Integrating Low-Rank and Group-Sparse Structures for Robust Multi-Task Learning

Jianhui Chen, Jiayu Zhou, Jieping Ye
Computer Science and Engineering, Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, Tempe, AZ 85287
{Jianhui.Chen, Jiayu.Zhou, Jieping.Ye}@asu.edu

ABSTRACT

Multi-task learning (MTL) aims at improving the generalization performance by utilizing the intrinsic relationships among multiple related tasks. A key assumption in most MTL algorithms is that all tasks are related, which, however, may not be the case in many real-world applications. In this paper, we propose a robust multi-task learning (RMTL) algorithm which learns multiple tasks simultaneously as well as identifies the irrelevant (outlier) tasks. Specifically, the proposed RMTL algorithm captures the task relationships using a low-rank structure, and simultaneously identifies the outlier tasks using a group-sparse structure. The proposed RMTL algorithm is formulated as a non-smooth convex (unconstrained) optimization problem. We propose to adopt the accelerated proximal method (APM) for solving such an optimization problem. The key component in APM is the computation of the proximal operator, which can be shown to admit an analytic solution. We also theoretically analyze the effectiveness of the RMTL algorithm. In particular, we derive a key property of the optimal solution to RMTL; moreover, based on this key property, we establish a theoretical bound for characterizing the learning performance of RMTL. Our experimental results on benchmark data sets demonstrate the effectiveness and efficiency of the proposed algorithm.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - Data Mining

General Terms

Algorithms

Keywords

Multi-task learning, robust, low-rank patterns, group-sparsity

1. INTRODUCTION

Multi-task learning (MTL) aims at improving generalization performance by utilizing the intrinsic relationships among multiple tasks. One key assumption in most of the existing MTL algorithms is that all tasks are correlated via a certain structure, which, for example, includes hidden units in neural networks [7], a common prior in a hierarchical Bayesian model [3, 29, 35, 37], parameters in Gaussian process covariance [17], kernels and regularizations [10], and common feature representation [1, 9, 2, 27, 16, 34]. Under such an assumption, the knowledge learned from one task is transferable to the other tasks and learning multiple tasks simultaneously generally leads to improved performance.

In many real-world applications involving multiple tasks, it is usually the case that a group of tasks are related while some other tasks are irrelevant to such a group. Simply pooling all tasks together and learning them simultaneously under a presumed structure may degrade the overall learning performance. It is thus desirable to identify irrelevant (outlier) tasks in the development of the multi-task learning algorithms. Learning multiple tasks under this setting is usually referred to as robust multi-task learning [36].

Recently robust multi-task learning has received increasing attention in the areas of data mining and machine learning. In [30, 33, 13], the task clustering (TC) approach is proposed for discovering the common structures in multiple learning tasks. The main idea behind the TC algorithms is to cluster similar tasks into different groups and constrain the tasks from the same group to share the same model representation or parameters. In [36, 38], multivariate student t-processes and their generalization are proposed for distinguishing good tasks from noisy or outlier tasks. The t-processes-based MTL algorithms model the relationship of multiple tasks using a task covariance matrix and they are robust by nature as t-process is implicitly an infinite Gaussian mixture. In [26, 14], the block-sparse structures ($\ell_{1,\infty}$-norm or $\ell_{2,1}$-nrom) are employed to extract essential features shared across the tasks and hence improve the robustness of the learning algorithms.

In this paper, we propose a robust multi-task learning (RMTL) algorithm which learns multiple tasks simultaneously as well as identifies the irrelevant (outlier) tasks. Specifically, our proposed RMTL algorithm captures the relationship of multiple related tasks using a low-rank structure and meanwhile identifies the outlier tasks using a group-sparse structure. The proposed RMTL algorithm is formulated as a non-smooth convex (unconstrained) optimization problem in which the least squares loss is regularized by a nonnegative linear combination of the trace norm and the $\ell_{1,2}$-norm. The optimization problems involving the trace norm and the $\ell_{1,2}$-norm can be routinely reformulated as semi-definite programs or second-order cone programs, both of which, however, are not scalable to large-scale data. We propose to adopt the accelerated proximal method (APM) for solving the proposed RMTL formulation efficiently. One key component in applying APM for solving RMTL is the computation of the associated proximal op-

KDD'11, August 21–24, 2011, San Diego, California, USA.

erator, which is a non-smooth optimization problem involving two optimization variables. The associated proximal operator can be shown to admit an analytic solution. We also conduct theoretical analysis on the performance bound of the composite regularization in RMTL. We first present key properties of the optimal solution to RMTL (Lemma 4.3). We then present an assumption associated with the prescribed training samples and the geometric structures of the matrices of interest; based on this assumption, we derive a performance bound for the combined regularization for multi-task regression (Theorem 4.1). We conduct simulations on benchmark data sets to demonstrate the effectiveness and efficiency of the proposed algorithm.

Notation Denote $\mathbb{N}_m = \{1, \cdots, m\}$. For any $A = [a_1, \cdots, a_m] \in \mathbb{R}^{d\times m}$, let $a_i \in \mathbb{R}^d$ be the i-th column of A; denote by $\|a_i\|_2$ the ℓ_2-norm of a_i; let $\|A\|_{\infty,2} = \|a_j\|_2$, where $j = \arg\max_i \|a_i\|_2$; let $\|A\|_{1,2} = \sum_{i=1}^m \|a_i\|_2$; let $\{\sigma_i(A)\}_{i=1}^r$ be the set of non-zero singular values in non-increasing order, where $r = \text{rank}(A)$; denote by $\|A\|_2 = \sigma_1(A)$ and by $\|A\|_* = \sum_{i=1}^r \sigma_i(A)$ the operator norm and the trace norm of A, respectively.

2. ROBUST MTL FRAMEWORK

Assume that we are given m (regression) learning tasks. Each task is associated with a set of training data

$$\{(x_1^i, y_1^i), \cdots, (x_{n_i}^i, y_{n_i}^i)\} \subset \mathbb{R}^d \times \mathbb{R},\ i \in \mathbb{N}_m, \tag{1}$$

and a linear predictive function f_i as

$$f_i(x_j^i) = w_i^T x_j^i \approx y_j^i,\ x_j^i \in \mathbb{R}^d,\ y_j^i \in \mathbb{R}, \tag{2}$$

where i and j index the task and the training sample respectively, w_i is the weight vector, n_i and d denote the training sample size and the feature dimensionality respectively.

We consider the multi-task learning setting where multiple tasks are divided into two groups, i.e., the related tasks group and the irrelevant (outlier) tasks group. We consider a composite structure which couples the related tasks using a low-rank structure and identifies the outlier tasks using a group-sparse structure. Denote the transformation matrix of the m tasks by $W = [w_1, \cdots, w_m] \in \mathbb{R}^{d\times m}$. Specifically, W is given by the direct summation of a low-rank matrix $L = [l_1, \cdots, l_m] \in \mathbb{R}^{d\times m}$ (of a smaller set of basis factors), and a group-sparse (column-sparse) matrix $S = [s_1, \cdots, s_m] \in \mathbb{R}^{d\times m}$ (of zero-vectors in the columns). The weight vector of the i-th task can be expressed as

$$w_i = l_i + s_i,\ l_i \in \mathbb{R}^d,\ s_i \in \mathbb{R}^d,\ i \in \mathbb{N}_m, \tag{3}$$

where l_i and s_i are from the aforementioned low-rank structure and the group-sparse structure, respectively.

We propose a robust multi-task learning formulation (RMTL) to learn multiple tasks simultaneously as well as identify the irrelevant outlier tasks. Mathematically, RMTL is formulated as

$$\min_{L,S}\ \mathcal{L}\left((l_i + s_i)^T x_j^i, y_j^i\right) + \alpha\|L\|_* + \beta\|S\|_{1,2}, \tag{4}$$

where the trace norm regularization term encourages the desirable low-rank structure in the matrix L (for coupling the related tasks), and the $\ell_{1,2}$-norm regularization term induces the desirable group-sparse structure in the matrix S (for identifying the outlier tasks), α and β are non-negative trade-off parameters, and $\mathcal{L}(\cdot,\cdot)$ represents the commonly used least squares loss function. Note that the empirical evaluation of the (averaged) least square loss of the m tasks over the prescribed training data can be expressed as

$$\mathcal{L}\left((l_i + s_i)^T x_j^i, y_j^i\right) = \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{1}{mn_i}\left((l_i + s_i)^T x_j^i - y_j^i\right)^2. \tag{5}$$

Our motivation behind the proposed RMTL formulation in Eq. (4) is as follows: if the i-th task is from the related tasks group, s_i is expected to be a zero-vector and hence w_i obeys the specified low-rank structure constraint; on the other hand, if the i-th task is from the outlier tasks group, s_i is expected to be non-zero and w_i is equal to a direct sum of l_i and the non-zero s_i.

The RMTL formulation in Eq. (4) is an unconstrained convex optimization problem with a non-smooth objective function. Such a problem is difficult to solve directly due to the non-smoothness in the trace norm and the $\ell_{1,2}$-norm regularization terms.

The proposed RMTL formulation in Eq. (4) subsumes several representative algorithms as special cases. As $\beta \to +\infty$, RMTL is degenerated into

$$\min_L\ \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{1}{mn_i}\left(l_i^T x_j^i - y_j^i\right)^2 + \alpha\|L\|_*. \tag{6}$$

The formulation in Eq. (6) is essentially the least squares regression with trace norm regularization, in which multiple learning tasks are coupled via a low-rank structure. On the other hand, as $\alpha \to \infty$, RMTL is degenerated into

$$\min_S\ \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{1}{mn_i}\left(s_i^T x_j^i - y_j^i\right)^2 + \beta\|S\|_{1,2}. \tag{7}$$

The formulation in Eq. (7) is essentially a variant of the ridge regression with the smooth term $\sum_{i=1}^m \|s_i\|^2$ replaced by the non-smooth term $\sum_{i=1}^m \|s_i\|$. In such a formulation, the multiple tasks are decoupled and each task can be learned (optimized) via

$$\min_{s_i}\ \frac{1}{mn_i}\sum_{j=1}^{n_i}\left(s_i^T x_j^i - y_j^i\right)^2 + \beta\|s_i\|_2.$$

Note that similar low-rank and group-sparse structures are studied from a different perspective in [32, 12], which focus on decomposing a given data matrix into a unique sum of a low-rank structure and a column-sparse structure and providing a theoretical guarantee for existence and uniqueness of the decomposition.

3. ACCELERATED PROXIMAL METHOD

In this section, we consider to solve the RMTL formulation in Eq. (4) using the accelerated proximal method (APM) [4, 24, 25]. APM has attracted extensive attentions in the machine learning and data mining communities [20, 19, 15, 8, 18, 21] due to its optimal convergence rate among all first-order techniques and its ability of dealing with large-scale non-smooth optimization problems. Note that in this paper, we focus on discussing the key ingredient of APM, i.e, the proximal operator and its efficient computation; the detailed description of APM can be found in [4, 24, 25].

3.1 Proximal Operator

For the optimization problem in Eq. (4), we symbolically denote its variables by

$$Z = \begin{bmatrix} L \\ S \end{bmatrix},\ L \in \mathbb{R}^{d\times m},\ S \in \mathbb{R}^{d\times m},$$

and denote the smooth and non-smooth components of its objective function respectively by

$$f(Z) = \mathcal{L}\left((l_i + s_i)^T x_j^i, y_j^i\right),\ g(Z) = \alpha\|L\|_* + \beta\|S\|_{1,2}. \tag{8}$$

To solve Eq. (4), APM maintains two sequences of variables: a feasible solution sequence $\{Z_k\}$ and a searching point sequence $\{\widehat{Z}_k\}$. The general scheme of APM can be described as below: at the k-th iteration of APM, the solution point Z_{k+1} can be computed via

$$Z_{k+1} = \arg\min_{Z} \frac{\gamma_k}{2} \left\| Z - \left(\widehat{Z}_k - \frac{1}{\gamma_k} \nabla f(\widehat{Z}_k) \right) \right\|_F^2 + g(Z), \quad (9)$$

where $\widehat{Z}_k$ denotes a searching point constructed from a linear combination of Z_k and Z_{k-1} from previous iterations, and $\nabla f(\widehat{Z}_k)$ denotes the derivative of the smooth component $f(\cdot)$ in Eq. (8) at $\widehat{Z}_k$, γ_k specifies the step size which can be appropriately determined by iteratively increasing its value until the inequality

$$f(Z_{k+1}) \leq f(\widehat{Z}_k) + \langle \nabla f(\widehat{Z}_k), Z_{k+1} - \widehat{Z}_k \rangle + \frac{\gamma_k}{2} \|Z_{k+1} - \widehat{Z}_k\|_F^2, \quad (10)$$

is satisfied. The procedure in Eq. (9) is commonly referred to as the proximal operator [23]. The efficient computation of the proximal operator is critical for the practical convergence of APM, as it is involved in each iteration of the APM algorithm.

3.2 Proximal Operator Computation

For the optimization problem in Eq. (4), its proximal operator can be expressed as an optimization problem of the general form

$$\min_{L_z, S_z} \quad \|L_z - L_{\hat{z}}\|_F^2 + \|S_z - S_{\hat{z}}\|_F^2 + \hat{\alpha}\|L_z\|_* + \hat{\beta}\|S_z\|_{1,2}, \quad (11)$$

where $\hat{\alpha} = \frac{2\alpha}{\gamma_k}$ and $\hat{\beta} = \frac{2\beta}{\gamma_k}$. It can be easily verified that the optimization of L_z and S_z in Eq. (11) are decoupled. Moreover, the optimal solution to Eq. (11) admits an analytic form as presented below.

Computation of L_z The optimal L_z to Eq. (11) can be obtained by solving the following optimization problem:

$$\min_{L_z} \quad \|L_z - L_{\hat{z}}\|_F^2 + \hat{\alpha}\|L_z\|_*. \quad (12)$$

The computation procedure above is equal to the matrix shrinkage operator discussed in [6, 11]. In essence it applies soft-thresholding to the non-zero singular values [28] of $L_{\hat{z}}$ as summarized in the following theorem.

THEOREM 3.1. *Given an arbitrary $L_{\hat{z}}$ in Eq. (12), let rank$(L_{\hat{z}}) = r$ and denote the singular value decomposition (SVD) of $L_{\hat{z}}$ in the reduced form as*

$$L_{\hat{z}} = U_{\hat{z}} \Sigma_{\hat{z}} V_{\hat{z}}^T, \ \ \Sigma_{\hat{z}} = diag\left(\{\sigma_i\}_{i=1}^r\right)$$

where, $U_{\hat{z}} \in \mathbb{R}^{d\times r}$ and $V_{\hat{z}} \in \mathbb{R}^{m\times r}$ consist of orthonormal columns, $\Sigma_{\hat{z}} \in \mathbb{R}^{r\times r}$ is diagonal, and $\{\sigma_i\}_{i=1}^r$ represent the non-zero singular values. Then the optimal L_z^* to Eq. (12) is given by

$$L_z^* = U_{\hat{z}}\, diag\left(\left\{ \sigma_i - \frac{1}{2}\hat{\alpha} \right\}_+ \right) V_{\hat{z}}^T,$$

where $\{e\}_+ = \max(e, 0)$.

The dominating cost in solving Eq. (12) lies in the compact SVD operation on the matrix $L_{\hat{z}} \in \mathbb{R}^{d\times m}$ ($m \ll d$ in general MTL settings).

Computation of S_z The optimal S_z to Eq. (11) can be obtained by solving the following optimization problem:

$$\min_{S_z} \quad \|S_z - S_{\hat{z}}\|_F^2 + \hat{\beta}\|S_z\|_{1,2}. \quad (13)$$

It can be easily verified that in Eq. (13) the column vectors of S_z can be optimized separately. Specifically, each vector of the optimal S_z to Eq. (13) can be obtained via solving a subproblem in the form

$$\min_{s} \quad \|s - \hat{s}\|_2^2 + \hat{\beta}\|s\|_2. \quad (14)$$

It can be verified that the optimization problem above admits an analytic solution [19] as summarized in the following lemma.

LEMMA 3.1. *Let s^* be the optimal solution to the optimization problem in Eq. (14). Then s^* is given by*

$$s^* = \begin{cases} \hat{s}\left(1 - \frac{\hat{\beta}}{2\|\hat{s}\|_2}\right) & \|\hat{s}\|_2 > \frac{\hat{\beta}}{2} \\ 0 & 0 \leq \|\hat{s}\|_2 \leq \frac{\hat{\beta}}{2} \end{cases}.$$

The computation cost of solving Eq. (13) is relatively small compared to the cost of solving Eq. (12).

4. THEORETICAL ANALYSIS

In this section, we derive a performance bound for the proposed RMTL formulation in Eq. (4). This performance bound can be used to theoretically evaluate how well the integration of the low-rank structure and the group-sparse structure can estimate the multiple tasks (the ground truth of the linear predictive functions). Note that in the following analysis, for simplicity we assume that the training sample sizes for all tasks are the same; the derivation below can be easily extended to the setting where the training sample size for each task is different.

Assume that the linear predictive function associated with the i-th task satisfies

$$y_j^i = f_i(x_j^i) + \delta_{ij} = w_i^T x_j^i + \delta_{ij},\ i \in \mathbb{N}_m,\ j \in \mathbb{N}_n, \quad (15)$$

where $\{(x_j^i, y_j^i)\}$ are the training data pairs of the i-th task, and $\delta_{ij} \sim \mathcal{N}(0, \sigma_\delta^2)$ is a stochastic noise variable. For the i-th task, denote its training data matrix X_i and its label vector y_i respectively by

$$X_i = [x_1^i, \cdots, x_n^i] \in \mathbb{R}^{d\times n}, y_i = [y_1^i, \cdots, y_n^i]^T \in \mathbb{R}^n, i \in \mathbb{N}_m. \quad (16)$$

Denote the empirical evaluation of the i-th task f_i over the training data $\{x_j^i\}$ and the associated noise vector δ_i respectively by

$$\hat{f}_i = [f_i(x_1^i), \cdots, f_i(x_n^i)]^T \in \mathbb{R}^n,\ \delta_i = [\delta_{i1}, \cdots, \delta_{in}]^T \in \mathbb{R}^n. \quad (17)$$

It follows that Eq. (15) can be expressed in a compact form as

$$y_i = \hat{f}_i + \delta_i,\ i \in \mathbb{N}_m. \quad (18)$$

Moreover, the optimization problem in Eq. (4) can be rewritten as

$$(\widehat{L}_z, \widehat{S}_z)$$
$$= \arg\min_{L,S} \frac{1}{mn} \sum_{i=1}^m \|X_i^T(l_i + s_i) - y_i\|_2^2 + \alpha\|L\|_* + \beta\|S\|_{1,2}, \quad (19)$$

where $\widehat{L}_z = [\hat{l}_1, \cdots, \hat{l}_m]$ and $\widehat{S}_z = [\hat{s}_1, \cdots, \hat{s}_m]$ are the optimal solution pair obtained via solving Eq. (19).

4.1 Basic Properties of the Optimal Solution

We present some basic properties of the optimal solution pair defined in Eq. (19); these properties are important building blocks for our following theoretical analysis. We first define two operators, namely $\mathcal{Q}$ and its complement $\mathcal{Q}_\perp$, on an arbitrary matrix pair (of the same size), based on Lemma 3.4 in [28].

LEMMA 4.1. *Given any L and $\widehat{L}$ of the same size $d \times m$, let rank$(L) = r \leq \min(d, m)$ and denote the SVD of L as*

$$L = U \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} V^T,$$

where $U \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{m\times m}$ are orthogonal, and $\Sigma \in \mathbb{R}^{r\times r}$ is diagonal consisting of the non-zero singular values on its main diagonal. Let

$$U^T(\widehat{L} - L)V = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix},$$

where $M_{11} \in \mathbb{R}^{r\times r}$, $M_{12} \in \mathbb{R}^{r\times(m-r)}$, $M_{21} \in \mathbb{R}^{(d-r)\times r}$, and $M_{22} \in \mathbb{R}^{(d-r)\times(m-r)}$. Define $\mathcal{Q}$ and $\mathcal{Q}_\perp$ on $\widehat{L} - L$ as

$$\mathcal{Q}(\widehat{L}-L) = U \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & \mathbf{0} \end{bmatrix} V^T, \mathcal{Q}_\perp(\widehat{L}-L) = U \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & M_{22} \end{bmatrix} V^T.$$

Then rank$(\mathcal{Q}(\widehat{L}-L)) \leq 2r$, $L\mathcal{Q}_\perp^T(\widehat{L}-L) = L^T\mathcal{Q}_\perp(\widehat{L}-L) = 0$.

The results in Lemma 4.1 imply a condition under which the trace norm on a matrix pair is additive. From Lemma 4.1 we can verify

$$\|L + \mathcal{Q}_\perp(\widehat{L} - L)\|_* = \|L\|_* + \|\mathcal{Q}_\perp(\widehat{L} - L)\|_*, \tag{20}$$

for arbitrary L and $\widehat{L}$ of the same size. As a direct consequence of Lemma 4.1, we derive a bound on the trace norm of the matrices of interest as summarized below (the detailed proof is provided in the Appendix).

COROLLARY 4.1. *For an arbitrary matrix pair $\widehat{L}$ and L, the following inequality holds*

$$\|\widehat{L} - L\|_* + \|L\|_* - \|\widehat{L}\|_* \leq 2\|\mathcal{Q}(\widehat{L} - L)\|_*.$$

Analogous to the bound on the trace norm derived in Corollary 4.1, we derive a bound on the $\ell_{1,2}$-norm of the matrices of interest. Denote by $\mathcal{C}(S)$ the set of indices corresponding to the non-zero columns of the matrix S as

$$\mathcal{C}(S) = \{i : s_i \neq 0,\ i \in \mathbb{N}_m\}, \tag{21}$$

and by $\mathcal{C}_\perp(S)$ the associated complement (the set of indices corresponding to the zero columns). Denote by $\widehat{S}_{\mathcal{C}(S)}$ the matrix of the same columns as $\widehat{S}$ on the index set $\mathcal{C}(S)$ and of zero columns on the index set $\mathcal{C}_\perp(S)$, i.e., $\widehat{S}_{\mathcal{C}(S)} = [\tilde{s}_1, \cdots, \tilde{s}_m]$, where $\tilde{s}_i = \hat{s}_i$ if $i \in \mathcal{C}(S)$ and $\tilde{s}_i = \mathbf{0}$ if $i \in \mathcal{C}_\perp(S)$. The bound on the $\ell_{1,2}$-norm is summarized below (the detailed proof is provided in the Appendix).

LEMMA 4.2. *Given a matrix pair S and $\widehat{S}$ of the same size, the following inequality holds*

$$\|\widehat{S} - S\|_{1,2} + \|S\|_{1,2} - \|\widehat{S}\|_{1,2} \leq 2\|(\widehat{S} - S)_{\mathcal{C}(S)}\|_{1,2}. \tag{22}$$

We now present some important properties of the optimal solution in Eq. (19) as summarized in the following lemma.

LEMMA 4.3. *Consider the optimization problem in Eq. (19) for $m \geq 2$ and $n, d \geq 1$. Let X_i and y_i be defined in Eq. (16), and $\hat{f}_i$ and δ_i be defined in Eq. (17). Assume that all diagonal elements of the matrix $X_iX_i^T$ are equal to 1 (features are normalized). Take the regularization parameters α and β as*

$$\frac{\alpha}{\sqrt{m}}, \beta \geq \lambda,\ \lambda = \frac{2\sigma_\delta}{nm}\sqrt{d + t}, \tag{23}$$

where $t > 0$ is a universal constant. Then with probability of at least $1 - m\exp\left(-\frac{1}{2}\left(t - d\log\left(1 + \frac{t}{d}\right)\right)\right)$, for a global minimizer $\widehat{L}_z, \widehat{S}_z$ in Eq. (19) and any $L, S \in \mathbb{R}^{d\times m}$, we have

$$\frac{1}{nm}\sum_{i=1}^m \|X_i^T(\hat{l}_i + \hat{s}_i) - \hat{f}_i\|_2^2 \leq \frac{1}{nm}\sum_{i=1}^m \|X_i^T(l_i + s_i) - \hat{f}_i\|_2^2 + \alpha\|\mathcal{Q}(\widehat{L}_z - L)\| + \beta\|(\widehat{S}_z - S)_{\mathcal{C}(S)}\|_{1,2}, \tag{24}$$

where $\hat{l}_i$ and $\hat{s}_i$ (l_i and s_i) are the i-th columns of $\widehat{L}_z$ and $\widehat{S}_z$ (L and S), respectively.

PROOF. From the definition of $(\widehat{L}_z, \widehat{S}_z)$ in Eq. (19), we have

$$\frac{1}{nm}\sum_{i=1}^m \|X_i^T(\hat{l}_i + \hat{s}_i) - y_i\|_2^2 \leq \frac{1}{nm}\sum_{i=1}^m \|X_i^T(l_i + s_i) - y_i\|_2^2 \alpha\|L\|_* + \beta\|S\|_{1,2} - \alpha\|\widehat{L}_z\|_* - \beta\|\widehat{S}_z\|_{1,2}.$$

By substituting Eq. (18) into the inequality above and rearranging all terms, we have

$$\frac{1}{nm}\sum_{i=1}^m \|X_i^T(\hat{l}_i + \hat{s}_i) - \hat{f}_i\|_2^2 \leq \frac{1}{nm}\sum_{i=1}^m \|X_i^T(l_i + s_i) - \hat{f}_i\|_2^2 + \alpha(\|L\|_* - \|\widehat{L}_z\|_*) + \beta(\|S\|_{1,2} - \|\widehat{S}_z\|_{1,2}) + \frac{2}{nm}\sum_{i=1}^m \langle \hat{l}_i - l_i, X_i\delta_i\rangle + \frac{2}{nm}\sum_{i=1}^m \langle \hat{s}_i - s_i, X_i\delta_i\rangle. \tag{25}$$

Next we compute upper bounds for the terms $\frac{2}{nm}\sum_{i=1}^m \langle \hat{l}_i - l_i, X_i\delta_i\rangle$ and $\frac{2}{nm}\sum_{i=1}^m \langle \hat{s}_i - s_i, X_i\delta_i\rangle$ in Eq. (25), respectively. Define a set of random events $\{\mathcal{A}_i\}$ as

$$\mathcal{A}_i = \left\{\frac{2}{nm}\|X_i\delta_i\|_2 \leq \lambda\right\},\ \forall i \in \mathbb{N}_m.$$

For each $\mathcal{A}_i$, define a set of random variables $\{v_{ij}\}$ as

$$v_{ij} = \frac{1}{\sigma_\delta}\sum_{k=1}^n x_{jk}^i\delta_{ik},\ j \in \mathbb{N}_d,$$

where x_{jk}^i denotes the (j,k)-th entry of the data matrix X_i. Since all diagonal elements of the matrix $X_iX_i^T$ are equal to 1, it can be shown that $\{v_{i1}, v_{i2}, \cdots, v_{id}\}$ are i.i.d. Gaussian variables obeying $\mathcal{N}(0,1)$ (Lemma 1 in the Appendix). We can also verify that $\sum_{j=1}^d v_{ij}^2$ is a chi-squared random variable with d degrees of freedom. Moreover taking λ as in Eq. (23), we have

$$\Pr\left(\frac{2}{nm}\|X_i\delta_i\|_2 > \lambda\right) = \Pr\left(\sum_{j=1}^d\left(\sum_{k=1}^n x_{jk}^i\delta_{ik}\right)^2 \geq \frac{\lambda^2n^2m^2}{4}\right) = \Pr\left(\sum_{j=1}^d v_{ij}^2 \geq d + t\right) \leq \exp\left(-\frac{1}{2}\mu_d^2(t)\right),$$

where $\mu_d(t) = \sqrt{t - d\log\left(1 + \frac{t}{d}\right)}$ $(t > 0)$, and the last inequality above follows from a concentration inequality (Lemma 2 in the Appendix). Let $\mathcal{A} = \bigcap_{i=1}^m \mathcal{A}_i$. Denote by $\mathcal{A}_i^c$ the complement of each event $\mathcal{A}_i$. It follows that

$$\Pr(\mathcal{A}) \geq 1 - \Pr\left(\bigcup_{i=1}^m \mathcal{A}_i^c\right) \geq 1 - m\exp\left(-\frac{1}{2}\mu_d^2(t)\right).$$

Under the event $\mathcal{A}$, we derive a bound on the term $\frac{2}{nm}\sum_{i=1}^m\langle\hat{l}_i - l_i, X_i\delta_i\rangle$ as

$$\frac{2}{nm}\sum_{i=1}^m\langle\hat{l}_i - l_i, X_i\delta_i\rangle \le \frac{2}{nm}\sum_{i=1}^m\|\hat{l}_i - l_i\|_2\|X_i\delta_i\|_2 \le \lambda\sum_{i=1}^m\|\hat{l}_i - l_i\|_2 \le \alpha\|\widehat{L}_z - L\|_*, \quad (26)$$

where the first inequality above follows from Cauchy-Schwarz inequality and the second inequality follows from

$$\sum_{i=1}^m\|\hat{l}_i - l_i\|_2 \le \sqrt{m\sum_{i=1}^m\|\hat{l}_i - l_i\|_2^2} = \sqrt{m}\|\widehat{L}_z - L\|_F \le \sqrt{m}\|\widehat{L}_z - L\|_*.$$

Similarly under $\mathcal{A}$, we also derive a bound on the term $\frac{2}{nm}\sum_{i=1}^m\langle\hat{s}_i - s_i, X_i f_i\rangle$ as

$$\frac{2}{nm}\sum_{i=1}^m\langle\hat{s}_i - s_i, X_i\delta_i\rangle \le \frac{2}{nm}\sum_{i=1}^m\|\hat{s}_i - s_i\|_2\|X_i\delta_i\|_2 \le \beta\|\widehat{S}_z - S\|_{1,2}. \quad (27)$$

Moreover we bound the right side of Eq.(25) using the results from Eqs. (26) and (27). It follows that

$$\frac{1}{nm}\sum_{i=1}^m\|X_i^T(\hat{l}_i+\hat{s}_i)-\hat{f}_i\|_2^2 \le \frac{1}{nm}\sum_{i=1}^m\|X_i^T(l_i+s_i)-\hat{f}_i\|_2^2+ \alpha(\|\widehat{L}_z-L\|_*+\|L\|_*-\|\widehat{L}_z\|_*)+\beta(\|\widehat{S}_z-S\|_{1,2}+\|S\|_{1,2}-\|\widehat{S}_z\|_{1,2}).$$

Finally by applying Corollary 4.1 and Lemma 4.2 together with the inequality above, we complete the proof. □

4.2 Performance Bound

We present a performance bound of the proposed RMTL formulation in Eq. (19). This bound measures how well the multi-task learning scheme (via the integration of the low-rank structure and the $\ell_{1,2}$-norm structure) can estimate the linear predictive functions in Eq. (15).

We begin with some notations. Let $X \in \mathbb{R}^{md\times mn}$ be a block-diagonal matrix with its i-th block formed by the matrix $X_i \in \mathbb{R}^{d\times n}$ $(i \in \mathbb{N}_m)$. Define a diagonalization operator $\mathcal{D}$ on an arbitrary $\Omega = [\omega_1, \omega_2, \cdots, \omega_m] \in \mathbb{R}^{d\times m}$: $\mathcal{D}(\Omega) \in \mathbb{R}^{md\times m}$ is a block diagonal matrix with its i-th block formed by the column vector $\omega_i \in \mathbb{R}^d$. Let $\mathcal{F} = [\hat{f}_1, \cdots, \hat{f}_m]$, where $\hat{f}_i$ is defined in Eq. (17). Therefore we can rewrite Eq. (24) in a compact form as

$$\frac{1}{T}\|X^T\mathcal{D}(\widehat{L}_z+\widehat{S}_z)-\mathcal{D}(\mathcal{F})\|_F^2 \le \frac{1}{T}\|X^T\mathcal{D}(L+S)-\mathcal{D}(\mathcal{F})\|_F^2 + \alpha\|\mathcal{Q}(\widehat{L}_z - L)\|_* + \beta\|(\widehat{S}_z - S)_{\mathcal{C}(S)}\|_{1,2}, \quad (28)$$

where $T = nm$. We next introduce our assumption over a restricted set. The assumption is associated with training data X and the geometric structure of the matrices of interest.

ASSUMPTION 4.1. *For a matrix pair Γ_L and Γ_S of size d by m, let $s \le \min(d, m)$ and $q \le m$. We assume that there exist constants $\kappa_1(s)$ and $\kappa_2(q)$ such that*

$$\kappa_1(s) \triangleq \min_{\Gamma_L,\Gamma_S\in\mathcal{R}(s,q)}\frac{\|X\mathcal{D}(\Gamma_L+\Gamma_S)\|_F}{\sqrt{T}\|\mathcal{Q}(\Gamma_L)\|_*} > 0, \quad (29)$$

$$\kappa_2(q) \triangleq \min_{\Gamma_L,\Gamma_S\in\mathcal{R}(s,q)}\frac{\|X\mathcal{D}(\Gamma_L+\Gamma_S)\|_F}{\sqrt{T}\|(\Gamma_S)_{\mathcal{C}(S)}\|_{1,2}} > 0, \quad (30)$$

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

$$\mathcal{R}(s,q) = \Big\{\Gamma_L, \Gamma_S \in \mathbb{R}^{d\times m} \,|\, \Gamma_L \neq 0,\ \Gamma_S \neq 0,\ \textit{rank}(\mathcal{Q}(\Gamma_L)) \le s,\ |\mathcal{C}(\Gamma_S)| \le q\Big\},$$

and $\mathcal{C}(\cdot)$ is defined in Eq. (21), and $|\widehat{C}|$ denotes the number of elements in the set $\widehat{C}$.

The assumption in Eqs. (29) and (30) can be implied by several sufficient conditions as in [5]. Due to the space constraint, the details are omitted. Note that similar assumptions are used in [22] for deriving a certain performance bound for a different multi-task learning formulation.

We present the performance bound of the RMTL formulation in the following theorem.

THEOREM 4.1. *Consider the optimization problem in Eq. (19) for $m \ge 2$ and $n, d \ge 1$. Take the regularization parameters α and β as in Eq. (23). Then with probability of at least $1 - m\exp\left(-\frac{1}{2}\left(t - d\log\left(1+\frac{t}{d}\right)\right)\right)$, for a global minimizer $\widehat{L}_z, \widehat{S}_z$ in Eq. (19), we have*

$$\frac{1}{T}\|X\mathcal{D}(\widehat{L}_z+\widehat{S}_z)-\mathcal{D}(\mathcal{F})\|_F^2 \le (1+\epsilon)\inf_{L,S}\frac{1}{T}\|X\mathcal{D}(L+S)-\mathcal{D}(\mathcal{F})\|_F^2 + \mathcal{E}(\epsilon)\left(\frac{\alpha^2}{\kappa_1^2(2r)} + \frac{\beta^2}{\kappa_2^2(c)}\right), \quad (31)$$

where inf *is taken over all $L, S \in \mathbb{R}^{d\times m}$ with rank$(L) \le r$ and $|\mathcal{C}(S)| \le c$, and $\mathcal{E}(\epsilon) > 0$ is a constant depending only on ϵ.*

PROOF. Denote $\Gamma_L = \widehat{L}_z - L$ and $\Gamma_S = \widehat{S}_z - S$. It follows from Eq. (28) that

$$\frac{1}{T}\|X^T\mathcal{D}(\widehat{L}_z+\widehat{S}_z)-\mathcal{D}(\mathcal{F})\|_F^2 \le \frac{1}{T}\|X^T\mathcal{D}(L+S)-\mathcal{D}(\mathcal{F})\|_F^2 + \alpha\|\mathcal{Q}(\Gamma_L)\|_* + \beta\|(\Gamma_S)_{\mathcal{C}(S)}\|_{1,2}. \quad (32)$$

Given $\mathcal{Q}(\Gamma_L) \le 2r$ (from Lemma 4.1) and $|\mathcal{C}(S)| \le c$, we derive upper bounds on $\alpha\|\mathcal{Q}(\Gamma_L)\|_*$ and $\beta\|(\Gamma_S)_{\mathcal{C}(S)})\|_{1,2}$ over the restrict set $\mathcal{R}(2r, c)$ based on Assumptions 4.1, respectively. It follows from Eq. (29) in Assumption 4.1 that

$$2\alpha\|\mathcal{Q}(\Gamma_L)\|_* \le \frac{2\alpha}{\kappa_1(2r)\sqrt{T}}\|X\mathcal{D}(\Gamma_L+\Gamma_S)\|_F \le \frac{2\alpha}{\kappa_1(2r)\sqrt{T}}\Big(\|X\mathcal{D}(\widehat{L}_z+\widehat{S}_z)-\mathcal{D}(\mathcal{F})\|_F + \|X\mathcal{D}(L+S)-\mathcal{D}(\mathcal{F})\|_F\Big) \le \frac{\alpha^2\tau}{\kappa_1^2(2r)} + \frac{1}{\tau T}\|X\mathcal{D}(\widehat{L}_z+\widehat{S}_z)-\mathcal{D}(\mathcal{F})\|_F^2 + \frac{\alpha^2\tau}{\kappa_1^2(2r)} + \frac{1}{\tau T}\|X\mathcal{D}(L+S)-\mathcal{D}(\mathcal{F})\|_F^2, \quad (33)$$

where the last inequality above follows from $2ab \le a^2\tau + b^2\frac{1}{\tau}$ for $\tau > 0$. Similarly, we have

$$2\beta\|(\Gamma_S)_{\mathcal{C}(S)}\|_{1,2} \le \frac{\beta^2\tau}{\kappa_2^2(c)} + \frac{1}{\tau T}\|X\mathcal{D}(\widehat{L}_z+\widehat{S}_z)-\mathcal{D}(\mathcal{F})\|_F^2 + \frac{\beta^2\tau}{\kappa_2^2(c)} + \frac{1}{\tau T}\|X\mathcal{D}(L+S)-\mathcal{D}(\mathcal{F})\|_F^2. \quad (34)$$

Substituting Eqs. (33) and (34) into Eq. (32) and setting $\tau = 2 + \frac{4}{\epsilon}$,

we obtain

$$\frac{1}{T}\|X\mathcal{D}(\widehat{L}_z + \widehat{S}_z) - \mathcal{D}(\mathcal{F})\|_F^2$$
$$\leq \frac{\tau+2}{\tau-2}\|X\mathcal{D}(L+S) - \mathcal{D}(\mathcal{F})\|_F^2 + \frac{2\tau^2}{\tau-2}\left(\frac{\alpha^2}{\kappa_1^2(2r)} + \frac{\beta^2}{\kappa_2^2(c)}\right)$$
$$= (1+\epsilon)\|X\mathcal{D}(L+S) - \mathcal{D}(\mathcal{F})\|_F^2 + \mathcal{E}(\epsilon)\left(\frac{\alpha^2}{\kappa_1^2(2r)} + \frac{\beta^2}{\kappa_2^2(c)}\right),$$

where $\mathcal{E}(\epsilon) = \epsilon(\frac{1}{2} + \frac{1}{\epsilon})^2$. This completes the proof. □

The performance bound described in Eq. (31) can be refined by choosing specific values for the regularization parameters α and β: it can be verified that the component $\frac{\alpha^2}{\kappa_1^2(2r)} + \frac{\beta^2}{\kappa_2^2(c)}$ is minimized if α and β are chosen to be proportional to $\kappa_1^2(2r)$ and $\kappa_2^2(c)$, respectively.

5. EXPERIMENTS

In this section, we evaluate the proposed RMTL formulation in Eq. (4) in comparison with other representative algorithms for multi-task learning; we also conduct numerical studies on the APM algorithm in comparison with the commonly used proximal method (PM) [25, 24] for solving RMTL. All algorithms are implemented in Matlab. Note that for numerical accuracy consideration, we solve the RMLT formulation with its objective function multiplied by nm, where m and n correspond to the task number and the sum of the sample sizes for all tasks, respectively.

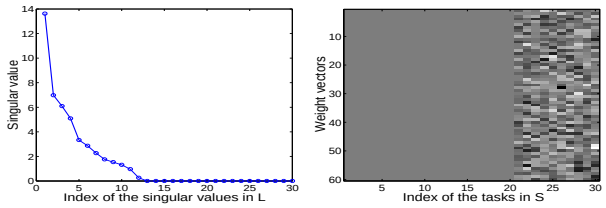

Figure 1: Demonstration of the extracted low-rank and group structures: the left plot shows the singular values of the low-rank component L (the last 18 singular values are zero); the right plot demonstrates the structure of the group-sparse component S (the first 20 columns are zero-vectors). In the right plot the grey area corresponds to the pixels of zero-value.

5.1 Demonstration of Extracted Structures

We apply the RMTL algorithm on a synthetic data set and then demonstrate the extracted low-rank and group-sparse structures. The synthetic data is constructed as follows: set the task number $m = 30$, the size of the training samples for each task $n_i = 50$, and the feature dimensionality of the training samples $d = 60$; generate the entries of the training data $X_i \in \mathbb{R}^{d\times n_i}$ (for the i-th task) randomly from the distribution $\mathcal{N}(0, 25)$; generate the entries in the low-rank component L (of size $d \times m$) randomly from $\mathcal{N}(0, 16)$ and then set its smallest 20 singular values at 0; generate the entries in the group-sparse component S (of size $d \times m$) randomly from $\mathcal{N}(0, 20)$ and then set its first 20 columns as zero-vectors; construct the response (target) vector of each task as $y_i = X_i^T(L+S) + \delta_i \in \mathbb{R}^{n_i}$ $(i \in \mathbb{N}_m)$, where each entry in the vector δ_i is randomly generated from $\mathcal{N}(0, 1)$. Under this experimental setting, we construct 20 related tasks as well as 10 outlier tasks, where each task is associated with 50 training samples of feature dimensionality 60.

In Figure 1, we present the low-rank component L and the group-sparse component S obtained by solving RMTL with $\alpha = 50$ and $\beta = 10$. From the left plot of Figure 1, we can observe that the matrix L (of size 60×50) has 12 non-zero singular values; this result is consistent with our problem setting of using a low-rank structure to capture the tasks relationship. From the right plot of Figure 1, we can observe that the first 20 columns (corresponding to the related tasks) in S are zero vectors, while the last 10 columns (corresponding to the outlier tasks) are non-zero vectors. The results in Figure 1 empirically demonstrate the effectiveness of RMTL.

5.2 Performance Evaluation of RMTL

We evaluate the RMTL algorithm on multi-task regression problems in comparison with other representative algorithms including ridge regression (Ridge), least squares with ℓ_1-norm regularization (Lasso), least squares with trace norm regularization (TraceNorm), least squares with low-rank and sparse structures regularization (Sparse-LowRank) [8], and convex multi-task feature learning (CMTL) [2]. The normalized mean squared error (nMSE) and the averaged mean squared error (aMSE) are employed as the regression performance measures as used in previous studies [2, 38]. Note that nMSE is defined as the mean squared error (MSE) divided by the variance of the target vector; aMSE is defined as MSE divided by the squared norm of the target vector. We adopt APM to solve RMTL and terminate APM when the relative change of the objective values in two successive iterations is smaller than 10^{-5}. We use the School data[1] and the SARCOS data[2] for the experiments.

The School data consists of the exam scores of 15362 students from 139 secondary schools; each student is described by 27 attributes such as gender and ethnic group. The exam score prediction of the students can be cast into a multi-task regression (learning) problem: we are given 139 tasks (schools), where each task has a different number of samples (students) and each sample has 27 features (attributes). We randomly select 10%, 20%, and 30% of the samples (from each task) to form the training set and use the rest of the samples as the test set. The experimental results averaged over 15 random repetitions are presented in Table 1. From the presented results, we have the following observations: (1) RMTL outperforms all other competing algorithms in terms of nMSE and aMSE; (2) the multi-task learning algorithms (TraceNorm, Sparse-LowRank, CMTL, and RMTL) outperform the single-task learning algorithms (Ridge and Lasso) in terms of both nMSE and aMSE; (3) the performance of CMTL is similar to that of TraceNorm; this result may be due to the use of similar penalty terms in CMTL and TraceNorm.

The SARCOS data is collected for an inverse dynamics prediction problem for a seven degrees-of-freedom anthropomorphic robot arm. This data consists of 48933 observations corresponding to 7 joint torques; each of the observations is described by 21 features including 7 joint positions, 7 joint velocities, and 7 joint accelerations. Our goal is to construct mappings from each observation to 7 joint torques. We randomly select $50, 100, 150$ observations to form 3 training sets and accordingly randomly select 5000 observations to form 3 test sets. The experimental results averaged over 15 random repetitions are presented in Table 2. From the experimental results, we have the following observations: (1) RMTL performs better than or compares competitively to all other competing algorithms in terms of both nMSE and aMSE; (2) the multi-task learning algorithms (TraceNorm, Sparse-LowRank, CMTL,

[1] http://www.cs.ucl.ac.uk/staff/A.Argyriou/code/
[2] http://www.gaussianprocess.org/gpml/data/

Table 1: Performance comparison of the six competing algorithms in terms of the normalized MSE (nMSE) and the averaged MSE (aMSE) with standard deviation using the School data. All parameters of the six methods are determined via cross-validation and the reported regression performance is averaged over 15 random repetitions. Note that a smaller value of nMSE and aMSE represents better regression performance.

Measure	training ratio	Ridge	Lasso	TraceNorm	Sparse-LowRank	CMTL	Robust MTL
nMSE	10%	1.0398 ± 0.0038	1.0261 ± 0.0132	0.9359 ± 0.0370	0.9175 ± 0.0261	0.9413 ± 0.0021	$\mathbf{0.9130 \pm 0.0039}$
	20%	0.8773 ± 0.0043	0.8754 ± 0.0194	0.8211 ± 0.0032	0.8126 ± 0.0132	0.8327 ± 0.0039	$\mathbf{0.8055 \pm 0.0103}$
	30%	0.8171 ± 0.0090	0.8144 ± 0.0091	0.7870 ± 0.0012	0.7657 ± 0.0091	0.7922 ± 0.0052	$\mathbf{0.7600 \pm 0.0032}$
aMSE	10%	0.2713 ± 0.0023	0.2682 ± 0.0036	0.2504 ± 0.0102	0.2419 ± 0.0081	0.2552 ± 0.0032	$\mathbf{0.2330 \pm 0.0018}$
	20%	0.2303 ± 0.0003	0.2289 ± 0.0051	0.2156 ± 0.0015	0.2114 ± 0.0041	0.2131 ± 0.0071	$\mathbf{0.2018 \pm 0.0025}$
	30%	0.2165 ± 0.0021	0.2137 ± 0.0012	0.2089 ± 0.0012	0.2011 ± 0.0022	0.1922 ± 0.0102	$\mathbf{0.1822 \pm 0.0014}$

Table 2: Performance comparison of the six competing algorithms in terms of nMSE and aMSE with standard deviation using the SARCOS data.The experimental setting is similar to the one described in Table 1.

Measure	training size	Ridge	Lasso	TraceNorm	Sparse-LowRank	CMTL	Robust MTL
nMSE	50	0.2454 ± 0.0260	0.2337 ± 0.0180	0.2257 ± 0.0065	0.2127 ± 0.0033	0.2192 ± 0.0016	$\mathbf{0.2123 \pm 0.0038}$
	100	0.1821 ± 0.0142	0.1616 ± 0.0027	0.1531 ± 0.0017	0.1495 ± 0.0023	0.1568 ± 0.0037	$\mathbf{0.1456 \pm 0.0138}$
	150	0.1501 ± 0.0054	0.1469 ± 0.0028	0.1318 ± 0.0053	$\mathbf{0.1236 \pm 0.0004}$	0.1301 ± 0.0034	0.1245 ± 0.0015
aMSE	50	0.1330 ± 0.0143	0.1228 ± 0.0083	0.1122 ± 0.0064	0.1073 ± 0.0026	0.1156 ± 0.0011	$\mathbf{0.0982 \pm 0.0026}$
	100	0.1053 ± 0.0096	0.0907 ± 0.0023	0.0805 ± 0.0026	0.0793 ± 0.0047	0.0852 ± 0.0013	$\mathbf{0.0737 \pm 0.0083}$
	150	0.0846 ± 0.0045	0.0822 ± 0.0014	0.0772 ± 0.0023	$\mathbf{0.0661 \pm 0.0062}$	0.0755 ± 0.0025	0.0674 ± 0.0014

and RMTL) outperform the single-task learning algorithms (Ridge and Lasso) in terms of both nMSE and aMSE. We also observe that Sparse-LowRank has a similar performance to RMTL. In Sparse-LowRank, incoherent low-rank and (ℓ_1-norm based) sparse structures [8] are used to capture the task relatedness as well as identify discriminative features for each task. These results imply that allowing each task to independently select discriminative features may improve the robustness of the algorithm.

5.3 Sensitivity Studies on RMTL

We conduct a sensitivity study on the proposed RMTL formulation. In particular, we study how the regularization parameters and the training sample size affect the regression performance of RMTL in terms of nMSE and aMSE, respectively.

Effect of the Regularization Parameters For this experiment, we randomly select 10% of the School data as the training set and use the rest of the data as the test set. By fixing $\beta = 100$ as well as varying the value of α in α-value set, i.e., $[50 : 50 : 500]$, we study how the parameter α affects the regression performance of RMTL. Similarly, by fixing $\alpha = 150$ as well as varying the value of β in β-value set of $[20 : 5 : 115]$, we study how the parameter β affects the regression performance of RMTL. In Figure 2, we present the regression performance (averaged over 15 random repetitions) of RMTL in terms of nMSE (1st and 3rd plots) and aMSE (2nd and 4th plots) for each pair of (α, β). From Figure 2, we can observe that both nMSE and aMSE change with different settings of (α, β); we can also observe that the best performance of RMTL for a fixed α (or a fixed β) is obtained by setting β in the middle of β-value set (or setting the value of α in the middle of α-value set).

Effect of the Training Ratio For this experiment, we randomly select $\{10\%, 20\%, \cdots, 80\%\}$ of the School data as the training set and use the rest of the data as the test set. We study how the the training sample size (in terms of the training ratio) affects the regression performance of RMTL. Note that the regularization parameters α and β are determined via double cross-validation. The experimental results are presented in Figure 3. We can observe that by increasing the training ratio, both the nMSE and aMSE decrease; this result is consist with our expectation that more training data will lead to more accurate predictive model and hence better generalization performance.

5.4 Numerical Studies on APM

We conduct numerical studies on APM in comparison with PM for solving RMTL in terms of the computation time (in seconds) and the iteration number. We randomly select 10% of the School data for the following experiments. The experimental setting is described as follows: we stop PM when the change of the objective value in two successive iterations is smaller than 10^{-i} and record the attained objective value; such a value is then used as the stopping criterion in APM, that is, we stop APM when the attained objective value in APM is equal to or smaller than the one previously obtained from PM; we vary the stopping criterion of PM in the set $\{10^{-i}\}_{i=1}^{6}$ and record the required computation time and iteration number for both PM and APM. From the experimental results presented in Figure 4, we have the following observations: (1) APM requires less computation time and iteration number than PM for attaining the same objective value; (2) both APM and PM require more computation time and a larger iteration number if the stopping criterion is set as a smaller value (higher accuracy).

6. CONCLUSION

In this paper, we propose a robust multi-task learning (RMTL) algorithm which learns multiple tasks simultaneously as well as identifies the outlier tasks. The proposed RMTL algorithm captures the task relationships using a low-rank structure, and simultaneously identifies the outlier tasks using a group-sparse structure. RMTL is formulated as a non-smooth convex (unconstrained) optimization problem in which the least square loss is regularized by a combination of the trace norm regularization and the $\ell_{1,2}$-norm regularization. We propose to adopt the accelerated proximal method (APM) for solving this optimization problem and develop efficient algorithms for computing the associated proximal operator. We also conduct a theoretical analysis on the proposed RMTL formulation. In particular, we derive a key property of the optimal solution to RMTL; based on the key property, we establish

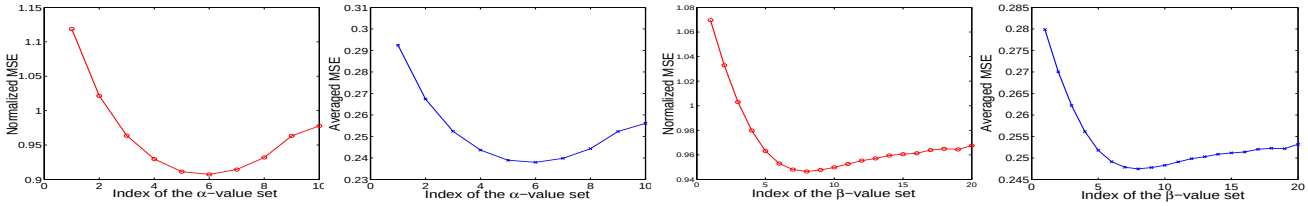

Figure 2: Sensitivity study on RMTL: study the effect of the parameters α and β in terms of nMSE (1st and 3rd plots) and aMSE (2nd and 4th plots), respectively. For the first two plots, we set $\beta = 100$ and vary α in the α-value set $[50 : 50 : 500]$; for the last two plots, we set $\beta = 150$ and vary β in the β-value set $[20 : 5 : 115]$.

a theoretical performance bound to characterize the learning performance of RMTL. Our experimental results on benchmark data sets demonstrate the effectiveness and efficiency of the proposed algorithms. In the future, we plan to apply the proposed RMTL to other real world applications such as Alzheimer's disease study and *Drosophila* gene expression images analysis.

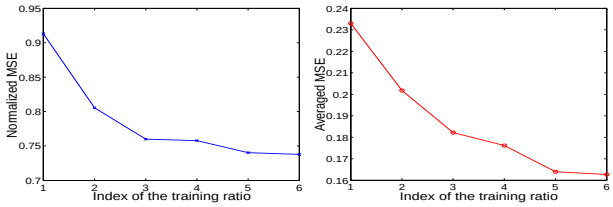

Figure 3: Sensitivity study on RMTL: study the effect of the training ratio in terms of nMSE (left plot) and aMSE (right plot), respectively. The regularization parameters are determined via double cross-validation. The i-th coordinate on the x-axis corresponds to the training ratio $i \times 10\%$.

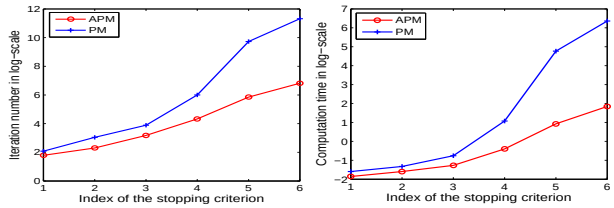

Figure 4: Computation cost comparison of APM and PM in terms of the iteration number (left plot) and the computation time in seconds (right plot) for solving RMTL. The i-th coordinate on the x-axis corresponds to the stopping criterion 10^{-i}.

Acknowledgment

This research is sponsored in part by NSF IIS-0953662 and NSF CCF-1025177.

APPENDIX

Proof of Corollary 4.1

PROOF. From Lemma 4.1, we have

$$\widehat{L} - L = \mathcal{Q}(\widehat{L} - L) + \mathcal{Q}_\perp(\widehat{L} - L)$$

for any matrix pair L and $\widehat{L}$. It follows that

$$\begin{aligned} \|\widehat{L}\|_* &= \|L + \mathcal{Q}(\widehat{L} - L) + \mathcal{Q}_\perp(\widehat{L} - L)\|_* \\ &\geq \|L + \mathcal{Q}_\perp(\widehat{L} - L)\|_* - \|\mathcal{Q}(\widehat{L} - L)\|_* \\ &= \|L\|_* + \|\mathcal{Q}_\perp(\widehat{L} - L)\|_* - \|\mathcal{Q}(\widehat{L} - L)\|_*, \end{aligned}$$

where the inequality above follows from the triangle inequality and the last equality above follows from Eq. (20). Moreover,

$$\begin{aligned} &\|\widehat{L} - L\|_* + \|L\|_* - \|\widehat{L}\|_* \\ &\leq \|\widehat{L} - L\|_* + \|L\|_* - \Big(\|L\|_* + \|\mathcal{Q}_\perp(\widehat{L} - L)\|_* - \|\mathcal{Q}(\widehat{L} - L)\|_*\Big) \\ &\leq 2\|\mathcal{Q}(\widehat{L} - L)\|_*. \end{aligned}$$

We complete the proof of this corollary. □

Proof of Lemma 4.2

PROOF. From the definition of $\mathcal{C}(S)$ in Eq. (21), we have

$$S_{\mathcal{C}_\perp(S)} = \mathbf{0}, \;\; \|(\widehat{S} - S)_{\mathcal{C}_\perp(S)}\|_{1,2} = \|\widehat{S}_{\mathcal{C}_\perp(S)}\|_{1,2}.$$

It follows that

$$\begin{aligned} &\|(\widehat{S} - S)_{\mathcal{C}_\perp(S)}\|_{1,2} + \|S\|_{1,2} - \|\widehat{S}\|_{1,2} \\ = \;& \|\widehat{S}_{\mathcal{C}_\perp(S)}\|_{1,2} + \|S\|_{1,2} - \|\widehat{S}\|_{1,2} \\ = \;& \|S_{\mathcal{C}(S)}\|_{1,2} - \|\widehat{S}_{\mathcal{C}(S)}\|_{1,2} \\ \leq \;& \|(S - \widehat{S})_{\mathcal{C}(S)}\|_{1,2} = \|(\widehat{S} - S)_{\mathcal{C}(S)}\|_{1,2}. \end{aligned}$$

By substituting the equation above into the left side of Eq. (22), we complete the proof of this lemma. □

LEMMA 1. *Let $\delta_1, \delta_2, \cdots, \delta_n$ be a random sample of size n from the Gaussian distribution $\mathcal{N}(0, \sigma)$. Let $x_1, x_2, \cdots, x_n$ satisfy $x_1^2 + x_2^2 + \cdots + x_n^2 = 1$. Denote a random variable v as*

$$v = \frac{1}{\sigma} \sum_{i=1}^{n} x_i \delta_i.$$

Then v obeys the Gaussian distribution $\mathcal{N}(0, 1)$.

PROOF. Since $\{\delta_i\}$ are mutually independent, the mean of the random variable v can be computed as

$$\mathbb{E}(v) = \mathbb{E}\left(\frac{1}{\sigma} \sum_{i=1}^{n} x_i \delta_i\right) = \frac{1}{\sigma} \sum_{i=1}^{n} x_i \mathbb{E}\left(\delta_i\right) = 0.$$

Similarly, the variance of v can be computed

$$\mathbb{E}\left(v - \mathbb{E}(v)\right)^2 = \mathbb{E}\left(\frac{1}{\sigma^2} \sum_{i=1}^{n} x_i^2 \delta_i^2\right) = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i^2 \mathbb{E}\left(\delta_i^2\right) = 1,$$

where the first equality follows from $\mathbb{E}(\delta_i\delta_j) = 0$ $(i \neq j)$. Using the fact that the sum of Gaussian random variables is Gaussian distributed, we complete the proof of this lemma. □

LEMMA 2. *Let $\mathcal{X}_p^2$ be a chi-squared random variable with p degrees of freedom. Then*

$$\Pr\left(\mathcal{X}_p^2 \geq p + \pi\right) \leq \exp\left(-\frac{1}{2}\left(\pi - p\log\left(1 + \frac{\pi}{p}\right)\right)\right), \pi > 0.$$

PROOF. From Theorem 4.1 in [31], we approximate the chi-square distribution using a normal distribution as

$$\Pr\left(\mathcal{X}_p^2 \geq q\right) \leq \Pr\left(\mathcal{N}_{0,1} \geq z_p(q)\right),\ q > p,$$

where $\mathcal{N}_{0,1} \sim \mathcal{N}(0,1)$ and $z_p(q) = \sqrt{q - p - p\log\left(\frac{q}{p}\right)}$. It is known that for $x \sim \mathcal{N}(0,1)$, the inequality $\Pr(x \geq t) \leq \exp(-\frac{t^2}{2})$ holds. Therefore we have

$$\Pr\left(\mathcal{X}_p^2 \geq q\right) \leq \exp\left(-\frac{1}{2}z_p^2(q)\right).$$

By substituting $q = p + \pi$ $(\pi > 0)$ into the inequality above, we complete the proof of this lemma. □

A. REFERENCES

[1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[3] B. Bakker and T. Heskes. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4:83–99, 2003.

[4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Science*, 2:183–202, 2009.

[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

[6] J.-F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[7] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[8] J. Chen, J. Liu, and J. Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. In *KDD*, 2010.

[9] J. Chen, L. Tang, J. Liu, and J. Ye. A convex formulation for learning shared structures from multiple tasks. In *ICML*, 2009.

[10] T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

[11] D. Goldfarb and S. Ma. Convergence of fixed point continuation algorithms for matrix rank minimization. *Submitted to Foundations of Computational Mathematics*.

[12] D. Hsu, S. Kakade, and T. Zhang. Robust matrix decomposition with outliers. arXiv:1011.1518, 2010.

[13] L. Jacob, F. Bach, and J.-P. Vert. Clustered multi-task learning: A convex formulation. In *NIPS*, 2008.

[14] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *NIPS*, 2010.

[15] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *ICML*, pages 457–464, 2009.

[16] S. Kim and E. P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML*, 2010.

[17] N. D. Lawrence and J. C. Platt. Learning to learn with the informative vector machine. In *ICML*, 2004.

[18] J. Liu, J. Chen, and J. Ye. Large-scale sparse logistic regression. In *KDD*, 2009.

[19] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l2,1-norm minimization. In *UAI*, pages 339–348, 2009.

[20] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.

[21] J. Liu, L. Yuan, and J. Ye. An efficient algorithm for a class of fused lasso problems. In *KDD*, 2010.

[22] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. In *COLT*, 2008.

[23] J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.

[24] A. Nemirovski. *Efficient Methods in Convex Programming*. Lecture Notes, 1995.

[25] Y. Nesterov. *Introductory Lectures on Convex Programming*. Lecture Notes, 1998.

[26] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint l21-norms minimization. In *NIPS*, 2010.

[27] G. Obozinski, B. Taskar, and M. I. Jordan. Multi-task feature selection. In *Technical report, Dept. of Statistics, UC Berkeley*, 2006.

[28] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Review*, (3):471–501, 2010.

[29] A. Schwaighofer, V. Tresp, and K. Yu. Learning gaussian process kernels via hierarchical bayes. In *NIPS*, 2004.

[30] S. Thrun and J. O'Sullivan. Discovering structure in multiple learning tasks: The TC algorithm. In *ICML*, 1996.

[31] D. L. Wallace. Bounds on normal approximations to student's and the chi-square distributions. *Annals of Mathematical Statistics*, 30(4):1121–1130, 1959.

[32] H. Xu, C. Caramanis, and S. Sanghavi. Robust pca via outlier pursuit. In *NIPS*, 2010.

[33] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, 8:35–63, 2007.

[34] X. Yang, S. Kim, and E. P. Xing. Heterogeneous multitask learning with joint sparsity constraints. In *NIPS*, 2009.

[35] K. Yu, V. Tresp, and A. Schwaighofer. Learning gaussian processes from multiple tasks. In *ICML*, 2005.

[36] S. Yu, V. Tresp, and K. Yu. Robust multi-task learning with *t*-processes. In *ICML*, 2007.

[37] J. Zhang, Z. Ghahramani, and Y. Yang. Learning multiple related tasks using latent independent component analysis. In *NIPS*, 2005.

[38] Y. Zhang and D.-Y. Yeung. Multi-task learning using generalized t process. In *AISTATS*, 2010.