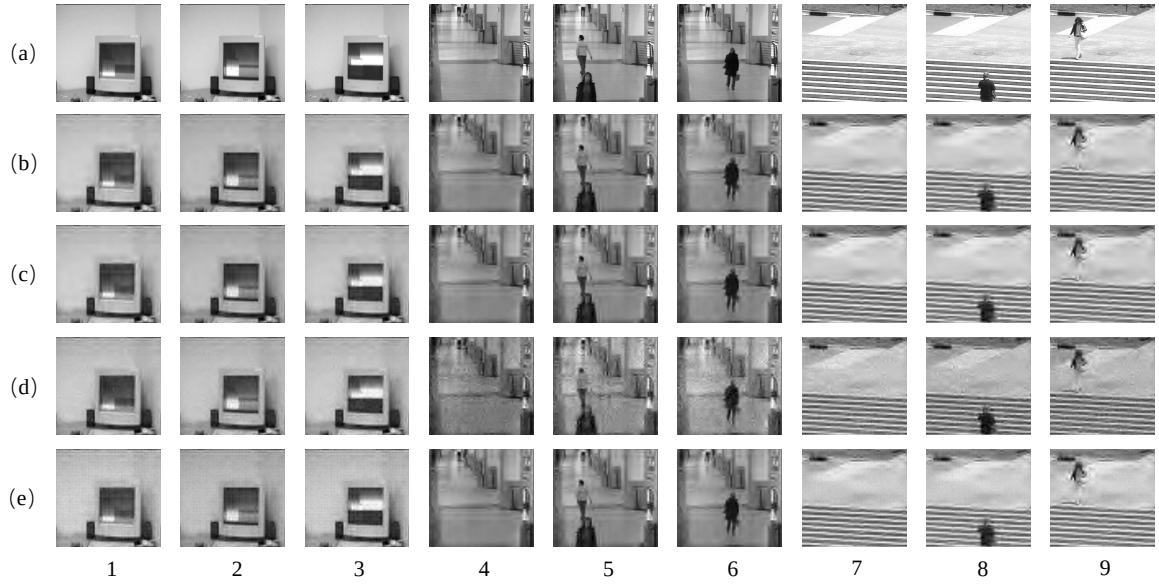
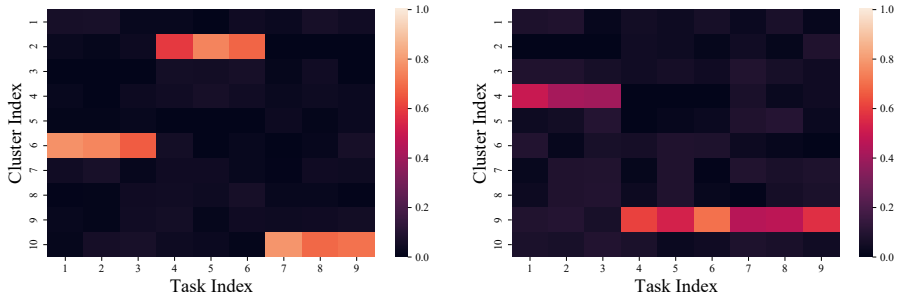


Figure 7. Reconstructed images from video images. (a) Linear, (b) EP-SBCS, (c) MT-AMP, (d) MT-CS, (e)WBDOA. Digits below the figure denotes the task index.

the proposed method converges faster and is more stable under different situations. Finally, Figs. 6(a)–(b) show the probability heatmap results of clustering in the proposed method and the WBDOA method respectively. While both methods are capable to automatically and correctly cluster all tasks, the proposed method offers lower NMSE than the WBDOA method, as shown in Table I.

2) Still Images from Video Sequence: In this example, experiments on 9 snapshots of 3 different scenes are used for performance comparison, and thus 9 tasks in total are considered. The first three snapshots, which correspond to computer scenes, are referred to task 1 to task 3; the three snapshots in the mall scenes are regarded as task 4 to task 6; and the remaining three snapshots are considered to be task 7 to task 9. The same stick-breaking truncation level of $L = 10$ is assumed. All the images are of size 240×256 and the sensing matrix is constructed in the same manner as in the above simulation experiments. A hybrid CS scheme [20] for image reconstruction is considered here, with a coarsest scale $j_0 = 3$ and a finest scale $j_1 = 6$ on the "Daubechies 8" wavelets. Coarse scale



(a) (b)

Figure 8. Probability heatmap results of clustering for video images. (a) EP-SBCS. (b) WBDOA.

coefficients denote the j_0^2 wavelet transform coefficients in the upper-left square and fine-scale coefficients denote the upper-left $j_1^2 - j_0^2$ coefficients except coarse-scale coefficients. Only fine-scale coefficients are reconstructed with no compression applied in the coarse scale, i.e., coarse scale coefficients a re r etained. T he reconstructed images using linear reconstruction with all 4096 measurements are shown in Fig. 7(a) which is the best possible performance. Figs. 7(b)–(e) show the reconstructed results for the EP-SBCS, MT-AMP, WBDOA, and MT-CS, respectively, where the number of measurements is $P = 1717$ for each task. Note that both MT-AMP and MT-CS methods are used to recover images over the same scene classified m anually a s o ne c luster i n a dvance. T able II shows the full performance comparison. Again, the proposed EP-SBCS method reduces the reconstruction error compared to the other three methods shown in Table II, which is consistent with the conclusions drawn from the previous examples. Fig. 8(a) shows the probability heatmap for EP-SBCS in which all snapshots are clustered correctly according to different scenes. The probability heatmap for WBDOA is shown in Fig. 8(b), where tasks 4–6 and 7–9 are clustered into one class. Therefore, the WBDOA method leads to a larger NMSE than that of the proposed EP-SBCS method due to the incorrect clustering, as evident in Table II.

3) Face Image Sequences: In this example, more realistic scenario is used for the performance comparisons. Unlike the still image based experiments above from the diverse scenes above, we take the face image sequences from the extended Yale Face Database B into consideration [47], [48]. 9 images of the identical human subject under different illumination conditions are used. All the images are cropped to size 168×192 and the original images are shown as Figs. 9(a). The stick-breaking truncation level $L = 10$ is still assumed. Note that all images are captured from the same subject and highly correlated, and thus both MT-AMP and MT-CS method are used to recover images by sharing the identical sparse pattern across all the tasks. Figs. 9(b)–(e) show the reconstructed results for the EP-SBCS, MT-AMP, WBDOA, and MT-CS methods, respectively. Their full performances are shown in Table III. It is found that the proposed EP-SBCS method has the lowest NMSE across all the tasks, compared to the other three methods shown in Table III, which is consistent with the conclusions drawn from the previous examples. Fig. 10(a) shows the probability heatmap for the EP-SBCS method, and it is observed that the proposed method classifies images into three classes. It is reasonable that tasks 1–4 are clustered into one class due to relatively bright illumination conditions, and the tasks 7–9 are also classified into one class due to their dark illumination conditions.

Figure 9. Reconstructed images from face image sequences. (a) Origin, (b) EP-SBCS, (c) MT-AMP, (d) WBDOA, (e)MT-CS. Digits below the figure denotes the task index.

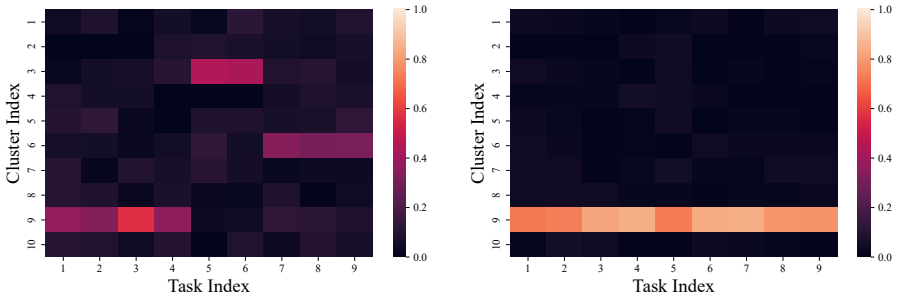
However, it seems that the WBDOA method recovers images by sharing the identical sparse pattern across images, and the probability heatmap in the shown in Fig. 10(b). The reason could be that all the tasks are highly correlated, and thus are classified into one class in the WBDOA method. Therefore, the proposed EP-SBCS method leads to a smaller NMSE than that of WBDOA method due to the appropriate structure clustering, as evident in Table III.

Table III
PERFORMANCE OF RECONSTRUCTED FACE IMAGE SEQUENCE.

	NMSE									Time(s)
	Task 1	Task 2	Task 3	Taks 4	Task 5	Task 6	Task 7	Task 8	Task 9	Average
EP-SBCS	**0.0669**	**0.0554**	**0.0506**	**0.0382**	**0.0415**	**0.0499**	**0.0406**	**0.0401**	**0.0406**	33.6842
MT-AMP	0.0684	0.0711	0.0697	0.0718	0.0636	0.0727	0.0594	0.0526	0.0589	46.7693
WBDOA	0.1965	0.1955	0.1943	0.1860	0.1907	0.1834	0.1724	0.1836	0.1698	67.9951
MT-CS	0.2324	0.2513	0.2471	0.2525	0.2490	0.2423	0.2591	0.2560	0.2572	**20.7707**

V. CONCLUSION

This paper developed a hierarchical Bayesian framework to analyze the problem of simultaneously inferring both the underlying structures and clusters of multiple related signals, leading to enhanced sparse reconstruction performance. Specifically, by using Gaussian process, spike-and-slab priors were extended to encode the inner statistical correlation within each task. To model the clustering mechanisms among tasks, the DP priors were employed on the support of each task. This represents a nonparametric approach where the number of clusters is not set *a priori* but is rather inferred from the data set. Within this framework, an inference algorithm based on the expectation propagation scheme was developed which re-estimates each factor with the remaining of other



(a) (b)

Figure 10. Probability heatmap results of clustering for face image sequences. (a) EP-SBCS. (b) WBDOA.

factors sequentially and iteratively. By applying the EP framework, the approximate full posterior of the hierarchical Bayesian model with spike-and-slab priors became analytically tractable and retained the sufficient complexity to capture the critical characteristics of the true density. This EP-based algorithm was compared with the state-of-the-art algorithms including MT-AMP, MT-CS, and WBDOA methods, and improved reconstruction performance for multi-task problems was verified. The advantages of the proposed algorithm were more pronounced in recovering multiple tasks with different sparse patterns. The proposed method automatically clustered signals accounting for the statistical correlations inherent to multi-task learning, regardless of whether these correlations were within individual tasks or across tasks of each cluster. In so doing, it reduced the number of measurements required for any task and improved the accuracy of the reconstructed signals.

REFERENCES

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

[2] M. G. Amin and F. Ahmad, "Compressive sensing for through-the-wall radar imaging," *Journal of Electronic Imaging*, vol. 22, no. 3, p. 030901, 2013.

[3] Q. Wu, Y. D. Zhang, M. G. Amin, and B. Himed, "High-resolution passive SAR imaging exploiting structured bayesian compressive sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 8, pp. 1484–1497, 2015.

[4] Q. Wu, Y. D. Zhang, M. G. Amin, and B. Himed, "Multi-static passive SAR imaging based on Bayesian compressive sensing," in *SPIE Compressive Sensing III*, vol. 9109, 2014, p. 910902.

[5] M. Carlin, P. Rocca, G. Oliveri, F. Viani, and A. Massa, "Directions-of-arrival estimation through Bayesian compressive sensing strategies," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 7, pp. 3828–3838, 2013.

[6] S. Qin, Y. D. Zhang, and M. G. Amin, "Generalized coprime array configurations for direction-of-arrival estimation," *IEEE Transactions on Signal Processing*, vol. 63, no. 6, pp. 1377–1390, 2015.

[7] L. Wang, L. Zhao, G. Bi, C. Wan, L. Zhang, and H. Zhang, "Novel wideband DOA estimation based on sparse Bayesian learning with Dirichlet process priors," *IEEE Transactions on Signal Processing*, vol. 64, no. 2, pp. 275–289, 2015.

[8] Y. Wiaux, L. Jacques, G. Puy, A. M. M. Scaife, and P. Vandergheynst, "Compressed sensing imaging techniques for radio interferometry," *Monthly Notices of the Royal Astronomical Society*, vol. 395, no. 3, pp. 1733–1742, 05 2009.

[9] S. Liu, Y. D. Zhang, and T. Shan, "Detection of weak astronomical signals with frequency-hopping interference suppression," *Digital Signal Processing*, vol. 72, pp. 1 – 8, 2018.

[10] S. Zhang, Y. Gu, C. Won, and Y. D. Zhang, "Dimension-reduced radio astronomical imaging based on sparse reconstruction," in *2018 IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2018, pp. 470–474.

[11] Y. D. Zhang, M. G. Amin, and B. Himed, "Reduced interference time-frequency representations and sparse reconstruction of undersampled data," in *21st European Signal Processing Conference (EUSIPCO 2013)*, 2013, pp. 1–5.

[12] E. Sejdić, I. Orović, and S. Stanković, "Compressive sensing meets time–frequency: an overview of recent advances in time–frequency processing of sparse signals," *Digital signal processing*, vol. 77, pp. 22–35, 2018.

[13] S. Zhang and Y. D. Zhang, "Low-rank hankel matrix completion for robust time-frequency analysis," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6171–6186, 2020.

[14] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[15] J. Tropp, A. C. Gilbert *et al.*, "Signal recovery from partial information via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[16] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit," *IEEE transactions on Information Theory*, vol. 58, no. 2, pp. 1094–1121, 2012.

[17] T. Park and G. Casella, "The Bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[18] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on signal processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[19] L. Yu, H. Sun, J.-P. Barbot, and G. Zheng, "Bayesian compressive sensing for cluster structured sparse signals," *Signal Processing*, vol. 92, no. 1, pp. 259–269, 2012.

[20] S. Ji, D. Dunson, and L. Carin, "Multitask compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 92–106, 2008.

[21] X.-T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4349–4360, 2012.

[22] Q. Wu, Y. D. Zhang, M. G. Amin, and B. Himed, "Complex multitask Bayesian compressive sensing," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3375–3379.

[23] Y. Qi, D. Liu, D. Dunson, and L. Carin, "Multi-task compressive sensing with Dirichlet process priors," in *Proceedings of 25th International Conference on Machine Learning*, 2008, pp. 768–775.

[24] Z. Zhang, T.-P. Jung, S. Makeig, Z. Pi, and B. D. Rao, "Spatiotemporal sparse bayesian learning with applications to compressed sensing of multichannel physiological signals," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 22, no. 6, pp. 1186–1197, 2014.

[25] L. He and L. Carin, "Exploiting structure in wavelet-based bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 9, pp. 3488–3497, 2009.

[26] D. M. Blei, M. I. Jordan *et al.*, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–143, 2006.

[27] S. Baillet, J. C. Mosher, and R. M. Leahy, "Electromagnetic brain mapping," *IEEE Signal processing magazine*, vol. 18, no. 6, pp. 14–30, 2001.

[28] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[29] M. R. Andersen, O. Winther, and L. K. Hansen, "Bayesian inference for structured spike and slab priors," in *Advances in Neural Information Processing Systems*, 2014, pp. 1745–1753.

[30] M. R. Andersen, A. Vehtari, O. Winther, and L. K. Hansen, "Bayesian inference for spatio-temporal spike-and-slab priors," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 5076–5133, 2017.

[31] J. Liu, Q. Wu, and M. Amin, "Multi-task Bayesian compressive sensing exploiting signal structures," *Signal Processing*, vol. 178, p. 107804, 2021.

[32] Y. Wang, D. Wipf, J.-M. Yun, W. Chen, and I. Wassell, "Clustered sparse bayesian learning," in *Uncertainty in Artificial Intelligence (UAI)*, July 2015.

[33] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the american statistical association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[34] Y. Whye Teh, M. Jordan, M. Beal, and D. Blei, "Sharing clusters among related groups: Hierarchical dirichlet processes," in *NIPS04 Proceedings of the 17th International Conference on Neural Information Processing Systems*, 2004, pp. 1385–1392.

[35] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 2012, vol. 118.

[36] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[37] W. Fan and N. Bouguila, "Infinite Dirichlet mixture models learning via expectation propagation," *Advances in Data Analysis and Classification*, vol. 7, no. 4, pp. 465–489, 2013.

[38] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, "Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 130–144, 2011.

[39] Y. W. Teh, "Dirichlet processes," in *Encyclopedia of Machine Learning*. Springer, 2010.

[40] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, pp. 639–650, 1994.

[41] D. Görür and C. E. Rasmussen, "Dirichlet process gaussian mixture models: Choice of the base distribution," *Journal of Computer Science and Technology*, vol. 25, no. 4, pp. 653–664, 2010.

[42] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[43] T. Minka, "Estimating a Dirichlet distribution," Tech. Rep., 2000.

[44] T. Minka *et al.*, "Divergence measures and message passing," Technical report, Microsoft Research, Tech. Rep., 2005.

[45] J. Liu, Q. Wu, and Y. D. Zhang, "Multi-task adaptive matching pursuit for sparse signal recovery exploiting signal structures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 4998–5002.

[46] Y. Lecun, "The MNIST database of handwritten digits," *http://yann.lecun.com/exdb/mnist/*, 1998.

[47] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[48] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.