Safe Screening for Multi-Task Feature Learning with Multiple Data Matrices

Jie Wang[1] JWANGUMI@UMICH.EDU
Jieping Ye[1,2] JPYE@UMICH.EDU
[1]Department of Computational Medicine and Bioinformatics, University of Michigan, MI 48109 USA
[2]Department of Electrical Engineering and Computer Science, University of Michigan, MI 48109 USA

Abstract

Multi-task feature learning (MTFL) is a powerful technique in boosting the predictive performance by learning multiple related classification/regression/clustering tasks simultaneously. However, solving the MTFL problem remains challenging when the feature dimension is extremely large. In this paper, we propose a novel screening rule—that is based on the **d**ual **p**rojection onto **c**onvex sets (DPC)—to quickly identify the *inactive features*—that have zero coefficients in the solution vectors across all tasks. One of the appealing features of DPC is that: it is *safe* in the sense that the detected inactive features are guaranteed to have zero coefficients in the solution vectors across all tasks. Thus, by removing the inactive features from the training phase, we may have substantial savings in the computational cost and memory usage *without sacrificing accuracy*. To the best of our knowledge, it is the *first* screening rule that is applicable to sparse models with *multiple* data matrices. A key challenge in deriving DPC is to solve a nonconvex problem. We show that we can solve for the global optimum efficiently via a properly chosen parametrization of the constraint set. Moreover, DPC has very low computational cost and can be integrated with any existing solvers. We have evaluated the proposed DPC rule on both synthetic and real data sets. The experiments indicate that DPC is very effective in identifying the inactive features—especially for high dimensional data—which leads to a speedup up to several orders of magnitude.

Proceedings of the 31st International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

1. Introduction

Empirical studies have shown that learning multiple related tasks (MTL) simultaneously often provides superior predictive performance relative to learning each tasks independently (Ando & Zhang, 2005; Argyriou et al., 2008; Bakker & Heskes, 2003; Evgeniou et al., 2005; Zhang et al., 2006; Chen et al., 2013). This observation also has solid theoretical foundations (Ando & Zhang, 2005; Baxter, 2000; Ben-David & Schuller, 2003; Caruana, 1997), especially when the training sample size is small for each task. One popular MTL method especially for high-dimensional data is multi-task feature learning (MTFL), which uses the group Lasso penalty to ensure that all tasks select a common set of features (Argyriou et al., 2007). MTFL has found great success in many real-world applications including but not limited to: breast cancer classification (Zhang et al., 2010), disease progression prediction (Zhou et al., 2012), gene data analysis (Kim & Xing, 2009), and neural semantic basis discovery (Liu et al., 2009a). A major issue in MTFL—that is of great practical importance—is to develop efficient solvers (Liu et al., 2009b; Sra, 2012; Wang et al., 2013a; Gong et al., 2014). However, it remains challenging to apply the MTFL models to large-scale problems.

The idea of *screening* has been shown to be very effective in scaling the data and improving the efficiency of many popular sparse models, e.g., Lasso (El Ghaoui et al., 2012; Wang et al., 2013b; Wang et al.; Xiang et al., 2011; Tibshirani et al., 2012), nonnegative Lasso (Wang & Ye, 2014), group Lasso (Wang et al., 2013b; Wang et al.; Tibshirani et al., 2012), mixed-norm regression (Wang et al., 2013a), ℓ_1-regularized logistic regression (Wang et al., 2014b), sparse-group Lasso (Wang & Ye, 2014), support vector machine (SVM) (Ogawa et al., 2013; Wang et al., 2014a), and least absolute deviations (LAD) (Wang et al., 2014a). Essentially, screening aims to quickly identify the zero components in the solution vectors such that the corresponding features—called *inactive features* (e.g., Lasso)—or data samples—called *non-support vectors* (e.g., SVM)—can be removed from the optimization. Therefore, the size of

the data matrix and the number of variables to be computed can be significantly reduced, which may lead to substantial savings in the computational cost and memory usage *without sacrificing accuracy*. Compared to the solvers without screening, the speedup gained by the screening methods can be several orders of magnitude.

However, we note that all the existing screening methods are only applicable to sparse models with a *single* data matrix. Therefore, motivated by the challenges posed by large-scale data and the promising performance of existing screening methods, we propose a novel framework for developing effective and efficient screening rules for a popular MTFL model via the **d**ual **p**rojection onto **c**onvex sets (DPC). The framework of DPC extends the state-of-the-art screening rule, called EDPP (Wang et al.), for the standard Lasso problem (Tibshirani, 1996)—that assumes a *single* data matrix—to a popular MTFL model—that involves *multiple* data matrices across different tasks. To the best of our knowledge, DPC is the *first* screening rule that is applicable to sparse models with *multiple* data matrices.

The DPC screening rule detects the inactive features by maximizing a convex function over a convex set containing the dual optimal solution, which is a nonconvex problem. To find the region containing the dual optimal solution, we show that the corresponding dual problem can be formulated as a *projection* problem—which admits many desirable geometric properties—by utilizing the *bilinearity* of the *inner product*. Then, by a carefully chosen parameterization of the constraint set, we transform the nonconvex problem to a quadratic programming problem over one quadratic constraint (QP1QC) (Gay, 1981), which can be solved for the global optimum efficiently. Experiments on both synthetic and real data sets indicate that the speedup gained by DPC can be orders of magnitude. Moreover, DPC shows better performance as the feature dimension increases, which makes it a very competitive candidate for the applications of very high-dimensional data.

We organize the rest of this paper as follows. In Section 2, we briefly review some basics of a popular MTFL model. Then, we derive the dual problem in Section 3. Based on an indepth analysis of the geometric properties of the dual problem and the dual feasible set, we present the proposed DPC screening rule in Section 4. In Section 5, we evaluate the DPC rule on both synthetic and real data sets. We conclude this paper in Section 6. Please refer to the journal version for proofs not included in the main text.

Notation: Denote the ℓ_2 norm by $\|\cdot\|$. For $\mathbf{x} \in \mathbb{R}^n$, let its ith component be x_i, and the diagonal matrix with the entries of $\mathbf{x}$ on the main diagonal be $\mathrm{diag}(\mathbf{x})$. For a set of positive integers $\{N_t : t = 1, \ldots, T, \sum_{t=1}^T N_t = N\}$, we denote the tth subvector of $\mathbf{x} \in \mathbb{R}^N$ by $\mathbf{x}_t$ such that $\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_T^T)^T$, where $\mathbf{x}_t \in \mathbb{R}^{N_t}$ for $t = 1, \ldots, T$. For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we use $\langle \mathbf{x}, \mathbf{y} \rangle$ and $\mathbf{x}^T\mathbf{y}$ interchangeably to denote the inner product. For a matrix $M \in \mathbb{R}^{m\times n}$, let $\mathbf{m}^i$, $\mathbf{m}_j$, and m_{ij} be its ith row, j^{th} column and $(i,j)^{th}$ entry, respectively. We define the $(2,1)$-norm of M by $\|M\|_{2,1} = \sum_{i=1}^m \|\mathbf{m}^i\|$. For two matrices $A, B \in \mathbb{R}^{m\times n}$, we define their inner product by $\langle A, B\rangle = \mathrm{tr}(A^TB)$. Let I be the identity matrix. For a convex function $f(\cdot)$, let $\partial f(\cdot)$ be its subdifferential. For a vector $\mathbf{x}$ and a convex set $\mathcal{C}$, the *projection operator* is:

$$\mathrm{P}_{\mathcal{C}}(\mathbf{x}) := \mathrm{argmin}_{\mathbf{y}\in\mathcal{C}} \ \tfrac{1}{2}\|\mathbf{y} - \mathbf{x}\|.$$

2. Basics

In this section, we briefly review some basics of a popular MTFL model and mention several equivalent formulations.

Suppose that we have T learning tasks $\{(X_t, \mathbf{y}_t) : t = 1, \ldots, T\}$, where $X_t \in \mathbb{R}^{N_t\times d}$ is the data matrix of the tth task with N_t samples and d features, and $\mathbf{y}_t \in \mathbb{R}^{N_t}$ is the corresponding response vector. A widely used MTFL model (Argyriou et al., 2007) takes the form of

$$\min_{W\in\mathbb{R}^{d\times T}} \sum\nolimits_{t=1}^T \tfrac{1}{2}\|\mathbf{y}_t - X_t\mathbf{w}_t\|^2 + \lambda\|W\|_{2,1}, \quad (1)$$

where $\mathbf{w}_t \in \mathbb{R}^d$ is the weight vector of the tth task and $W = (\mathbf{w}_1, \ldots, \mathbf{w}_T)$. Because the $\|\cdot\|_{2,1}$-norm induces sparsity on the rows of W, the weight vectors across all tasks share the same sparse pattern. We note that the model in (1) is equivalent to several other popular MTFL models.

The first example introduces a positive weight parameter ρ_t for $t = 1, \ldots, T$ to each term in the loss function:

$$\min_{W\in\mathbb{R}^{d\times T}} \sum\nolimits_{t=1}^T \tfrac{1}{2\rho_t}\|\mathbf{y}_t - X_t\mathbf{w}_t\|^2 + \lambda\|W\|_{2,1},$$

which reduces to (1) by setting $\widetilde{\mathbf{y}}_t = \frac{\mathbf{y}_t}{\sqrt{\rho_t}}$ and $\widetilde{X}_t = \frac{X_t}{\sqrt{\rho_t}}$.

The second example introduces another regularizer to (1):

$$\min_{W\in\mathbb{R}^{d\times T}} \sum\nolimits_{t=1}^T \tfrac{1}{2}\|\mathbf{y}_t - X_t\mathbf{w}_t\|^2 + \lambda\|W\|_{2,1} + \rho\|W\|_F^2,$$

where ρ is a positive parameter and $\|\cdot\|_F$ is the Frobenius norm. Let $I \in \mathbb{R}^{d\times d}$ be the identity matrix and $\mathbf{0}$ be the d-dimensional vector with all zero entries. By letting

$$\bar{X}_t = (X_t^T, \sqrt{2\rho_t}I)^T, \ \ \bar{\mathbf{y}}_t = (\mathbf{y}_t^T, \mathbf{0}^T)^T, t = 1, \ldots, T,$$

we can also simplify the above MTFL model to (1).

In this paper, we focus on developing the DPC screening rule for the MTFL model in (1).

3. The Dual Problem

In this section, we show that we can formulate the dual problem of the MTFL model in (1) as a projection problem by utilizing the bilinearity of the inner product.

We first introduce a new set of variables:

$$\mathbf{z}_t = \mathbf{y}_t - X_t \mathbf{w}_t,\ t = 1, \ldots, T. \tag{2}$$

Then, the MTFL model in (1) can be written as

$$\min_{W,\mathbf{z}} \sum\nolimits_{t=1}^{T} \tfrac{1}{2}\|\mathbf{z}_t\|^2 + \lambda \|W\|_{2,1}, \tag{3}$$
$$\text{s.t.} \quad \mathbf{z}_t = \mathbf{y}_t - X_t \mathbf{w}_t,\ t = 1, \ldots, T.$$

Let $\lambda\theta \in \mathbb{R}^N$ be the vector of Lagrangian multipliers. Then, the Lagrangian of (1) is

$$L(W, \mathbf{z}; \theta) = \sum\nolimits_{t=1}^{T} \tfrac{1}{2}\|\mathbf{z}_t\|^2 + \lambda \|W\|_{2,1} + \lambda \sum\nolimits_{t=1}^{T} \langle \theta_t, \mathbf{y}_t - X_t \mathbf{w}_t - \mathbf{z}_t \rangle. \tag{4}$$

To get the dual problem, we need to minimize $L(W, \mathbf{z}; \theta)$ over W and $\mathbf{z}$. We can see that

$$0 = \nabla_{\mathbf{z}} L(W, \mathbf{z}; \theta) \Rightarrow \operatorname{argmin}_{\mathbf{z}} L(W, \mathbf{z}; \theta) = \lambda\theta. \tag{5}$$

For notational convenience, let

$$f(W) = \lambda \|W\|_{2,1} - \lambda \sum\nolimits_{t=1}^{T} \langle \theta_t, X_t \mathbf{w}_t \rangle.$$

Thus, to minimize $L(W, \mathbf{z}; \theta)$ with respect to W, it is equivalent to minimize $f(W)$, i.e.,

$$\{\hat{W} : 0 \in \partial_W L(\hat{W}, \mathbf{z}; \theta)\} = \{\hat{W} : 0 \in \partial f(\hat{W})\}.$$

By the *bilinearity of the inner product*, we can decouple $f(W)$ into a set of independent subproblems. Indeed, we can rewrite the second term of $f(W)$ as

$$\sum\nolimits_{t=1}^{T} \langle \theta_t, X_t \mathbf{w}_t \rangle = \sum\nolimits_{t=1}^{T} \langle X_t^T \theta_t, \mathbf{w}_t \rangle = \langle M, W \rangle, \tag{6}$$

where $M = (X_1^T \theta_1, \ldots, X_T^T \theta_T)$. Eq. (6) expresses $\langle M, W \rangle$ by the sum of the inner products of the corresponding columns. By the bilinearity of the inner product, we can also express $\langle M, W \rangle$ by the sum of the inner products of the corresponding rows:

$$\sum\nolimits_{t=1}^{T} \langle \theta_t, X_t \mathbf{w}_t \rangle = \langle M, W \rangle = \sum\nolimits_{\ell=1}^{d} \langle \mathbf{m}^\ell, \mathbf{w}^\ell \rangle. \tag{7}$$

Denote the jth column of X_t by $\mathbf{x}_j^{(t)}$. We can see that

$$\mathbf{m}^\ell = (\langle \mathbf{x}_\ell^{(1)}, \theta_1 \rangle, \langle \mathbf{x}_\ell^{(2)}, \theta_2 \rangle, \ldots, \langle \mathbf{x}_\ell^{(T)}, \theta_T \rangle). \tag{8}$$

Moreover, as $\|W\|_{2,1} = \sum_{\ell=1}^{d} \|\mathbf{w}^\ell\|$, Eqs. (7) implies that:

$$f(W) = \lambda \sum\nolimits_{\ell=1}^{d} f^{(\ell)}(\mathbf{w}^\ell),$$

where $f^{(\ell)}(\mathbf{w}^\ell) = \|\mathbf{w}^\ell\| - \langle \mathbf{m}^\ell, \mathbf{w}^\ell \rangle$. Thus, to minimize $f(W)$, we can minimize each $f^{(\ell)}(\mathbf{w}^\ell)$ separately. The subdifferential counterpart of the Fermat's rule (Bauschke & Combettes, 2011), i.e., $0 \in \partial f^{(\ell)}(\hat{\mathbf{w}}^\ell)$, yields:

$$\mathbf{m}^\ell \in \begin{cases} \hat{\mathbf{w}}^\ell / \|\hat{\mathbf{w}}^\ell\|, & \text{if } \hat{\mathbf{w}}^\ell \neq 0, \\ \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| \leq 1\}, & \text{if } \hat{\mathbf{w}}^\ell = 0, \end{cases} \tag{9}$$

where $\hat{\mathbf{w}}^\ell$ is the minimizer of $f^{(\ell)}(\cdot)$.

We note that Eq. (9) implies $\|\mathbf{m}^\ell\| \leq 1$. If this is not the case, then $f^\ell(\cdot)$ is not lower bounded (see the supplements for discussions), i.e., $\min_{\mathbf{w}^\ell} f^\ell(\mathbf{w}^\ell) = -\infty$. Thus, by Eqs. (5) and (9), the dual function is

$$q(\theta) = \min\nolimits_{W,\mathbf{z}} L(W, \mathbf{z}; \theta) = \begin{cases} -\frac{\lambda^2}{2}\|\theta\|^2 + \lambda\langle \theta, \mathbf{y} \rangle, & \|\mathbf{m}^\ell\| \leq 1,\ \forall\, \ell \in \{1, \ldots, d\}, \\ -\infty, & \text{otherwise.} \end{cases} \tag{10}$$

Maximizing $q(\theta)$ yields the dual problem of (1) as follows:

$$\max_{\theta} \ \tfrac{1}{2}\|\mathbf{y}\|^2 - \tfrac{\lambda^2}{2}\left\|\tfrac{\mathbf{y}}{\lambda} - \theta\right\|^2, \tag{11}$$
$$\text{s.t.} \sum\nolimits_{t=1}^{T} \langle \mathbf{x}_\ell^{(t)}, \theta_t \rangle^2 \leq 1,\ \ell = 1, \ldots, d.$$

It is evident that the problem in (11) is equivalent to

$$\min_{\theta} \ \tfrac{1}{2}\left\|\tfrac{\mathbf{y}}{\lambda} - \theta\right\|^2, \tag{12}$$
$$\text{s.t.} \sum\nolimits_{t=1}^{T} \langle \mathbf{x}_\ell^{(t)}, \theta_t \rangle^2 \leq 1,\ \ell = 1, \ldots, d.$$

In view of (12), it is indeed a projection problem. Let $\mathcal{F}$ be the feasible set of (12). Then, the optimal solution of (12), denoted by $\theta^*(\lambda)$, is the projection of $\mathbf{y}/\lambda$ onto $\mathcal{F}$, namely,

$$\theta^*(\lambda) = \mathrm{P}_{\mathcal{F}}\left(\tfrac{\mathbf{y}}{\lambda}\right). \tag{13}$$

4. The DPC Rule

In this section, we present the proposed DPC screening rule for the MTFL model in (1). Inspired by the Karush-Kuhn-Tucker (KKT) conditions (Güler, 2010), in Section 4.1, we first present the general guidelines. The most challenging part lies in two folds: 1) we need to estimate the dual optimal solution as accurately as possible; 2) we need to solve a nonconvex optimization problem. In Section 4.2, we give an accurate estimation of the dual optimal solution based on the geometric properties of the projection operators. Then, in Section 4.3, we show that we can efficiently solve for the global optimum to the nonconvex problem. We present the DPC rule for the MTFL model (1) in Section 4.4.

4.1. Guidelines for Developing DPC

We present the general guidelines to develop screening rules for the MTFL model (1) via the KKT conditions.

Let $W^*(\lambda) = (\mathbf{w}_1^*(\lambda), \ldots, \mathbf{w}_T^*(\lambda))$ be the optimal solution (1). By Eqs. (2), (5) and (9), the KKT conditions are:

$$\mathbf{y}_t = X_t \mathbf{w}_t^*(\lambda) + \lambda \theta_t^*(\lambda),\ t = 1, \ldots, T, \tag{14}$$

$$g_\ell(\theta^*(\lambda)) \in \begin{cases} 1, & \text{if } (\mathbf{w}^\ell)^*(\lambda) \neq 0, \\ [-1, 1], & \text{if } (\mathbf{w}^\ell)^*(\lambda) = 0, \end{cases} \ \ell = 1, \ldots, d. \tag{15}$$

where $(\mathbf{w}^\ell)^*(\lambda)$ is the ℓth row of $W^*(\lambda)$, and

$$g_\ell(\theta) = \sum\nolimits_{t=1}^{T} \langle \mathbf{x}_\ell^{(t)}, \theta_t \rangle^2,\ \ell = 1, \ldots, d. \tag{16}$$

For $\ell = 1, \ldots, d$, Eq. (15) yields

$$g_\ell(\theta^*(\lambda)) < 1 \Rightarrow (\mathbf{w}^\ell)^*(\lambda) = 0. \tag{R}$$

The rule in (**R**) provides a method to identify the rows in $W^*(\lambda)$ that have *only* zero entries. However, (**R**) is not applicable to real applications, as it assumes knowledge of $\theta^*(\lambda)$, and solving the dual problem (12) could be as expensive as solving the primal problem (1). Inspired by SAFE (El Ghaoui et al., 2012), we can first estimate a set $\mathbf{\Theta}$ that contains $\theta^*(\lambda)$, and then relax (**R**) as follows:

$$\max\nolimits_{\theta \in \mathbf{\Theta}} g_\ell(\theta) < 1 \Rightarrow (\mathbf{w}^\ell)^*(\lambda) = 0,\ \ell = 1, \ldots, d. \tag{R*}$$

Therefore, to develop a screening rule for the MTFL model in (1), (**R***) implies that: 1) we need to estimate a region $\mathbf{\Theta}$—that turns out to be a ball (please refer to Section 4.2)—containing $\theta^*(\lambda)$; 2) we need to solve the maximization problem—that turns out to be nonconvex (please refer to Section 4.3)—on the left hand side of (**R***).

4.2. Estimation of the Dual Optimal Solution

Based on the geometric properties of the dual problem (12) that is a projection problem, we first derive the closed form solutions of the primal and dual problems for specific values of λ in Section 4.2.1, and then give an accurate estimation of $\theta^*(\lambda)$ for the general cases in Section 4.2.2.

4.2.1. Closed form solutions

The primal and dual optimal solutions $W^*(\lambda)$ and $\theta^*(\lambda)$ are generally unknown. However, when the value of λ is sufficiently large, we expect that $W^*(\lambda) = 0$, and $\theta^*(\lambda) = \frac{\mathbf{y}}{\lambda}$ by Eq. (14). The following theorem confirms this.

Theorem 1. *For the MTFL model in (1), let*

$$\lambda_{\max} = \max_{\ell = 1, \ldots, d} \sqrt{\sum\nolimits_{t=1}^{T} \langle \mathbf{x}_\ell^{(t)}, \mathbf{y} \rangle^2}. \tag{17}$$

Then, the following statements are equivalent:

$$\tfrac{\mathbf{y}}{\lambda} \in \mathcal{F} \Leftrightarrow \theta^*(\lambda) = \tfrac{\mathbf{y}}{\lambda} \Leftrightarrow W^*(\lambda) = 0 \Leftrightarrow \lambda \geq \lambda_{\max}.$$

Remark 1. *Theorem 1 indicates that: both the primal and dual optimal solutions of the MTFL model (1) admit closed form solutions for $\lambda \geq \lambda_{\max}$. Thus, we will focus on the cases with $\lambda \in (0, \lambda_{\max})$ in the rest of this paper.*

4.2.2. The general cases

Theorem 1 gives a closed form solution of $\theta^*(\lambda)$ for $\lambda \geq \lambda_{\max}$. Therefore, we can estimate $\theta^*(\lambda)$ with $\lambda < \lambda_{\max}$ in terms of a known $\theta^*(\lambda_0)$. Specifically, we can simply set $\lambda_0 = \lambda_{\max}$ and utilize the result $\theta^*(\lambda_{\max}) = \mathbf{y}/\lambda_{\max}$. To make this paper self-contained, we first review some geometric properties of projection operators.

Theorem 2. (Ruszczyński, 2006) *Let $\mathcal{C}$ be a nonempty closed convex set. Then, for any point $\bar{\mathbf{u}}$, we have*

$$\mathbf{u} = \mathbf{P}_{\mathcal{C}}(\bar{\mathbf{u}}) \Leftrightarrow \bar{\mathbf{u}} - \mathbf{u} \in N_{\mathcal{C}}(\mathbf{u}),$$

where $N_{\mathcal{C}}(\mathbf{u}) = \{\mathbf{v} : \langle \mathbf{v}, \mathbf{u}' - \mathbf{u} \rangle \leq 0,\ \forall \mathbf{u}' \in \mathcal{C}\}$ is called the normal cone to $\mathcal{C}$ at $\mathbf{u} \in \mathcal{C}$.

Another useful property of the projection operator in estimating $\theta^*(\lambda)$ is the so-called *firmly nonexpansiveness*.

Theorem 3. (Bauschke & Combettes, 2011) *Let $\mathcal{C}$ be a nonempty closed convex subset of a Hilbert space $\mathcal{H}$. The projection operator with respect to $\mathcal{C}$ is firmly nonexpansive, namely, for any $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{H}$,*

$$\|\mathrm{P}_{\mathcal{C}}(\mathbf{u}_1) - \mathrm{P}_{\mathcal{C}}(\mathbf{u}_2)\|^2 + \|(I - \mathrm{P}_{\mathcal{C}})(\mathbf{u}_1) - (I - \mathrm{P}_{\mathcal{C}})(\mathbf{u}_2)\|^2 \leq \|\mathbf{u}_1 - \mathbf{u}_2\|^2. \tag{18}$$

The firmly nonexpansiveness of projection operators leads to the following useful result.

Corollary 4. *Let $\mathcal{C}$ be a nonempty closed convex subset of a Hilbert space $\mathcal{H}$ and $0 \in \mathcal{C}$. For any $\mathbf{u} \in \mathcal{H}$, we have:*

1. $\|\mathrm{P}_{\mathcal{C}}(\mathbf{u})\|^2 + \|\mathbf{u} - \mathrm{P}_{\mathcal{C}}(\mathbf{u})\|^2 \leq \|\mathbf{u}\|^2$.
2. $\langle \mathbf{u}, \mathbf{u} - \mathrm{P}_{\mathcal{C}}(\mathbf{u}) \rangle \geq 0$.

Remark 2. *Part 1 of* Corollary 4 *indicates that: if a closed convex set $\mathcal{C}$ contains the origin, then, for any point $\mathbf{u}$, the norm of its projection with respect to $\mathcal{C}$ is upper bounded by the norm of $\|\mathbf{u}\|$. The second part is a useful consequence of the first part and plays a crucial role in the estimation of the dual optimal solution (see Theorem 5).*

We are now ready to present an accurate estimation of the dual optimal solution $\theta^*(\lambda)$.

Theorem 5. *For the MTFL model in (1), suppose that $\theta^*(\lambda_0)$ is known with $\lambda_0 \in (0, \lambda_{\max}]$. Let g_ℓ be given by* Eq. (16) *for $\ell = 1, \ldots, d$, and*

$$\ell_* \in \left\{ \mathrm{argmax}_{\ell = 1, \ldots, d}\ g_\ell(\mathbf{y}) \right\}. \tag{19}$$

For any $\lambda \in (0, \lambda_0)$, we define

$$\mathbf{n}(\lambda_0) = \begin{cases} \frac{\mathbf{y}}{\lambda_0} - \theta^*(\lambda_0), & \text{if } \ \lambda_0 \in (0, \lambda_{\max}), \\ \nabla g_{\ell_*}\left(\frac{\mathbf{y}}{\lambda_{\max}}\right), & \text{if } \ \lambda_0 = \lambda_{\max}. \end{cases} \tag{20}$$

$$\mathbf{r}(\lambda, \lambda_0) = \tfrac{\mathbf{y}}{\lambda} - \theta^*(\lambda_0), \tag{21}$$

$$\mathbf{r}^{\perp}(\lambda, \lambda_0) = \mathbf{r}(\lambda, \lambda_0) - \frac{\langle \mathbf{n}(\lambda_0), \mathbf{r}(\lambda, \lambda_0) \rangle}{\|\mathbf{n}(\lambda_0)\|^2} \mathbf{n}(\lambda_0). \tag{22}$$

Then, the following holds:

1. $\mathbf{n}(\lambda) \in N_{\mathcal{F}}(\theta^*(\lambda))$,
2. $\langle \mathbf{y}, \mathbf{n}(\lambda_0)\rangle \geq 0$,
3. $\langle \mathbf{r}(\lambda, \lambda_0), \mathbf{n}(\lambda_0)\rangle \geq 0$,
4. $\left\|\theta^*(\lambda) - \left(\theta^*(\lambda_0) + \frac{1}{2}\mathbf{r}^{\perp}(\lambda, \lambda_0)\right)\right\| \leq \frac{1}{2}\|\mathbf{r}^{\perp}(\lambda, \lambda_0)\|$.

Consider Theorem 5. Part 1 characterizes $\theta^*(\lambda)$ via the normal cone. Parts 2 and 3 illustrate key geometric identities that lead to the accurate estimation of $\theta^*(\lambda)$ in part 4 (see supplement for details).

Remark 3. *The estimation of the dual optimal solution in DPC and EDPP* (Wang et al.)*—that is for Lasso—are both based on the geometric properties of the projection operators. Thus, the formulas of the estimation in Theorem 5 are similar to that of EDPP. However, we note that the estimations in DPC and EDPP are determined by the completely different geometric structures of the corresponding dual feasible sets. Problem* (12) *implies that the dual feasible set of the MTFL model* (1) *is much more complicated than that of Lasso—which is a polytope* (*the intersection of a set of closed half spaces*)*. Therefore, the estimation of the dual optimal solution in DPC is much more challenging than that of EDPP, e.g., we need to find a vector in the normal cone to the dual feasible set at* $\mathbf{y}/\lambda_{\max}$ [*see* $\mathbf{n}(\lambda_{\max})$].

For notational convenience, let

$$\mathbf{o}(\lambda, \lambda_0) = \theta^*(\lambda_0) + \frac{1}{2}\mathbf{r}^{\perp}(\lambda, \lambda_0). \tag{23}$$

Theorem 5 implies that $\theta^*(\lambda)$ lies in the ball:

$$\mathbf{\Theta}(\lambda, \lambda_0) = \left\{\theta : \|\theta - \mathbf{o}(\lambda, \lambda_0)\| \leq \frac{1}{2}\|\mathbf{r}^{\perp}(\lambda, \lambda_0)\|\right\}. \tag{24}$$

4.3. Solving the Nonconvex Problem

In this section, we solve the optimization problem in ($\mathbf{R}^*$) with $\mathbf{\Theta}$ given by $\mathbf{\Theta}(\lambda, \lambda_0)$ [see Eq. (24)], namely,

$$s_\ell(\lambda, \lambda_0) = \max_{\theta \in \mathbf{\Theta}(\lambda, \lambda_0)} \left\{ g_\ell(\theta) = \sum\nolimits_{t=1}^{T} \langle \mathbf{x}_\ell^{(t)}, \theta_t\rangle^2 \right\}. \tag{25}$$

Although $g_\ell(\cdot)$ and $\mathbf{\Theta}(\lambda, \lambda_0)$ are convex, problem (25) is nonconvex, as it is a maximization problem. However, we can efficiently solve for the *global* optimal solutions to (25) by transforming it to a QP1PC via a parametrization of the constraint set. We first cite the following result.

Theorem 6. (Gay, 1981) *Let H be a symmetric matrix and D be a positive definite matrix. Consider*

$$\min_{\|D\mathbf{u}\| \leq \Delta} \psi(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T H \mathbf{u} + \mathbf{q}^T \mathbf{u}, \tag{26}$$

where $\Delta > 0$. Then, $\mathbf{u}^*$ minimizes $\psi(\mathbf{u})$ over the constraint set if and only if there exists $\alpha^* \geq 0$—that is unique—such that $(H + \alpha^* D^T D)\mathbf{u}^*$ is positive semidefinite,

$$(H + \alpha^* D^T D)\mathbf{u}^* = -\mathbf{q}, \tag{27}$$

$$\|D\mathbf{u}^*\| = \Delta, \text{ if } \alpha^* > 0. \tag{28}$$

We are now ready to solve for $s_\ell(\lambda, \lambda_0)$.

Theorem 7. *Let $\mathbf{o} = \mathbf{o}(\lambda, \lambda_0)$ and $\mathbf{u}^*$ be the optimal solution of problem* (26) *with $\Delta = \frac{1}{2}\|\mathbf{r}^{\perp}(\lambda, \lambda_0)\|$, $D = I$,*
$H = -\operatorname{diag}(2\|\mathbf{x}\|_\ell^{(1)}, \ldots, 2\|\mathbf{x}\|_\ell^{(T)})$,
$\mathbf{q} = -\left(2\|\mathbf{x}_\ell^{(1)}\||\langle \mathbf{x}_\ell^{(1)}, \mathbf{o}_1\rangle|, \ldots, 2\|\mathbf{x}_\ell^{(T)}\||\langle \mathbf{x}_\ell^{(T)}, \mathbf{o}_T\rangle|\right)^T$,
namely, there exists a $\alpha^* \geq 0$ such that α^* and $\mathbf{u}^*$ solve Eqs. (27) *and* (28)*. Let*

$$\rho_\ell = \max\nolimits_{t=1,\ldots,T} \|\mathbf{x}_\ell^{(t)}\|, \quad \mathcal{I}_\ell = \left\{t_* : \|\mathbf{x}_\ell^{(t_*)}\| = \rho_\ell\right\}.$$

Then, the following hold:

1. *α^* is unique, and $\alpha^* \geq 2\rho_\ell$.*
2. *We define $\bar{\mathbf{u}} \in \mathbb{R}^T$ by*

$$\bar{u}_t = \begin{cases} -q_t/(h_{tt} + 2\rho_\ell), & \text{if } t \notin \mathcal{I}_\ell, \\ 0, & \text{otherwise}. \end{cases}$$

Then, we have

$$\alpha^* \in \begin{cases} 2\rho_\ell, & \text{if } \|\bar{\mathbf{u}}\| \leq \Delta, \text{ and } \langle \mathbf{x}_\ell^{(t_*)}, \mathbf{o}_{t_*}\rangle = 0, \text{ for } t_* \in \mathcal{I}_\ell, \\ (2\rho_\ell, \infty), & \text{otherwise}. \end{cases}$$

3. *Let $\mathcal{V} = \{\mathbf{v} \in \mathbb{R}^T : v_t = 0 \text{ for } t \notin \mathcal{I}_\ell, \|\bar{\mathbf{u}} + \mathbf{v}\| = \Delta\}$. Then, we have*

$$\mathbf{u}^* \in \begin{cases} \bar{\mathbf{u}} + \mathbf{v},\ \mathbf{v} \in \mathcal{V}, & \text{if } \alpha^* = 2\rho_\ell, \\ -(H + \alpha^* I)^{-1}\mathbf{q}, & \text{otherwise}. \end{cases}$$

4. *The maximum value of problem* (25) *is given by*

$$s_\ell(\lambda, \lambda_0) = \textstyle\sum_{t=1}^{T} \langle \mathbf{x}_\ell^{(t)}, \mathbf{o}_t\rangle^2 + \frac{\alpha^*}{2}\Delta^2 - \frac{1}{2}\mathbf{q}^T\mathbf{u}^*.$$

Proof. We first transform problem (25) to a QP1PC by a parameterization of $\mathbf{\Theta}(\lambda, \lambda_0)$:

$$\mathbf{\Theta}(\lambda, \lambda_0) = \left\{ \begin{pmatrix} \mathbf{o}_1 + u_1\theta_1 \\ \vdots \\ \mathbf{o}_T + u_T\theta_T \end{pmatrix} : \|\mathbf{u}\| \leq r, \|\theta_t\| \leq 1, , t = 1, \ldots, T \right\},$$

where $\mathbf{u} = (u_1, \ldots, u_T)^T$. We define

$$h_\ell(\mathbf{u}, \theta) = g_\ell\left(\begin{pmatrix} \mathbf{o}_1 + u_1\theta_1 \\ \vdots \\ \mathbf{o}_T + u_T\theta_T \end{pmatrix}\right).$$

Thus, problem (25) becomes

$$s_\ell(\lambda, \lambda_0) = \max_{\|\mathbf{u}\| \leq \Delta} \left\{ \max_{\{\theta : \|\theta_t\| \leq 1, t = 1, \ldots, T\}} h_\ell(\mathbf{u}, \theta) \right\}.$$

By the Cauchy-Schwartz inequality, for a fixed $\mathbf{u}$, we have

$$\phi(\mathbf{u}) = \max_{\{\theta:\|\theta_t\|\leq 1, t=1,\ldots,T\}} h_\ell(\mathbf{u},\theta)$$
$$= \sum_{t=1}^T u_t^2\|\mathbf{x}_\ell^{(t)}\|^2 + 2|u_t|\|\mathbf{x}_\ell^{(t)}\||\langle \mathbf{x}_\ell^{(t)},\mathbf{o}_t\rangle| + \langle \mathbf{x}_\ell^{(t)},\mathbf{o}_t\rangle^2.$$

Let $-\psi(\mathbf{u}) = \sum_{t=1}^T u_t^2\|\mathbf{x}_\ell^{(t)}\|^2 + 2u_t\|\mathbf{x}_\ell^{(t)}\||\langle \mathbf{x}_\ell^{(t)},\mathbf{o}_t\rangle|$. We can see that

$$\max_{\|\mathbf{u}\|\leq r} \phi(\mathbf{u}) = \max_{\|\mathbf{u}\|\leq r} -\psi(\mathbf{u}) + \sum_{t=1}^T \langle \mathbf{x}_\ell^{(t)},\mathbf{o}_t\rangle^2.$$

Thus, problem (25) becomes

$$s_\ell(\lambda,\lambda_0) = -\min_{\|\mathbf{u}\|\leq r} \psi(\mathbf{u}) + \sum_{t=1}^T \langle \mathbf{x}_\ell^{(t)},\mathbf{o}_t\rangle^2.$$

Therefore, to solve (25), it suffices to solve problem (26) with Δ, D, H, and $\mathbf{q}$ as in the theorem.

The statement follows immediately from Theorem 6. □

Remark 4. *To develop the DPC rule, ($\mathbf{R}^*$) implies that we only need the maximum value of problem (25). Thus, Theorem 6 does not show the global optimal solutions. However, in view of the proof, we can easily compute the global optimal solutions in terms of α^* and $\mathbf{u}^*$.*

Computing α^* and $\mathbf{u}^*$ Consider Theorem 7. If $\|\bar{\mathbf{u}}\| \leq \Delta$ and $\langle \mathbf{x}_\ell^{(t_*)},\mathbf{o}_{t_*}\rangle = 0$ for $t_* \in \mathcal{I}_\ell$, then α^* and $\mathbf{u}^*$ admit closed form solutions. Otherwise, α^* is strictly larger than $2\rho_\ell$, which implies that $H + \alpha^* I$ is positive definite and invertible. If this is the case, we apply Newton's method (Gay, 1981) to find α^* as follows. Let

$$\varphi(\alpha) = \|(H+\alpha I)^{-1}\mathbf{q}\|^{-1} - \Delta^{-1}.$$

Because $\varphi(\cdot)$ is strictly increasing on $(2\rho_\ell,\infty)$, α^* is the unique root of $\varphi(\cdot)$ on $(2\rho_\ell,\infty)$. Let $\alpha_0 = 2\rho_\ell$. Then, the kth iteration of Newton's method to solve $\varphi(\alpha^*) = 0$ is:

$$\mathbf{u}_k = -(H+\alpha_{k-1}I)^{-1}\mathbf{q}, \tag{29}$$

$$\alpha_k = \alpha_{k-1} + \|\mathbf{u}_k\|^2 \frac{\|\mathbf{u}_k\| - \Delta}{\Delta \mathbf{u}_k^T (H+\alpha_{k-1}I)^{-1}\mathbf{u}_k}. \tag{30}$$

As pointed out by Moré & Sorensen (1983), Newton's method is very efficient to find α^* as $\varphi(\alpha)$ is almost linear on $(2\rho_\ell,\infty)$. Our experiments indicates that five iterations usually leads to an accuracy higher than 10^{-15}.

4.4. The Proposed DPC Rule

As implied by $\mathbf{R}^*$, we present the proposed screening rule, DPC, for the MTFL model (1) in the following theorem.

Theorem 8. *For the MTFL model (1), suppose that $\theta^*(\lambda_0)$ is known with $\lambda_0 \in (0,\lambda_{\max}]$. Then, we have*

$$s_\ell(\lambda,\lambda_0) < 1 \Rightarrow (\mathbf{w}^\ell)^*(\lambda) = 0,\ \lambda \in (0,\lambda_0),$$

where $s_\ell(\lambda,\lambda_0)$ is given by Theorem 7.

In real applications, the optimal parameter value of λ is generally unknown. Commonly used approaches to determine an appropriate value of λ, such as cross validation and stability selection, need to solve the MTFL model over a grid of tuning parameter values $\lambda_1 > \lambda_2 > \ldots > \lambda_{\mathcal{K}}$, which is very time consuming. Inspired by the ideas of Strong Rule (Tibshirani et al., 2012) and SAFE (El Ghaoui et al., 2012), we develop the sequential version of DPC. Specifically, suppose that the optimal solution $W^*(\lambda_k)$ is known. Then, we apply DPC to identify the inactive features of MTFL model (1) at λ_{k+1} via $W^*(\lambda_k)$. We repeat this process until all $W^*(\lambda_k)$, $k = 1,\ldots,\mathcal{K}$ are computed.

Corollary 9. DPC *For the MTFL model (1), suppose that we are given a sequence of parameter values $\lambda_{\max} = \lambda_0 > \lambda_1 > \ldots > \lambda_{\mathcal{K}}$. Then, for any $k = 1,2,\ldots,\mathcal{K}-1$, if $W^*(\lambda_k)$ is known, we have*

$$s_\ell(\lambda_{k+1},\lambda_k) < 1 \Rightarrow (\mathbf{w}^\ell)^*(\lambda_{k+1}) = 0,$$

where $s_\ell(\lambda,\lambda_0)$ is given by Theorem 7.

We omit the proof of Corollary 9 as it is a direct application of Theorem 8.

5. Experiments

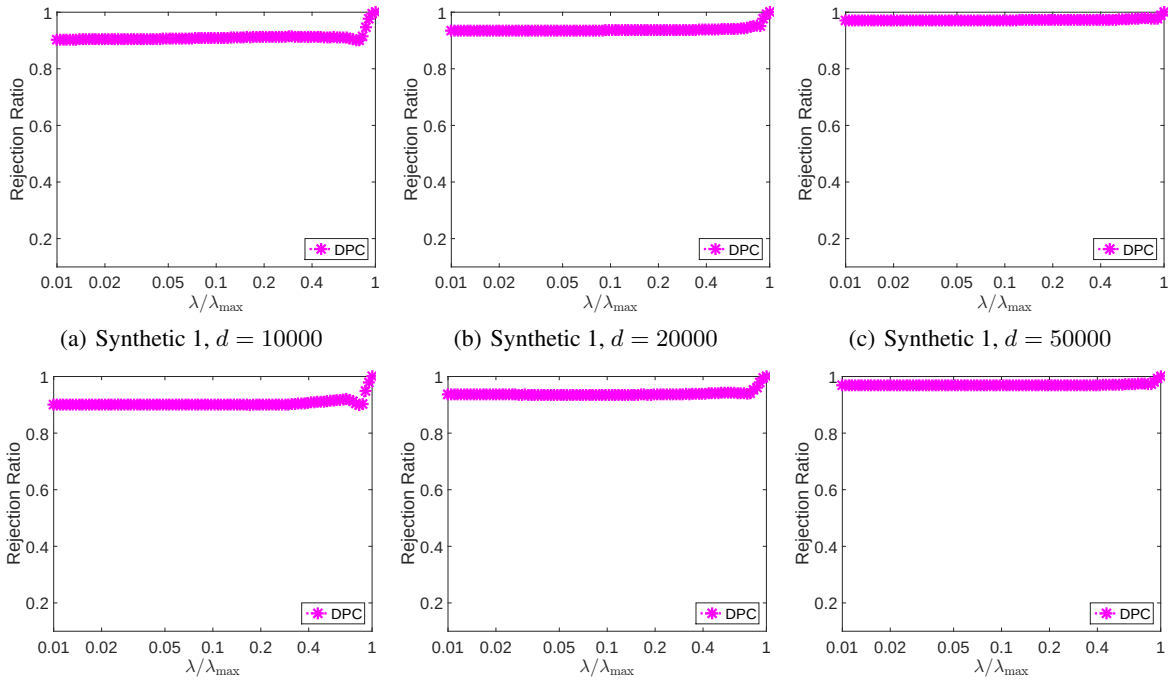
We evaluate DPC on both synthetic and real data sets. To measure the performance of DPC, we report the *rejection ratio*, namely, the ratio of the number of inactive features identified by DPC to the actual number of inactive features. We also report the *speedup*, i.e., the ratio of the running time of solver without screening to the running time of solver with DPC. The solver is from the SLEP package (Liu et al., 2009c). For each data set, we solve the MTFL model in (1) along a sequence of 100 tuning parameter values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}$ from 1.0 to 0.01. We only evaluate DPC since no existing screening rule is applicable for the MTFL model in (1).

5.1. Synthetic Studies

We perform experiments on two synthetic data sets, called Synthetic 1 and Synthetic 2, that are commonly used in the literature (Tibshirani et al., 2012; Zou & Hastie, 2005). Both synthetic 1 and Synthetic 2 have 50 tasks. Each task contains 50 samples. For $t = 1,\ldots,50$, the true model is

$$\mathbf{y}_t = \mathbf{X}_t\mathbf{w}_t^* + 0.01\epsilon,\ \epsilon \sim N(0,1).$$

For Synthetic 1, the entries of each data matrix $\mathbf{X}_t$ are i.i.d. standard Gaussian with pairwise correlation zero, i.e., $\mathrm{corr}\left(\mathbf{x}_i^{(t)},\mathbf{x}_j^{(t)}\right) = 0$. For Synthetic 2, the entries of each data matrix $\mathbf{X}_t$ are drawn from i.i.d. standard Gaussian with pairwise correlation $0.5^{|i-j|}$, i.e., $\mathrm{corr}\left(\mathbf{x}_i^{(t)},\mathbf{x}_j^{(t)}\right) = 0.5^{|i-j|}$. To construct $\mathbf{w}_t^*$, we first randomly select 10% of the features. Then, the corresponding components of $\mathbf{w}_t^*$



(a) Synthetic 1, $d = 10000$ (b) Synthetic 1, $d = 20000$ (c) Synthetic 1, $d = 50000$

(d) Synthetic 2, $d = 10000$ (e) Synthetic 2, $d = 20000$ (f) Synthetic 2, $d = 50000$

Figure 1. Rejection ratios of DPC on two synthetic data sets with different feature dimensions.

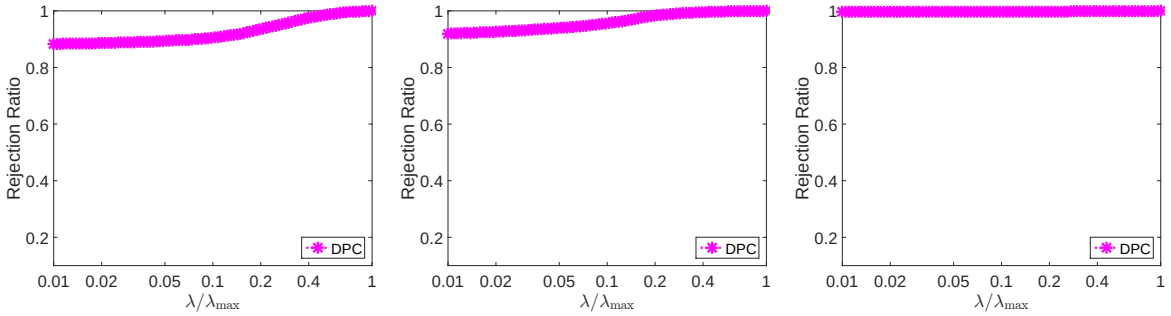
are populated from a standard Gaussian, and the remaining ones are set to 0. For both Synthetic 1 and Synthetic 2, we set the feature dimension to 10000, 20000, and 50000, respectively. For each setting, we run 20 trials and report the average performance in Fig. 1 and Table 1. Fig. 1 shows the rejection ratios of DPC on Synthetic 1 and Synthetic 2. For all the six settings, the rejection ratios of DPC are higher than 90%, even for small parameter values. This demonstrates one of the advantages of DPC, as previous empirical studies (El Ghaoui et al., 2012; Tibshirani et al., 2012; Wang et al.) indicate that the capability of screening rules in identifying inactive features usually decreases as the parameter value decreases. Moreover, Fig. 1 also shows that as the feature dimension increases, the rejection ratios of DPC become higher—that is very close to 1. This implies that the potential capability of DPC in identifying the inactive features on high-dimensional data sets would be even more significant.

Table 1 presents the running time of the solver with and without DPC. The speedup is very significant, which is up to 60 times. Take Synthetic 1 for example. When the feature dimension is 50000, the solver without DPC takes about 40.68 hours to solve problem (1) at 100 paramater values. In contrast, combined with DPC, the solver only takes less than one hour to solve the same 100 problems—which leads to a speedup about 60 times. Table 1 also shows that the computational cost of DPC is very low—which is negligible compared to that of the solver without screening. Moreover, as the rejection ratios of DPC increases with feature dimension growth (see Fig. 1), Table 1 shows that the speedup by DPC increases as well.

5.2. Experiments on Real Data Sets

We perform experiments on three real data sets: 1) the TDT2 text data set (Cai et al., 2009); 2) the animal data set (Lampert et al., 2009); 3) the Alzheimers Disease Neuroimaging Initiative (ADNI) data set (`http://adni.loni.usc.edu/`).

The Animal Data Set The data set consists of 30475 images of 50 animals classes. By following the experiment settings in (Kang et al., 2011), we choose 20 animal classes in the data set: antelope, grizzly-bear, killer-whale, beaver, Dalmatian, Persiancat, horse, german- shepherd, blue-whale, Siamese-cat, skunk, ox, tiger, hippopotamus, leopard, moose, spidermonkey, humpback-whale, elephant, and gorilla. We construct 20 tasks, each of which is a classification task of one type of animal against all the others. For the tth task, we first randomly select 30 samples from the tth class as the positive samples; and then we randomly select 30 samples from all the other classes as the negative samples. We utilize all the seven sets of features kindly provided by Lampert et al. (2009): color histogram features, local self-similarity features, PyramidHOG (PHOG) features, SIFT features, colorSIFT features, SURF features, and DECAF features. Thus, each image is represented by a 15036-dimensional vectors. Hence, the data matrix X_t of the tth task is of 60×15036, where $t = 1, \ldots, 20$.



(a) Animal, $d = 15036$ (b) TDT2, $d = 24262$ (c) ADNI, $d = 504095$

Figure 2. Rejection ratios of DPC on three real data sets.

Table 1. Running time (in minutes) for solving the MTFL model (1) along a sequence of 100 tuning parameter values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}$ from 1.0 to 0.01 by (a): the solver (Liu et al., 2009c) without screening (see the third column); (b): the solver with DPC (see the fifth column).

	d	solver	DPC	DPC+solver	**speedup**
Synthetic 1	10000	405.75	0.7	28.12	**14.43**
	20000	913.70	1.36	37.02	**24.68**
	50000	2441.57	3.50	42.08	**58.03**
Synthetic 2	10000	406.85	0.70	29.28	**13.89**
	20000	906.09	1.37	36.66	**24.72**
	50000	2435.38	3.46	44.78	**54.39**
Animal	15036	311.71	0.47	16.36	**19.05**
TDT2	24262	958.66	1.87	44.11	**21.74**
ADNI	504095	9625.58	21.13	35.34	**272.37**

The TDT2 Data Set The original data set contains 9394 documents of 30 categories. Each document is represented by a 36771-dimensional vector. Similar to the Animal data set, we construct 30 tasks, each of which is a classification task of one category against all the others (Amit et al., 2007). Also, for the tth task, we first randomly select 50 samples from the tth category as the positive samples, and then we randomly select 50 samples from all the other categories as the negative samples. Moreover, we remove the features that have only zero entries, thus leaving us 24262 features. Hence, the data matrix X_t of the tth task is of 100×24262, where $t = 1, \ldots, 30$.

The ADNI Data Set The data set consists of 747 patients with 504095 single nucleotide polymorphisms (SNPs), and the volume of 93 brain regions for each patient. We first randomly select 20 brain regions. Then, for each region, we randomly select 50 patients, and utilize the corresponding SNPs data as the data matrix and the volumes of that brain region as the response. Thus, we have 20 tasks, each of which is a regression task. The data matrix X_t of the tth task is of 50×504095, where $t = 1, \ldots, 20$.

Fig. 2 shows the rejection ratios of DPC—that are above 90%—on the aforementioned three real data sets. In particular, the rejection ratios of DPC on the ADNI data set are higher than 99% at the 100 parameter values. Table 1 shows that the resulting speedup is very significant—that is up to 270 times. We note that the feature dimension of the ADNI data set is more than *half million*. Without screening, Table 1 shows that the solver takes about *seven days* (approximately *one week*) to compute the MTFL model (1) at 100 parameter values. However, integrated with the DPC screening rule, the solver computes the 100 solutions in about *half an hour*. The experiments again indicate that DPC provides better performance (in terms of rejection ratios and speedup) for higher dimensional data sets.

6. Conclusion

In this paper, we propose a novel screening method for the MTFL model in (1), called DPC. The DPC screening rule is based on an indepth analysis of the geometric properties of the dual problem and the dual feasible set. To the best of our knowledge, DPC is the first screening rule that is applicable to sparse models with multiple data matrices. DPC is *safe* in the sense that the identified features by DPC are guaranteed to have zero coefficients in the solution vectors across all tasks. Experiments on synthetic and real data sets demonstrate that DPC is very effective in identifying the inactive features, which leads to a substantial savings in computational cost and memory usage *without sacrificing accuracy*. Moreover, DPC is more effective as the feature dimension increases, which makes DPC a very competitive candidate for the applications of very high-dimensional data. We plan to extend DPC to more general MTFL models, e.g., the MTFL models with multiple regularizers.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (IIS-0953662, IIS-1421057, and IIS-1421100) and the National Institutes of Health (R01 LM010730 and U54 EB020403).

References

Amit, Y., Fink, M., Srebro, N., and Ullman, S. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th Annual International Conference on Machine Learning*, 2007.

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

Argyriou, A., Evgeniou., T., and Pontil, M. Multi-task feature learning. In *Advances in neural information processing systems*, 2007.

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

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

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

Baxter, J. A model for inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

Ben-David, S. and Schuller, R. Exploiting task relatedness for multiple task learning. In *Proceedings of Computational Learning Theory*, 2003.

Cai, Deng, Wang, Xuanhui, and He, Xiaofei. Probabilistic dyadic data analysis with local and global consistency. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 105–112, 2009.

Caruana, R. Multitask learning. *Machine Learning*, 28: 41–75, 1997.

Chen, J., Tang, L., Liu, J., and Ye, J. A convex formulation for learning shared structures from multiple tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1025–1038, 2013.

El Ghaoui, L., Viallon, V., and Rabbani, T. Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization*, 8:667–698, 2012.

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

Gay, D. Computing optimal locally constrained steps. *SIAM Journal on Scientific and Statistical Computing*, 1981.

Gong, P., Zhou, J., Fan, W., and Ye, J. Efficient multi-task feature learning with calibration. In *International Conference On Knowledge Discovery and Data Mining*, 2014.

Güler, O. *Foundations of Optimization*. Springer, 2010.

Kang, Z., Grauman, K., and Sha, F. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th Annual International Conference on Machine Learning*, 2011.

Kim, S. and Xing, E. Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

Lampert, C., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class atttribute transfer. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

Liu, H., Palatucci, M., and Zhang, J. Bsparsity coordinate descent procedures for the multi-task with applications to neural semantic basis discovery. In *International Conference on Machine Learning*, 2009a.

Liu, J., Ji, S., and Ye, J. Multi-task feature learning with efficient $\ell_{2,1}$-norm minimization. In *The* 25th *Conference on Uncertainty in Artificial Intelligence*, 2009b.

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

Moré, J. and Sorensen, D. Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, 1983.

Ogawa, K., Suzuki, Y., and Takeuchi, I. Safe screening of non-support vectors in pathwise SVM computation. In *Proceedings of the 30th Annual International Conference on Machine Learning*, 2013.

Ruszczyński, A. *Nonlinear Optimization*. Princeton University Press, 2006.

Sra, S. Fast projections onto mixed-norm balls with applications. *Data Mining and Knowledge Discovery*, 2012.

Tibshirani, R. Regression shringkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society Series B*, 74:245–266, 2012.

Wang, J. and Ye, J. Two-layer feature reduction for sparse-group lasso via decomposition of convex sets. In *Advances in Neural Information Processing Systems*, 2014.

Wang, J., Wonka, P., and Ye, J. Lasso screening rules via dual polytope projection. *Journal of Machine Learning Research,* to appear.

Wang, J., Liu, J., and Ye, J. Efficient mixed-norm regularization: Algorithms and safe screening methods. *arXiv:1307.4156*, 2013a.

Wang, J., Zhou, J., Wonka, P., and Ye, J. Lasso screening rules via dual polytope projection. In *Advances in Neural Information Processing Systems*, 2013b.

Wang, J., Wonka, P., and Ye, J. Scaling SVM and least absolute deviations via exact data reduction. In *Proceedings of the 31th Annual International Conference on Machine Learning*, 2014a.

Wang, J., Zhou, J., Liu, J., Wonka, P., and Ye, J. A safe screening rule for sparse logistic regression. In *Advances in Neural Information Processing Systems*, 2014b.

Xiang, Z. J., Xu, H., and Ramadge, P. J. Learning sparse representation of high dimensional data on large scale dictionaries. In *Advances in Neural Information Processing Systems*, 2011.

Zhang, J., Ghahramani, Z., and Yang, Y. Learning multiple related tasks using latent independent component analysis. In *Advances in Neural Information Processing Systems*, 2006.

Zhang, Y., Yeung, D., and Xu, Q. Probabilistic multi-task feature selection. In *Advances in Neural Information Processing Systems*, 2010.

Zhou, J., Liu, J., Narayan, V., and Ye, J. Modeling disease progression via fused sparse group lasso. In *International Conference On Knowledge Discovery and Data Mining*, 2012.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67:301–320, 2005.