MULTI-TASK ADAPTIVE MATCHING PURSUIT FOR SPARSE SIGNAL RECOVERY EXPLOITING SIGNAL STRUCTURES

Jiahao Liu[†]*, Qisong Wu*[†] *and Yimin D. Zhang*[‡]

[†]Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education,
Southeast University, Nanjing, 210096, China
[‡] Department of Electrical and Computer Engineering,
Temple University, Philadelphia, PA, 19122, USA

ABSTRACT

Multi-task compressive sensing is a framework that, by leveraging the useful information contained in multiple tasks, significantly reduces the number of measurements required for sparse signal recovery and achieves improved sparse reconstruction performance of all tasks. In this paper, a novel multi-task adaptive matching pursuit (MT-AMP) algorithm based on a hierarchical Bayesian model is proposed with the exploitation of both the group structure across different tasks and the intra-group correlation, yielding an effective means to simultaneously perform sparse recovery as well as learn the statistical inter-task and intra-group relationships. Experimental results using both synthetic data and real data sets demonstrate the superiorities of the proposed method over existing state-of-the-art algorithms.

Index Terms— Compressive sensing, multi-task learning, structured spike-and-slab prior, sparse recovery, intra-group correlation

1. INTRODUCTION

Sparse signal recovery and compressive sensing (CS) have attracted significant attention in recent years [1]. CS techniques are capable to recover signals from a small number of measurement samples with a high probability, given that the signals are sparse or can be sparsely represented in some domain. They have been widely used in many applications, such as radar imaging [2, 3], radio astronomy [4–6], dictionary learning [7, 8], and image classification [9].

Advances in sensing technology have facilitated easy acquisition of multiple different measurements of the same underlying physical phenomena. For example, in face recognition or action recognition we may have different views of a person's face captured under different illumination conditions or with different facial postures [10]. In automatic target recognition, multiple synthetic aperture radar (SAR) views are acquired [11]. It has been shown that recovering multiple related tasks simultaneously, rather than individually, often significantly improve performance relative to recovering each task independently [12]. This is the case, for example, when only a few data per task are available, so that there is an advantage in "pooling" together data across many related tasks. A typical multi-task CS model can be described as [13],

$$\mathbf{y}_t = \mathbf{A}_t\mathbf{x}_t + \mathbf{n}_t, \;\; t \in \{1, \cdots, D\}, \tag{1}$$

where $\mathbf{y}_t \in \mathbb{R}^M$ denotes the measurement vector at time t, and $\mathbf{n}_t$ is an unknown zero-mean Gaussian noise vector. $\mathbf{A}_t \in \mathbb{R}^{M\times L}$ is a sensing dictionary matrix at time t with $M \ll L$, and without loss of generality, all atoms of $\{\mathbf{A}_t\}_{t=1}^D$ are normalized, i.e., the lth atom in the tth sensing dictionary matrix $\|\mathbf{a}_{lt}\|_2^2 = 1$ for $l = \{1, \cdots, L\}$ and $t \in \{1, \cdots, D\}$, where $\mathbf{a}_{lt}$ is the lth column of matrix $\mathbf{A}_t$. In addition, $\mathbf{x}_t \in \mathbb{R}^{L\times 1}$ is the required reconstructed sparse vector with K non-zero entries at time t. In this model, $\mathbf{x}_t$ has the same or a similar sparsity support, i.e., the respective positions of the non-zero entries are similar for different t. Denote x_{lt} as the lth elment of $\mathbf{x}_t$. Then, the above group sparsity suggests that $\mathbf{x}_{l\cdot} = [x_{l1}, \cdots, x_{lD}]$, which aligns the lth element across all D tasks, shares the same sparsity pattern, and the values of these elements are often generally dependent or correlated due to the correlation of observed measurements.

A number of algorithms have been proposed for the reconstruction of group sparse signals. These algorithms include greedy-based algorithms, such as block-OMP (BOMP) [14], and basis pursuit-based ones, such as group Lasso [15]. Bayesian approaches form a different class of sparse signal reconstruction algorithms, which generally yield improved performance. Sparse Bayesian learning algorithms provide effective solutions to a large class of problems based on a nonparametric Bayesian framework, and thus have the capability of inferring the sparsity parameter and avoiding the nuisance parameters [16–18]. The multi-task compressive sensing (MT-CS) algorithm [17] adopts a hierarchical Bayesian model for multi-task recovery with group structure and a more general approach complex multi-task Bayesian compressive sensing (CMT-BCS) [19] has been proposed for the recovery of complex signals. The clustered multi-task Bayesian compressive sensing algorithm further uses the intra-task dependency to improve the reconstruction performance [20].

In this paper, a novel multi-task adaptive matching pursuit method is proposed that exploits the signals structures for sparse signal reconstruction in the hierarchical Bayesian framework. We first extend a spike-and-slab prior to form a generalized multitask model and induce the relationship between the tasks based on a hierarchical model. Motivated by the kernel technique in the correlation learning model [21, 22], we place a Toeplitz matrix to learn the correlation between the elements within each group. A novel greedy-based inference approach is then proposed for the non-convex optimization model induced by the extended spike-and-slab prior. Since the hierarchical Bayesian model allows the estimation of the prior and the correlation parameters in an unsupervised manner,

The work of J. Liu and Q. Wu was supported in part by Contacts 61701109,11704069, 11674057 and 11574048 with National Natural Science Foundation, in part by Contract BK20160701 with National Natural Science Foundation of Jiangsu Province. The work of Y. D. Zhang was supported in part by the National Science Foundation under Grant No. AST-1547420. Correspondence email: qisong.wu@seu.edu.cn.

the proposed algorithm has the capability of automatically inferring the sparsity and learning both the group structure across tasks and the intra-group correlation.

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 the variable w given a and b. $\mathcal{N}(w|a,b)$ denotes that random variable w follows a Gaussian distribution with mean a and variance b, $\mathbf{I}_M$ denotes the $M \times M$ identity matrix, and $\mathbb{I}(w=0)$ stands for an indicator function, which equals to 1 if and only if $w=0$, and is 0 otherwise. We use $\mathbf{x}_{i\cdot}$ to denote the ith row of matrix $\mathbf{X}$, and $(.)^T$ denotes the transpose of a matrix or vector.

2. THE PROPOSED MODEL

2.1. Generative Model

In this subsection, we illustrate the proposed generative model. Without loss of generality, the measurement vectors follow the Gaussian distribution, i.e.,

$$\mathbf{Y}|\mathbf{X},\sigma^2 \sim \prod_{t=1}^{D} \mathcal{N}(\mathbf{A}_t\mathbf{x}_t, \sigma^2\mathbf{I}_M), \tag{2}$$

where $\mathbf{Y} = [\mathbf{y}_1,\cdots,\mathbf{y}_t,\cdots,\mathbf{y}_D] \in \mathbb{R}^{M\times D}$ is an observed measurement matrix, $\mathbf{X} = [\mathbf{x}_1,\cdots,\mathbf{x}_t,\cdots,\mathbf{x}_D] \in \mathbb{R}^{L\times D}$ denotes the required reconstructed sparse matrix across D tasks, and σ^2 represents the noise variance.

To encourage the group sparsity across all tasks and model the correlation between the elements within each group, we place an extended spike-and-slab prior with the structure exploitation, expressed as,

$$\mathbf{x}_{l\cdot} \sim \gamma_l\mathcal{N}(0,\sigma^2\lambda^{-1}\mathbf{B}) + (1-\gamma_l)\mathbb{I}(\mathbf{x}_{l\cdot}=\mathbf{0}), \tag{3}$$

where λ is a scalar parameter related to the noise precision, and γ_l is a binary indicator variable, which follows the Bernoulli distribution. This prior implies that all the elements in the lth group share the identical prior γ_l. When the value of γ_l is 1, all the elements in the lth group are nonzero. On the other hand, when the value of γ_l is 0, all the elements in this group will be zero. Therefore, we have the Bernoulli distribution imposed on the vector $\boldsymbol{\gamma} = [\gamma_1,\cdots,\gamma_L]^T$ as

$$\boldsymbol{\gamma}|\boldsymbol{\kappa} \sim \prod_{l=1}^{L} \text{Bernoulli}(\gamma_l|\kappa_l), \tag{4}$$

where $\boldsymbol{\kappa} = [\kappa_1,\cdots,\kappa_L]^T$ is a hyperparameter vector.

It has been shown that in many applications the multiple tasks are often correlated, and the signal sparse reconstruction performance can be significantly improved by exploiting this correlation [23–25]. In addition, in many applications stronger correlation exists between tasks nearby and weaker correlation exists between tasks far away. Inspired by the kernel technique in the correlation learning model [21, 22], we define matrix $\mathbf{B} \in \mathbb{R}^{D\times D}$ as a positive definite kernel matrix to capture the correlation of elements within each group. To avoid overfitting, all L groups share the identical $\mathbf{B}$. A first-order auto-regressive (AR) process is often sufficient to model intra-group correlation. In this case, a Toeplitz kernel matrix is imposed to model the correlation between elements within each group,

$$\text{Toeplitz}([1,c,\cdots,c^{D-1}]) = \begin{bmatrix} 1 & c & \cdots & c^{D-1} \\ \vdots & & & \vdots \\ c^{D-1} & c^{D-2} & \cdots & 1 \end{bmatrix},$$

where $0 \le c < 1$ is the AR coefficient. Notice that the above kernel matrix is real and symmetric, and all its entries take values within $[0,1]$. The diagonal entries take the value of unity, and the values of the off-diagonal elements decrease, depending on their respective distance from the main diagonal. When the scalar c is zero, this kernel matrix reduces to the identity matrix, and thus it implies that all the elements are independent. On the other hand, when c approaches to 1, it shows a strong correlation between elements within groups. In the proposed model, the intra-group correlation will be automatically learned by estimating the AR coefficient c.

Compared with the approaches in [26] and [27], the above generative model is more general because it takes the underlying group structure into consideration and exploits intra-group correlation learning to improve the reconstruction performance.

2.2. Optimization Problem

According to the generative model above, the posterior probability distribution of latent random variables $\mathbf{X}, \boldsymbol{\gamma}$ and c can be expressed by

$$f(\mathbf{X},\boldsymbol{\gamma},c|\mathbf{Y},\lambda,\sigma^2,\boldsymbol{\kappa}) \propto f(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma},\sigma^2)f(\mathbf{X}|\boldsymbol{\gamma},\sigma^2,\lambda,c)f(\boldsymbol{\gamma}|\boldsymbol{\kappa}). \tag{5}$$

A maximum a posteriori (MAP) estimation is performed, and the optimal values of $\mathbf{X}^*, \boldsymbol{\gamma}^*$ and c^* are given by [28]:

$$(\mathbf{X}^*,\boldsymbol{\gamma}^*,c^*) = \arg\max_{\mathbf{X},\boldsymbol{\gamma},c} f(\mathbf{X},\boldsymbol{\gamma},c \mid \mathbf{Y},\boldsymbol{\kappa},\lambda,\sigma^2). \tag{6}$$

The above optimization problem is equivalent to the following minimization problem,

$$(\mathbf{X}^*,\ \boldsymbol{\gamma}^*,\ c^*) = \arg\min_{\mathbf{X},\boldsymbol{\gamma},c} L(\mathbf{X},\ \boldsymbol{\gamma},\ c), \tag{7}$$

where

$$\begin{aligned} L(\mathbf{X},\ \boldsymbol{\gamma},\ c) &= \sum_{t=1}^{D} \|\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t\|_2^2 \\ &+ \lambda\sum_{l=1}^{L} \mathbf{x}_{l\cdot}\mathbf{B}^{-1}\mathbf{x}_{l\cdot}^T + \sum_{l=1}^{L}\rho_l\gamma_l, \end{aligned} \tag{8}$$

and $\rho_l = \sigma^2 \ln\left((2\pi)^D|\sigma^2\lambda^{-1}\mathbf{B}|\frac{(1-\kappa_l)^2}{\kappa_l^2}\right)$.

Because it involves a binary indicator variable $\boldsymbol{\gamma}$, this optimization problem is non-convex and thus cannot be effectively solved using conventional convex optimization algorithms. We propose a greedy-based multi-task adaptive matching pursuit (MT-AMP) algorithm to handle this non-convex problem and the alternative minimization scheme [29] is adopted, as illustrated in Section 3.

3. MULTI-TASK ADAPTIVE MATCHING PURSUIT

3.1. Signal Recovery: Optimization for $\mathbf{X}$ and $\boldsymbol{\gamma}$

In this subsection, given the parameter c, the optimization of $\mathbf{X}$ and $\boldsymbol{\gamma}$ can be expressed as

$$\begin{aligned} (\mathbf{X}^*,\ \boldsymbol{\gamma}^*) &= \arg\min_{\mathbf{X},\boldsymbol{\gamma}} \sum_{t=1}^{D} \|\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t\|_2^2 \\ &+ \lambda\sum_{l=1}^{L} \mathbf{x}_{l\cdot}\mathbf{B}^{-1}\mathbf{x}_{l\cdot}^T + \sum_{l=1}^{L}\rho_l\gamma_l. \end{aligned} \tag{9}$$

A greedy-based MT-AMP is proposed to effectively solve this problem by adding elements into or removing elements from the support set of $\mathbf{X}$. It is clear that, given the support set of $\mathbf{X}$, i.e., $\mathcal{S} = \{l : \gamma_l \neq 0\}$ for $l \in \{1, \cdots, L\}$, the optimization problem in Eq. (9) will reduce to

$$\mathbf{X}^{\mathcal{S}} = \arg\min_{\mathbf{X}^{\mathcal{S}}} \sum_{t=1}^{D} \|\mathbf{y}_t - \mathbf{A}_t^{\mathcal{S}}\mathbf{x}_t^{\mathcal{S}}\|_2^2 + \lambda \sum_{l\in\mathcal{S}} \mathbf{x}_{l\cdot}\mathbf{B}^{-1}\mathbf{x}_{l\cdot}^T, \quad (10)$$

where $\mathbf{X}^{\mathcal{S}}$ ($\mathbf{A}_t^{\mathcal{S}}$) is a sub-matrix of $\mathbf{X}$ ($\mathbf{A}_t$) composed by the rows (columns) of $\mathbf{X}$ ($\mathbf{A}_t$) indexed by $\mathcal{S}$. $\mathbf{x}_t^{\mathcal{S}}$ is the vector containing only the active elements of $\mathbf{x}_t$ which are indexed by $\mathcal{S}$. This optimization problem can be easily solved using conventional convex optimization algorithms. The bijection $\mathcal{S} \leftrightarrow \mathbf{X}^{\mathcal{S}}$ implies that solving Eq. (9) is equivalent to finding the support set $\mathcal{S}$. This prompts us to utilize a greedy-based method to find the support set $\mathcal{S}$ and then solve the problem Eq.(10). In particular, we update $\mathcal{S}$ at each iteration by either adding one of the unselected indices into $\mathcal{S}$ or removing one of the existing indices from $\mathcal{S}$. For a given $\mathcal{S}$, we define

$$\mathbf{r}_t^{\mathcal{S}} = \mathbf{y}_t - \mathbf{A}_t^{\mathcal{S}}\mathbf{x}_t^{\mathcal{S}}, \quad \theta_{\mathcal{S}} = \sum_{l\in\mathcal{S}} \rho_l \quad (11)$$

and

$$g(\mathcal{S}) = \min_{\mathbf{X}^{\mathcal{S}}} \sum_{t=1}^{D} \|\mathbf{y}_t - \mathbf{A}_t^{\mathcal{S}}\mathbf{x}_t^{\mathcal{S}}\|_2^2 + \lambda \sum_{l\in\mathcal{S}} \mathbf{x}_{l\cdot}\mathbf{B}^{-1}\mathbf{x}_{l\cdot}^T + \theta_{\mathcal{S}}. \quad (12)$$

At each step, the choice of the index and the action of adding/removing elements are decided by computing the two values $U_{\mathcal{S}}$ and $V_{\mathcal{S}}$, where

$$U_{\mathcal{S}} = \min_{l\notin\mathcal{S}} g(\mathcal{S}\cup\{l\}) - g(\mathcal{S}), \quad (13)$$

is the reduction in the cost function if adding one of the unselected indices into $\mathcal{S}$, and

$$V_{\mathcal{S}} = \min_{j\in\mathcal{S}} g(\mathcal{S}\backslash\{j\}) - g(\mathcal{S}) \quad (14)$$

is the reduction of cost function if removing one of the existing indices from $\mathcal{S}$. If both $U_{\mathcal{S}}$ and $V_{\mathcal{S}}$ are greater than or equal to 0, we stop the algorithm since no further improvement is obtained and this suggests the algorithm has converged. Otherwise, we compare $U_{\mathcal{S}}$ and $V_{\mathcal{S}}$ to update $\mathcal{S}$ by adding l to $\mathcal{S}$ (if $U_{\mathcal{S}} < V_{\mathcal{S}}$) or removing j from $\mathcal{S}$ (if $U_{\mathcal{S}} > V_{\mathcal{S}}$), whichever further reduces the cost function.

Nevertheless, the cost of precisely computing the values of $U_{\mathcal{S}}$ and $V_{\mathcal{S}}$ is very expensive, making this idea hardly practical. To solve this problem, instead of directly estimating the values of $U_{\mathcal{S}}$ and $V_{\mathcal{S}}$, we compute their upper bounds, respectively denoted as $\overline{U}_{\mathcal{S}}$ and $\overline{V}_{\mathcal{S}}$, to reduce the computational cost. In particular, the values of $\overline{U}_{\mathcal{S}}$ and $\overline{V}_{\mathcal{S}}$ can be computed by

$$\begin{aligned}\overline{U}_{\mathcal{S}} &= \min_{l\notin\mathcal{S}}\{\rho_l - \phi_l^T(\mathbf{I}_D + \lambda\mathbf{B}^{-1})^{-1}\phi_l\},\\ \overline{V}_{\mathcal{S}} &= \min_{j\in\mathcal{S}}\{2\mathbf{x}_{j\cdot}\phi_j + \mathbf{x}_{j\cdot}(\mathbf{I}_D - \lambda\mathbf{B}^{-1})\mathbf{x}_{j\cdot}^T - \rho_j\},\end{aligned}$$

where $\phi_i = [\mathbf{a}_{i1}^T\mathbf{r}_1^{\mathcal{S}}, \cdots, \mathbf{a}_{iD}^T\mathbf{r}_D^{\mathcal{S}}]^T$. Based on these approximated values, an updated $\mathcal{S}$ is acquired. Given the updated $\mathcal{S}$, the new values of $\mathbf{X}^{\mathcal{S}}$ and $\mathbf{r}_t^{\mathcal{S}}$ can be calculated precisely before moving to the new iteration.

Besides $\overline{U}_{\mathcal{S}}$ and $\overline{V}_{\mathcal{S}}$, the initialization of $\mathcal{S}$ can also significantly influence the convergence of MT-AMP. A beneficial initialization guidance is provided in the following proposition.

Proposition 1. *If $\rho_l < 0$, then $l \in \mathcal{S}_{opt}$, where $\mathcal{S}_{opt}$ is the optimal support set.*

Proof. Assume that $l \notin \mathcal{S}_{opt}$ and $\rho_l < 0$, then

$$\begin{aligned} g(\mathcal{S}_{opt}\cup\ l) &\leq \sum_{t=1}^{D} \|\mathbf{r}_t^{\mathcal{S}_{opt}} - \mathbf{a}_{lt}\mathbf{x}_{lt}\|_2^2 + \lambda\mathbf{x}_{l\cdot}\mathbf{B}^{-1}\mathbf{x}_{l\cdot}^T\\ &+ \lambda\sum_{i\in\mathcal{S}_{opt}} \mathbf{x}_{i\cdot}\mathbf{B}^{-1}\mathbf{x}_{i\cdot}^T + \theta_{\mathcal{S}_{opt}} + \rho_l\\ &= g(\mathcal{S}_{opt}) + \mathbf{x}_{l\cdot}\mathbf{x}_{l\cdot}^T - 2\sum_{t=1}^{D}\mathbf{a}_{lt}^T\mathbf{r}_t^{\mathcal{S}_{opt}}x_{lt}\\ &+ \lambda\mathbf{x}_{l\cdot}\mathbf{B}^{-1}\mathbf{x}_{l\cdot}^T + \rho_l. \end{aligned}$$

Let

$$p(\mathbf{x}_{l\cdot}) = \mathbf{x}_{l\cdot}\mathbf{x}_{l\cdot}^T - 2\sum_{d=1}^{D}\mathbf{a}_{ld}^T\mathbf{r}_d^{\mathcal{S}_{opt}}x_{ld} + \lambda\mathbf{x}_{l\cdot}\mathbf{B}^{-1}\mathbf{x}_{l\cdot}^T + \rho_l, \quad (15)$$

where $p(\cdot) : \mathbb{R}^{1\times D} \to \mathbb{R}$. As the matrix $\mathbf{B}$ is a positive definite matrix, it is obvious that as all the elements of $\mathbf{x}_{l\cdot}$ approaches to positive infinity, $p(\mathbf{x}_{l\cdot})$ will equal to positive infinity. In addition, $p(\mathbf{0}) = \rho_l < 0$ and $p(\mathbf{x}_{l\cdot})$ is continuous. Thus there exists $\widehat{\mathbf{x}}_{l\cdot}$ whose all elements are nonzero such that $\rho_l < p(\widehat{\mathbf{x}}_{l\cdot}) < 0$. Then we can obtain

$$g(\mathcal{S}_{opt}\cup l) \leq g(\mathcal{S}_{opt}) + p(\widehat{\mathbf{x}}_{l\cdot}) \leq g(\mathcal{S}_{opt}), \quad (16)$$

which suggests that adding l into $\mathcal{S}_{opt}$ can decrease the cost function. However, it is contradict to the assumption $l \notin \mathcal{S}_{opt}$. This implies that if $\rho_l < 0$, then $l \in \mathcal{S}_{opt}$. □

According to Proposition 1 we can initialize $\mathcal{S}_0 = \{l : \rho_l < 0\}$.

3.2. Intra-Group Correlation Learning: Optimization for c

Given $\mathbf{X}^{\mathcal{S}}$ and $\mathcal{S}$, the updating step of c is performed at each iteration. We empirically calculate the value of c by $c \triangleq \frac{m_1}{m_0}$ [22], where m_0 is the average of elements along the main diagonal of matrix $\mathbf{B}$, whereas m_1 represents that along the main sub-diagonal. The learning rule of matrix $\mathbf{B}$ is obtained by setting $\frac{\partial L(\mathbf{X},\boldsymbol{\gamma},c)}{\partial\mathbf{B}} = 0$, which results in the following closed-form solution

$$\mathbf{B} = \frac{\lambda\sum_{l\in\mathcal{S}}\mathbf{x}_{l\cdot}^T\mathbf{x}_{l\cdot}}{\sigma^2\sum_{l\in\mathcal{S}}\gamma_l}. \quad (17)$$

Then parameter c can be updated using the new matrix $\mathbf{B}$. Therefore, the intra-group correlation will be automatically learned and estimated.

3.3. Summary of the Proposed Algorithm

The entire optimization procedure is summarized in Algorithm 1. Since the updating step of $\overline{U}_{\mathcal{S}}$ and $\overline{V}_{\mathcal{S}}$ guarantees the reduction of objective function after each iteration, the algorithm can converge within finite steps.

Algorithm 1 Multi-Task Adaptive Matching Pursuit

Input: $\mathbf{A}, \mathbf{y}, \lambda, \sigma^2$.

Initialize c and $\mathcal{S}$
while true **do**
  Update $\mathbf{X}^{\mathcal{S}}$ by Eq. (10)
  Calculate: $[\overline{U}_{\mathcal{S}}, l]$ and $[\overline{V}_{\mathcal{S}}, j]$
  if $\min(\overline{U}_{\mathcal{S}}, \overline{V}_{\mathcal{S}}) > 0$ **then** break while
  else if $\overline{U}_{\mathcal{S}} < \overline{V}_{\mathcal{S}}$ **then** $\mathcal{S} = \mathcal{S} \cup \{l\}$
  else $\mathcal{S} = \mathcal{S} \backslash \{j\}$
  end if
  Update $\mathbf{B}$ by Eq. (17)
  Update $c : c = \frac{m_1}{m_0}$
end while

Output: $\mathbf{X}$ and c.

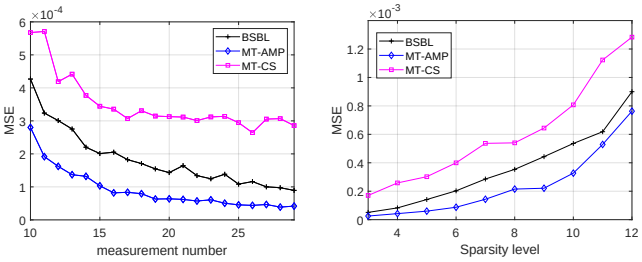

Fig. 1. Performance comparison. (a) MSE versus the number of measurements. (b) MSE versus the sparsity level.

4. SIMULATIONS AND EXPERIMENTS

In this section, experiments on both synthetic data and real data sets are performed to verify the effectiveness of the proposed method. In the simulations, $\lambda = 0.001, \sigma = 0.1$ and all the elements of vector $\boldsymbol{\kappa}$ are set as 0.5 in the entire experiments so that the binary variable vector γ is active in the initialization. Two competitive state-of-the-art methods, including MT-CS [17] and block sparse Bayesian learning (BSBL) [22], are compared, and the mean square error (MSE) is used as the performance index. All experimental results were obtained by average of 100 trials.

4.1. One-Dimensional Synthetic Data

In this subsection, the following parameters are used in the synthetic data: The length of each task is $L = 64$ with 5 nonzero elements, $D = 8$ tasks share the identical nonzero positions, but take random values following a Gaussian distribution with zero mean and the variance of 1. The measurement matrices $\mathbf{A}_t \in \mathbb{R}^{20\times 64}$ for $t = \{1, \cdots, D\}$ are randomly generated Gaussian matrices. With loss of generality, a Gaussian noise with zero mean is considered, and the signal-to-noise ratio (SNR) is 3dB.

The performance comparisons versus the number of measurements and the sparsity level are shown in Fig. 1(a) and Fig. 1(b), respectively. It is observed that the reconstructed MSE decreases with the increase of the measurement number, and the proposed method has the smallest MSE across all the level of measurement number in Fig. 1(a), compared to these in MT-CS and BSBL. On the other hand, performance comparison versus the sparsity level is also shown in Fig. 1(b). It is clear that the proposed method achieves the best reconstruction performance, which benefits from the spike-and-slab prior with the structure exploitation.

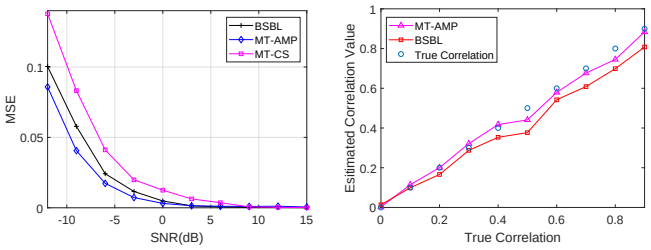

Fig. 2. Performance comparison. (a) MSE versus the input SNR. (b) Comparison of the estimated correlation values.

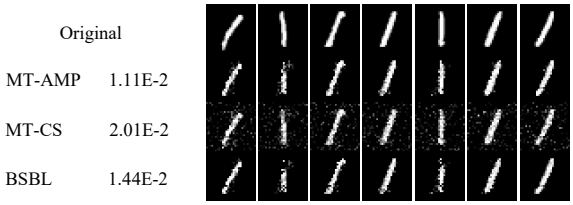

Fig. 3. Reconstructed images from MNIST data set. The numbers appeared next to each method is the average MSE.

Fig. 2(a) shows the performance comparison versus SNR. It is observed that the MSE decreases with the increase of SNR, and the MSE obtained by the proposed method is less than these in other two methods, and thus is more robust to SNR. To verify the accuracy of intra-group correlation learning of the proposed method, we show the estimated value of correlation parameter c, as shown in Fig. 2(b). It can be found that the estimated values of MT-AMP are closer to the true values, compared to the estimated value in the BSBL method with the true value of the correlation varying from 0 to 0.9.

4.2. Real Image Recovery

In this subsection, real data sets of handwritten digit images from the MNIST data set [30] are used for performance comparison. The reconstructed results using 7 tasks are shown in Fig. 3. All original images have the size 28×28. The experiment parameter setting is similar to the first set of examples, and the number of measurements is 350. It is observed that the sparsity support and values across 7 tasks are highly correlated. As shown in Fig. 3, the images recovered by the MT-AMP are closest to the original images and have the smallest MSE by exploiting the group structure across tasks and intra-group correlation learning, compared to those obtained by the MT-CS and BSBL methods.

5. CONCLUSION

In this paper, a novel multi-task adaptive matching pursuit (MT-AMP) method has been proposed for sparse recovery in a hierarchical Bayesian framework. To encourage the group sparsity, we impose an extended spike-and-slab prior to model the group structure across all the tasks. In addition, a Toeplitz matrix structure is used to model the correlation between the elements within groups. According to the proposed generative model, a greedy based adaptive matching pursuit is then proposed to perform the inference for this non-convex optimization problem. Simulations and experimental results demonstrate the superiority of the proposed algorithm over other existing state-of-the-art algorithms.

6. REFERENCES

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

[2] 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.

[3] Q. Wu, Y. D. Zhang, F. Ahmad, and M. G. Amin, "Compressive-sensing-based high-resolution polarimetric through-the-wall radar imaging exploiting target characteristics," *IEEE Antennas and Wireless Propagation Letters*, vol. 14, pp. 1043–1047, 2015.

[4] 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, 2009.

[5] 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.

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

[7] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 24–38, Jan. 2016.

[8] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten, "Dictionary learning for sparse representation: A novel approach," *IEEE Signal Processing Letters*, vol. 20, no. 12, pp. 1195–1198, 2013.

[9] H. S. Mousavi, U. Srinivas, V. Monga, Y. Suo, M. Dao, and T. D. Tran, "Multi-task image classification via collaborative, hierarchical spike-and-slab priors," in *Proceedings of IEEE International Conference on Image Processing*, 2014, pp. 4236–4240.

[10] X. 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.

[11] H. Zhang, N. M. Nasrabadi, Y. Zhang, and T. S. Huang, "Multi-view automatic target recognition using joint sparse representation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 3, pp. 2481–2497, 2012.

[12] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *Journal of Machine Learning Research*, vol. 6, no. Nov, pp. 1817–1853, 2005.

[13] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[14] L. Jacob, G. Obozinski, and J. P. Vert, "Group Lasso with overlap and graph Lasso," in *Proceedings of International Conference on Machine Learning*, Montreal, Quebec, Canada, June 2009, pp. 433–440.

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

[16] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, no. Jun, pp. 211–244, 2001.

[17] S. Ji, D. Dunson, and L. Carin, "Multitask compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 92–106, 2009.

[18] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Bayesian compressive sensing using laplace priors.," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 53–63, 2009.

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

[20] Q. Wu, Y. D. Zhang, M. G. Amin, and B. Himed, "Mutli-task Bayesian compressive sensing exploiting intra-task dependency," *IEEE Signal Processing Letters*, vol. 22, no. 4, pp. 430–434, 2015.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York: Springer, 2006.

[22] Z. Zhang and B. D. Rao, "Extension of SBL algorithms for the recovery of block sparse signals with intra-block correlation," *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 2009–2015, 2013.

[23] J. Zhang, Z. Ghahramani, and Y. Yang, "Learning multiple related tasks using latent independent component analysis," in *Proceedings of Advances in Neural Information Processing Systems*, 2006, pp. 1585–1592.

[24] B. Bakker and T. Heskes, "Task clustering and gating for Bayesian multitask learning," *Journal of Machine Learning Research*, vol. 4, no. May, pp. 83–99, 2003.

[25] S. Ben-David and R. Schuller, "Exploiting task relatedness for multiple task learning," in *Learning Theory and Kernel Machines*, pp. 567–580. Springer, 2003.

[26] T. H. Vu, H. S. Mousavi, and V. Monga, "Adaptive matching pursuit for sparse signal recovery," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 4331–4335.

[27] H. S. Mousavi, V. Monga, and T. D. Tran, "Iterative convex refinement for sparse recovery," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1903–1907, 2015.

[28] N. M. Nasrabadi, "Pattern recognition and machine learning," *Journal of Electronic Imaging*, vol. 16, no. 4, pp. 049901, 2007.

[29] I. Csiszar, "Information geonetry and alternating minimization procedures," *Statistics & Decisions*, vol. 1, pp. 205–237, 1984.

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