On Better Exploring and Exploiting Task Relationships in Multitask Learning: Joint Model and Feature Learning

Ya Li, Xinmei Tian, *Member, IEEE*, Tongliang Liu, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Multitask learning (MTL) aims to learn multiple tasks simultaneously through the interdependence between different tasks. The way to measure the relatedness between tasks is always a popular issue. There are mainly two ways to measure relatedness between tasks: common parameters sharing and common features sharing across different tasks. However, these two types of relatedness are mainly learned independently, leading to a loss of information. In this paper, we propose a new strategy to measure the relatedness that jointly learns shared parameters and shared feature representations. The objective of our proposed method is to transform the features of different tasks into a common feature space in which the tasks are closely related and the shared parameters can be better optimized. We give a detailed introduction to our proposed MTL method. Additionally, an alternating algorithm is introduced to optimize the nonconvex objection. A theoretical bound is given to demonstrate that the relatedness between tasks can be better measured by our proposed MTL algorithm. We conduct various experiments to verify the superiority of the proposed joint model and feature MTL method.**

***Index Terms*— Feature learning, multitask learning (MTL), parameter sharing.**

I. Introduction

SINGLE-TASK learning learns different tasks separately and ignores the intrinsic relatedness between different tasks. Multitask learning (MTL) can get rid of this drawback by jointly measuring the interdependence between different tasks. The performance of all tasks is can be improved with additional information provided by the relationship between tasks. Considering the merits of MTL, it has been applied to various research areas, e.g., Web image search [1], video tracking [2], disease prediction [3], and relative attributes learning [4].

MTL assumes that tasks have some intrinsic relatedness. Consequently, the proper measurement of task relatedness will benefit the learning of tasks and improve the performance of each other. Conversely, improper relatedness measurement introduces noise and degrades the performance. Recently, researchers have given substantial attention to measuring task relatedness. Existing algorithms mainly use two methods to measure the relatedness between tasks: shared common models/parameters [5]–[10] and shared common feature representations [11]–[17]. MTL of sharing common models/parameters (multitask model learning) makes the assumption that the models of different tasks have something in common in their parameters. MTL of sharing common feature representations (multitask feature learning) assumes that related tasks share a subset of common features.

Both multitask model learning and multitask feature learning suffer from their own defects. They only consider one aspect of task relatedness. The relatedness is directly captured in the original feature space in multitask model learning. However, considering the noise and complexity of features in real-world data sets, task relatedness measured by the original features may not be obvious. As a result, the performance of multitask model learning may degrade. Multitask feature learning tackles this drawback by learning a common subset of feature representations. However, it ignores the relatedness between model parameters. We develop a new multitask model and feature joint learning method in this paper that can successfully explore and exploit task relatedness. Our model learns a set of common features shared by different with which the relatedness between tasks is maximized. Consequently, the common models can be better measured jointly.

The objective function of the proposed method is formulated as a nonconvex problem and an alternating algorithm is proposed to optimize it. Additionally, we present sound theoretical analyses to prove the better ability of measuring task relatedness with our joint model and feature learning method. Various experimental results are reported to demonstrate the effectiveness of our proposed method, especially on tasks with shared features or shared models.

The remainder of this paper is organized as follows. In Section II, we briefly review previous related works in MTL. In Section III, we give a detailed derivation and optimization algorithm of our proposed method. Section IV derives a theoretical error bound to demonstrate

Manuscript received January 31, 2016; revised September 22, 2016 and March 24, 2017; accepted March 24, 2017. Date of publication April 17, 2017; date of current version April 16, 2018. This work was supported in part by the 973 Project under Grant 2015CB351803, in part by NSFC under Grant 61572451 and Grant 61390514, in part by the Youth Innovation Promotion Association CAS under Grant CX2100060016, in part by the Fok Ying Tung Education Foundation under Grant WF2100060004, and in part by the Fundamental Research Funds for the Central Universities WK2100060011, and in part by the Australian Research Council under Project FT-130101457, Project DP-140102164, and Project LP-150100671. *(Corresponding Author: Xinmei Tian.)*

Y. Li and X. Tian are with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application Systems, University of Science and Technology of China, Hefei 230027, China (e-mail: muziyiye@mail.ustc.edu.cn; xinmei@ustc.edu.cn).

T. Liu and D. Tao are with the UBTech Sydney Artificial Intelligence Institute and the School of Information Technologies in the Faculty of Engineering and Information Technologies at The University of Sydney, Sydney, NSW 2008, Australia (e-mail: tongliang.liu@sydney.edu.au; dacheng.tao@sydney.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2690683

the merits of our proposed algorithm. Experimental results are reported in Section V with conclusions and future work given in Section VI.

II. Related Work

In recent years, researchers have paid much attention to MTL. Compared with single-task learning, its effectiveness has been demonstrated through the theoretical analysis in many works [18]–[22]. For example, a novel inductive bias learning method was proposed by Baxter [18]. This paper derived explicit bounds, demonstrating that learning multiple-related tasks within an environment potentially achieves substantially better generalization than does learning a single task. Ben-David and Schuller [19] proposed a useful concept of task relatedness to derive a better generalization of error bounds. Maurer *et al.* [20] applied the dictionary learning and sparse coding to MTL and introduced a generalization bound by measuring the hypothesis complexity. Ando and Zhang [21] made assumption that all tasks shared a common structure and showed a reliable estimation of shared parameters between tasks when the number of tasks was large.

With more extensive applications of MTL, some single-task learning algorithms have been extended to an MTL framework. For example, some works extended the Bayesian method into MTL methods with the assumption that the models of tasks are indeed related [23]. Hierarchical Bayesian models can be learned by sharing parameters as hyperparameters at a high level. The relatedness between tasks can also be measured by deep neural networks, such as sharing nodes or layers of the network [24]. As one of the most popular single-task learning methods, a support vector machine (SVM) has been studied in many MTL works [5], [12], [25]–[30]. Jebara [12] proposed an MTL method using maximum entropy discrimination based on the large-margin SVM. Zhu *et al.* [25] proposed an infinite latent SVM for MTL. It combines the large-margin idea with a nonparametric Bayesian model to discover the latent features for MTL.

The most difficult aspect of MTL is simultaneously measuring the relatedness between tasks and keeping the individual characteristics. Multitask model learning and multitask feature learning are two main categories of MTL methods. For multitask model learning, Xue *et al.* [6] proposed two different forms of MTL problem using a Dirichlet process-based statistical model and developed efficient algorithms to solve the proposed methods. Evgeniou and Pontil [5] introduced an MTL model by minimizing a regularized objection similar to SVMs. This paper assumed that all tasks shared a mean hyperplane with a particular offset on their own. A nonparametric Bayesian model was proposed by Rai and Daume [8] to capture task relatedness under the assumption that parameters shared a latent subspace. The dimensionality of the subspace was automatically inferred by the proposed model. For the category of multitask feature learning, Argyriou *et al.* [11] developed a convex MTL method for learning shared features between tasks. The learned common features were regularized by a L21-norm to control the dimensionality of the latent feature space. Jebara [12] proposed a general MTL framework using large-margin classifiers. Three scenarios were discussed: multitask feature learning, multitask kernel combination, and graphical multitask model [12]. To improve the efficiency of MTL on high-dimensional problems, a novel MTL method was proposed by learning low-dimensional features of tasks jointly [13].

Recently, the defects of measuring task relatedness in traditional MTL methods have been widely discussed. The assumptions that all tasks are related through sharing common parameters or common features are usually not suitable for real-world MTL problems. Considering the defects of such assumptions, a number of works [31]–[34] have been proposed to improve the performance of MTL. For example, Kang *et al.* [31] learned a shared feature representation across tasks while simultaneously clustering the tasks into different groups. Chen *et al.* [33] proposed a robust MTL method that learned multiple tasks jointly while simultaneously finding outlier tasks. Another robust multitask feature learning method was proposed by Gong *et al.* [32]. This model was similar to the method in [33]. This paper decomposed the weight matrix into two components and imposed the group Lasso penalty on both components. The group Lasso penalty was imposed on the row of the first component for capturing the shared features between relevant tasks, and the same group Lasso penalty was imposed on the column of the second component to find the outlier tasks. Another work [34] proposed a dirty model for MTL by utilizing an idea similar to [32] and [33]. The model used both block-sparse regularization and elementwise sparsity regularization to capture the true features used for each task. Block-sparse regularization learned the shared features across tasks, and elementwise regularization guaranteed that some features were used for some tasks but not all. These works can be divided into two categories: task clustering and outlier task finding.

However, these works only consider one aspect of task relatedness: either shared features or shared parameters. In this paper, we consider the shared features and shared parameters simultaneously to overcome the problems in the existing MTL methods. The relatedness can be better modeled in our MTL framework, especially when both feature relatedness and model relatedness exist between the tasks.

III. Multitask Model and Feature Joint Learning

We introduce our newly proposed MTL method specifically in this section. We first show the objective optimization problem and then convert the nonconvex problem into a convex formulation. An efficient optimization algorithm is given at the end of this section.

The idea of our proposed method is shown in Fig. 1. There are three related tasks in the original feature space. However, the interdependence between them is not as strong as assumed in MTL due to the noise and complexity of feature representation. It may lead to bad performances of MTL in the original feature space. In this paper, we transform the original feature space into a new feature space, in which different tasks are tightly related and are possible to share a common hyperplane a_0. The specific characteristic of task t is represented by an offset a_t.

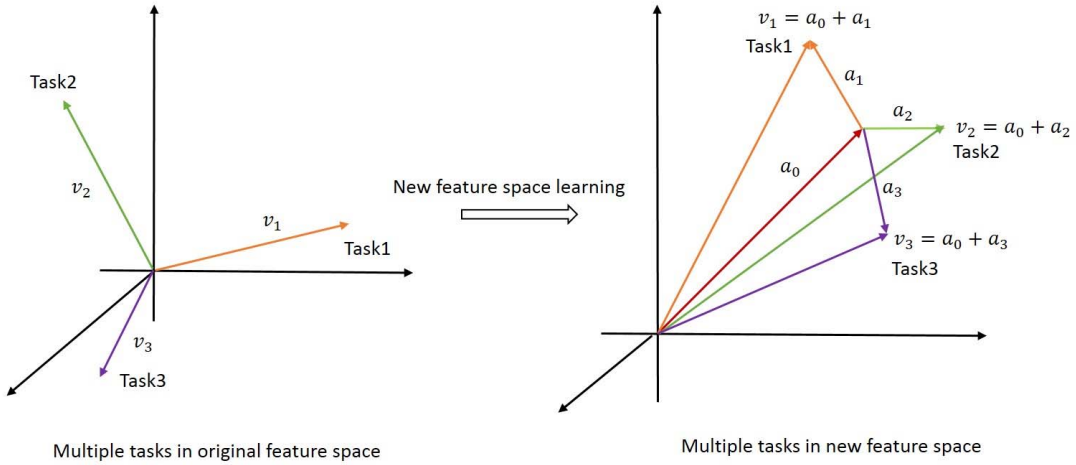


Fig. 1. Framework of our proposed MTL method. In the original feature space, tasks have weak relatedness. We aim to map the data into a new space, and therefore, all tasks can be more closely related and share a common hyperplane in this new feature space.

A. Nonconvex Objective

Suppose we have T different tasks. Each task t is related to a data set D_t that can be formulated as follows:

$$D_t = \{(x_{t1}, y_{t1}), (x_{t2}, y_{t2}), \ldots, (x_{tm_t}, y_{tm_t})\}$$

where m_t is the number of data samples in task t. $x_{ti} \in (R)^d$ and $y_{ti} \in \mathbb{R}$ are the corresponding feature representation and output of sample i in task t. In this paper, we consider to learn T different linear functions to predict the output, given the input feature representation in each task

$$f_t(x_{ti}) = v_t^T x_{ti} \approx y_{ti} \tag{1}$$

where $t \in \{1, 2, \ldots, T\}$. Single-task learning methods treat these T linear functions as separate tasks and just utilize the data information from each task. Consequently, it ignores the interdependence between tasks, which may provide more valuable information about the distribution of training data. Considering the drawbacks of the single-task learning algorithm, MTL is proposed to uncover the relatedness between tasks and gains performance improvement of all tasks. The improvement is expected to be obvious especially with a small amount of training data. The relatedness between tasks can provide more additional information in such situation.

Considering the weak interdependence between tasks in the original feature space, we propose to learn a feature mapping matrix $U \in \mathbb{R}^{d \times d}$

$$f_t(x_{ti}) = \langle a_t + a_0, U^T x_{ti} \rangle. \tag{2}$$

The weight v_t is decomposed into two components: a_0 and a_t. a_0 is the shared central hyperplane in the new feature space and a_t represents the offset of task t to maintain its own characteristic. The learned feature mapping matrix U is supposed to guarantee the assumption that all tasks share a central hyperplane with an offset in the new feature space. With the above formulation, our objective function of MTL can be formulated as follows:

$$\min_{V,a_0,U} \sum_{t=1}^{T} \sum_{i=1}^{m_t} l(y_{ti}, \langle v_t, U^T x_{ti} \rangle) + \frac{\gamma}{T} \|V - a_0 * \mathbf{1}\|_{2,1}^2 + \beta \|a_0\|_2^2 \tag{3}$$

where $V = [v_1, v_2, \ldots, v_T]$ and $\mathbf{1}$ represents a vector of all ones. Noting that $v_t = a_t + a_0$, we can reformulate problem (3) with $A = [a_1, a_2, \ldots, a_T]$ as

$$\min_{A,a_0,U} \sum_{t=1}^{T} \sum_{i=1}^{m_t} l(y_{ti}, \langle a_t + a_0, U^T x_{ti} \rangle) + \frac{\gamma}{T} \|A\|_{2,1}^2 + \beta \|a_0\|_2^2. \tag{4}$$

The third regularization term in problem (4) denotes the square of the L2-norm of vector a_0 and it aims to measure the smoothness and complexity of the central hyperplane. The second regularization term is the square of the L21-norm of matrix A, which can be explicitly expressed as $\|A\|_{2,1} = (\sum_{i=1}^{d} \|a^i\|_2)$. a^i denotes the ith row of matrix A. The L21-norm guarantees that all tasks share a subset of common features and the sparsity of shared features. The first term is the loss function, which measures the error between ground truth and predicted results.

There are three main differences between our proposed formulation and the formulation proposed in [11]. First, the learning ability of feature mapping matrix U has some limitation due to its orthogonal property. However, such limitations are ignored in [11]. It is more reasonable to share a subset of common features around a_0 instead of the fixed original point. The proposed method can well prevent the limitation of orthogonal matrix U by selecting features around a more robust point a_0. Second, our method considers the task relatedness of both features and model parameters. However, the method proposed in [11] just uncovered the shared common features across tasks leading to loss of information between related models. These tasks were treated independently when learning their model parameters in the learned new feature

space. Third, it is more challenging to solve an optimization problem learning both of common features and common model parameters.

The proposed objective function is nonconvex. To briefly show the nonconvexity of the problem, we give a counter example. Assuming that all the variables are scalars, it is easy to show that the proposed objective is nonconvex. It is usually difficult to get an optimal solution of a nonconvex objective. Instead, we convert the nonconvex objective into an equivalent convex problem. And an alternating algorithm is proposed to solve it in Section III-C.

B. Conversion to an Equivalent Convex Optimization Problem

For simple optimization, the nonconvex optimization problem (4) is converted into an equivalent convex problem in this section.

Theorem 1: The nonconvex problem (4) can be equivalently converted into a convex optimization problem as follows:

$$\begin{aligned}\min_{W,w_0,D}\ & \sum_{t=1}^{T}\sum_{i=1}^{m_t} l(y_{ti},\langle w_t+w_0,x_{ti}\rangle)\\ &+\frac{\gamma}{T}\sum_{t=1}^{T}\langle w_t, D^+w_t\rangle+\beta\langle w_0,w_0\rangle\\ \text{s.t. }& \text{trace}(D)\le 1, \text{range}(W)\subseteq\text{range}(D), D\in S_+^d. \end{aligned}\tag{5}$$

If $(\hat{W},\hat{w_0},\hat{D})$ is an optimal solution of convex problem (5), the corresponding optimal solution $(\hat{A},\hat{a_0},\hat{U})$ of nonconvex problem (4) can be formulated as $\hat{A}=\hat{U}^T\hat{W}$, $\hat{a_0}=\hat{U}^T\hat{w_0}$ and the columns of $\hat{U}$ form an orthonormal basis of eigenvectors of $\hat{D}$. Additionally, if $(\hat{A},\hat{a_0},\hat{U})$ forms an optimal solution of nonconvex problem (4), the corresponding optimal solution of convex problem (5) can be formulated as $\hat{W}=\hat{U}\hat{A}$, $\hat{w_0}=\hat{U}\hat{a_0}$, and $\hat{D}=\hat{U}Diag((\|\hat{a}^i\|_2)/(\|\hat{A}\|_{2,1}))_{i=1}^d\hat{U}^T$.

Note that $trace(D)=\sum_{i=1}^d D_{ii}$ and $D\in S_+^d$ indicates that D is a positive semidefinite symmetric matrix. $range(W)$ represents a set of vectors $\{x\in\mathbb{R}^n : x=Wz$, for some $z\in\mathbb{R}^T\}$. D^+ denotes the pseudoinverse of matrix D. $Diag(a_0)_{i=1}^d$ is a diagonal matrix and the vector a_0 forms the diagonal elements.

To show the convexity of problem (5), an additional function is introduced as $f:\mathbb{R}^d\times S^d\to\mathbb{R}\bigcup\{+\infty\}$, which can be explicitly formulated as

$$f(w,D)=\begin{cases} w^TD^+w & \text{if } D\in S_+^d \text{ and } w\in\text{range}(D)\\ +\infty. \end{cases}\tag{6}$$

With the additional function, problem (5) is equal to minimizing the sum of T additional functions plus the loss term and the term $\langle w_0,w_0\rangle$ in problem (5), subjected to the trace constraints. Its rightness can be guaranteed by the equality between the T constraints $w_t\in range(D)$ and the constraint $range(W)\in range(D)$. The loss term in problem (5) is the sum of loss function l, which is convex for (w_t,w_0) and a linear map, and therefore, it is convex. Additionally, the term $\langle w_0,w_0\rangle$ and the trace constraint are also convex. To show the convexity of problem (5), it is sufficient to show that f is convex. The details of f being convex can be found in [11].

C. Optimization Algorithm

An alternating optimization algorithm is proposed to solve problem (5) in this section. Then, the final optimal solution of problem (4) can be obtained according to Theorem 1.

We first optimize problem (5) with respect to parameters (W,w_0) by fixing matrix D. This optimization problem can be separated into T different tasks with a fix D in [11]. Comparing with the optimization of the objective in [11], the optimization of our newly proposed objective function is more challenging because of the shared parameter w_0. It cannot be viewed as T independent optimization problems. Our objective can be formulated as follows:

$$\begin{aligned}\min_{W,w_0}\ & \sum_{t=1}^{T}\sum_{i=1}^{m_t} l(y_{ti},\langle w_t+w_0,x_{ti}\rangle)\\ &+\frac{\gamma}{T}\sum_{t=1}^{T}\langle w_t, D^+w_t\rangle+\beta\langle w_0,w_0\rangle,\\ \text{s.t. }& \text{trace}(D)\le 1, \text{range}(W)\subseteq\text{range}(D),\quad D\in S_+^d. \end{aligned}\tag{7}$$

The loss function used in this paper is a least square loss, which is the same as that used in previous works. To solve problem (7), we introduce some additional variables. Note that $X_t=[x_{t1},x_{t2},\ldots,x_{tm_t}]\in\mathbb{R}^{d\times m_t}$ which denotes a data matrix of task t and the corresponding output of task t is represented as $Y_t=[y_{t1},y_{t2},\ldots,y_{tm_t}]^T\in\mathbb{R}^{m_t}$. M denotes the sum of amount of data points from all T tasks

$$M=m_1+m_2+\cdots+m_T.$$

Let $X=b\text{diag}(X_1,X_2,\ldots,X_T)\in\mathbb{R}^{dT\times M}$ and $Y=[Y_1^T,Y_2^T,\ldots,Y_T^T]^T\in\mathbb{R}^M$. $b\text{diag}\{X_1,X_2,\ldots,X_T\}$ denotes a block diagonal matrix and its diagonal entries are data from the T tasks. Y denotes the outputs of all data belonging to the T different tasks. Let $D_0=b\text{diag}(\underbrace{D,D,\ldots,D}_{T})\in\mathbb{R}^{dT\times M}$, $W_0=[\underbrace{w_0^T,w_0^T,\ldots,w_0^T}_{T}]^T\in\mathbb{R}^{dT}$, and $W_1=[w_1^T,w_2^T,\ldots,w_T^T]^T\in\mathbb{R}^{dT}$.

Problem (7) can be reformulated as

$$\min_{W_1,W_0}\ \|Y-X^T(W_1+W_0)\|_2^2+\frac{\gamma}{T}W_1^TD_0^+W_1+\beta w_0^Tw_0.\tag{8}$$

Note that $W_0=I_0\times w_0$ with $I_0=[\underbrace{I,I,\ldots,I}_{T}]^T\in\mathbb{R}^{dT\times d}$ and $I\in\mathbb{R}^{d\times d}$ denotes an identity matrix. We can reformulate problem (8) as an L2-norm regularized regression problem with some additional variables. Note that $Z_1=\sqrt{(\gamma/T)}(D_0^+)^{(1/2)}W_1$, and let $Z_2=\sqrt{\beta}w_0$. Then, $W_1=\sqrt{(T/\gamma)}(D_0^+)^{-(1/2)}Z_1$ and $W_0=\sqrt{(1/\beta)}I_0Z_2$. $(D_0^+)^{(1/2)}=b\text{diag}(\underbrace{(D^+)^{(1/2)},(D^+)^{(1/2)},\ldots,(D^+)^{(1/2)}}_{T})$ and $(D_0^+)^{-(1/2)}=b\text{diag}(\underbrace{(D^+)^{-(1/2)},(D^+)^{-(1/2)},\ldots,(D^+)^{-(1/2)}}_{T})$.

We have

$$\frac{\gamma}{T}W_1^T D_0^+ W_1 + \beta w_0^T w_0 = \left[Z_1^T, Z_2^T\right]\left[Z_1^T, Z_2^T\right]^T = Z^T Z$$

$$W_1 + W_0 = \left[\sqrt{\frac{T}{\gamma}}\left(D_0^+\right)^{-\frac{1}{2}}, \sqrt{\frac{1}{\beta}}I_0\right]\left[Z_1^T, Z_2^T\right]^T = MZ. \tag{9}$$

$M = [\sqrt{(T/\gamma)}(D_0^+)^{-(1/2)}, \sqrt{(1/\beta)}I_0]$ and $Z = [Z_1^T, Z_2^T]$. Consequently, the above problem is reformulated as the following standard L2-norm regularized problem:

$$\min_Z \|Y - X^T MZ\|_2^2 + Z^T Z. \tag{10}$$

The solution can be explicitly expressed as

$$Z = (M^T XX^T M + I)^{-1} M^T XY. \tag{11}$$

Additionally, we need to optimize problem (5) with respect to matrix D by fixing parameters (W, w_0). The objective can be simply formulated as

$$\min_D \sum_{t=1}^T \langle w_t, D^+ w_t\rangle$$
$$\text{s.t. } D \in S_+^d, \text{trace}(D) \le 1, \text{range}(W) \subseteq \text{range}(D). \tag{12}$$

The optimal solution is explicitly shown as (the details can be found in [11])

$$\hat{D} = \frac{(WW^T)^{\frac{1}{2}}}{\text{trace}(WW^T)^{\frac{1}{2}}}. \tag{13}$$

IV. Theoretical Analysis

For better understanding the merits of our method, a generalization bound of the nonconvex problem (4) is analyzed in this section. We first reformulate the problem by converting the two soft constraints $(\gamma/T)\|A\|_{2,1}^2$ and $\beta\|a_0\|_2^2$ into hard ones

$$\min_{a_t, a_0, U, \varepsilon_1, \varepsilon_2} \sum_{t=1}^T \sum_{i=1}^{m_t} l(y_{ti}, \langle a_t + a_0, U^T x_{ti}\rangle) + \varepsilon_1 + \varepsilon_2,$$
$$\text{s.t. } \gamma \frac{1}{T}\|A\|_{2,1}^2 \le \varepsilon_1$$
$$\beta\|a_0\|_2^2 \le \varepsilon_2. \tag{14}$$

The demonstration of the equality between problem (14) and problem (4) can be found in [35], and ε_1 and ε_2 are of order $\mathcal{O}(1)$. Let $\varepsilon_1 = \varepsilon_2 = \mathcal{O}(1)$; then, the above problem becomes

$$\min_{a_t, a_0, U} \sum_{t=1}^T \sum_{i=1}^{m_t} l(y_{ti}, \langle a_t + a_0, U^T x_{ti}\rangle)$$
$$\text{s.t. } \|A\|_{2,1}^2 \le \mathcal{O}\left(\frac{T}{\gamma}\right)$$
$$\|a_0\|_2^2 \le \mathcal{O}\left(\frac{1}{\beta}\right). \tag{15}$$

Consequently, we analyze the problem with hard constraints instead. We derive a generalization bound of the proposed problem following a similar way to that of [36] by setting $\varepsilon = 1$:

$$\min_{a_t, a_0, U} \sum_{t=1}^T \sum_{i=1}^{m_t} l(y_{ti}, \langle a_t + a_0, U^T x_{ti}\rangle)$$
$$\text{s.t. } \|A\|_{2,1}^2 \le \frac{T}{\gamma}$$
$$\|a_0\|_2^2 \le \frac{1}{\beta}. \tag{16}$$

We assume that the loss function l satisfies the following Lipschitz-like condition, to upper bound the generalization error.

Definition 1: A loss function l is c-admissible with respect to the hypothesis class H if there exists a $c \in \mathbb{R}_+$, where $\mathbb{R}_+$ denotes the set of nonnegative real numbers, such that for any two hypotheses $h, h' \in H$ and example $(x, y) \in \mathcal{X} \times \mathbb{R}$, the following inequality holds:

$$|l(y, h(x)) - l(y, h'(x))| \le c|h(x) - h'(x)|.$$

We can have Theorem 2.

Theorem 2: Suppose B is the upper bound of loss function l, i.e., $l(y, f(x)) \le B$, and the loss function l is c-admissible corresponding to the linear function class. For any optimal solution (A, a_0, U) of problem (4), by replacing the hard constraints $\|A\|_{2,1}^2 \le (T/\gamma)$ and $\|a_0\|_2^2 \le (1/\beta)$ with the soft constraints $\gamma(1/T)\|A\|_{2,1}^2$ and $\beta\|a_0\|_2^2$, and for any $\delta > 0$, we have the following results with the probability of at least $1 - \delta$:

$$E_x \sum_{t=1}^T \sum_{i=1}^{m_t} l(y_{ti}, \langle a_t + a_0, U^T x_{ti}\rangle) - \sum_{t=1}^T \sum_{i=1}^{m_t} l(y_{ti}, \langle a_t + a_0, U^T x_{ti}\rangle)$$
$$\le 2c\left(\sqrt{\frac{T}{\gamma}} + \sqrt{\frac{1}{\beta}}\right)\sqrt{\sum_{t=1}^T m_t S(X_t)} + 3B\sqrt{\frac{\sum_{t=1}^T m_t \ln\left(\frac{2}{\delta}\right)}{2}}$$

where $S(X_t) = \text{tr}(\hat{\Sigma}(x_t)) = (1/m_t)\sum_{i=1}^{m_t}\|x_{ti}\|_2^2$ is the empirical covariance for the observations of the tth task. Letting $m_1 = \ldots = m_T = m$ and $\|x_t\|_2 \le r, t = 1, \ldots, T$, with a probability of at least $1 - \delta$, we have

$$\frac{1}{T}\sum_{t=1}^T E_x l(y_t, \langle a_t + a_0, U^T x_t\rangle) - \frac{1}{T}\sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m l(y_{ti}, \langle a_t + a_0, U^T x_{ti}\rangle)$$
$$\le \frac{2cr}{\sqrt{\gamma m}} + \frac{2cr}{\sqrt{\beta m T}} + 3B\sqrt{\frac{\ln(2/\delta)}{2mT}}.$$

Remark 1: The first term $(2cr)/(\sqrt{\gamma m})$ in Theorem 2 is the generalization bound related to the learning of matrix A and the second term $(2cr)/(\sqrt{\beta m T})$ corresponding to a_0. This theoretical result demonstrates that the learning of shared hyperplane a_0 is of order $\mathcal{O}(\sqrt{1/mT})$ and it can be better learned with more tasks. Consequently, our proposed multitask

TABLE I
EXPERIMENTAL RESULTS COMPARISON ON SCHOOL DATA SET

Measure	Training ratio	L2-R	L1-R	TraceMT	LowRankMT	CMTL	SLMT	MTDirty	MTMF
nMSE	10%	1.0398 ± 0.0038	1.0261 ± 0.0132	0.9359 ± 0.0370	0.9175 ± 0.0261	0.9413 ± 0.0021	0.9130 ± 0.0039	0.9543 ± 0.0129	$\mathbf{0.7783 \pm 0.0082}$
	20%	0.8773 ± 0.0043	0.8754 ± 0.0194	0.8211 ± 0.0032	0.8126 ± 0.0132	0.8327 ± 0.0039	0.8055 ± 0.0103	0.8396 ± 0.0142	$\mathbf{0.7432 \pm 0.0045}$
	30%	0.8171 ± 0.0090	0.8144 ± 0.0091	0.7870 ± 0.0012	0.7657 ± 0.0091	0.7922 ± 0.0052	0.7600 ± 0.0032	0.7985 ± 0.0053	$\mathbf{0.7299 \pm 0.0064}$
aMSE	10%	$0.2713 \pm +0.0023$	0.2682 ± 0.0036	0.2504 ± 0.0102	0.2419 ± 0.0081	0.2552 ± 0.0032	0.2330 ± 0.0018	0.2327 ± 0.0031	$\mathbf{0.1898 \pm 0.0018}$
	20%	0.2303 ± 0.0003	0.2289 ± 0.0051	0.2156 ± 0.0015	0.2114 ± 0.0041	0.2131 ± 0.0071	0.2018 ± 0.0025	0.2048 ± 0.0036	$\mathbf{0.1813 \pm 0.0010}$
	30%	0.2156 ± 0.0021	0.2137 ± 0.0012	0.2089 ± 0.0012	0.2011 ± 0.0022	0.1922 ± 0.0102	0.1822 ± 0.0014	0.1943 ± 0.0016	$\mathbf{0.1776 \pm 0.0019}$

joint learning method can perform better than single-task learning methods. Additionally, a_0 is encouraged to be larger with the constraints of $\|A\|_{2,1}$ and the utility of feature mapping matrix U. Thus, the generalization bound of our proposed method has a faster convergence than the method proposed in [11], which demonstrates the efficiency of our method.

The proof of Theorem 2 is given in the Appendix.

V. EXPERIMENTS

We show various experimental results and analyses to demonstrate the effectiveness of our proposed MTL method in this section. The comparison with several state-of-the-art MTL algorithms further supports the merits of our multitask model and feature joint learning methods [multi-task model and feature (MTMF)]. We compare our MTMF with two single-task learning methods—L2-norm regularized regression (L2-R) and L1-norm regularized regression (L1-R), as well as five state-of-the-art MTL algorithms, including trace norm regularized multitask learning, low-rank regularized multitask learning with sparse structure [37], convex multitask feature learning (CMTL) [11], multitask learning with a dirty model [34], and group sparse and low-rank regularized robust MTL [33]. These five MTL algorithms are the representative methods of MTL, and the performance of them has been demonstrated to be promising on various data sets. The comparison with these methods can sufficiently demonstrate the effectiveness of our proposed MTMF. The data sets used in our experiments are School data set,[1] SARCOS data set,[2] Isolet data set,[3] and MNIST data set.[4]

A. School Data Set

This data set was collected to evaluate the effectiveness of schools by Inner London Education Authority. It consists of 139 related tasks to predict the examination scores of students from 139 secondary schools. The information of each student is encoded into a binary feature vector of 27 dimensions. There are totally 15 362 samples. Single-task learning methods, such as L1-R and L2-R, learn these 139 tasks independently using their own data. All MTL methods aim to improve the performance of these 139 tasks by uncovering the relatedness between the tasks. The experimental settings follow the previous works to fairly compare their performance.

Different ratios (10%, 20%, and 30%) of training samples are randomly selected for training, and the rest of samples are split into validation and test set. Consider the randomness of selection which may cause large variations in the results, we repeat all selections ten times. All parameters are selected via the validation set. For all the methods, the performance is evaluated by the average mean squared error (aMSE) and normalized mean squared error (nMSE), which have been used in [32] and [33]. The aMSE can be calculated through dividing the mean squared error by the variance of target vector, and the nMSE can be calculated through dividing the mean squared error by the squared norm of target vector.

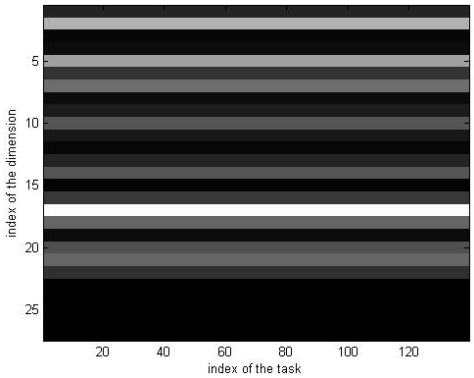


Fig. 2. Visualization of matrix A_0 learned on School data set.

Table I gives the performance of all the methods on School data set. From Table I, we can conclude that all MTL methods can well uncover the relationships between tasks and improve the performance comparing with single-task learning methods. Another observation is that our proposed method performs the best with different training ratios. The improvement is especially obvious with a small amount of training samples, which indicates the success of our method to learn a new feature space and the strong ability of discovering latent relatedness between the tasks.

To analyze the properties of learned weight matrix $A_0 = [\underbrace{a_0, a_0, \ldots, a_0}_{T}]$ and A, we visualize them in Figs. 2 and 3. The results are obtained using 20% of the training samples. The zero values are denoted as black pixels in the figures. Most of the pixels in Fig. 3 are black, which reveal the sparsity of the learned matrix A. A small subset of the features are shared across tasks corresponding to the 15 nonzero rows of matrix A. From Fig. 2, we observe that A_0 is also a sparse matrix. However, the features not used in matrix A appear in A_0, which means that MTMF can better utilize the information of the features. If we only use A, all the tasks are forced to share some of the features without the utilization of other features. The relatedness between the

[1] http://ttic.uchicago.edu/ argyriou/code/.
[2] http://www.gaussianprocess.org/gpml/data/.
[3] https://archive.ics.uci.edu/ml/datasets/ISOLET
[4] http://yann.lecun.com/exdb/mnist/

TABLE II
EXPERIMENTAL RESULTS COMPARISON ON SARCOS DATA SET

Measure	Training size	L2-R	L1-R	TraceMT	LowRankMT	CMTL	SLMT	MTDirty	MTMF
nMSE	50	0.2454 ± 0.0260	0.2337 ± 0.0180	0.2257 ± 0.0065	0.2127 ± 0.0033	0.2192 ± 0.0016	0.2123 ± 0.0038	0.1742 ± 0.0178	$\mathbf{0.1640 \pm 0.0208}$
	100	0.1821 ± 0.0142	0.1616 ± 0.0027	0.1531 ± 0.0017	0.1495 ± 0.0023	0.1568 ± 0.0037	0.1456 ± 0.0138	0.1274 ± 0.0060	$\mathbf{0.1155 \pm 0.0215}$
	150	0.1501 ± 0.0054	0.1469 ± 0.0028	0.1318 ± 0.0053	0.1236 ± 0.0004	0.1301 ± 0.0034	0.1245 ± 0.0015	0.1129 ± 0.0039	$\mathbf{0.1057 \pm 0.0043}$
aMSE	50	0.1330 ± 0.0143	0.1228 ± 0.0083	0.1122 ± 0.0064	0.1073 ± 0.0026	0.1156 ± 0.0011	0.0982 ± 0.0026	0.0625 ± 0.0063	$\mathbf{0.0588 \pm 0.0074}$
	100	0.1053 ± 0.0096	0.0907 ± 0.0023	0.0805 ± 0.0026	0.0793 ± 0.0047	0.0852 ± 0.0013	0.0737 ± 0.0083	0.0458 ± 0.0021	$\mathbf{0.0415 \pm 0.0023}$
	150	0.0846 ± 0.0045	0.0822 ± 0.0014	0.0772 ± 0.0023	0.0661 ± 0.0062	0.0755 ± 0.0025	0.0674 ± 0.0014	0.0405 ± 0.0011	$\mathbf{0.0379 \pm 0.0012}$

TABLE III
EXPERIMENTAL RESULTS COMPARISON ON ISOLET DATA SET

Measure	Training ratio	TraceMT	LowRankMT	CMTL	SLMT	MTDirty	MTMF
nMSE	15%	0.6044 ± 0.0154	0.6307 ± 0.0058	0.7000 ± 0.0106	0.5987 ± 0.0092	0.6764 ± 0.0112	$\mathbf{0.5691 \pm 0.0082}$
	20%	0.5705 ± 0.0069	0.6166 ± 0.0093	0.6491 ± 0.0108	0.5741 ± 0.0078	0.6344 ± 0.0182	$\mathbf{0.5526 \pm 0.0046}$
	25%	0.5622 ± 0.0086	0.6011 ± 0.0165	0.6288 ± 0.0049	0.5635 ± 0.0087	0.6212 ± 0.0299	$\mathbf{0.5498 \pm 0.0090}$
aMSE	15%	0.1424 ± 0.0035	0.1486 ± 0.0019	0.1650 ± 0.0029	0.1411 ± 0.0024	0.1594 ± 0.0029	$\mathbf{0.1314 \pm 0.0019}$
	20%	0.1343 ± 0.0015	0.1452 ± 0.0022	0.1528 ± 0.0025	0.1352 ± 0.0017	0.1494 ± 0.0043	$\mathbf{0.1301 \pm 0.0012}$
	25%	0.1321 ± 0.0025	0.1412 ± 0.0042	0.1477 ± 0.0017	0.1324 ± 0.0025	0.1459 ± 0.0076	$\mathbf{0.1292 \pm 0.0025}$

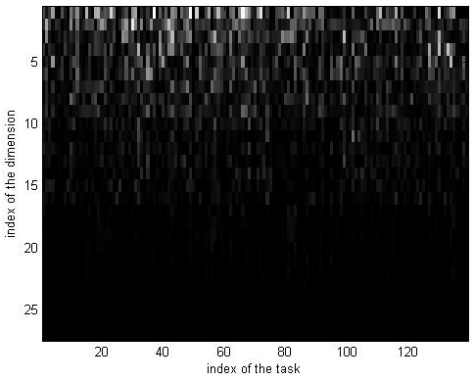


Fig. 3. Visualization of matrix A learned on School data set.

tasks becomes closer than they really are. A_0 helps all the tasks utilize more information from the features that are not shared through the matrix A. This is one of the reasons that our MTMF outperforms the CMTL.

B. SARCOS Data Set

This data set is used to learn the inverse dynamic of an SARCOS anthropomorphic robot arm. It aims to predict the seven joint torques using the provided 48 933 samples described by a feature vector of 21 dimensions. In this experiment, we have seven tasks corresponding to predict these seven joint torques. Different amounts of samples (50, 100, 150) are randomly selected as training data and 5000 samples are selected correspondingly as the validation set and test set. The best parameters are selected on the validation set for all the methods. Consider the randomness of selection, we repeat all experiments 15 times and average performance is reported.

The comparison of the experimental results between different methods is shown in Table II. Similar conclusions can be made to those of experiments on School data set. Our proposed method can consistently outperform all other algorithms, and all MTL methods perform better than the two single-task learning methods. This further demonstrates the merits of MTL and effectiveness of our method compared with other MTL methods.

C. Isolet Data Set

In this section, we conduct experiments on Isolet data set from the UCI repository. It consists of 7797 pronunciation samples of the 26 English letters from 150 speakers. These speakers are split into five groups corresponding to five different tasks. The goal of each task is to predict the labels (1–26) of letters according to the pronunciation. In the experiment, the labels of English letters are treated as the regression values following the same setup as used in [38]. Different ratios (15%, 20%, and 25%) of samples are randomly selected as the training data, and the rest is split into the validation set and test set. All the experiments are repeated ten times, and the best parameters are selected on the validation set. We first reduce the dimensionality of the data to 100 using principal component analysis (PCA).

The performance is reported in Table III. Note that the two single-task learning methods L2-R and L1-R are not tested on Isolet data set because of the bad performance on School and SARCOS data sets. Our proposed MTL method outperforms other baselines obviously on this data set, which proves the robustness of our method on various applications.

D. MNIST Data Set

We further study the effectiveness of our approach on a handwritten digit recognition data set: MNIST. This data set is composed of 60 000 training examples and 10 000 test examples. There are ten different handwritten digit numbers, corresponding to ten different binary classification tasks. Multiway classification is treated as an MTL problem, where each task is a classification task of one digit against all the other digits [31], [39]. We randomly select 500, 1000, and 1500 examples (50, 100, and 150 examples are selected from each digit number) from the 60 000 training samples as training set and 1000 samples from the test samples to form the test sets. The dimensionality of images is reduced to 64 using PCA. All the experiments are repeated 20 times, and mean average precision is reported.

The results on this data set are shown in Table IV. We compare our MTMF method with five other multitask regression learning methods. The results show that our proposed method

TABLE IV

EXPERIMENTAL RESULTS COMPARISON ON MNIST

Training size	50	100	150
TraceMT	0.8088 ± 0.0118	0.8297 ± 0.0114	0.8382 ± 0.0111
LowRankMT	0.7483 ± 0.0260	0.8088 ± 0.0181	0.8289 ± 0.0192
CMTL	0.8091 ± 0.0108	0.8343 ± 0.0124	0.8391 ± 0.0115
SLMT	0.7578 ± 0.0165	0.8144 ± 0.0160	0.8264 ± 0.0147
MTDirty	0.7955 ± 0.0131	0.8152 ± 0.0128	0.8202 ± 0.0191
MTMF	$\mathbf{0.8180 \pm 0.0125}$	$\mathbf{0.8394 \pm 0.0144}$	$\mathbf{0.8484 \pm 0.0111}$

TABLE V

p-VALUES BETWEEN OUR PROPOSED METHOD AND THE NEXT BEST METHOD ON ALL THE DATA SETS

Index Number	School dataset	SARCOS dataset	Isolet dataset	MNIST dataset
1	3.47×10^{-5}	0.5963	1.47×10^{-10}	4.34×10^{-5}
2	2.86×10^{-8}	0.4447	2.87×10^{-6}	2.59×10^{-2}
3	2.64×10^{-6}	0.4245	1.51×10^{-6}	1.33×10^{-5}
4	3.89×10^{-6}	0.5923	2.42×10^{-10}	-
5	2.26×10^{-9}	0.4436	2.76×10^{-6}	-
6	9.28×10^{-5}	0.4244	1.62×10^{-6}	-

outperforms the other five MTL methods on the MNIST data set.

E. Analysis on p-Values

To further demonstrate that the proposed method is indeed statistically significantly better than the next best method, we present the p-values between our proposed method and the next best method in Table V. Table V includes six groups of experiments on the School data set (nMSE: 10%, 20%, and 30%; aMSE: 10%, 20%, and 30%), SARCOS data set (nMSE: 50, 100, and 150; aMSE: 50, 100, and 150), Isolet data set (nMSE: 15%, 20%, and 25%; aMSE: 15%, 20%, and 25%), and three groups of experiments on the MNIST data set (AP: 50, 100, and 150). We index the experiments for all the data sets from 1 to 6. From Table V, our proposed method significantly outperforms the next best methods on the School data set, Isolet data set, and MNIST data set, as the p-values are substantially smaller than 0.05. On the SARCOS data set, our method does not perform significantly better than the next best method. However, the proposed method performs much better than all other methods.

The main reason that our proposed method outperforms other MTL methods is that our proposed method considers the shared features and shared parameters simultaneously. Therefore, our proposed method can perform better if the data have both feature relatedness and model relatedness. Additionally, we can balance the importance between feature relatedness and model relatedness through tradeoff parameters γ and β. Thus, our model can degenerate to just share feature representations or share model. Consequently, our proposed model is more robust to various data.

F. Sensitivity Analysis on MTMF

In this section, we conduct the experiments to analyze the sensitivity of our proposed MTMF method. We will mainly discuss how the regularization parameters γ and β and the training size affect the performance of our MTMF method. All the experiments are conducted on the School data set.

1) Analysis of the Training Ratio: In these experiments, we randomly select 10%, 20%, 30%, 40%, 50%, and 60% of the data as the training sets and use the remaining data as the test sets. We study how the training size affects the performance of MTMF. The experiments are repeated ten times, and the regularization parameters (γ, β) are selected through validation. The results are shown in Fig. 4. We can conclude that the proposed method outperforms the other methods significantly with consistent increase of training ratio. It is also found that the performance of MTL algorithms improves quicker when having a small amount of training samples and that the performance improves only slightly when the training ratio reaches a high level. It is consistent with the learning ability of MTL. The relatedness between different tasks can provide more information to each task especially when the amount of training data is small. This results in a rapid increase in performance. However, the contribution of information from other tasks will decrease when task itself has sufficient training samples, which leads to a smaller increase in performance.

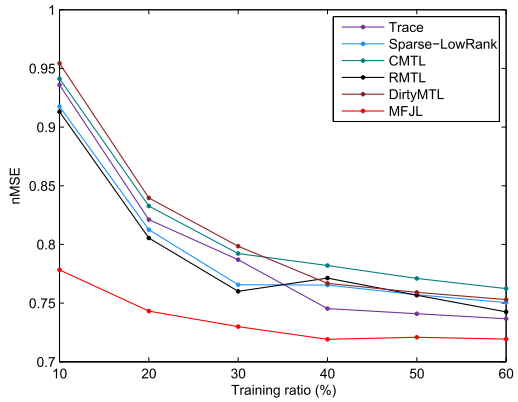


Fig. 4. Sensitivity analysis on training size.

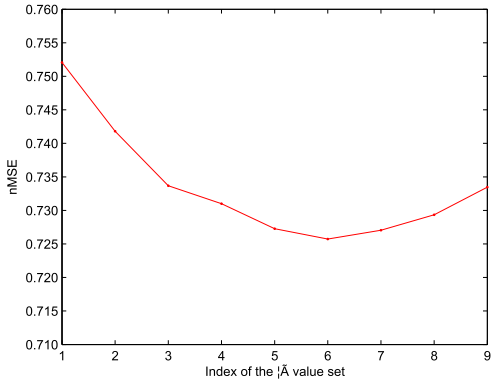


Fig. 5. Sensitivity analysis on γ.

2) Analysis of the Regularization Parameters: We conduct the experiments on the School data set to analyze the sensitivity of the two regularization parameters. We randomly select 20% of the data as the training set and the remaining data as the test set. For the sensitivity analysis of the parameter γ, we fix $\beta = 1$ and vary the value of γ as $\{1, 10, 100, 200, 500, 1000, 2000, 3000, 5000\}$. For the parameter β, we fix $\gamma = 100$ and vary the value of β as $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 50, 100\}$. The results are shown in Figs. 5 and 6. In Fig. 5, we can see that the best performance by MTMF is obtained by setting $\gamma = 1000$ when $\beta = 1$ is fixed. From Fig. 6, we see that the best performance

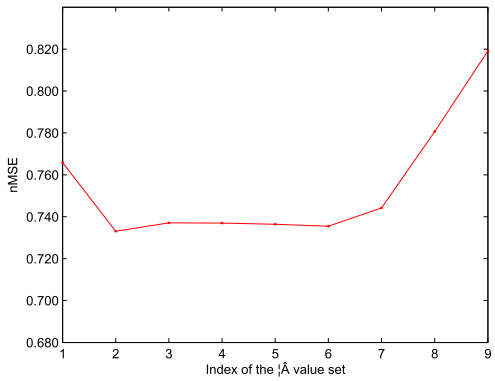


Fig. 6. Sensitivity analysis on β.

by MTMF is obtained by setting the value of β as a small value. Additionally, the performance of MTMF changes slightly when the value of β is in the range of $[10^{-4}, 1]$. In general, MTMF performs well when the ratio (γ/β) reaches a relatively high value (approximately 1000). This means that only a few features will be shared across tasks and that the central hyperplane a_0 will play an important role.

VI. Conclusion and Future Work

In this paper, we summarize the defects of the traditional MTL methods and propose a novel MTL framework, which learns shared latent feature representation and shared parameters jointly. The proposed method is introduced in detail, and a new algorithm for optimizing the nonconvex problem is proposed. Additionally, we theoretically demonstrate the merits of the proposed method compared with single-task learning and its strong ability to measure the relatedness between tasks. We conduct various experiments on four MTL data sets, and the results have demonstrated the effectiveness of the proposed method.

In the future, we consider to extend the multitask model and feature joint learning method into a more general framework. In this paper, the learned feature mapping matrix U is an orthogonal matrix. It may be more efficient if the orthogonal matrix U is replaced by a common matrix. Additionally, we make assumptions that all tasks share a common parameter, which is not suitable for some real-world cases. Considering this, we will attempt to automatically learn the relatedness between tasks and not make assumptions about the relatedness.

Appendix
Proof of Theorem 2

Before we provide the proof of Theorem 2, we need to introduce some used tools. We first give an introduction to the concentration inequality [40], which is better known as Hoeffding's inequality.

Theorem 3: Let $x_1, \ldots, x_n$ be independent random variables with the range $[a_i, b_i]$ for $i = 1, \ldots, n$. Let $S_n = \sum_{i=1}^{n} x_i$. Then, for any $\epsilon > 0$, the following inequalities hold:

$$\Pr\{S_n - ES_n \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$
$$\Pr\{ES_n - S_n \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

We then introduce the Rademacher complexity [41], which is suitable to derive dimensionality-independent generalization bounds.

Definition 2: Let $X = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ be an independent distributed sample, and let F be a function class on $\mathcal{X}$. Let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher variables, which are uniformly distributed in $\{-1, 1\}$. The empirical Rademacher complexity is defined as

$$\mathfrak{R}_n(F) = E_\sigma \sup_{f \in F} \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i).$$

The Rademacher complexity is defined as

$$\mathfrak{R}(F) = E_x \mathfrak{R}_n(F).$$

According to the symmetric distribution property of random variables, the following theorem [42] holds.

Theorem 4: Let

$$\Phi(X) = \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (E_x f(x) - f(x_i)).$$

Then

$$E_x \Phi(X) \leq \mathfrak{R}(F).$$

Combining Theorem 4 and Hoeffding's inequality, we have the following.

Theorem 5 [42]: Let F be an $[a, b]$-valued function class on $\mathcal{X}$, and $X = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$. For any $\delta > 0$, with a probability of at least $1 - \delta$, we have

$$\sup_{f \in F}\left(E_x f(x) - \frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) \leq \mathfrak{R}(F) + (b-a)\sqrt{\frac{\ln(1/\delta)}{2n}}$$

or

$$\sup_{f \in F}\left(E_x f(x) - \frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) \leq \mathfrak{R}_n(F) + 3(b-a)\sqrt{\frac{\ln(2/\delta)}{2n}}$$

The following property of Rademacher complexity [41] will help to construct the upper bound.

Lemma 1: If $\phi : \mathbb{R} \to \mathbb{R}$ is Lipschitz with constant L and satisfies $\phi(0) = 0$, then

$$\mathfrak{R}_n(\phi \circ F) \leq 2L\mathfrak{R}_n(F).$$

Lemma 2: Let

$$\mathfrak{R}_n(l \circ (A, a_0, U)) = 2E_\sigma \sup_{a_t, a_0, U} \sum_{t=1}^{T}\sum_{i=1}^{m_t} \sigma_{ti} l(y_{ti}, \langle a_t + a_0, U^T x_{ti}\rangle)$$

where σ_{ti} are Rademacher variables indexed by $t = 1, \ldots, T$ and $i = 1, \ldots, m_t$. We have

$$\mathfrak{R}_n(l \circ (A, a_0, U)) \leq 2c\left(\sqrt{\frac{T}{\gamma}} + \sqrt{\frac{1}{\beta}}\right)\sqrt{\sum_{t=1}^{T} m_t S(X_t)}$$

where $S(X_t) = \mathrm{tr}(\hat{\Sigma}(x_t)) = \frac{1}{m_t}\sum_{i=1}^{m_t} \|x_{ti}\|_2^2$ is the empirical covariance for the observations of the tth task.

Proof: We have

$$
\begin{aligned}
&\mathfrak{R}_n(l \circ (A, a_0, U)) \\
&= 2E_\sigma \sup_{a_t, a_0, U} \sum_{t=1}^{T} \sum_{i=1}^{m_t} \sigma_{ti} l(y_{ti}, \langle a_t + a_0, U^T x_{ti} \rangle) \\
&= 2E_\sigma \sup_{a_t, a_0, U} \sum_{t=1}^{T} \sum_{i=1}^{m_t} \sigma_{ti} l(y_{ti}, \langle U(a_t + a_0), x_{ti} \rangle) \\
&\qquad \text{(Using Lemma 1)} \\
&\leq 2cE_\sigma \sup_{a_t, a_0, U} \sum_{t=1}^{T} \sum_{i=1}^{m_t} \sigma_{ti} \langle U(a_t + a_0), x_{ti} \rangle \\
&\leq 2cE_\sigma \sup_{a_t, U} \sum_{t=1}^{T} \sum_{i=1}^{m_t} \sigma_{ti} \langle Ua_t, x_{ti} \rangle \\
&\quad + 2cE_\sigma \sup_{a_0, U} \sum_{t=1}^{T} \sum_{i=1}^{m_t} \sigma_{ti} \langle Ua_0, x_{ti} \rangle \\
&= 2cE_\sigma \sup_{a_t, U} \sum_{t=1}^{T} \left\langle Ua_t, \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\rangle \\
&\quad + 2cE_\sigma \sup_{a_0, U} \left\langle Ua_0, \sum_{t=1}^{T} \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\rangle \\
&\qquad \text{(Using Hölder's inequality)} \\
&\leq 2cE_\sigma \sup_{a_t, U} \sqrt{\sum_{t=1}^{T} \|Ua_t\|_2^2} \sqrt{\sum_{t=1}^{T} \left\| \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\|_2^2} \\
&\quad + 2cE_\sigma \sup_{a_0, U} \|Ua_0\|_2 \left\| \sum_{t=1}^{T} \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\|_2 \\
&\qquad \text{(Since } U^T U = I\text{)} \\
&= 2cE_\sigma \sup_{a_t} \sqrt{\sum_{t=1}^{T} \|a_t\|_2^2} \sqrt{\sum_{t=1}^{T} \left\| \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\|_2^2} \\
&\quad + 2cE_\sigma \sup_{a_0} \|a_0\|_2 \left\| \sum_{t=1}^{T} \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\|_2 \\
&\leq 2cE_\sigma \sup_{a_t} \|A\|_{2,1} \sqrt{\sum_{t=1}^{T} \left\| \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\|_2^2} \\
&\quad + 2cE_\sigma \sup_{a_0} \|a_0\|_2 \left\| \sum_{t=1}^{T} \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\|_2 \\
&\qquad \left(\text{Since } \|A\|_{2,1}^2 \leq \frac{T}{\gamma}, \|a_0\|_2^2 \leq \frac{1}{\beta} \right) \\
&\leq \frac{2c\sqrt{T}}{\sqrt{\gamma}} E_\sigma \sqrt{\sum_{t=1}^{T} \left\| \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\|_2^2} + \frac{2c}{\sqrt{\beta}} E_\sigma \left\| \sum_{t=1}^{T} \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\|_2 \\
&= \frac{2c\sqrt{T}}{\sqrt{\gamma}} E_\sigma \sqrt{\sum_{t=1}^{T} \left\| \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\|_2^2} + \frac{2c}{\sqrt{\beta}} E_\sigma \sqrt{\left\| \sum_{t=1}^{T} \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\|_2^2} \\
&\qquad \text{(Since the sqrt function is concave)} \\
&\leq \frac{2c\sqrt{T}}{\sqrt{\gamma}} \sqrt{E_\sigma \sum_{t=1}^{T} \left\| \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\|_2^2} \\
&\quad + \frac{2c}{\sqrt{\beta}} \sqrt{E_\sigma \left\| \sum_{t=1}^{T} \sum_{i=1}^{m_t} \sigma_{ti} x_{ti} \right\|_2^2} \\
&\qquad \text{(Since } \sigma_{ti} \text{ are independent and } E\sigma_{ti} = 0, E\sigma_{ti}^2 = 1\text{)} \\
&\leq \frac{2c\sqrt{T}}{\sqrt{\gamma}} \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{m_t} \|x_{ti}\|_2^2} + \frac{2c}{\sqrt{\beta}} \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{m_t} |x_{ti}\|_2^2} \\
&\leq 2c \left(\sqrt{\frac{T}{\gamma}} + \sqrt{\frac{1}{\beta}} \right) \sqrt{\sum_{t=1}^{T} m_t S(X_t)}.
\end{aligned}
$$

■

Theorem 2 follows by combining Theorem 5 and Lemma 2.

REFERENCES

[1] X. Wang, C. Zhang, and Z. Zhang, "Boosted multi-task learning for face verification with applications to Web image and video search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 142–149.

[2] X. Mei, Z. Hong, D. Prokhorov, and D. Tao, "Robust multitask multiview tracking in videos," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 11, pp. 2874–2890, Nov. 2015.

[3] D. Zhang and D. Shen, "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease," *Neuroimage*, vol. 59, no. 2, pp. 895–907, 2012.

[4] L. Chen, Q. Zhang, and B. Li, "Predicting multiple attributes via relative multi-task learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1027–1034.

[5] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 109–117.

[6] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram, "Multi-task learning for classification with Dirichlet process priors," *J. Mach. Learn. Res.*, vol. 8, pp. 35–63, May 2007.

[7] K. Yu, V. Tresp, and A. Schwaighofer, "Learning Gaussian processes from multiple tasks," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 1012–1019.

[8] P. Rai and H. Daume, "Infinite predictor subspace models for multitask learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 613–620.

[9] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l 2, 1-norm minimization," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009, pp. 339–348.

[10] S.-I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller, "Learning a meta-level prior for feature relevance from multiple related tasks," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 489–496.

[11] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, 2008.

[12] T. Jebara, "Multitask sparsity via maximum entropy discrimination," *J. Mach. Learn. Res.*, vol. 12, no. 1, pp. 75–110, 2011.

[13] M. Lapin, B. Schiele, and M. Hein, "Scalable multitask representation learning for scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1434–1441.

[14] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. Roy. Statist. Soc., Ser. B, Statist. Methodol.*, vol. 70, no. 1, pp. 53–71, 2008.

[15] G. Obozinski, B. Taskar, and M. I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Statist. Comput.*, vol. 20, no. 2, pp. 231–252, Apr. 2010.

[16] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[17] C. Hou, F. Nie, D. Yi, and Y. Wu, "Feature selection via joint embedding learning and sparse regression," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2011, vol. 22. no. 1, p. 1324.

[18] J. Baxter. (Jun. 2011). "A model of inductive bias learning." [Online]. Available: https://arxiv.org/abs/1106.0245

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

[20] A. Maurer, M. Pontil, and B. Romera-Paredes. (Sep. 2012). "Sparse coding for multitask and transfer learning." [Online]. Available: https://arxiv.org/abs/1209.0738
[21] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, Nov. 2005.
[22] C. Li, M. Georgiopoulos, and G. C. Anagnostopoulos, "Multitask classification hypothesis space with improved generalization bounds," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1468–1479, Jul. 2015.
[23] T. Heskes, "Empirical Bayes for learning to learn," in *Proc. ICML*, 2000, pp. 367–374.
[24] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
[25] J. Zhu, N. Chen, and E. P. Xing, "Infinite latent svm for classification and multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1620–1628.
[26] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," *J. Mach. Learn. Res.*, vol. 6, no. 12, pp. 615–637, Apr. 2005.
[27] F. Cai and V. Cherkassky, "Generalized SMO algorithm for SVM-based multitask learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 997–1003, Jun. 2012.
[28] C. Li, M. Georgiopoulos, and G. C. Anagnostopoulos, "A unifying framework for typical multitask multiple kernel learning problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1287–1297, Jul. 2014.
[29] Y. Li, X. Tian, M. Song, and D. Tao, "Multi-task proximal support vector machine," *Pattern Recognit.*, vol. 48, no. 10, pp. 3249–3257, 2015.
[30] C. Li, M. Georgiopoulos, and G. C. Anagnostopoulos, "Pareto-path multitask multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 51–61, Jan. 2015.
[31] Z. Kang, K. Grauman, and F. Sha, "Learning with whom to share in multi-task feature learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 521–528.
[32] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 895–903.
[33] J. Chen, J. Zhou, and J. Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 42–50.
[34] A. Jalali, P. Ravikumar, and S. Sanghavi, "A dirty model for multiple sparse regression," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 7947–7968, Dec. 2013.
[35] D. Vainsencher, S. Mannor, and A. M. Bruckstein, "The sample complexity of dictionary learning," *J. Mach. Learn. Res.*, vol. 12, pp. 3259–3281, Feb. 2011.
[36] N. Mehta and A. G. Gray, "Sparsity-based generalization bounds for predictive sparse coding," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 36–44.
[37] J. Chen, J. Liu, and J. Ye, "Learning incoherent sparse and low-rank patterns from multiple tasks," *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 4, p. 22, 2012.
[38] P. Gong, J. Ye, and C.-S. Zhang, "Multi-stage multi-task feature learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1988–1996.
[39] Y. Amit, M. Fink, and N. Srebro, "Uncovering shared structures in multiclass classification," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 17–24.
[40] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.
[41] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Mar. 2003.
[42] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2012.

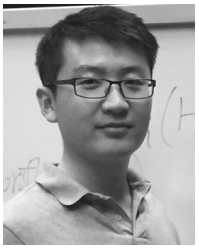
Ya Li received the B.S. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2013. He is currently pursuing the Ph.D. degree with USTC.

His current research interests include machine learning.

Xinmei Tian (M'13) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

She is currently an Associate Professor with the CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System, University of Science and Technology of China. Her current research interests include multimedia information retrieval and machine learning.

Dr. Tian received the Excellent Doctoral Dissertation of Chinese Academy of Sciences Award in 2012 and the Nomination of the National Excellent Doctoral Dissertation Award in 2013.



Tongliang Liu received the B.Eng. degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, and the Ph.D. degree from the University of Technology Sydney, Ultimo, NSW, Australia.

He is currently a Lecturer with the School of Information Technologies and the Faculty of Engineering and Information Technologies, and a Core Member with the UBTech Sydney AI Institute, The University of Sydney, Sydney, NSW, Australia. He has authored or co-authored over 20 research papers, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the TRANSACTIONS ON IMAGE PROCESSING, the International Conference on Machine Learning, and Knowledge Discovery and Data Mining. His current research interests include statistical learning theory, computer vision, and optimization.

Dacheng Tao (F'15) was a Professor of Computer Science and the Director of the Centre for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW, Australia. He is currently a Professor of Computer Science and ARC Future Fellow with the School of Information Technologies and the Faculty of Engineering and Information Technologies, and the Founding Director of the UBTech Sydney Artificial Intelligence Institute, The University of Sydney, Sydney, NSW, Australia. He mainly applies statistics and mathematics to artificial intelligence and data science. His research results have expounded in one monograph and over 500 publications at prestigious journals and prominent conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the TRANSACTIONS ON IMAGE PROCESSING, the *Journal of Machine Learning Research*, the International Journal of Computer Vision, the Neural Information Processing Systems, the Conference on Information and Knowledge Management, the International Conference on Machine Learning, the Conference on Computer Vision and Pattern Recognition, the International Conference on Computer Vision, the European Conference on Computer Vision, the International Conference on Artificial Intelligence and Statistics, the International Conference on Data Mining (ICDM), and the ACM Special Interest Group on Knowledge Discovery and Data Mining. His current research interests include computer vision, data science, image processing, machine learning, and video surveillance.

Dr. Tao is a fellow of OSA, IAPR, and SPIE. He received several best paper awards, such as the best theory/algorithm paper runner up award in the IEEE ICDM'07, the Best Student Paper Award in the IEEE ICDM'13, and the 2014 ICDM 10-year highest impact paper award. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award, and the 2015 UTS Vice-Chancellor's Medal for Exceptional Research.