Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Flexible multi-task learning with latent task grouping

Shi Zhong [a], Jian Pu [b], Yu-Gang Jiang [a], Rui Feng [a], Xiangyang Xue [a]

[a] Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China
[b] Institute of Neuroscience, Chinese Academy of Sciences, Shanghai, China

ARTICLE INFO

Article history:
Received 20 October 2015
Received in revised form
21 December 2015
Accepted 27 December 2015
Communicated by Peng Cui

Keywords:
Multi-task learning
Group structure
Regularization

ABSTRACT

In multi-task learning, using task grouping structure has been shown to be effective in preventing inappropriate knowledge transfer among unrelated tasks. However, the group structure often has to be predetermined using prior knowledge or heuristics, which has no theoretical guarantee and could lead to unsatisfactory learning performance. In this paper, we present a *flexible* multi-task learning framework to identify *latent* grouping structures under agnostic settings, where the prior of the latent subspace is unknown to the learner. In particular, we relax the latent subspace to be full rank, while imposing sparsity and orthogonality on the representation coefficients of target models. As a result, the target models still lie on a low dimensional subspace spanned by the selected basis tasks, and the structure of the latent task subspace is fully determined by the data. The final learning process is formulated as a joint optimization procedure over both the latent space and the target models. Besides providing proofs of theoretical guarantee on learning performance, we also conduct empirical evaluations on both synthetic and real data. Experimental results and comparisons with competing approaches corroborate the effectiveness of the proposed method.

1. Introduction

Multi-task learning (MTL) aims to train prediction models for a series of tasks jointly and simultaneously. Through exploiting the commonality among multiple tasks, MTL has been shown to be more effective than independently training each single task. In an ideal scenario where all the tasks are related, the identified commonality can normally help improve the generalization performance significantly. However, in realistic applications, the hidden structures among multiple learning tasks can be very complicated. For instance, the learning tasks could consist of several disjoint or partially overlapped task groups as well as some outlier tasks. In the case that some unrelated tasks are mixed together, simply sharing commonality among all the tasks will certainly decrease the learning performance, and such a phenomenon is thus called negative transfer [1].

One way to avoid the negative transfer is to organize related tasks in clusters, namely *task grouping* [2], and knowledge transfer is performed only within each group. Briefly speaking, there exist two levels of task grouping. The first level groups the learning tasks in an explicit manner, where one often assumes that the prediction models of the tasks within the same group share certain commonalities, such as similar structures, parameters, or priors [3–7]. However, such an explicit task grouping strategy tends to be over rigorous, as it often results in disjoint sharing of commonality among all the tasks. On the contrary, the second level, implicit task grouping, was proposed as a valuable option since it can reveal the hidden relationships among the learning tasks [8–12]. For example, Kang et al. [11] proposed a subspace based regularization framework to identify disjoint task groups, where the tasks within each group are assumed to lie in a low-dimensional space. Realizing the limitation of the disjoint task grouping, Kumar and Daumé III [12] further proposed a subspace based task grouping strategy that allows tasks from different groups to overlap by having common basis tasks, namely Grouping and Overlap in MTL (GO-MTL). However, the determination of the number of hidden basis tasks remains unsolved in principle, and prior works often rely on heuristics or empirical validation, which holds no theoretical guarantee.

Motivated by Kumar and Daumé III [12], in this paper we propose a flexible MTL (FMTL) paradigm to identify the task grouping and overlap without imposing any specific structure assumptions, e.g., the number of latent basis tasks. Similar to [12], we assume that the model parameters $\{\mathbf{w}_l\}_{l=1}^{L}$ of L learning tasks reside in a latent subspace spanned by a set of unknown basis tasks $\mathbf{M} = \mathbf{m}_1, \cdots, \mathbf{m}_k, \cdots, \mathbf{m}_L$, where $\mathbf{m}_k \in \mathbb{R}^d$ is the model parameter for the k-th basis task and d is the feature dimension. More specifically, we use a latent factor model to factorize the target model into the latent subspace and the corresponding representation as

E-mail addresses: zhongshi@fudan.edu.cn (S. Zhong), jianpu@fudan.edu.cn (J. Pu), ygj@fudan.edu.cn (Y.-G. Jiang), fengrui@fudan.edu.cn (R. Feng), xyxue@fudan.edu.cn (X. Xue).

http://dx.doi.org/10.1016/j.neucom.2015.12.092
0925-2312/

$\mathbf{w}_l = \mathbf{M}\mathbf{s}_l$. Instead of predetermining the size of latent basis tasks and constraining the subspace to be low rank [12], we use a full rank subspace and introduce two regularization terms to the corresponding representation matrix $\mathbf{S} = \mathbf{s}_1, \cdots, \mathbf{s}_l, \cdots, \mathbf{s}_L$ of the learning tasks. The first regularization term enforces $\mathbf{S}$ to be row sparse that encourages the related tasks to share a subset of basis tasks. The second column-orthogonality regularization term supplies as a complement of the row-sparsity term, which prohibits unrelated tasks to share basis tasks. Finally, we formulate the learning procedure as an optimization problem over two variables, i.e., the latent basis tasks $\mathbf{M}$ and the target model $\{\mathbf{w}_l\}_{l=1}^{L}$. Since the optimization over the latent tasks can be solved analytically, the original problem can be reformed as a *convex* minimization problem over the transformed target model that can be efficiently solved using the accelerated proximal gradient method. We show that our proposed FMTL method holds theoretical guarantee of the performance bound. Extensive experiments are conducted to validate the effectiveness of our method in both regression and classification problems, and results demonstrate that our method outperforms several recent MTL methods.

The remainder of the paper is organized as follows. Section 2 gives a brief review of related works. Section 3 introduces our new formulation of FMTL with latent task grouping. In Section 4, we elaborate the optimization strategy with detailed analysis. We discuss theoretical performance bound in Section 3, and provide empirical studies and comparisons with representative MTL algorithms in Section 6. Finally, Section 7 concludes this paper.

2. Related work

Due to practical needs in many applications, significant efforts have been paid to the design of MTL algorithms. *Model commonality* has been regarded as one of the key ingredients for joint model training. Many works focused on exploiting structure commonality of multiple learning tasks, such as low rank subspace sharing [3,13] and feature set sharing [8,14–19]. In addition, parameter commonality aims to identify the shared parameters across different tasks. Depending on the form of the used models, the shared parameters can be the hidden units in neural networks [4], the priors in hierarchical Bayesian models [5,20–22], the parameters in Gaussian process covariance [23], the feature mapping matrices [24], graph induced structures [25,26] and even the similarity metrics [7,6]. However, these methods solely considering model commonality may suffer from unsatisfactory learning performance since they neglected the fact that some tasks may be unrelated, which is often true in real applications.

To avoid the adverse effect incurred by unrelated tasks, one effective solution is to organize the tasks into groups, namely *task grouping* where the commonality is mainly shared within each group. Thrun and O'Sullivan [2] proposed to mutually measure the relatedness of tasks and select sharing information, which is regarded as one of the pioneer works for task grouping. Jacob et al. [9] developed a similar idea by imposing a cluster norm penalty and formulating the learning procedure as a convex optimization problem. Kang et al. [11] presented a disjoint task grouping method, where they assumed that the commonality sharing only occurs within each task group. By imposing a sparse inducing penalty, Kumar and Daumé III [12] further proposed to group the tasks in a low dimensional subspace using the latent factor model, where the tasks from different groups can partially share a subset of basis tasks. However, the number of latent basis tasks has to be determined by empirical validation or heuristics. Realizing the importance of inferring the "right" number of latent tasks, Passos et al. [27] and Gupta et al. [28] employed nonparametric Bayesian methods to infer the number of latent tasks. However, these Bayesian inference based methods have no guarantee of convergence rate and could suffer from a local optimum.

Finally, identifying *outlier tasks* has also been investigated in some recent works, where one assumes that the major task groups are peppered with some irrelevant outlier tasks. Hence, extracting those outlier tasks can help further alleviate the impact from the negative transfer. A decomposition scheme was utilized to partition the model into a group task component and an outlier task component [29–32].

3. Formulation

Assuming that we are given L tasks associated with training data $\{(\mathbf{X}_1, \{\mathbf{y}_1), \cdots, (\mathbf{X}_L, \mathbf{y}_L)\}$, where $\mathbf{X}_l \in \mathbb{R}^{d \times n_l}$ and $\mathbf{y}_l \in \mathbb{R}^{n_l}$ are the input and output of n_l training instances for the l-th task. For a typical linear regression or classification problem, the prediction function for the l-th task is usually expressed by $f(\mathbf{X}, \mathbf{w}_l) = f(\mathbf{X}^\top \mathbf{w}_l)$, where $\mathbf{w}_l$ is the parameter vector of the target model. We stack all the parameter vectors of the L tasks to obtain a target parameter matrix $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_L] \in \mathbb{R}^{d \times L}$.

Following Kumar and Daumé III [12], we use a latent model to factorize the target matrix into two matrices as:

$$\mathbf{W} = \mathbf{M}\mathbf{S}, \tag{1}$$

where each column of $\mathbf{M}$ represents a latent task, and each column of $\mathbf{S} = \mathbf{s}_1, \cdots, \mathbf{s}_L$ is the representation of each target model using the latent tasks: $\mathbf{w}_l = \mathbf{M}\mathbf{s}_l$. In the latent factor model proposed by [12], the latent task subspace is set to be low rank, i.e., $\mathbf{M} \in \mathbb{R}^{d \times k}$ with $k < \min(d, L)$. Hence, the rank of $\mathbf{W}$ is less than or equal to k, which reflects the hidden grouping structure of the tasks. However, as mentioned earlier, the rank of $\mathbf{M}$, i.e., the number of basis tasks has to be predetermined empirically based on prior information or heuristics, which has no theoretical guarantee.

As the objective is to obtain a low rank parameter matrix $\mathbf{W}$ to reveal the grouping structure of the learning tasks, we enforce the representation matrix $\mathbf{S}$ to exhibit the low rank structure, while relaxing the latent subspace $\mathbf{M}$ to be a full rank matrix, i.e., $\mathbf{M} \in \mathbb{R}^{d \times d}$. In particular, we impose two structure regularization terms of the representation matrix $\mathbf{S}$. The first is a ℓ_{21}-norm regularization term, which introduces *row-sparsity* on $\mathbf{S}$ matrix that encourages related tasks to share a subset of basis tasks. The second term is *column-orthogonality* that prevents unrelated tasks from sharing common basis. Formally, we formulate our FMTL objective as:

$$\min_{\mathbf{M},\mathbf{S}} \sum_{l=1}^{L} \mathcal{L}(f(\mathbf{X}^\top \mathbf{M}\mathbf{s}_l), \mathbf{y}_l) + \alpha \|\mathbf{S}\|_{2,1} + \beta \|\mathbf{S}^\top \mathbf{S}\|_F^2,$$
$$\text{subject to : } \mathbf{M}^\top \mathbf{M} = \mathbf{I}_{d \times d}. \tag{2}$$

The first component $\mathcal{L}(f(\mathbf{X}^\top \mathbf{M}\mathbf{s}_l)$ is a preselected loss function on the training set. For regression and classification problems, the squared loss and logistic loss are typically used, respectively:

$$\mathcal{L}(f(\mathbf{X}_l^\top \mathbf{M}\mathbf{s}_l), \mathbf{y}_l) = (\mathbf{X}_l^\top \mathbf{M}\mathbf{s}_l - \mathbf{y}_l)^2$$
$$\mathcal{L}(f(\mathbf{X}_l^\top \mathbf{M}\mathbf{s}_l), \mathbf{y}_l) = \log\left(1 + \exp(-\mathbf{y}_l \mathbf{X}_l^\top \mathbf{M}\mathbf{s}_l)\right).$$

The second and third components represent the two types of structure regularization terms on the representation matrix of the target model in the latent subspace, where the coefficients α and β weigh the contribution from each term. The constraint $\mathbf{M}^\top \mathbf{M} = \mathbf{I}_{d \times d}$ is used to ensure that the latent basis tasks $\mathbf{M}$ are orthogonal and form a subspace in $\mathbb{R}^d$.

Note that the $\ell_{2,1}$ norm as a structural penalty forces $\mathbf{S}$ to be a row sparse matrix, which is equivalent to selecting a subset of basis tasks to represent the target model $\mathbf{W}$. The column-orthogonal regularization term is employed to penalize the

Please cite this article as: S. Zhong, et al., Flexible multi-task learning with latent task grouping, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2015.12.092

sharing of basis tasks among unrelated tasks, similar to the work by Romera-Paredes et al. [33], where orthogonal regularization was used to reach the same goal. Such a hybrid form of structure regularization in our objective can help derive the hidden structure of the multiple learning tasks. The row-sparsity helps reveal the underlying low-rank grouping characteristics without empirically setting the dimensionality of the latent subspace. The column-orthogonality shrinks the correlation strength between unrelated tasks to be zero, while maintaining partial overlap between tasks across different groups.

Since $\mathbf{M}$ is orthogonal and full rank, we can easily derive the representation matrix as $\mathbf{S}=\mathbf{M}^{-1}\mathbf{W}=\mathbf{M}^{\top}\mathbf{W}$, and $\mathbf{S}^{\top}\mathbf{S}=\mathbf{W}^{\top}\mathbf{W}$. Substituting them into Eq. (A.1), we have

$$\min_{\mathbf{M},\mathbf{W}} \sum_{l=1}^{L} \mathcal{L}(f(\mathbf{X}^{\top}\mathbf{w}_l),\mathbf{y}_l)+\alpha\|\mathbf{M}^{\top}\mathbf{W}\|_{2,1}+\beta\|\mathbf{W}^{\top}\mathbf{W}\|_F^2,$$
$$\text{subject to}: \mathbf{M}^{\top}\mathbf{M}=\mathbf{I}_{d\times d}. \tag{3}$$

The above objective explicitly derives the target model $\mathbf{W}$ and the latent subspace $\mathbf{M}$ simultaneously, though implicitly imposing the representation matrix with certain structure characteristics. Note that if we set $\beta=0$, the above objective reduces to a special case which only ensures task grouping in the subspace spanned by the latent basis tasks, which is equivalent to the subspace learning based MTL by [8]. However, the column-orthogonality term is critical to avoid negative transfer.

4. Optimization and analysis

Since the minimization problem in Eq. (3) is over two variables $\mathbf{W}$ and $\mathbf{M}$, a standard strategy is to employ alternative optimization scheme to derive the optimal solution [12,34]. In this section, we first show that the objective of the proposed FMTL can be reformulated as a minimization problem over a single variable $\mathbf{W}$ through implicitly applying a similar alternating minimization strategy. Then, we propose to use an accelerated gradient method to efficiently tackle the optimization problem.

We consider the optimization over the latent subspace $\mathbf{M}$ with a fixed model $\mathbf{W}$. Then the minimization problem in Eq. (3) can be rewritten as:

$$\min_{\mathbf{M}} \|\mathbf{M}^{\top}\mathbf{W}\|_{2,1},$$
$$\text{subject to}: \mathbf{M}^{\top}\mathbf{M}=\mathbf{I}_{d\times d}. \tag{4}$$

Motivated by Argyriou et al. [8], the above optimization can be addressed using the following theorem:

Theorem 1. *The trace norm is the minimal $\|\cdot\|_{2,1}$ norm over all possible orthonormal bases. For any matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$,*

$$\|\mathbf{A}\|_{*}=\min_{\mathbf{U}}\|\mathbf{U}\mathbf{A}\|_{2,1}, \tag{5}$$

where $\mathbf{U}$ is chosen as an eigenbasis of matrix $\mathbf{A}\mathbf{A}^{\top}$ and it satisfies $\mathbf{U}^{\top}\mathbf{U}=\mathbf{I}_{n\times n}$.

Based on the above theorem, the optimal solution of $\mathbf{M}$ in Eq. (4) is an eigenbasis of the matrix $\mathbf{W}\mathbf{W}^{\top}$ and the minimal value of the cost function is $\|\mathbf{W}\|_{*}$. Substituting $\|\mathbf{M}^{\top}\mathbf{W}\|_{2,1}$ in Eq. (3) by its optimal value $\|\mathbf{W}\|_{*}$, we transform the constrained *bivariate* minimization problem into an unconstrained *univariate* minimization problem as:

$$\min_{\mathbf{W}} \sum_{l=1}^{L} \mathcal{L}(f(\mathbf{X}^{\top}\mathbf{w}_l),\mathbf{y}_l)+\alpha\|\mathbf{W}\|_{*}+\beta\|\mathbf{W}^{\top}\mathbf{W}\|_F^2. \tag{6}$$

Next we introduce an efficient accelerated proximal gradient method to derive the optimal $\mathbf{W}$ iteratively. Note that solving the above minimization problem with respect to $\mathbf{W}$ is equivalent to using the alternating scheme to solve the original objective in Eq. (3), except that we do not need to explicitly compute the intermediate value of $\mathbf{M}$ during the iteration procedure.

Since the above cost function is convex and nonsmooth due to the trace norm, we separate the cost function into a smooth term $g(\mathbf{W})$ and a nonsmooth term $h(\mathbf{W})$[1]:

$$g(\mathbf{W})=\sum_{l=1}^{L}\mathcal{L}(f(\mathbf{X}^{\top}\mathbf{w}_l),\mathbf{y}_l)+\beta\|\mathbf{W}^{\top}\mathbf{W}\|_F^2,$$
$$h(\mathbf{W})=\alpha\|\mathbf{W}\|_{*},$$

and the optimal solution is to minimize the following cost:

$$\mathbf{W}^{*}=\arg\min_{\mathbf{W}} g(\mathbf{W})+h(\mathbf{W}). \tag{7}$$

Then solving Eq. (7) using the proximal gradient method is written as:

$$\mathbf{W}^{k+1}=\mathrm{Prox}_{h,\lambda^k}(\mathbf{W}^k-\lambda^k\nabla g(\mathbf{W})), \tag{8}$$

where λ^k is the step length, which can be computed by line search. The proximal operator $\mathrm{Prox}_{g,\lambda}(\cdot)$ on the function $g(\cdot)$ with the parameter λ is defined as:

$$\mathrm{Prox}_{g,\lambda}(\mathbf{C})=\min_{\mathbf{W}}\left(g(\mathbf{W})+\frac{1}{2\lambda}\|\mathbf{W}-\mathbf{C}\|_F^2\right), \tag{9}$$

where the matrix variable $\mathbf{C}$ has the same size as $\mathbf{W}$. In particular, the proximal operator on the trace norm function can be solved through performing soft-thresholding on the singular values of the matrix $\mathbf{C}$. This is summarized in the following theorem [35]:

Theorem 2. *Let $\mathbf{C}\in\mathbb{R}^{d\times L}$ and the singular value decomposition of $\mathbf{C}$ is represented as $\mathbf{C}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$, where $\mathbf{U}\in\mathbb{R}^{d\times r}$ and $\mathbf{V}\in\mathbb{R}^{L\times r}$ have orthonormal columns, $\mathbf{\Sigma}\in\mathbb{R}^{r\times r}$ is diagonal, $r=rank(\mathbf{C})$. Then*

$$\mathrm{Prox}_{\|\cdot\|_{*},\lambda}(\mathbf{C})=\mathbf{U}\mathbf{\Sigma}_{\lambda}\mathbf{V}^{\top}, \tag{10}$$

where $\mathbf{\Sigma}_{\lambda}$ is diagonal with $(\mathbf{\Sigma}_{\lambda})_{ii}=\max(0,\mathbf{\Sigma}_{ii}-\lambda)$.

The proximal gradient descent method in Eq. (8) clearly suggests an iterative procedure to derive the optimal target model $\mathbf{W}$, and each single update process can be efficiently done using Theorem 2. Algorithm 1 summarizes the proposed FMTL using the proximal gradient method with extrapolation to ensure a quadratic rate of convergence, which is also called *accelerated proximal gradient method* [36]. As the objective function in Eq. (6) is convex, the final algorithm is formulated as a single loop to obtain the global convergence.

Algorithm 1. Flexible Multi-Task Learning (FMTL) with Latent Task Grouping.

Require: $\mathbf{X}_l$: data matrix of the lth task;
$\mathbf{y}_l$: response of the lth task.

1: Set both the initial target matrix $\mathbf{W}^0$ and auxiliary matrix $\mathbf{C}^1$ to zero matrices.
2: **while** not converged **do**
3: Evaluate the gradient descent: $\mathbf{G}^k=\mathbf{C}^k-\lambda^k\nabla f(\mathbf{C}^k)$;
4: Determine the step length λ^k by line search;
5: Decompose the matrix $\mathbf{C}^k$ using SVD: $\mathbf{C}^k=\mathbf{U}^k\mathbf{\Sigma}^k\mathbf{V}^k$.
6: Compute the target matrix $\mathbf{W}$ by soft-thresholding the singular values: $\mathbf{W}^k=\mathbf{U}^k\mathbf{\Sigma}_{\lambda^k}^k\mathbf{V}^k$.
7: $\mathbf{C}^k=\mathbf{W}^k+\frac{k-1}{k+2}(\mathbf{W}^k-\mathbf{W}^{k-1})$.
8: Update the iteration counter: $k=k+1$.
9: **end while**

[1] Without loss of generality, we assume the empirical loss function is smooth.

Please cite this article as: S. Zhong, et al., Flexible multi-task learning with latent task grouping, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2015.12.092

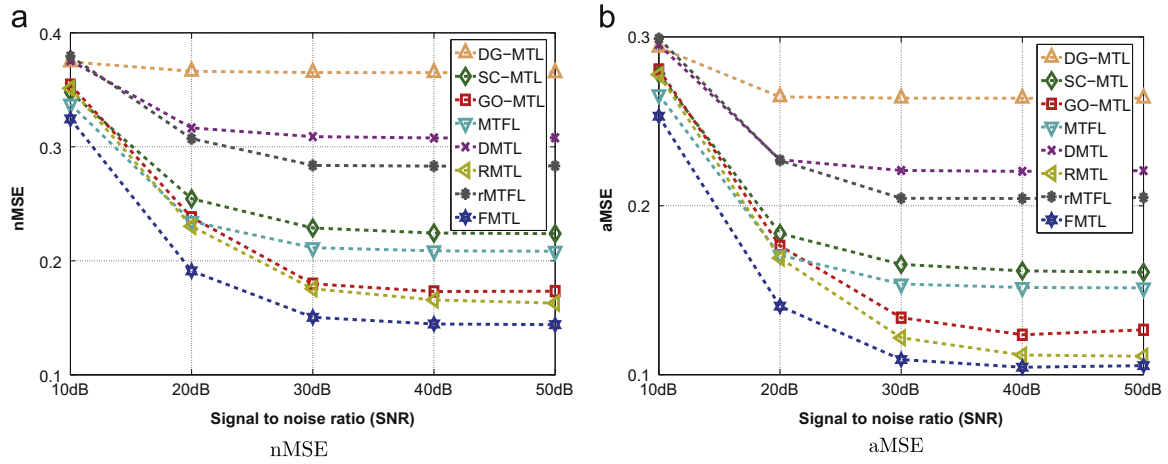


Fig. 1. Regression performance of eight MTL methods, including the proposed FMTL, on synthetic data under different SNRs ranging from 10 dB to 50 dB. (a) nMSE. (b) aMSE.

Note that the computational complexity of SVD (line 4 in Algorithm 1) is $\mathcal{O}(d^2L+L^3)$. APG requires $\mathcal{O}(\sqrt{1/\epsilon})$ iterations to achieve ϵ error of the optimal value. Therefore, the total computational complexity of the proposed FMTL method is $\mathcal{O}(\sqrt{(d^2L+L^3)/\epsilon})$.

5. Performance bound

In this section we provide performance bounds of our method. The squared loss is used in this analysis.

We use $\mathcal{I}(\mathbf{S})$ and $\mathcal{I}^c(\mathbf{S})$ to represent the sets of indices for the nonzero and zero rows of matrix $\mathbf{S}$, respectively, which are defined as

$$\mathcal{I}(\mathbf{S})=\{i:\|\mathbf{s}_{i\cdot}\|_1\neq 0\},$$
$$\mathcal{I}^c(\mathbf{S})=\{i:\|\mathbf{s}_{i\cdot}\|_1=0\}.$$

Then, we make the following valid assumption about the training data and the target model:

Assumption 1. For a matrix $\mathbf{\Gamma}\in\mathbb{R}^{d\times L}$, let $s\leq d$. We assume that there exist nonnegative constants $\kappa_1(s)$ and $\kappa_2(s)$ such that

$$\kappa_1(s)=\min_{\mathbf{\Gamma}\in\mathcal{R}(s)}\frac{\|\mathbf{X}^\top vec(\mathbf{M}\mathbf{\Gamma})\|}{\sqrt{dL}\|\mathbf{\Gamma}_{\mathcal{I}(\mathbf{S})}\|_F},$$
$$\kappa_2(s)=\min_{\mathbf{\Gamma}\in\mathcal{R}(s)}\frac{\|\mathbf{X}^\top vec(\mathbf{M}\mathbf{\Gamma})\|}{\sqrt{dL}\|\mathbf{\Gamma}^\top\mathbf{\Gamma}\|_F},$$

Here we define a vectorization operator $vec(\cdot)$ over an arbitrary matrix $\mathbf{Z}$ as $vec(\mathbf{Z})=[\mathbf{z}_1^\top,\cdots,\mathbf{z}_l^\top,\cdots,\mathbf{z}_L^\top]^\top$, where $\mathbf{z}_l$ is the l-th column vector of $\mathbf{Z}$. The restricted set $\mathcal{R}(s)$ is defined as:

$$\mathcal{R}(s)=\{\mathbf{\Gamma}\in\mathbb{R}^{d\times L}:\mathbf{\Gamma}\neq\mathbf{0},|\mathcal{I}(\mathbf{S})|\leq s,$$
$$\|\mathbf{\Gamma}_{\mathcal{I}^c(\mathbf{S})}\|_{2,1}\leq\alpha_1\|\mathbf{\Gamma}_{\mathcal{I}(\mathbf{S})}\|_{2,1}\}$$

where $|\mathcal{I}(\mathbf{S})|$ is the cardinality of the set $\mathcal{I}(\mathbf{S})$, i.e., the number of nonzero columns in $\mathbf{S}$. The submatrix $\mathbf{\Gamma}_{\mathcal{I}^c(\mathbf{S})}$ is obtained by selecting a subset of rows from $\mathbf{\Gamma}$ with the row indices as $\mathcal{I}^c(\mathbf{S})$.

Note that the above assumption about $\kappa_1(s)$ is a generalized version of the so-called *restricted eigenvalue assumption* [37], which ensures that the ratio between the empirical loss over training data and the row-sparsity penalty is lower bounded. Similar assumptions of the restricted eigenvalue have been used in previous works for analyzing the performance of other MTL models [38,30,31]. In addition, we impose the assumption of $\kappa_2(s)$, which associates with the training data matrix and the column-orthogonality structure of the representation matrix.

Based on the above assumptions, the following theorem gives performance bounds to measure how well FMTL can approximate the true $\mathbf{S}$ matrix defined in Eq. (A.1).

Theorem 3. *Let $\hat{\mathbf{S}}$ be the optimal solution of Eq. (A.1) for $L\geq 2$ and $n,d\geq 1$, and $\mathbf{S}^*$ be the oracle solution, $\gamma=\|\mathbf{S}^*\|_F$. The regularization parameters α is chosen as*

$$\alpha\geq\frac{2\sigma}{nL}\sqrt{dL+b},$$

here b is a positive scalar. Denote that $\eta=\frac{\sqrt{s}}{\kappa_1(s)}$ and $\xi=\left(1-\beta\left(\frac{1}{\kappa_2(s)}+\frac{2\gamma}{\kappa_1}\right)^2\right)^{-1}$, then under Assumption 2, the following results hold with a probability of at least $1-\exp(-\frac{1}{2}(b-dL\log(1+\frac{b}{dL})))$

$$\frac{1}{nL}\|\mathbf{X}^\top vec(\mathbf{M}\hat{\mathbf{S}})-vec(\mathbf{Y})\|^2\leq\alpha^2\eta^2\xi^2,$$

$$\|\hat{\mathbf{S}}-\mathbf{S}^*\|_{2,1}\leq\alpha(\alpha_1+1)\eta^2\xi,$$

$$\|(\hat{\mathbf{S}}^\top\hat{\mathbf{S}}-\mathbf{S}^{*\top}\mathbf{S}^*)\|_F\leq\alpha\left(\frac{1}{\kappa_2(s)}+\frac{2\gamma}{\kappa_1}\right)\eta\xi.$$

The above theorem presents an important theoretical guarantee of the performance of our FMTL. In particular, the first inequality measures the bound of the empirical loss. The second and third inequalities bound the representation matrix $\mathbf{S}$ in terms of $\ell_{2,1}$ norm and Frobenius norm, respectively.[2] Variables η and ξ are the effects induced by the row sparse penalty and the column-orthogonality penalty, respectively. Note that comparing with the bound as an addition form in [31], our performance bound is represented as a product form of η and ξ. In other words, if the data fits the model well, it may lead to small η and ξ and our bounds tend to be fairly tight.

[2] Detailed proof of Theorem 3 can be found in the supplementary material.

Please cite this article as: S. Zhong, et al., Flexible multi-task learning with latent task grouping, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2015.12.092

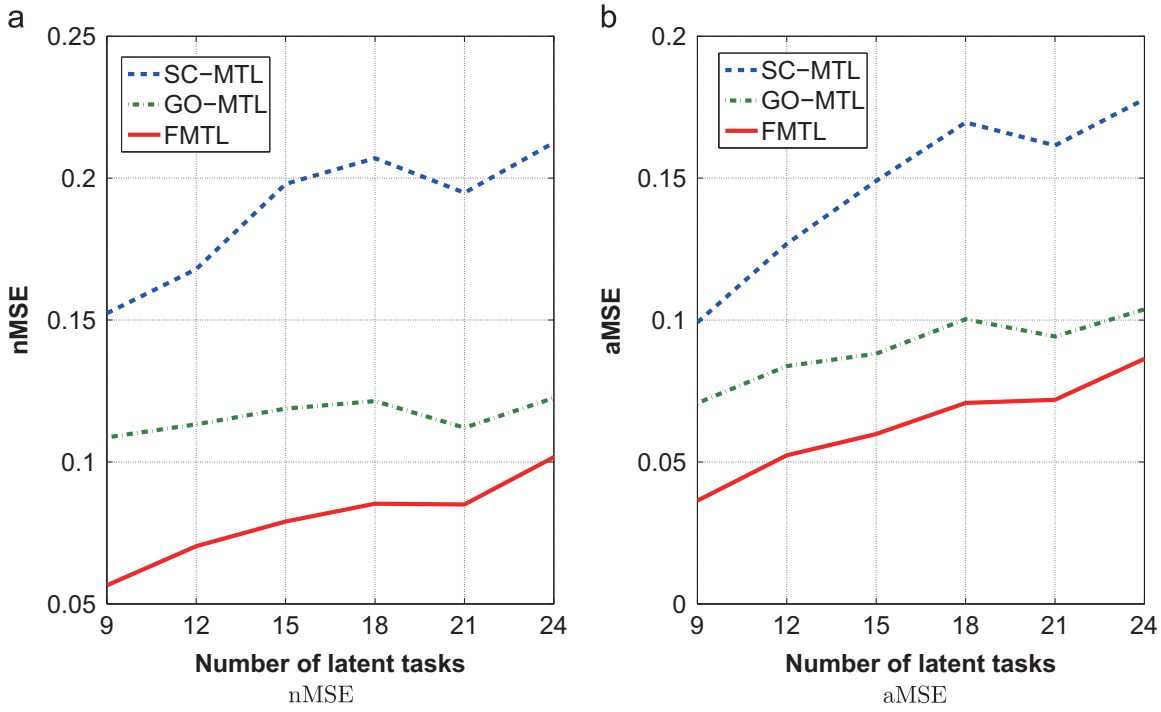


Fig. 2. Comparison of three latent subspace-based MTL methods on synthetic data with different number of latent tasks. (a) nMSE. (b) aMSE.

6. Experiments

In this section, we conduct extensive experiments on both regression and classification problems to demonstrate the effectiveness of the FMTL method. For the regression problem, we employ the logistic loss in the cost function for training and measure the error using normalized mean squared error (nMSE) and averaged mean squared error (aMSE), which have been frequently used in the literature [30,39–41]. For the classification problem, we use the squared loss in the cost function and evaluate the prediction performance by the area under the ROC curve (AUC). We adopt one synthetic dataset and five real-world datasets, including the SARCOS dataset for regression and four image datasets for classification (the Olivetti Faces, the Animal, the MNIST, and the USPS datasets). We use cross-validation to identify the optimal parameters for all the compared methods. Performance is reported through averaging results over 20 random trials.

We compare with several representative MTL methods, including multi-task feature learning (MTFL) [8], disjoint-group MTL (DG-MTL) [11], the GO-MTL [12] and sparse-coding MTL (SC-MTL) [34]. These methods can be generalized to both the classification and the regression tasks. In addition, we also include several decomposition-based methods for the regression problem: dirty model for MTL (DMTL) [29], robust MTL (RMTL) [30], and robust multi-task feature learning (rMTFL) [31]. Note that in the comparative study we mainly compare with those methods that can identify general task structures, e.g., grouping, overlap and outliers. We discuss experimental results in the following.

6.1. Synthetic data

We first use synthetic data for performance evaluation and comparison, which is generated using the follows steps: (1) generate a random orthonormal matrix to form the latent subspace $\mathbf{M}$; (2) for each task group, randomly select a set of active features to form the latent representation matrix $\mathbf{S}$; (3) compute the target model using the factorization equation as $\mathbf{W}=\mathbf{MS}$; (4) for each task, the feature set $\mathbf{X}_l$ is obtained from Gaussian distribution, and the response $\mathbf{y}_l$ is computed by $\mathbf{y}=\mathbf{X}_l\mathbf{w}_l+\boldsymbol{\epsilon}_l$, where $\boldsymbol{\epsilon}_l$ indicates the Gaussian noise with a certain SNR level.

The data is partitioned into a training set (50%), a validation set (20%) and a test set (30%). Fig. 1 shows the regression performance measured by nMSE and aMSE, where we can see that our FMTL method outperforms all the competing methods with significant margins. With increasing SNR, all the methods tend to offer more accurate results, and the performance gain from FMTL is consistent under all the SNR levels.

Note that the proposed FMTL method is closely related to SC-MTL and GO-MTL in terms of the latent factorization of the target model. In particular, GO-MTL needs to predetermine the number of latent basis tasks, while our method learns the latent basis structure directly from the data. We specifically design an experiment to compare the performance sensitivity of these three methods with respect to various latent structures. Fig. 2 shows the performance with different numbers of latent basis tasks, ranging from 9 to 24. With an increasing number of latent tasks, the performance tends to decrease since more parameters need to be estimated with the same amount of training data. Clearly, FMTL provides significantly lower error rates.

We also visualize the recovered latent representation matrix $\mathbf{S}$ in Fig. 3. As GO-MTL requires to set the number of latent tasks, we report three trials with under-estimation, exact estimation and over-estimation of this number. From the figure, we can observe two phenomena. First, FMTL has significantly better recovery than SC-MTL on the zero entries, which is because of the shrinking effect from the column-orthogonal term. Second, for the nonzero entries, FMTL achieves a much more accurate estimation than the GO-MTL, which is due to the use of the imposed row-sparsity that explicitly help recover the grouping structure.

6.2. Real data

Next we discuss results on real data. We first conduct regression experiments using the SARCOS data, which was generated from an inverse dynamics prediction system that contains seven degrees-of-freedom anthropomorphic robot arm. This dataset has

Please cite this article as: S. Zhong, et al., Flexible multi-task learning with latent task grouping, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2015.12.092

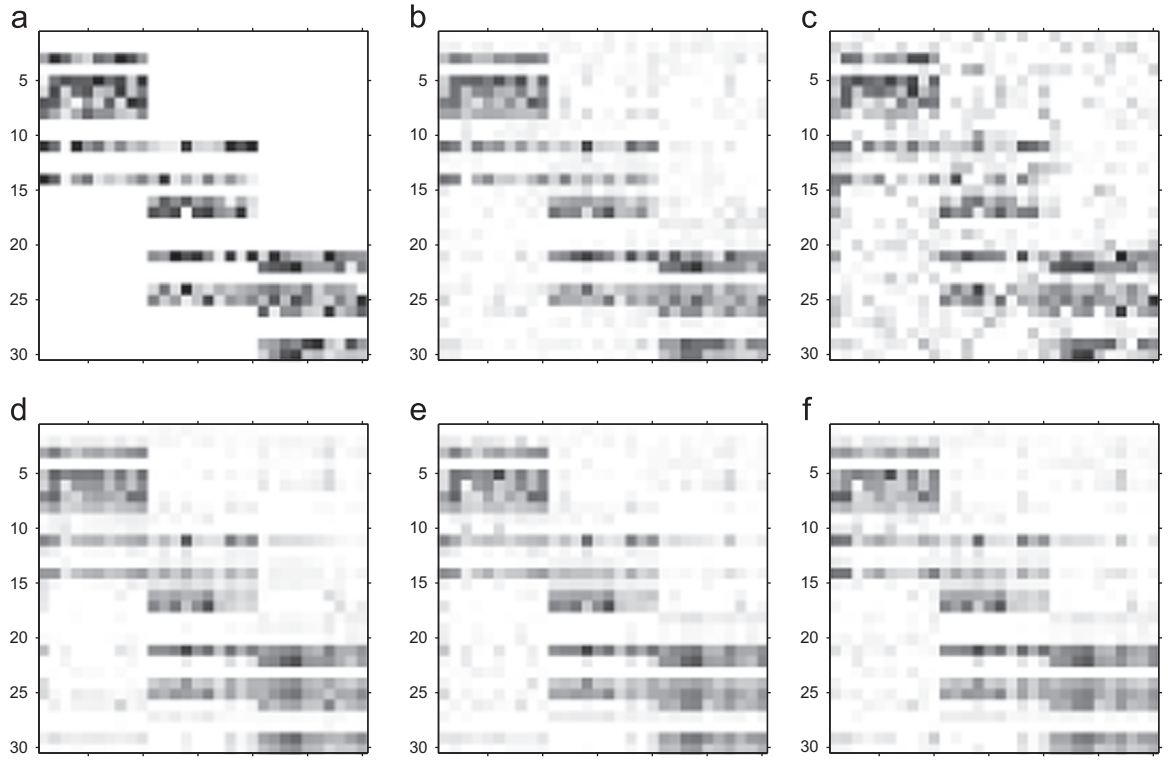


Fig. 3. The recovered latent representation matrix **S** using different methods: (a) ground-truth, (b) FMTL, (c) SC-MTL, (d) GO-MTL with $k=10$, (e) GO-MTL with $k=16$, and (f) GO-MTL with $k=30$.

Table 1
Comparison of regression performance in terms of nMSE and aMSE on the SARCOS dataset. FMTL achieves the best results under all the evaluated settings.

Measure	Training #	DG-MTL	SC-MTL	GO-MTL	MTFL	DMTL	RMTL	rMTFL	FMTL
nMSE	50	0.0692 ± 0.0076	0.0819 ± 0.0067	0.0669 ± 0.0064	0.0627 ± 0.0068	0.0629 ± 0.0048	0.0700 ± 0.0056	0.0644 ± 0.0068	**0.0604** ± 0.0066
	100	0.0494 ± 0.0047	0.0640 ± 0.0042	0.0511 ± 0.0026	0.0447 ± 0.0022	0.0469 ± 0.0038	0.0488 ± 0.0042	0.0474 ± 0.0040	**0.0444** ± 0.0020
	150	0.0431 ± 0.0024	0.0594 ± 0.0025	0.0470 ± 0.0013	0.0402 ± 0.0015	0.0422 ± 0.0019	0.0435 ± 0.0019	0.0472 ± 0.0186	**0.0401** ± 0.0015
aMSE	50	0.0655 ± 0.0072	0.0775 ± 0.0063	0.0634 ± 0.0060	0.0593 ± 0.0064	0.0596 ± 0.0045	0.0663 ± 0.0053	0.0610 ± 0.0064	**0.0572** ± 0.0062
	100	0.0468 ± 0.0044	0.0606 ± 0.0039	0.0484 ± 0.0024	0.0423 ± 0.0021	0.0444 ± 0.0036	0.0462 ± 0.0040	0.0449 ± 0.004038	**0.0421** ± 0.0019
	150	0.0408 ± 0.0022	0.0562 ± 0.0023	0.0445 ± 0.0012	0.0381 ± 0.0015	0.0399 ± 0.0018	0.0412 ± 0.0018	0.0447 ± 0.0176	**0.0379** ± 0.0014

a total of 48,933 observations corresponding to 7 joint torques. Each observation is represented by a 21-dimensional feature, including 7 joint positions, 7 joint velocities and 7 joint accelerations. Following the settings of Chen et al. [30], we randomly select 50, 100 and 150 observations to form 3 training sets, and use 200 and 5000 observations as validation and test sets.

The regression performance measured by nMSE and aMSE is summarized in Table 1. Our proposed FMTL achieves the best performance across all the compared methods with lower error rates and smaller variances. Especially, FMTL significantly outperforms the other two recent latent subspace-based methods, SC-MTL [34] and GO-MTL [12]. The relative performance gain is as high as 15% in some cases. The MTFL method, which can be considered as a special case of our method with the column-orthogonality term removed ($\beta=0$), produces the second best results. In the case of using a small set of training samples (e.g., 50), our method has much lower error rates than MTFL.

We further adopt four image datasets to evaluate the classification performance. The first dataset is the Olivetti Faces, which contains 40 classes and each class has 10 samples. We use PCA to reduce the feature dimension to 123 to retain 95% of the variance. For each task, we pick 1, 2 and 3 samples for training separately, and use 2 samples for validation and the rest for testing. The other three datasets, Animal, MNIST and USPS, have been popularly used in previous MTL literature [11]. The Animal dataset contains 20 animal categories and the original feature is reduced to 202-dimension using PCA. Both the MNIST and the USPS datasets are handwritten digits with 10 classes. Following the conventions in using these datasets, we also apply PCA to reduce the feature dimensions to 64 and 87, respectively. We adopt exactly the same data splitting strategy as Kang et al. [11] to generate the training, validation and test sets. To evaluate the performance using various sizes of training data, we randomly select 100, 200 and 300 training samples from the training set respectively, and keep the validation and test sets unchanged.

Please cite this article as: S. Zhong, et al., Flexible multi-task learning with latent task grouping, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2015.12.092

Table 2
Comparison of classification performance, measured by AUC, on the Olivetti Faces, Animal, MNIST and USPS datasets.

Dataset	Training #	DG-MTL	GO-MTL	SC-MTL	MTFL	FMTL
Olivetti Faces	1	0.5837 ± 0.0087	0.7939 ± 0.0356	0.9154 ± 0.0139	0.8741 ± 0.0183	**0.9321 ± 0.0117**
	2	0.6481 ± 0.0075	0.9137 ± 0.0219	**0.9602 ± 0.0068**	0.9413 ± 0.0105	0.9601 ± 0.0066
	3	0.7583 ± 0.0071	0.9254 ± 0.0157	**0.9815 ± 0.0035**	0.9779 ± 0.0032	0.9782 ± 0.0036
Animal	100	0.5947 ± 0.0117	0.6436 ± 0.0151	0.6428 ± 0.0162	0.6077 ± 0.0117	**0.6465 ± 0.0182**
	200	0.5331 ± 0.0125	0.6672 ± 0.0172	0.6740 ± 0.0089	0.6037 ± 0.0141	**0.6885 ± 0.0098**
	300	0.5993 ± 0.0122	0.6916 ± 0.0153	0.6934 ± 0.0127	0.6249 ± 0.0108	**0.7158 ± 0.0092**
MNIST	100	0.8753 ± 0.0151	0.9175 ± 0.0101	0.9011 ± 0.0121	0.9110 ± 0.0098	**0.9379 ± 0.0062**
	200	0.9050 ± 0.0092	0.9426 ± 0.0065	0.9246 ± 0.0074	0.9296 ± 0.0064	**0.9561 ± 0.0055**
	300	0.9272 ± 0.0040	0.9541 ± 0.0048	0.9364 ± 0.0032	0.9392 ± 0.0030	**0.9645 ± 0.0032**
USPS	100	0.8959 ± 0.0108	0.9355 ± 0.0079	0.9195 ± 0.0083	0.9242 ± 0.0079	**0.9510 ± 0.0054**
	200	0.9380 ± 0.0065	0.9615 ± 0.0060	0.9460 ± 0.0051	0.9499 ± 0.0050	**0.9712 ± 0.0038**
	300	0.9480 ± 0.0050	0.9693 ± 0.0033	0.9519 ± 0.0043	0.9542 ± 0.0046	**0.9762 ± 0.0023**

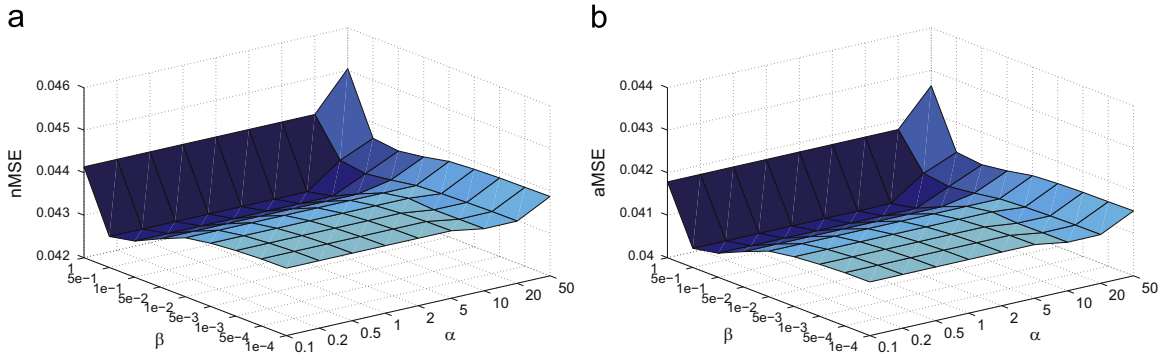


Fig. 4. Evaluation of parameter sensitivity of FMTL on SARCOS dataset: (a) nMSE, (b) aMSE.

The classification accuracies measured by AUC on these four datasets are shown in Table 2, where the proposed FMTL again achieves the best performance in most of the test cases. In particular, when given only one training data for each task on the face recognition problem, FMTL is significantly better than the other MTL approaches. This demonstrates that the FMTL method is very effective particularly when given sparse training samples since it can reveal the latent task structures to improve joint training. Even for the only two cases on which SC-MTL produces the highest accuracies, FMTL offers very close performance. These results clearly show that, besides in the regression tasks, the proposed FMTL also performs very strongly in classification problems.

We finally analyze parameter sensitivity for the proposed FMTL method, using the SARCOS dataset. Two key parameters of the proposed methods are α and β. The parameter α controls the task grouping penalty, while β controls orthogonality of irrelevant tasks. We use 30% data samples for training and the rest for testing. By varying α in [0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50] and varying β in $[1e^{-4}, 5e^{-4}, 1e^{-3}, 5e^{-3}, 1e^{-2}, 5e^{-2}, 1e^{-1}, 5e^{-1}, 1]$, we show the performance measured by nMSE and aMSE in Fig. 4. It is clear to see that the performance of both nMSE and aMSE achieve best when β is close to $5e^{-1}$ and α is in [0.1, 5].

7. Conclusion

We have proposed a flexible multi-task learning approach to identify the task grouping and overlap structures in a latent subspace. Different from most existing methods, we impose no assumptions on the structures of the basis tasks and derive the optimal latent subspace and the target models based on the structure fully learnt from the input data. Instead of constraining the latent subspace to be low rank, we impose a mixed structure penalties of row-sparsity and column-orthogonality on the representation of the target model. Such a hybrid structure regularization ensures that the target models are grouped into clusters through selecting a common set of basis tasks, and meanwhile avoids negative transfer by shrinking the correlations between less related tasks to be zero. Rigorous analysis of the performance bounds has been provided, and extensive experiments on both synthetic and real datasets have clearly validated the effectiveness of our approach. One useful future work is to extend the proposed approach to work on nonlinear problems, which are often seen in practical applications.

Acknowledgements

This work was supported in part by a National 863 Program (#2014AA015101), a grant from the NSF China (#61572134), and the EU FP7 QUICK project under Grant Agreement no. PIRSESGA-2013-612652.

Appendix A. Proof of Theorem 3

In this appendix, we present the proof of Theorem 3, which measures the performance bound of the following problem:

$$\min_{\mathbf{M},\mathbf{S}} \sum_{l=1}^{L} \mathcal{L}(f(\mathbf{X}^\top \mathbf{M}\mathbf{s}_l), \mathbf{y}_l) + \alpha \|\mathbf{S}\|_{2,1} + \beta \|\mathbf{S}^\top \mathbf{S}\|_F^2, \tag{A.1}$$

$$\text{subject to } : \mathbf{M}^\top \mathbf{M} = \mathbf{I}_{d\times d}. \tag{A.2}$$

First, we provide two lemmas originally presented in [31]:

Lemma 1. *For any matrix pair $\mathbf{S}, \hat{\mathbf{S}}$, we have the following inequality:*

$$\|\hat{\mathbf{S}}-\mathbf{S}\|_{2,1} + \|\mathbf{S}\|_{2,1} - \|\hat{\mathbf{S}}\|_{2,1} \le 2\|(\hat{\mathbf{S}}-\mathbf{S})_{\mathcal{I}(\mathbf{S})}\|_{2,1}. \tag{A.3}$$

Lemma 2. *Let $\chi^2(d)$ be a χ^2 random variable with d degrees of freedom. Then, for $\forall b > 0$, we have*

$$\Pr(\chi^2(d) \le d + b) > 1 - \exp\left(-\frac{1}{2}\left(b - d\log\left(1+\frac{b}{d}\right)\right)\right). \tag{A.4}$$

Then we present another lemma:

Lemma 3. *Let δ_i be i.i.d. random variables with $\delta_i \sim N(0,\sigma^2)$, matrix $\mathbf{X}$ is row normalized $\sum_j x_{ij}^2 = 1$, and $\sum_j m_j^2 = 1$. Then, we have*

$$v = \frac{1}{\sigma}\sum_k \sum_j m_k x_{kj} \delta_j \tag{A.5}$$

is a standard normal random variable.

Proof. Since $\delta_j \sim N(0,\sigma^2)$, v must be a normal random variable. Next, we show that the mean and variance of v are 0 and 1, respectively.

$$\mathbb{E}(v) = \mathbb{E}\left(\frac{1}{\sigma}\sum_k\sum_j m_k x_{kj}\delta_j\right) = \frac{1}{\sigma}\sum_k\sum_j m_k x_{kj}\mathbb{E}(\delta_j) = 0,$$

$$\mathbb{V}(v) = \mathbb{E}(v^2) = \frac{1}{\sigma^2}\sum_k m_k^2 \sum_j x_{kj}^2 \mathbb{E}(\delta_j^2) = 1. \square$$

Theorem 4. *Let $\hat{\mathbf{S}}$ be the optimal solution of Eq. (A.1) for $L \ge 2$ and $n, d \ge 1$. Assuming the data $\mathbf{X}_l^\top$ is normalized, and choosing the regularization parameter α as*

$$\alpha \ge \frac{2\sigma}{nL}\sqrt{dL+b}, \tag{A.6}$$

where b is a positive scalar. Then with probability at least $1-\exp(-\frac{1}{2}(b-dL\log(1+\frac{b}{dL})))$, we have

$$\sum_{l=1}^{L}\frac{1}{nL}\|\mathbf{X}_l^\top \mathbf{M}\hat{\mathbf{s}}_l - \mathbf{f}_l\|_2^2 \le \sum_{l=1}^{L}\frac{1}{nL}\|\mathbf{X}_l^\top M\mathbf{s}_l - \mathbf{f}_l\|_2^2 \tag{A.7}$$

$$+\alpha\|(\hat{\mathbf{S}}-\mathbf{S})_{\mathcal{I}(\mathbf{S})}\|_{2,1} + \beta\|\hat{\mathbf{S}}^\top\hat{\mathbf{S}} - \mathbf{S}^\top\mathbf{S}\|_F^2. \tag{A.8}$$

Proof. As $\hat{\mathbf{S}}$ is the optimal solution of Eq. (A.1), we have

$$\sum_{l=1}^{L}\frac{1}{nL}\|\mathbf{X}_l^\top \mathbf{M}\hat{s}_l - \mathbf{y}_l\|_2^2 \le \sum_{l=1}^{L}\frac{1}{nL}\|\mathbf{X}_l^\top \mathbf{M}\mathbf{s}_l - \mathbf{y}_l\|_2^2$$
$$+\alpha(\|\mathbf{S}\|_{2,1} - \|\hat{\mathbf{S}}\|_{2,1}) + \beta(\|\mathbf{S}^\top\mathbf{S}\|_F^2 - \|\hat{\mathbf{S}}^\top\hat{\mathbf{S}}\|_F^2). \tag{A.9}$$

Assuming the regression model is given by a linear representation plus Gaussian noise, that is

$$y_l = \mathbf{f}_l + \delta_l = \mathbf{X}_l^\top \mathbf{M}\mathbf{s}_l + \delta_l, \tag{A.10}$$

where δ_l is a vector and each entry $\delta_{li} \sim N(0,\sigma^2)$. Thus, we have

$$\sum_{l=1}^{L}\frac{1}{nL}\|\mathbf{X}_l^\top \mathbf{M}\hat{s}_l - \mathbf{f}_l\|_2^2 \tag{A.11}$$

$$\le \sum_{l=1}^{L}\frac{1}{nL}\|\mathbf{X}_l^\top \mathbf{M}\mathbf{s}_l - \mathbf{f}_l\|_2^2 + \alpha(\|\mathbf{S}\|_{2,1} - \|\hat{\mathbf{S}}\|_{2,1})$$
$$+\beta(\|\mathbf{S}^\top\mathbf{S}\|_F^2 - \|\hat{\mathbf{S}}^\top\hat{\mathbf{S}}\|_F^2) + \frac{2}{nL}\langle \mathbf{Z}, \hat{\mathbf{S}}-\mathbf{S}\rangle, \tag{A.12}$$

where $\mathbf{Z} = [\mathbf{M}^\top\mathbf{X}_1\delta_1, \cdots, \mathbf{M}^\top\mathbf{X}_L\delta_L] \in \mathbb{R}^{d\times L}$, and its (i,j)-th entry is given by

$$z_{ij} = \sum_{k=1}^{n}\sum_s m_{ki}x_{ks}\delta_{sj}. \tag{A.13}$$

Denote that

$$v_{ij} = \frac{1}{\sigma}z_{ij}. \tag{A.14}$$

As each data matrix $\mathbf{X}_l$ is normalized, according to Lemma 3, the v_{ij} are i.i.d standard normal variables and $v_{ij} \sim N(0,1)$. Thus

$$\frac{1}{\sigma}\|\mathbf{Z}\|_F^2 = \sum_i\sum_j v_{ij}^2 \tag{A.15}$$

is a χ^2 random variable with dL degrees of freedom, using Lemma 2 we have

$$\Pr\left(\frac{1}{nL}\|\mathbf{Z}\|_F \le \alpha\right) \le 1 - \exp\left(-\frac{1}{2}\left(b - dL\log\left(1+\frac{b}{dL}\right)\right)\right), \tag{A.16}$$

where $\alpha \ge \frac{\sigma}{nL}\sqrt{dL+b}$.

Thus, with probability at least $1-\exp(-\frac{1}{2}(b-dL\log(1+\frac{b}{dL})))$, we have

$$\frac{2}{nL}\langle \mathbf{Z}, \hat{\mathbf{S}}-\mathbf{S}\rangle \le \frac{2}{nL}\|\mathbf{Z}\|_F\|\hat{\mathbf{S}}-\mathbf{S}\|_F \tag{A.17}$$

$$\frac{2}{nL}\langle \mathbf{Z}, \hat{\mathbf{S}}-\mathbf{S}\rangle \le \alpha\|\hat{\mathbf{S}}-\mathbf{S}\|_F \tag{A.18}$$

$$\frac{2}{nL}\langle \mathbf{Z}, \hat{\mathbf{S}}-\mathbf{S}\rangle \le \alpha\|\hat{\mathbf{S}}-\mathbf{S}\|_{2,1}. \tag{A.19}$$

Therefore, using Lemma A.11 we have following inequality:

$$\alpha(\|\mathbf{S}\|_{2,1} - \|\hat{\mathbf{S}}\|_{2,1}) + \frac{2}{nL}\langle \mathbf{Z}, \hat{\mathbf{S}}-\mathbf{S}\rangle \le \alpha\|(\hat{\mathbf{S}}-\mathbf{S})_{\mathcal{I}(\mathbf{S})}\|_{2,1}. \tag{A.20}$$

In addition, the following triangle inequality always holds:

$$\|\mathbf{S}^\top\mathbf{S}\|_F^2 - \|\hat{\mathbf{S}}^\top\hat{\mathbf{S}}\|_F^2 \le \|\hat{\mathbf{S}}^\top\hat{\mathbf{S}} - \mathbf{S}^\top\mathbf{S}\|_F^2. \tag{A.21}$$

Substitute the above two inequalities into the inequality (A.11), we verify the theorem.□

Recall the assumption we made:

Assumption 2. For a matrix $\mathbf{\Gamma} \in \mathbb{R}^{d\times L}$, let $s \le d$. We assume that there exist constants $\kappa_1(s)$ and $\kappa_2(s)$ such that

$$\kappa_1(s) = \min_{\mathbf{\Gamma}\in\mathcal{R}(s)} \frac{\|\mathbf{X}^\top vec(\mathbf{M}\mathbf{\Gamma})\|}{\sqrt{dL}\|\mathbf{\Gamma}_{\mathcal{I}(\mathbf{S})}\|_F} > 0,$$

$$\kappa_2(s) = \min_{\mathbf{\Gamma}\in\mathcal{R}(s)} \frac{\|\mathbf{X}^\top vec(\mathbf{M}\mathbf{\Gamma})\|}{\sqrt{dL}\|\mathbf{\Gamma}^\top\mathbf{\Gamma}\|_F} > 0,$$

where the restricted set $\mathcal{R}(s)$ is defined as:

$$\mathcal{R}(s) = \{\mathbf{\Gamma} \in \mathbb{R}^{d\times L} : \mathbf{\Gamma} \ne \mathbf{0}, |\mathcal{I}(\mathbf{S})| \le s,$$
$$\|\mathbf{\Gamma}_{\mathcal{I}^c(\mathbf{S})}\|_{2,1} \le \alpha_1\|\mathbf{\Gamma}_{\mathcal{I}(\mathbf{S})}\|_{2,1}\}$$

where $|\mathcal{I}|$ counts the number of elements in the set $\mathcal{I}$.

Theorem 5. *Let $\hat{\mathbf{S}}$ be the optimal solution of Eq. (A.1) for $L \ge 2$ and $n, d \ge 1$, and $\mathbf{S}^*$ be the oracle solution, $\gamma = \|\mathbf{S}^*\|_F$. The regularization*

parameter α *is chosen as*

$$\alpha \geq \frac{2\sigma}{nL}\sqrt{dL+b},$$

where b *is a positive scalar. Then under Assumption 2, the follo wing results hold with a probability of at least* $1-\exp\left(-\frac{1}{2}(b-dL\log(1+\frac{b}{dL}))\right)$

$$\frac{1}{nL}\|\mathbf{X}^\top vec(\mathbf{M}\hat{\mathbf{S}})-vec(\mathbf{Y})\|^2 \leq \frac{\alpha^2 s}{\kappa_1^2(s)}\left(1-\beta\left(\frac{1}{\kappa_2(s)}+\frac{2\gamma}{\kappa_1}\right)^2\right)^{-2},$$

$$\|\hat{\mathbf{S}}-\mathbf{S}^*\|_{2,1} \leq \frac{\alpha\cdot s(\alpha_1+1)}{\kappa_1^2(s)}\left(1-\beta\left(\frac{1}{\kappa_2(s)}+\frac{2\gamma}{\kappa_1}\right)^2\right)^{-1},$$

$$\|(\hat{\mathbf{S}}^\top\hat{\mathbf{S}}-\mathbf{S}^{*\top}\mathbf{S}^*)\|_F \leq \frac{\alpha\sqrt{s}}{\kappa_1(s)}\left(\frac{1}{\kappa_2(s)}+\frac{2\gamma}{\kappa_1}\right)\left(1-\beta\left(\frac{1}{\kappa_2(s)}+\frac{2\gamma}{\kappa_1}\right)^2\right)^{-1}.$$

Proof. Using the Theorem 4 and setting $\mathbf{S}=\mathbf{S}^*$, we have

$$\frac{1}{nL}\|\mathbf{X}^\top\cdot vec(\mathbf{M}\hat{\mathbf{S}})-vec(\mathbf{F})\|^2 \tag{A.22}$$

$$\leq \alpha\|(\hat{\mathbf{S}}-\mathbf{S})_{\mathcal{I}(\mathbf{S})}\|_{2,1}+\beta\|\hat{\mathbf{S}}^\top\hat{\mathbf{S}}-\mathbf{S}^\top\mathbf{S}\|_F^2. \tag{A.23}$$

Under Assumption 2, we have

$$\|(\hat{\mathbf{S}}-\mathbf{S}^*)_{\mathcal{I}(\mathbf{S}^*)}\|_{2,1} \tag{A.24}$$

$$\leq \sqrt{s}\|(\hat{\mathbf{S}}-\mathbf{S}^*)_{\mathcal{I}(\mathbf{S}^*)}\|_F \tag{A.25}$$

$$\leq \frac{\sqrt{s}}{\kappa_1(s)\sqrt{nL}}\|\mathbf{X}^\top\cdot vec(\mathbf{M}\hat{\mathbf{S}})-vec(\mathbf{F})\|, \tag{A.26}$$

and

$$\|(\hat{\mathbf{S}}^\top\hat{\mathbf{S}}-\mathbf{S}^{*\top}\mathbf{S}^*)\|_F \tag{A.27}$$

$$=\|(\hat{\mathbf{S}}-\mathbf{S}^*)^\top(\hat{\mathbf{S}}-\mathbf{S}^*)+\mathbf{S}^{*\top}(\mathbf{S}^*-\hat{\mathbf{S}})+(\hat{\mathbf{S}}-\mathbf{S}^*)^\top\mathbf{S}^*\|_F \tag{A.28}$$

$$\leq\|(\hat{\mathbf{S}}-\mathbf{S}^*)^\top(\hat{\mathbf{S}}-\mathbf{S}^*)\|_F+\|\mathbf{S}^{*\top}(\mathbf{S}^*-\hat{\mathbf{S}})\|_F+\|(\hat{\mathbf{S}}-\mathbf{S}^*)^\top\mathbf{S}^*\|_F \tag{A.29}$$

$$\leq\|(\hat{\mathbf{S}}-\mathbf{S}^*)^\top(\hat{\mathbf{S}}-\mathbf{S}^*)\|_F+2\|\mathbf{S}^*\|_F\|(\mathbf{S}^*-\hat{\mathbf{S}})\|_F \tag{A.30}$$

$$\leq\frac{1}{\kappa_2(s)\sqrt{nL}}\|\mathbf{X}^\top\cdot vec(\mathbf{M}\hat{\mathbf{S}})-vec(\mathbf{F})\| +\frac{2\gamma}{\kappa_1\sqrt{nL}}\|\mathbf{X}^\top\cdot vec(\mathbf{M}\hat{\mathbf{S}})-vec(\mathbf{F})\| \tag{A.31}$$

$$=\left(\frac{1}{\kappa_2(s)\sqrt{nL}}+\frac{2\gamma}{\kappa_1\sqrt{nL}}\right)\|\mathbf{X}^\top\cdot vec(\mathbf{M}\hat{\mathbf{S}})-vec(\mathbf{F})\|. \tag{A.32}$$

Thus, we obtain

$$\|\mathbf{X}^\top vec(\mathbf{M}\hat{\mathbf{S}})-vec(\mathbf{Y})\| \leq \frac{\alpha\sqrt{s}\sqrt{nL}}{\kappa_1(s)}\left(1-\beta\left(\frac{1}{\kappa_2(s)}+\frac{2\gamma}{\kappa_1}\right)^2\right)^{-1}. \tag{A.33}$$

Assumption 2 also gives

$$\|\mathbf{\Gamma}_{\mathcal{I}^c(\mathbf{S})}\|_{2,1} \leq \alpha_1\|\mathbf{\Gamma}_{\mathcal{I}(\mathbf{S})}\|_{2,1}. \tag{A.34}$$

Setting $\Gamma=\hat{\mathbf{S}}-\mathbf{S}^*$, we obtain

$$\|\hat{\mathbf{S}}-\mathbf{S}^*\|_{2,1} \leq (\alpha_1+1)\|(\hat{\mathbf{S}}-\mathbf{S}^*)_{\mathcal{I}(\mathbf{S}^*)}\|_{2,1}. \tag{A.35}$$

Thus,

$$\|\hat{\mathbf{S}}-\mathbf{S}^*\|_{2,1} \leq \frac{\alpha\cdot s(\alpha_1+1)}{\kappa_1^2(s)}\left(1-\beta\left(\frac{1}{\kappa_2(s)}+\frac{2\gamma}{\kappa_1}\right)^2\right)^{-1}, \tag{A.36}$$

$$\|(\hat{\mathbf{S}}^\top\hat{\mathbf{S}}-\mathbf{S}^{*\top}\mathbf{S}^*)\|_F \leq \frac{\alpha\sqrt{s}}{\kappa_1(s)}\left(\frac{1}{\kappa_2(s)}+\frac{2\gamma}{\kappa_1}\right)\left(1-\beta\left(\frac{1}{\kappa_2(s)}+\frac{2\gamma}{\kappa_1}\right)^2\right)^{-1}.\square \tag{A.37}$$

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.neucom.2015.12.092.

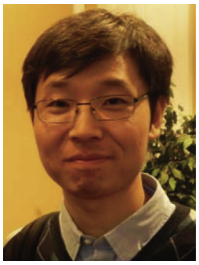
References

[1] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[2] S. Thrun, J. O'Sullivan, Clustering Learning Tasks and the Selective Cross-Task Transfer of Knowledge, Tech. Rep. CMU-CS-95-209, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, 1995.
[3] S. Negahban, M.J. Wainwright, Estimation of (near) low-rank matrices with noise and high-dimensional scaling, in: ICML, 2010.
[4] R. Caruana, Multitask learning, Mach. Learn. 28 (1) (1997) 41–75.
[5] B. Bakker, T. Heskes, Task clustering and gating for Bayesian multitask learning, J. Mach. Learn. Res. 4 (2003) 83–99.
[6] Y. Zhang, D.-Y. Yeung, Transfer metric learning by learning task relationships, in: SIGKDD, 2010.
[7] S. Parameswaran, K. Weinberger, Large margin multi-task metric learning, in: NIPS, 2010.
[8] A. Argyriou, T. Evgeniou, M. Pontil, Convex multi-task feature learning, Mach. Learn. 73 (3) (2008) 243–272.
[9] L. Jacob, F. Bach, J.-P. Vert, Clustered multi-task learning: a convex formulation, in: NIPS, 2008.
[10] Y. Xue, X. Liao, L. Carin, B. Krishnapuram, Multi-task learning for classification with Dirichlet process priors, J. Mach. Learn. Res. 8 (2007) 35–63.
[11] Z. Kang, K. Grauman, F. Sha, Learning with whom to share in multi-task feature learning, in: ICML, 2011.
[12] A. Kumar, H. Daumé III, Learning task grouping and overlap in multi-task learning, in: ICML, 2012.
[13] T.K. Pong, P. Tseng, S. Ji, J. Ye, Trace norm regularization: reformulations, algorithms, and multi-task learning, SIAM J. Optim. 20 (6) (2010) 3465–3489.
[14] S. Kim, E.P. Xing, Tree-guided group lasso for multi-task regression with structured sparsity, in: ICML, 2010.
[15] J. Liu, S. Ji, J. Ye, Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization, in: UAI, 2009.
[16] K. Lounici, M. Pontil, A.B. Tsybakov, S.A. van de Geer, Taking advantage of sparsity in multi-task learning, in: COLT, 2009.
[17] S. Negahban, M.J. Wainwright, Joint support recovery under high-dimensional scaling: benefits and perils of $\ell_{1,\infty}$-regularization, in: NIPS, 2008.
[18] X. Yang, S. Kim, E.P. Xing, Heterogeneous multitask learning with joint sparsity constraints, in: NIPS, 2009.
[19] Y. Zhang, D.-Y. Yeung, Q. Xu, Probabilistic multi-task feature selection, in: NIPS, 2010.
[20] A. Schwaighofer, V. Tresp, K. Yu, Learning Gaussian process kernels via hierarchical Bayes, in: NIPS, 2004.
[21] K. Yu, V. Tresp, A. Schwaighofer, Learning Gaussian processes from multiple tasks, in: ICML, 2005.
[22] J. Zhang, Z. Ghahramani, Y. Yang, Learning multiple related tasks using latent independent component analysis, in: NIPS, 2005.
[23] N.D. Lawrence, J.C. Platt, Learning to learn with the informative vector machine, in: ICML, 2004.
[24] R.K. Ando, T. Zhang, A Framework for learning predictive structures from multiple tasks and unlabeled data, J. Mach. Learn. Res. 6 (2005) 1817–1853.
[25] X. Chen, Q. Lin, S. Kim, J.G. Carbonell, E.P. Xing, Smoothing Proximal Gradient Method for General Structured Sparse Learning, CoRR abs/1202.3708.
[26] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, N. Sebe, No matter where you are: flexible graph-guided multi-task learning for multi-view head pose classification under target motion, in: ICCV, 2013.
[27] A. Passos, P. Rai, J. Wainer, H. Daumé III, Flexible modeling of latent task structures in multitask learning, in: ICML, 2012.
[28] S. Gupta, D. Phung, S. Venkatesh, Factorial multi-task learning : a Bayesian nonparametric approach, in: ICML, 2013.
[29] A. Jalali, P.D. Ravikumar, S. Sanghavi, C. Ruan, A dirty model for multi-task learning, in: NIPS, 2010.
[30] J. Chen, J. Zhou, J. Ye, Integrating low-rank and group-sparse structures for robust multi-task learning, in: SIGKDD, 2011.
[31] P. Gong, J. Ye, C. Zhang, Robust multi-task feature learning, in: SIGKDD, 2012.
[32] W. Zhong, J.T.-Y. Kwok, Convex multitask learning with flexible task clusters, in: ICML, 2012.
[33] B. Romera-Paredes, A. Argyriou, N. Berthouze, M. Pontil, Exploiting unrelated tasks in multi-task learning, in: AISTATS, 2012.

[34] A. Maurer, M. Pontil, B. Romera-Paredes, Sparse coding for multitask and transfer learning, in: ICML, 2012.
[35] J.-F. Cai, E.J. Cands, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (4) (2010) 1956–1982.
[36] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imag. Sci. 2 (1) (2009) 183–202.
[37] P.J. Bickel, Y. Ritov, A.B. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, Ann. Stat. 37 (4) (2009) 1705–1732.
[38] K. Lounici, M. Pontil, A.B. Tsybakov, S. van de Geer, Taking advantage of sparsity in multi-task learning, in: COLT, 2009.
[39] P. Gong, J. Ye, C. Zhang, Multi-stage multi-task feature learning, in: NIPS, 2012.
[40] J. Zhou, J. Chen, J. Ye, Clustered multi-task learning via alternating structure optimization, in: NIPS, 2011.
[41] Y. Zhang, D.-Y. Yeung, Multi-task learning using generalized t process, in: AISTATS, 2010.

Shi Zhong received the M.S. degree in Microelectronics from the School of Information Science and Technology, Fudan University, Shanghai, China, in 2003. He is currently working toward the Ph.D. degree in Computer Science at Fudan University. His research interests include multimedia and machine learning.

Jian Pu received the Ph.D. degree from Fudan University, Shanghai, China, in 2014. He is a Postdoctoral Researcher of Institute of Neuroscience, Chinese Academy of Sciences, Shanghai, China. His current research interests include machine learning, computer vision, and medical image computing.



Yu-Gang Jiang received the Ph.D. degree in computer science from the City University of Hong Kong, Kowloon, Hong Kong, in 2009. During 2008–2011, he was with the Department of Electrical Engineering, Columbia University, New York, NY, first year as a Visiting Scholar and later as a Post-Doctoral Research Scientist. He is currently an Associate Professor of computer science at Fudan University, Shanghai, China. His research interests include multimedia retrieval and computer vision.

Rui Feng received the Ph.D. degree from Shanghai Jiao Tong University, China, in 2003. He joined Fudan University, Shanghai, China, in 2003, where he is currently a Research Professor. His research interests include computer vision, multimedia and pattern recognition.

Xiangyang Xue received the B.S., M.S., and Ph.D. degrees in communication engineering from Xidian University, Xi'an, China, in 1989, 1992, and 1995, respectively. He joined the Department of Computer Science, Fudan University, Shanghai, China, in 1995, where he is currently a Professor. His current research interests include multimedia information processing and retrieval, pattern recognition and machine learning.

Please cite this article as: S. Zhong, et al., Flexible multi-task learning with latent task grouping, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2015.12.092