Safe Subspace Screening for Nuclear Norm Regularized Least Squares Problems

Abstract

Nuclear norm regularization has been shown very promising for pursing a low rank solution for matrix variable in various machine learning problems. Many efforts have been devoted to develop efficient algorithms for solving the optimization problem in nuclear norm regularization. Solving the problem for large-scale matrix variables, however, is still a challenging task since the complexity grows fast with the size of matrix variable. In this work, we propose a novel method called safe subspace screening (SSS), to improve the efficiency of the solver for nuclear norm regularized least squares problems. Motivated by the fact that the low rank solution can be represented by a few subspaces, the proposed method accurately discards a predominant percentage of inactive subspaces prior to solving the problem to reduce problem size. Consequently, a much smaller problem is required to solve, making it more efficient than optimizing the original problem. The proposed SSS is safe, in that its solution is identical to the solution from the solver. In addition, the proposed SSS can be used together with any existing nuclear norm solver since it is independent of the solver. We have evaluated the proposed SSS on several synthetic as well as real data sets. Extensive results show that the proposed SSS is very effective in inactive subspace screening and significantly improves the efficiency of existing solvers.

1. Introduction

To obtain a low rank matrix solution, many machine learning problems are formulated as minimizing nuclear norm regularized least squares problem (Yuan et al., 2007; Argyriou et al., 2008; Kang et al., 2011; Favaro et al., 2011). In the past several years, a number of efficient algorithms have been developed to solve the optimization problem raised by this formulation (Ji & Ye, 2009; Toh & Yuan, 2010; Jaggi & Sulovský, 2010; Mazumder et al., 2010; Shalev-Shwartz et al., 2011; Avron et al., 2012; Mishra et al., 2013; Hsieh & Olsen, 2014). Solving the problem for large-scale matrix variables, however, is still a challenging task since the computational complexity grows fast with the size of the matrix variable. On the other side, in many real applications, the size of matrix variable is becoming larger and larger in the big data era.

In the optimization of Lasso (Tibshirani, 1996), Ghaoui et al. lay the groundwork on safe screening method to identify the *features* that corresponding to zero coefficient in the solution and discard them prior to solving the optimization problem (Ghaoui et al., 2012). Their method has been further improved by a large body of work on screening performance (Xiang et al., 2011; Tibshirani et al., 2012; Wang et al., 2013; Liu et al., 2014) and extended to discard features for more general ℓ_1 norm regularized sparse problems (Wang et al., 2014b; Wang & Ye, 2014). In addition, the idea of screening has also been studied for discarding non-support *vectors* in the support vector machine (SVM) (Ogawa et al., 2013; Wang et al., 2014a) since there are only sparse support vectors used in the solution of SVM. Previous screening methods can be considered in two categories, one is safe screening method like (Ghaoui et al., 2012; Xiang et al., 2011; Wang et al., 2013; Ogawa et al., 2013), in which the discarded features are guaranteed to have zero coefficients in the solution, or vectors guaranteed to be non-support vectors. Another category is heuristic screening method such as strong rules (Tibshirani et al., 2012), sure independence screening (SIS) (Fan & Lv, 2008; Fan & Song, 2010). In these methods, since features are screened out by several heuristic criteria, some features corresponding to nonzero coefficients may be mistakenly discarded.

In this work, we propose a method called safe subspace screening (SSS) for discarding *subspaces* in nuclear norm regularized least squares problem. Suppose $\mathbf{W} \in \mathbb{R}^{d\times m}$ is the matrix variable, let us represent $\mathbf{W}$ as the sum of rank one matrices

$$\mathbf{W} = \sum_{i=1}^{d}\sum_{j=1}^{m} \mathbf{\Theta}_{ij}\mathbf{u}_i\mathbf{v}_j^T \tag{1}$$

where $\mathbf{\Theta} \in \mathbb{R}^{d\times m}$, $\{\mathbf{u}_i \in \mathbb{R}^d\}_{i=1}^d$ and $\{\mathbf{v}_j \in \mathbb{R}^m\}_{j=1}^m$ are orthogonal bases in $\mathbb{R}^{d\times d}$ and $\mathbb{R}^{m\times m}$, respectively. It is

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

easy to verify that any matrix in $\mathbb{R}^{d\times m}$ can be represented in this form as both $\{\mathbf{u}_i\}_{i=1}^d$ and $\{\mathbf{v}_j\}_{j=1}^m$ are orthogonal bases. Given $\mathbf{u}_i$ and $\mathbf{v}_j$, we aim to identify inactive subspaces that $\{\mathbf{u}_i\mathbf{v}_j^T|\mathbf{\Theta}_{ij}=0\}$ in the solution prior to solving the problem. This allows to solve an equivalent problem on a lower-dimensional subspace corresponding to $\mathbf{\Theta}_{ij}$ that are likely to be nonzero, thus reducing to a smaller problem and can be more efficiently solved.

Although nuclear norm can be considered as the ℓ_1 norm of singular values, a number of key differences between ℓ_1 norm and nuclear norm regularization make our work a nontrivial extension of previous feature screening works. Essentially, the feature screening rules for ℓ_1 norm regularization mainly make use of the Karush-Kuhn-Tucker (KKT) condition at the optimal solution. Specifically, the subgradient of ℓ_1 norm at zero and nonzero points have different ranges: $\{-1,1\}$ at nonzero points, and $[-1,1]$ at zero points. Therefore, one component of the solution will be zero if its subgradient belongs to $(-1,1)$ and it is a natural approach in such cases to discard the corresponding feature. Methods along this line, however, are not applicable for subspace screening because the subgradient of nuclear norm at both zero and nonzero $\mathbf{\Theta}_{ij}$ are $[-1,1]$ (Watson, 1992). Therefore, the subgradient at $\mathbf{\Theta}_{ij}$ can not be used to determine whether $\mathbf{\Theta}_{ij}$ in the solution is zero or not. More detailed technical derivation for this is provided in the Supplementary Materials. To address this problem, we propose a novel subspace screening rule by making use of the property of orthogonal subspaces. Specifically, one subspace will not appear in the solution and can be safely discarded if the solution is orthogonal to that subspace. In other words, for each subspace, we can evaluate the cosine for the angle between the solution and that subspace, and it can be screened out if the value is 0 meaning the solution is orthogonal to the subspace.

To utilize the aforementioned feature screening rule for identifying inactive features, we need to know the solution, which however is unknown before solving the problem. Therefore, previous feature screening methods usually construct a feasible set for the solution by using some prior knowledge. One common prior knowledge is that, for ℓ_1 norm regularization, there exists a particular regularization parameter which is the smallest one such that all elements of the solution to be zero. Although this also holds for nuclear norm as shown in Sec. 3.2, it is not surprising that this prior knowledge does not work well for subspace screening. In fact, the prior knowledge can not even identity any inactive subspace. The reason for that is, unlike the features that are fixed in feature screening, we need to choose $\{\mathbf{u}_i\}_{i=1}^d$ and $\{\mathbf{v}_j\}_{j=1}^m$ in subspace screening, which are quite important for the performance of subspace screening and can be chosen appropriately by utilizing the prior knowledge. On the other hand, if the same strategy as feature screening is adopted, the prior knowledge in this case is a zero matrix solution at that particular regularization parameter. Then, $\{\mathbf{u}_i\}_{i=1}^d$ and $\{\mathbf{v}_j\}_{j=1}^m$ can be only chosen as standard basis, which leads to $\mathbf{\Theta}=\mathbf{W}$. As we know, it is possible that a low rank $\mathbf{W}$ with all its elements being nonzero, then all elements of $\mathbf{\Theta}$ are also nonzero. In the proposed method, to provide more informative $\{\mathbf{u}_i\}_{i=1}^d$ and $\{\mathbf{v}_j\}_{j=1}^m$, we seek to utilize the solution at a very small regularization parameter, which can be easily obtained by exploiting a smart initialization strategy and it can provide a more appropriate choice for $\{\mathbf{u}_i\}_{i=1}^d$ and $\{\mathbf{v}_j\}_{j=1}^m$ as it has many singular vectors with nonzero singular values.

As the name indicates, the proposed method is safe in the sense that the discarded subspaces definitely do not appear in the solution. In addition, it can be used in conjunction with any existing nuclear norm solver as it is independent of the solver. To the best of our knowledge, the proposed method is the first work to identify and discard the subspaces that will not appear in the solution prior to solving the problem.

Notations: Throughout the paper, vectors and matrices will be denoted by lower and upper case boldface characters (e.g. $\mathbf{u}$ and $\mathbf{U}$), respectively. We use the notation $\mathbf{A}_{ij}$ to refer to the (i,j)th entry of $\mathbf{A}$. Moreover, the ith row and jth column of $\mathbf{A}$ are denoted by $\mathbf{A}_{i\cdot}$ and $\mathbf{A}_{\cdot j}$. Let $\|\cdot\|_2$ denote the Euclidean norm for a vector. For matrix norm, the Frobenius norm is denoted by $\|\cdot\|_F$. In addition, $\|\cdot\|_*$ and $\|\cdot\|_2$ denote the nuclear norm and spectral norm, respectively. The trace of a matrix is denoted by $\mathrm{Tr}\,[\cdot]$. $\mathbf{0}$ is used to denote a zero vector or matrix and its size is determined by the context. Let $\mathbf{I}$ denote an identity matrix with approximate size.

2. Motivation of Safe Subspace Screening

Specifically, we consider the following nuclear norm regularized least squares problem (Toh & Yuan, 2010)

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times m}} \frac{1}{2}\|\mathbf{X}\mathbf{W}-\mathbf{Y}\|_F^2+\lambda\|\mathbf{W}\|_* \tag{2}$$

where $\mathbf{X}\in\mathbb{R}^{n\times d}$ is the input data and $\mathbf{Y}\in\mathbb{R}^{n\times m}$ is the corresponding output, $\mathbf{W}\in\mathbb{R}^{d\times m}$ is the matrix variable, and λ is a regularization parameter. Many machine learning problems can be formulated as this form, e.g. multivariate learning regression (Lu et al., 2012), multi-task learning (Argyriou et al., 2008; Kang et al., 2011), subspace clustering (Favaro et al., 2011). Suppose we are given $\{\mathbf{u}_i\}_{i=1}^d$ and $\{\mathbf{v}_j\}_{j=1}^m$, substituting $\mathbf{W}$ in Eq. (1) into Eq. (2), we obtain the following equivalent problem

$$\min_{\mathbf{\Theta}\in\mathbb{R}^{d\times m}} \frac{1}{2}\left\|\mathbf{X}\sum_{i=1}^d\sum_{j=1}^m\mathbf{\Theta}_{ij}\mathbf{u}_i\mathbf{v}_j^T-\mathbf{Y}\right\|_F^2+\lambda\left\|\sum_{i=1}^d\sum_{j=1}^m\mathbf{\Theta}_{ij}\mathbf{u}_i\mathbf{v}_j^T\right\|_* \tag{3}$$

In the following, we use $\mathbf{W}^*_\lambda$ and $\mathbf{\Theta}^*_\lambda$ to denote the solutions to Eq. (2) and Eq. (3) when the value of regularization parameter is λ, respectively. It is easy to verify that $\{\mathbf{u}_i\mathbf{v}_j^T\}_{i=1,j=1}^{d,m}$ is orthogonal to each other. Therefore, for a particular subspace $\mathbf{u}_i\mathbf{v}_j^T$, the value of $(\mathbf{\Theta}^*_\lambda)_{ij}$ will be 0 if and only if

$$\left|\mathrm{Tr}\left[(\mathbf{W}^*_\lambda)^T\left(\mathbf{u}_i\mathbf{v}_j^T\right)\right]\right| = \left|\mathbf{u}_i^T\mathbf{W}^*_\lambda\mathbf{v}_j\right| = 0 \qquad (4)$$

since $\left(\mathbf{u}_i^T\mathbf{W}^*_\lambda\mathbf{v}_j\right)/\left\|\mathbf{W}^*_\lambda\right\|_F$ is the cosine of the angle between $\mathbf{W}^*_\lambda$ and $\mathbf{u}_i\mathbf{v}_j^T$. In other words, $\mathbf{u}_i\mathbf{v}_j^T$ can be safely discarded in the representation of $\mathbf{W}^*_\lambda$ and $(\mathbf{\Theta}^*_\lambda)_{ij}$ can be safely set as 0 even prior to optimizing Eq. (3). We only need to focus on $\mathbf{\Theta}^*_{ij}$ such that

$$\mathbf{u}_i^T\mathbf{W}^*_\lambda\mathbf{v}_j \neq 0 \qquad (5)$$

Let $\widehat{\mathbf{U}} = [\cdots, \mathbf{u}_i, \cdots]$ and $\widehat{\mathbf{V}} = [\cdots, \mathbf{v}_j, \cdots]$ be all the $\mathbf{u}_i$ and $\mathbf{v}_j$ that satisfy Eq. (5), respectively. Let $\widehat{\mathbf{U}}^\perp$ and $\widehat{\mathbf{V}}^\perp$ denote the set of $\mathbf{u}_i$ and $\mathbf{v}_j$ that do not appear in $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$, respectively. Based on these definitions, we can form a column basis $\widetilde{\mathbf{U}} = [\widehat{\mathbf{U}}, \widehat{\mathbf{U}}^\perp] \in \mathbb{R}^{d\times d}$ and row basis $\widetilde{\mathbf{V}} = [\widehat{\mathbf{V}}, \widehat{\mathbf{V}}^\perp] \in \mathbb{R}^{m\times m}$. Then, $\mathbf{W}$ can be re-parameterized as $\mathbf{W} = \widetilde{\mathbf{U}}\mathbf{\Theta}\widetilde{\mathbf{V}}^T$. By using this representation, Eq. (3) can be rewritten as

$$\min_{\mathbf{\Theta}\in\mathbb{R}^{d\times m}} \frac{1}{2}\left\|\mathbf{X}\widetilde{\mathbf{U}}\mathbf{\Theta}\widetilde{\mathbf{V}}^T - \mathbf{Y}\right\|_F^2 + \lambda\left\|\widetilde{\mathbf{U}}\mathbf{\Theta}\widetilde{\mathbf{V}}^T\right\|_* \qquad (6)$$

Suppose $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ have $\widehat{d}$ and $\widehat{m}$ columns, respectively. According to previous discussions, we only need to solve the $\widehat{d}\times\widehat{m}$ leading upper-left corner submatrix of $\mathbf{\Theta}$ since all other $\mathbf{\Theta}_{ij}$ corresponding to the subspaces can be safely discarded and their values are zero in the solution.

After applying safe subspace screening, the problem Eq. (6) reduces to the following equivalent problem

$$\min_{\widehat{\mathbf{\Theta}}\in\mathbb{R}^{\widehat{d}\times\widehat{m}}} \frac{1}{2}\left\|\mathbf{X}\widehat{\mathbf{U}}\widehat{\mathbf{\Theta}}\widehat{\mathbf{V}}^T - \mathbf{Y}\right\|_F^2 + \lambda\left\|\widehat{\mathbf{U}}\widehat{\mathbf{\Theta}}\widehat{\mathbf{V}}^T\right\|_* \qquad (7)$$

where $\widehat{\mathbf{\Theta}} = \mathbf{\Theta}_{1:\widehat{d},1:\widehat{m}} \in \mathbb{R}^{\widehat{d}\times\widehat{m}}$. Since both $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ are orthogonal bases, it implies $\|\widehat{\mathbf{U}}\widehat{\mathbf{\Theta}}\widehat{\mathbf{V}}^T\|_* = \|\widehat{\mathbf{\Theta}}\|_*$. Then the problem in Eq. (7) can be rewritten as

$$\min_{\widehat{\mathbf{\Theta}}\in\mathbb{R}^{\widehat{d}\times\widehat{m}}} \frac{1}{2}\left\|\mathbf{X}\widehat{\mathbf{U}}\widehat{\mathbf{\Theta}}\widehat{\mathbf{V}}^T - \mathbf{Y}\right\|_F^2 + \lambda\|\widehat{\mathbf{\Theta}}\|_* \qquad (8)$$

In Eq. (8), we only need to solve the optimization problem with a $\widehat{d}\times\widehat{m}$ matrix variable instead of $d\times m$ as in Eq. (2), leading to potentially substantial improvement in efficiency.

3. The Proposed Safe Subspace Screening

In this section, we present the details of the proposed safe subspace screening rule for the problem in Eq. (3).

3.1. Overview of the Proposed Method

To utilize the rule developed in Eq. (4) to identify inactive subspaces, we need the solution $\mathbf{W}^*_\lambda$, which is unknown prior to solving the Eq. (2). Therefore, we seek to construct a feasible set for $\mathbf{W}^*_\lambda$ and estimate the upper bound for $\left|\mathbf{u}_i^T\mathbf{W}^*_\lambda\mathbf{v}_j\right|$. In particular, the technique used to construct the feasible set is the so called variational inequality, which is a necessary condition for the optimal solution of a constrained optimization problem (Güler, 2010). Therefore, in Sec. 3.2, we first introduce the dual problem of Eq. (2) to obtain a constrained optimization problem. By using the relationship between primal and dual optimal solutions, the upper bound problem can be reformulated as a function of the dual optimal solution. Then, in Sec. 3.3, a feasible set is constructed for the dual optimal solution. For each pair of $\mathbf{u}_i$ and $\mathbf{v}_j$, Sec. 3.4 discusses how to estimate the upper bound over the feasible set. In fact, as we shall see, the upper bound problem has a closed form solution due to special structure of the objective function and constraints. The proposed safe subspace screening rule for Eq. (3) based on Eq. (4) is presented in Sec. 3.5. Due to space limitation, all technical derivations and proofs are provided in the Supplementary Materials.

3.2. The Dual Problem

The dual problem of Eq. (2) can be written as

$$\min \frac{1}{2}\left\|\mathbf{P} - \frac{\mathbf{Y}}{\lambda}\right\|_F^2 \quad s.t. \quad \left\|\mathbf{X}^T\mathbf{P}\right\|_2 \leq 1 \qquad (9)$$

where $\mathbf{P} \in \mathbb{R}^{n\times m}$ is the dual variable. Similarly, let $\mathbf{P}^*_\lambda$ denote the solution to Eq. (9) when the value of regularization parameter is λ. By using the KKT condition, we can establish the following relationship for the primal solution $\mathbf{W}^*_\lambda$ and the dual solution $\mathbf{P}^*_\lambda$

$$\lambda\mathbf{P}^*_\lambda = \mathbf{Y} - \mathbf{X}\mathbf{W}^*_\lambda \qquad (10)$$

According to this relationship, $\left|\mathbf{u}_i^T\mathbf{W}^*_\lambda\mathbf{v}_j\right|$ can be reformulated as

$$\left|\mathbf{u}_i\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{Y} - \lambda\mathbf{X}^T\mathbf{P}^*_\lambda\right)\right)\mathbf{v}_j\right| \qquad (11)$$

In addition, it is easy to verify that there exists a specific parameter value $\lambda_{\max}$ such that the primal optimal solution $\mathbf{W}^*_\lambda$ is $\mathbf{0}$ for any $\lambda \geq \lambda_{\max}$. According to Eq. (9) and Eq. (10), the $\lambda_{\max}$ can be analytically computed and $\lambda_{\max} = \left\|\mathbf{X}^T\mathbf{Y}\right\|_2$ which is the largest singular value (a.k.a. spectral norm) of $\mathbf{X}^T\mathbf{Y}$.

3.3. The Feasible Set of Dual Optimal Solution

In the following, we will make use of the variational inequality as in Lemma 1 to construct a feasible the dual optimal solution $\mathbf{P}^*_\lambda$.

Lemma 1. *(Güler, 2010) Let $\mathcal{G} \in \mathbb{R}^{d\times m}$ be a convex set and let f be a Gâteaux differentiable function on an open set containing $\mathcal{G}$. If $\mathbf{Z}^*$ is a local minimizer of f on $\mathcal{G}$, then*

$$\mathrm{Tr}[\nabla f(\mathbf{Z}^*)^T(\mathbf{Z}-\mathbf{Z}^*)] \geq 0,\ \forall \mathbf{Z} \in \mathcal{G} \tag{12}$$

As we can seen, to construct a feasible set for $\mathbf{Z}^*$ by Eq. (12), we need to find a known $\mathbf{Z}$ from $\mathcal{G}$. Therefore, to construct the feasible set for $\mathbf{P}^*_\lambda$ with $\lambda \in (0, \lambda_{\max})$, we assume that there exists another parameter λ_0 with $\lambda_0 \in (0, \lambda)$ and its dual solution $\mathbf{P}^*_{\lambda_0}$ is known. To make this assumption reasonable, we need to find an appropriate λ_0 such that its solution can be obtained trivially. Indeed, when λ_0 is close to zero, the solution $\mathbf{W}^*_{\lambda_0}$ can be easily obtained by using $\mathbf{W} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ as initialization, which is the solution at $\lambda = 0$. In addition, in many scenarios, the solution at λ_0 can be freely obtained. For instance, an appropriate value of λ for Eq. (2) needs to be determined since the optimal value of λ is generally unknown in real applications. Therefore, we usually need to solve Eq. (2) over a grid of regularization parameters $\lambda_1 < \lambda_2 < \cdots < \lambda_k$ and choose the optimal λ under certain criterion. After obtaining the solution $\mathbf{W}^*_{\lambda_{t-1}}$ at λ_{t-1}, it can be freely used to screen out inactive subspaces for Eq. (2) at λ_t.

Now, we describe how to construct a feasible set for the dual optimal solution $\mathbf{P}^*_\lambda$ by using the variational inequality. Since $\mathbf{P}^*_{\lambda_0}$ and $\mathbf{P}^*_\lambda$ are the solutions to Eq. (9) at λ_0 and λ, respectively, we can apply Lemma 1 to Eq. (9) and obtain

$$\mathrm{Tr}\left[\left(\mathbf{P}^*_{\lambda_0} - \frac{\mathbf{Y}}{\lambda_0}\right)^T\left(\mathbf{P}-\mathbf{P}^*_{\lambda_0}\right)\right] \geq 0 \tag{13}$$

$$\mathrm{Tr}\left[\left(\mathbf{P}^*_{\lambda} - \frac{\mathbf{Y}}{\lambda}\right)^T\left(\mathbf{P}-\mathbf{P}^*_{\lambda}\right)\right] \geq 0 \tag{14}$$

which holds for $\forall \mathbf{P} : \left\|\mathbf{X}^T\mathbf{P}\right\|_2 \leq 1$. Since $\mathbf{P} = \mathbf{P}^*_\lambda$ and $\mathbf{P} = \mathbf{P}^*_{\lambda_0}$ are also feasible for Eq. (13) and Eq. (14), respectively, substituting them into Eq. (13) and Eq. (14) leads to

$$\mathrm{Tr}\left[\left(\mathbf{P}^*_{\lambda_0} - \frac{\mathbf{Y}}{\lambda_0}\right)^T\left(\mathbf{P}^*_{\lambda}-\mathbf{P}^*_{\lambda_0}\right)\right] \geq 0 \tag{15}$$

$$\mathrm{Tr}\left[\left(\mathbf{P}^*_{\lambda} - \frac{\mathbf{Y}}{\lambda}\right)^T\left(\mathbf{P}^*_{\lambda_0}-\mathbf{P}^*_{\lambda}\right)\right] \geq 0 \tag{16}$$

From inequalities in Eq. (15) and Eq. (16), we obtain the feasible set for $\mathbf{P}^*_\lambda$

$$\mathcal{F}(\mathbf{P}^*_\lambda) = \left\{\mathbf{P} : \mathrm{Tr}\left[\left(\mathbf{P}^*_{\lambda_0} - \frac{\mathbf{Y}}{\lambda_0}\right)^T\left(\mathbf{P}-\mathbf{P}^*_{\lambda_0}\right)\right] \geq 0,\ \mathrm{Tr}\left[\left(\mathbf{P} - \frac{\mathbf{Y}}{\lambda}\right)^T\left(\mathbf{P}^*_{\lambda_0}-\mathbf{P}\right)\right] \geq 0\right\} \tag{17}$$

3.4. Estimating the Upper Bound

Given the feasible set $\mathcal{F}(\mathbf{P}^*_\lambda)$, we seek to estimate the upper bound of Eq. (11) over the feasible set for each pair of $\mathbf{u}_i$ and $\mathbf{v}_j$. Formally, we need to solve the following optimization problem

$$\max\ \left|\mathbf{u}_i^T\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{Y} - \lambda\mathbf{X}^T\mathbf{P}\right)\right)\mathbf{v}_j\right| \quad s.t.\ \mathbf{P} \in \mathcal{F}(\mathbf{P}^*_\lambda) \tag{18}$$

As mentioned before, the performance of subspace screening also relies on the choice of $\mathbf{u}_i$ and $\mathbf{v}_j$. In the proposed method, $\mathbf{u}_i$ and $\mathbf{v}_j$ are chosen as the singular vectors of $\mathbf{W}_{\lambda_0}$. Specifically, suppose the singular value decomposition (SVD) of $\mathbf{W}^*_{\lambda_0}$ is

$$\mathbf{W}^*_{\lambda_0} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{19}$$

Then, we let $\mathbf{u}_i = \mathbf{U}_{\cdot i}$ and $\mathbf{v}_j = \mathbf{V}_{\cdot j}$.

For reformulating the upper bound estimation problem in Eq. (18), we first introduce three variables

$$\mathbf{A} = \frac{\mathbf{Y}}{\lambda_0} - \mathbf{P}^*_{\lambda_0} = \frac{\mathbf{X}\mathbf{W}^*_{\lambda_0}}{\lambda_0} \tag{20}$$

$$\mathbf{B} = \frac{\mathbf{Y}}{\lambda} - \mathbf{P}^*_{\lambda_0} = \mathbf{A} + \left(\frac{\mathbf{Y}}{\lambda} - \frac{\mathbf{Y}}{\lambda_0}\right) \tag{21}$$

$$\mathbf{R} = 2\mathbf{P} - \left(\mathbf{P}^*_{\lambda_0} + \frac{\mathbf{Y}}{\lambda}\right) \tag{22}$$

where $\mathbf{A}$ can be considered as the scaled prediction based on $\mathbf{W}^*_{\lambda_0}$ by λ_0, and $\mathbf{B}$ is obtained by translating $\mathbf{A}$ with the difference between the scaled $\mathbf{Y}$ by λ_0 and λ. The following lemma shows that both $\mathbf{A}$ and $\mathbf{B}$ are nonzero matrices.

Lemma 2. *For any λ_0 and λ such that $0 < \lambda_0 < \lambda < \left\|\mathbf{X}^T\mathbf{Y}\right\|_2$, and $\mathbf{Y} \neq \mathbf{0}$, we have both $\mathbf{A} \neq \mathbf{0}$ and $\mathbf{B} \neq \mathbf{0}$.*

Next, we reformulate the upper bound problem in Eq. (18) by using the variables defined in Eq. (20), Eq. (21) and Eq. (22) and obtain the following equivalent problem

$$\max\ \frac{\lambda}{2}\left|\mathbf{u}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{B}\mathbf{v}_j - \mathbf{u}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{R}\mathbf{v}_j\right| \quad s.t.\ \mathrm{Tr}\left[\mathbf{A}^T(\mathbf{R}+\mathbf{B})\right] \leq 0, \|\mathbf{R}\|_F^2 \leq \|\mathbf{B}\|_F^2 \tag{23}$$

Let us define $\mathbf{S}_\mathrm{C} \in \mathbb{R}^{d\times m}$ and $\mathbf{S}_\mathrm{R} \in \mathbb{R}^{d\times m}$ such that

$$\mathbf{U}\mathbf{S}_\mathrm{C}\mathbf{V}^T = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{B}, \mathbf{U}\mathbf{S}_\mathrm{R}\mathbf{V}^T = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{R} \tag{24}$$

Then, the objective function in Eq. (23) can be further reformulated as

$$\frac{\lambda}{2}\left|\mathbf{u}_i^T\mathbf{U}\mathbf{S}_\mathrm{C}\mathbf{V}^T\mathbf{v}_j - \mathbf{u}_i^T\mathbf{U}\mathbf{S}_\mathrm{R}\mathbf{V}^T\mathbf{v}_j\right| = \frac{\lambda}{2}\left|(\mathbf{S}_\mathrm{C})_{ij} - (\mathbf{S}_\mathrm{R})_{ij}\right|$$

$$= \frac{\lambda}{2} \max\left((\mathbf{S}_{\mathrm{C}})_{ij} - (\mathbf{S}_{\mathrm{R}})_{ij}, -(\mathbf{S}_{\mathrm{C}})_{ij} + (\mathbf{S}_{\mathrm{R}})_{ij} \right) \quad (25)$$

which means we can solve the optimization problem by maximizing $-(\mathbf{S}_{\mathrm{R}})_{ij}$ and $(\mathbf{S}_{\mathrm{R}})_{ij}$ over the constraint set. They are further equivalent to minimizing $(\mathbf{S}_{\mathrm{R}})_{ij}$ and $-(\mathbf{S}_{\mathrm{R}})_{ij}$ over the constraint set, which can be unified as the following problem

$$\min e\,(\mathbf{S}_{\mathrm{R}})_{ij} \;\; s.t. \;\; \mathrm{Tr}\left[\mathbf{A}^T(\mathbf{R}+\mathbf{B})\right] \leq 0, \|\mathbf{R}\|_F^2 \leq \|\mathbf{B}\|_F^2 \quad (26)$$

where $e = \pm 1$. For convenience, we introduce a new matrix variable $\mathbf{D}$ defined as $\mathbf{D} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{U}$. According to Eq. (24), $(\mathbf{S}_{\mathrm{R}})_{ij}$ can be represented as $(\mathbf{S}_{\mathrm{R}})_{ij} = (\mathbf{D}_{\cdot i})^T \mathbf{R}\mathbf{V}_{\cdot j}$.

Eq. (26) should admit a closed form solution since the objective function is linear and the constrain set is the intersection of a linear and quadratic function (Bertsimas & Tsitsiklis, 1997). The following theorem provides the optimal solution for Eq. (26).

Theorem 1. *For any λ_0 and λ such that $0 < \lambda_0 < \lambda < \left\|\mathbf{X}^T\mathbf{Y}\right\|_2$, and both $\mathbf{X}$ and $\mathbf{Y}$ are not equal to $\mathbf{0}$. The optimal solution to Eq. (26) is*

$$(\mathbf{S}_R)_{ij} = -e\,\|\mathbf{D}_{\cdot i}\|_2 \|\mathbf{B}\|_F \quad (27)$$

if the following holds

$$\lambda_0 \,\mathrm{Tr}\left[\mathbf{A}^T\mathbf{B}\right] \|\mathbf{D}_{\cdot i}\|_2 \leq e\,\|\mathbf{B}\|_F \,\mathbf{\Sigma}_{ij} \quad (28)$$

otherwise

$$(\mathbf{S}_R)_{ij} = \frac{-e\mathbf{G}_{ij} - \mathrm{Tr}\left[\mathbf{A}^T\mathbf{B}\right]\mathbf{\Sigma}_{ij}}{\lambda_0 \|\mathbf{A}\|_F^2} \quad (29)$$

where $\mathbf{G}_{ij}$ is defined as

$$\sqrt{\left(\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2 - \left(\mathrm{Tr}\left[\mathbf{A}^T\mathbf{B}\right]\right)^2\right)\left(\lambda_0^2\|\mathbf{A}\|_F^2\|\mathbf{D}_{\cdot i}\|_2^2 - \mathbf{\Sigma}_{ij}^2\right)}$$

Since we have obtained the optimal value of Eq. (26), the upper bound of $\left|\mathbf{u}_i^T\mathbf{W}_\lambda^*\mathbf{v}_j\right|$ is also ready to obtain. Here, we use $\mathbf{\Phi} \in \mathbb{R}^{d\times m}$ and $\mathbf{\Psi} \in \mathbb{R}^{d\times m}$ to represent the upper bounds for all subspaces. Specifically, $\mathbf{\Phi}_{ij}$ and $\mathbf{\Psi}_{ij}$ denote the upper bounds of $-\mathbf{u}_i^T\mathbf{W}_\lambda^*\mathbf{v}_j$ and $\mathbf{u}_i^T\mathbf{W}_\lambda^*\mathbf{v}_j$, respectively. The values of $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are summarized in the following corollary.

Corollary 1. *For any λ_0 and λ such that $0 < \lambda_0 < \lambda < \left\|\mathbf{X}^T\mathbf{Y}\right\|_2$, and $\mathbf{Y} \neq \mathbf{0}$. We have*

$$\mathbf{\Phi}_{ij} = \begin{cases} 0.5\lambda\left(\|\mathbf{B}\|_F\|\mathbf{D}_{\cdot i}\|_2 - (\mathbf{S}_C)_{ij}\right) \\ \qquad \textit{if } \|\mathbf{B}\|_F\mathbf{\Sigma}_{ij} \leq -\lambda_0\,\mathrm{Tr}\left[\mathbf{A}^T\mathbf{B}\right]\|\mathbf{D}_{\cdot i}\|_2 \\ 0.5\lambda\left(\frac{\mathbf{G}_{ij}-\mathrm{Tr}\left[\mathbf{A}^T\mathbf{B}\right]\mathbf{\Sigma}_{ij}}{\lambda_0\|\mathbf{A}\|_F^2} - (\mathbf{S}_C)_{ij}\right) \\ \qquad \textit{otherwise} \end{cases}$$

$$\mathbf{\Psi}_{ij} = \begin{cases} 0.5\lambda\left(\|\mathbf{B}\|_F\|\mathbf{D}_{\cdot i}\|_2 + \mathbf{S}_{Cij}\right) \\ \qquad \textit{if } \|\mathbf{B}\|_F\mathbf{\Sigma}_{ij} \geq \lambda_0\,\mathrm{Tr}\left[\mathbf{A}^T\mathbf{B}\right]\|\mathbf{D}_{\cdot i}\|_2 \\ 0.5\lambda\left(\frac{\mathbf{G}_{ij}+\mathrm{Tr}\left[\mathbf{A}^T\mathbf{B}\right]\mathbf{\Sigma}_{ij}}{\lambda_0\|\mathbf{A}\|_F^2} + \mathbf{S}_{Cij}\right) \\ \qquad \textit{otherwise} \end{cases}$$

3.5. Safe Subspace Screening Rule

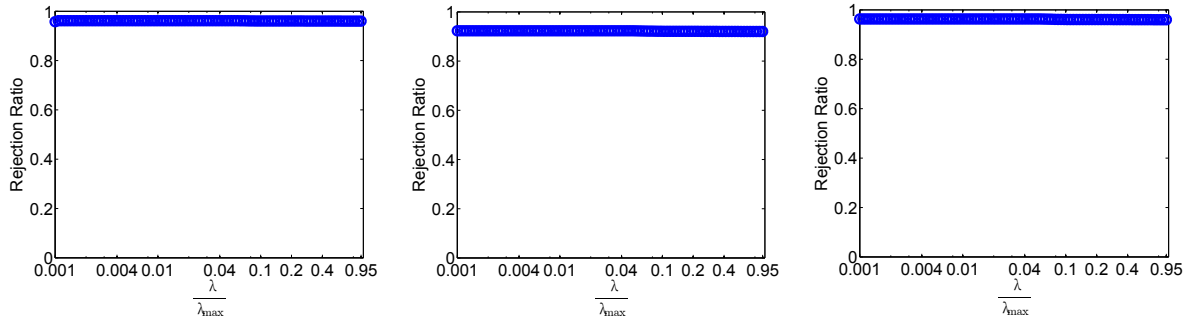
In view of Eq. (4), we are now ready to construct the safe subspace screening rule for Eq. (3). Let us introduce a new matrix $\mathbf{\Omega} \in \mathbb{R}^{d\times m}$ with its (i,j)th entry denoting the upper bound of $\left|\mathbf{u}_i^T\mathbf{W}_\lambda^*\mathbf{v}_j\right|$ meaning the value of $\mathbf{\Omega}_{ij}$ is $\max(\mathbf{\Phi}_{ij}, \mathbf{\Psi}_{ij})$. If $\mathbf{\Omega}_{ij} = 0$, it implies that both $-\mathbf{u}_i^T\mathbf{W}_\lambda^*\mathbf{v}_j$ and $\mathbf{u}_i^T\mathbf{W}_\lambda^*\mathbf{v}_j$ are equal to zero, then the value of $(\mathbf{\Theta}_\lambda^*)_{ij}$ must be zero and the subspace $\mathbf{u}_i\mathbf{v}_j^T$ can be discarded prior to solving Eq. (3). Formally, the proposed subspace screening method can be summarized in the following theorem.

Theorem 2. *For nuclear norm regularized least squares problem, suppose the solution $\mathbf{W}_{\lambda_0}^*$ is known and the SVD of $\mathbf{W}_{\lambda_0}^*$ as represented in Eq. (19). Let $\widehat{\mathbf{U}} = \mathbf{U}$, $\widehat{\mathbf{V}} = \mathbf{V}$, $\mathbf{u}_i = \mathbf{U}_{\cdot i}$ and $\mathbf{v}_j = \mathbf{V}_{\cdot j}$. For any $\lambda > \lambda_0$*
1. If $\lambda \geq \lambda_{max}$, then $\mathbf{W}_\lambda^* = \mathbf{0}$.
2. If $\lambda < \lambda_{max}$, for $1 \leq i \leq d$, if $\|\mathbf{\Omega}_{i\cdot}\|_\infty = 0$, then $(\mathbf{\Theta}_\lambda^*)_{i\cdot} = \mathbf{0}$ and $\widehat{\mathbf{U}}_{\cdot i}$ can be removed from $\widehat{\mathbf{U}}$. Similarly, for $1 \leq j \leq m$, if $\|\mathbf{\Omega}_{\cdot j}\|_\infty = 0$, then $(\mathbf{\Theta}_\lambda^*)_{\cdot j} = \mathbf{0}$ and $\widehat{\mathbf{V}}_{\cdot j}$ can be removed from $\widehat{\mathbf{V}}$. Then, solving Eq. (8) will get the identical result as optimizing Eq. (3).

4. Experiments

In this section, we perform experiments on several synthetic and real data sets to evaluate the performance of the proposed SSS. Since there is no existing method on safe subspace screening prior to solving the problem, we evaluate the proposed SSS by comparing the performance of the nuclear norm solver with SSS and without SSS. For the nuclear norm solver, we use the popular accelerated proximal gradient (APG) algorithm (Toh & Yuan, 2010; Ji & Ye, 2009). On each data set, we run the solver without and with SSS to optimize Eq. (2) along a sequence of 100 values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}$ from 0.001 to 0.95. To reduce statistical variability, all reported results are averaged over 10 trials. All experiments are performed on a workstation with Intel(R) Core(TM) i7-4930K 3.40 GHz CPU and 64G RAM

Suppose the 100 values of λ are indexed by $\lambda_t, 1 \leq t \leq 100$ in ascending order of value. In our experiments, the warm-start strategy is used for the solver. Specifically, for solving the optimization problem at λ_t with $t \geq 2$, the solution $\mathbf{W}_{\lambda_{t-1}}^*$ at λ_{t-1} will be used as the initialization. To solve the problem for the smallest regularization parameter



(a) Synthetic Set 1 (b) Synthetic Set 2 (c) Synthetic Set 3

Figure 1. The subspace rejection ratio of the proposed SSS on three synthetic data sets.

λ_1, we use the solution at $\lambda = 0$ that is $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ as initialization. In order to apply the proposed SSS for λ_1, we first solve the problem for a very small regularization parameter $\lambda_0 = (1e-6)\lambda_{\max}$ by using $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ as initialization.

Since the proposed SSS is safe, the solution obtained by the solver with SSS is the same as the solution directly from the solver. In other words, their predictive performances are the same to each other. To quantify the performance of the proposed method, similarly to (Wang et al., 2013), two measures are used in our experiments: (a) *subspace rejection ratio*: the ratio of the number of subspaces discarded by the proposed SSS to the total number of subspaces that can be safely discarded in the ground truth. More precisely, suppose the rank of ground truth is r, by using the notation in Sec. 2, we have

$$\text{subspace rejection ratio} = \frac{d \times m - \widehat{d} \times \widehat{m}}{d \times m - r^2}$$

(b) *speedup*: this value is the ratio of the computational time of the solver without the proposed SSS to the computational time of the solver with the proposed SSS.

4.1. Synthetic Data Sets

In this subsection, we evaluate the proposed method in the problem of multivariate linear regression on three synthetic data sets. Suppose the input $\mathbf{X} \in \mathbb{R}^{n\times d}$ is n samples with d-dimensional features for each and the output $\mathbf{Y} \in \mathbb{R}^{n\times m}$ is m responses for all samples, then it can be formulated as

$$\mathbf{Y} = \mathbf{X}\mathbf{W}^* + \mathbf{E}$$

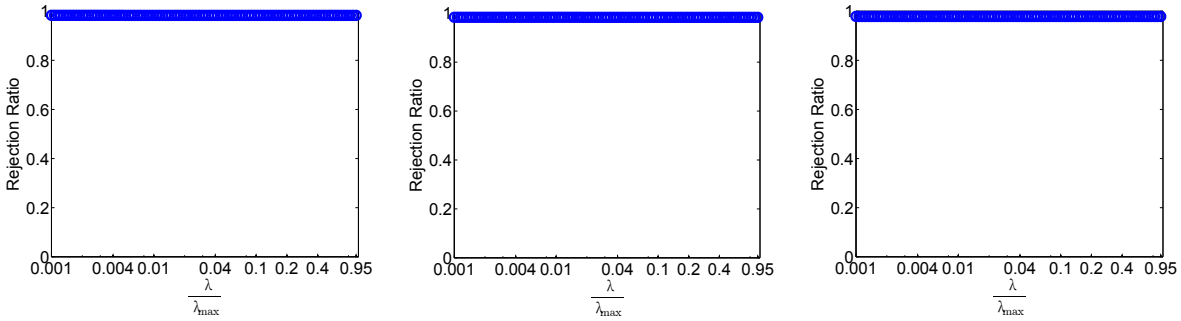
where $\mathbf{W}^* \in \mathbb{R}^{d\times m}$ is the model coefficient matrix and $\mathbf{E} \in \mathbb{R}^{n\times m}$ is the regression noise. To generate the synthetic data sets, we use a similar procedure as reported in (Jacob et al., 2008). Specifically, the ith observation is generated from a multivariate normal distribution $\mathbf{X}_{i\cdot} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the output of the jth response is obtained by $\mathbf{Y}_{ij} = \mathbf{X}_{i\cdot}\mathbf{W}_{\cdot j} + \mathcal{N}(0, 16)$. 200 samples are generated for each data set.

Table 1. Computational time (in minutes) for solving nuclear norm regularized least squares problem along a sequence of 100 parameter values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}$ from 0.001 to 0.95 on the three synthetic data sets by (a) "Solver" (solver without subspace screening); (b) "Solver with SSS" (solver in conjunction with the proposed SSS). "Prep." is the running time for solving the problem at λ_0. "SSS" is the total computational time used to perform the proposed subspace screening.

Data Set		Set 1	Set 2	Set 3
Solver		659.12	212.79	182.38
Solver with SSS	Prep.	2.27	0.60	0.64
	SSS	13.39	4.37	8.63
	Total	28.81	19.73	12.86
Speedup (times)		22.88	10.78	14.18

In data set 1, all $m = 5000$ models are assumed from 100 clusters each consisting of 50 models. All $d = 5000$ dimensions are randomly divided into 100 disjoint groups and each group is assigned to only one cluster. The coefficients for each model from a particular cluster are nonzero only for corresponding dimensions, and are zero for all other dimensions. For each cluster, a specific model coefficient is the cluster mean plus a model specific component $\mathcal{N}(\mathbf{0}, 4\mathbf{I})$. Data set 2 and data set 3 are the same as data set 1 except we change $d = 2500$ and $m = 2500$ for data set 2 and data set 3, respectively.

Fig. 1 shows the subspace rejection ratio of the proposed SSS on the three synthetic data sets. As observed, the proposed method consistently discards more than 90% inactive subspaces on all three data sets. Table 1 reports the computational time of the solver without or with the proposed SSS for solving the 100 nuclear norm regularized least squares problems, as well as the computational time used to perform the proposed SSS. Since most inactive subspaces have been screened out prior to solving the problem, the proposed SSS significantly improves the effi-



(a) PIE (b) MNIST (c) Yahoo Stock

Figure 2. The subspace rejection ratio of the proposed SSS on three real data sets.

ciency of the solver. The lowest speedup achieved by the proposed SSS on the three data sets is still up to 10.78. Moreover, as shown in the table, more significant improvement can be achieved for larger problem size. Especially, on the synthetic data set 1, the size of matrix variable is 5000×5000 and the solver spends 659.12 minutes to solve the 100 problems. In contrast, by enhancing the solver with the proposed SSS, only 28.81 minutes is used for the 100 problems, which leads to substantial saving in the computational time. The proposed SSS is not only effective in identifying inactive subsapces as shown in Fig. 1, but also efficient. As observed in Table 1, on the three data sets, the computational times of performing the proposed SSS are only 2.03%, 2.05% and 4.73% that of the solver without subspace screening. In addition, compared with the computational time of solver without subspace screening, the preparation procedure of the proposed SSS is also very efficient since it only occupies 0.34%, 0.28% and 0.35% on the three data sets, respectively.

4.2. Real Data Sets

In this subsection, we perform experiments on three real data sets to evaluate the performance of the proposed SSS. The details of the three data sets as follows.

PIE Face Image Data Set This data set used in this experiment consist of 11554 gray face images from 68 people, which were captured under various poses, illumination conditions and expressions (Sim et al., 2003; Cai et al., 2007). The size of each image is 32×32 pixels. We consider the subspace clustering task on it. Specifically, in each trial, we first randomly pick 70 images from each people and put them together as the dictionary $\mathbf{X}$. Then, another 70 images are picked from each people used as the target clustering subspace $\mathbf{Y}$. The feature dimension is reduced to 80 by performing PCA on the vectorized raw features. Then, then we get the dictionary $\mathbf{X} \in \mathbb{R}^{80\times 4760}$ and targeted clustering subspace $\mathbf{Y} \in \mathbb{R}^{80\times 4760}$. Therefore, we have $\mathbf{W} \in \mathbb{R}^{4760\times 4760}$.

MNIST Handwritten Digit Data Set This data set consists of $70,000$ grey images of scanned handwritten digits (LeCun et al., 1998). The sample sizes of training and testing are $60,000$ and $10,000$ respectively. We still consider a subspace clustering task. Specifically, in each trial, we randomly pick 600 images from training and testing for each digit to form the dictionary $\mathbf{X}$ and the target clustering subspace $\mathbf{Y}$, respectively. The feature dimension is reduced to 100 by performing PCA on the vectorized raw features. Finally, we obtain a dictionary $\mathbf{X} \in \mathbb{R}^{100\times 6000}$ and a target clustering subspace $\mathbf{Y} \in \mathbb{R}^{100\times 6000}$. Then, the problem is to learn $\mathbf{W} \in \mathbb{R}^{6000\times 6000}$.

Yahoo Stock Data In this data set, we consider the application of multivariate linear regression on the financial econometrics. Specifically, we aim to predict the future return of stock via multivariate linear regression by using the daily closing price. Let $\mathbf{y}_{t-1} \in \mathbb{R}^d$ and $\mathbf{y}_t \in \mathbb{R}^d$ denote the stock prices at day $(t-1)$ and t, respectively. Then, the problem can be formulated as $\mathbf{y}_t^T = \mathbf{y}_{t-1}^T \mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{d\times d}$ and we have $d = m$ in this case. To perform the experiment, in each trial, we download the daily closing prices for $m = 4676$ stocks during 101 days in 2013 form Yahoo Fiance. Then $\mathbf{X}$ and $\mathbf{Y}$ are formed as $\mathbf{X} = [\mathbf{y}_1 \cdots \mathbf{y}_{100}]^T \in \mathbb{R}^{100\times 4676}$ and $\mathbf{X} = [\mathbf{y}_2 \cdots \mathbf{y}_{101}]^T \in \mathbb{R}^{100\times 4676}$, which implies $\mathbf{W} \in \mathbb{R}^{4676\times 4676}$.

The subspace rejection ratios of the proposed SSS on the three real data sets are shown in Fig. 2. As observed, the proposed method is very effective on screening out inactive subspaces on real data sets in the sense that it successfully identifies more than 97% inactive subspaces on all three real data sets. As can be seen in Table 2, compared with results on synthetic data sets, the proposed SSS achieves better performance on real data sets in terms of speedup. Specifically, even the lowest speedup is up to 53.57 on the MNIST data set while it achieves around 80 speedup on both other two data sets. In addition, the computational time of performing the proposed SSS and running the preparation procedure are also much less than that of synthetic data

Table 2. Computational time (in minutes) for solving nuclear norm regularized least squares problem along a sequence of 100 parameter values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}$ from 0.001 to 0.95 on the three real data sets by (a) "Solver" (solver without subspace screening); (b) "Solver with SSS" (solver in conjunction with the proposed SSS). "Prep." is the running time for solving the problem at λ_0. "SSS" is the total computational time used to perform the proposed subspace screening.

Data Set		PIE	MNIST	Yahoo Stock
Solver		2395.54	2968.87	3075.09
Solver with SSS	Prep.	1.88	3.69	2.24
	SSS	11.21	22.04	10.93
	Total	31.22	55.42	37.26
Speedup (times)		76.72	53.57	82.53

sets. In particular, the percentage of computational time of the preparation procedure over that of the solver without subspace screening is 0.08%, 0.12% and 0.07% on three real data sets, respectively. Thus the time for preparation is quite negligible. Moreover, the largest value of percentage of performing the proposed SSS is 0.74% which shows that the proposed SSS is very efficient. One reason for the better performance of the proposed SSS on real data sets is that they are generally more complicated thus requiring more time for the solver to convergence. On the other hand, the proposed SSS only goes through the data once, whose computational time depends solely on the size of the matrix variable.

4.3. Comparison on Forward and Backward Solution Paths for the Solver

As mentioned at the beginning of this section, we can make use of the warm-start strategy to efficiently obtain the solutions for a sequence of value of λ. In our experiment, for a given $\lambda_t, 1 \leq t \leq 100$ in ascending ordering of value, we obtain the solution path by solving the problem from λ_1 to λ_{100}. We call this method as a forward solution path for solver. In contrast, there is an alternative method called backward solution path method, in which we solve the problem from λ_{100} to λ_1. In this method, we can only use $\mathbf{0}$ that is the solution of $\lambda_{\max}$ as initialization for λ_{100}. Intuitively, there is no clear theoretical proof as of which one is more efficient since the result should depend on the choice of λ_t. Here, we experimentally compare the performances of forward and backward solution paths. Specifically, we run the solver on the three synthetic data sets by using both the forward and backward methods and compare their computational time. The results are reported in Table 3 and they are averaged over 10 trials. As observed, the computational time of two paths on synthetic set 1 and set 3 are almost the same to each other, and the forward path is a little faster than the backward path.

Table 3. Computational time (in minutes) of forward and backward solution path for the solver on three synthetic data sets.

Data Set	Set 1	Set 2	Set 3
Forward	659.12	212.79	182.38
Backward	660.16	230.12	185.22

Table 4. Computational time (in minutes) for solving nuclear norm regularized least squares problem along a sequence of 100 parameter values of λ equally spaced on the logarithmic scale of $\lambda/\lambda_{\max}$ from 0.001 to 0.95 on the three synthetic data sets by (a) "ADMM" (ADMM without subspace screening); (b) "ADMM with SSS" (ADMM in conjunction with the proposed SSS). "Prep." is the running time for solving the problem at λ_0. "SSS" is the total computational time used to perform the proposed subspace screening.

Data Set		Set 1	Set 2	Set 3
ADMM		590.63	221.12	227.68
ADMM with SSS	Prep.	1.97	0.60	0.63
	SSS	13.40	4.43	9.74
	Total	26.05	16.93	14.58
Speedup (times)		22.67	13.06	15.61

4.4. Results of Safe Subspace Screening for ADMM

Further as we mentioned before, the proposed SSS can be used in conjunction with any nuclear norm solver. In this subsection, we evaluate the performance of the proposed SSS for another popular nuclear norm solver, i.e. the alternating direction method of multipliers (ADMM) (Boyd et al., 2011). Specifically, we perform experiments on the three synthetic data sets with the same setting as previous experiments except using ADMM as the solver here. The results are shown in Table 4. Compared with Table 1, the proposed SSS has shown similar improvements for ADMM as APG. This shows that the proposed SSS can extensively used to improve the efficiency of existing nuclear norm solvers.

5. Conclusions

In this work, we present a safe subspace screening method to improve the efficiency of the solver for nuclear norm regularized least squares problems. Essentially, the idea of subspace screening is to identify the subspaces that are orthogonal to the solution by using the convex optimization methods. The proposed method is able to effectively and efficiently discard inactive subspaces prior to solving the problem, thus greatly reducing the size of the optimization problem. Moreover, the proposed method can be used in conjunction with any nuclear norm solver since the it is independent of solver. Extensive experiments on three synthetic and three real data sets have shown that the proposed method significantly improves the efficiency of existing solvers.

References

Argyriou, Andreas, Evgeniou, Theodoros, and Pontil, Massimiliano. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

Avron, Haim, Kale, Satyen, Kasiviswanathan, Shiva Prasad, and Sindhwani, Vikas. Efficient and practical stochastic subgradient descent for nuclear norm regularization. In *Proc. ICML*, 2012.

Bertsimas, Dimitris and Tsitsiklis, John. *Introduction to Linear Optimization*. Athena Scientific Belmont, MA, 1997.

Boyd, Stephen P., Parikh, Neal, Chu, Eric, Peleato, Borja, and Eckstein, Jonathan. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Cai, Deng, He, Xiaofei, and Han, Jiawei. Efficient kernel discriminant analysis via spectral regression. In *Proc. ICDM*, 2007.

Fan, Jianqing and Lv, Jinchi. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70(5):849–911, 2008.

Fan, Jianqing and Song, Rui. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

Favaro, Paolo, Vidal, René, and Ravichandran, Avinash. A closed form solution to robust subspace estimation and clustering. In *Proc. CVPR*, 2011.

Ghaoui, Laurent El, Viallon, Vivian, and Rabbani, Tarek. Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization*, 8(4):667–698, 2012.

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

Hsieh, Cho-Jui and Olsen, Peder A. Nuclear norm minimization via active subspace selection. In *Proc. ICML*, 2014.

Jacob, Laurent, Bach, Francis, and Vert, Jean-Philippe. Clustered multi-task learning: A convex formulation. In *Proc. NIPS*, 2008.

Jaggi, Martin and Sulovský, Marek. A simple algorithm for nuclear norm regularized problems. In *Proc. ICML*, 2010.

Ji, Shuiwang and Ye, Jieping. An accelerated gradient method for trace norm minimization. In *Proc. ICML*, 2009.

Kang, Zhuoliang, Grauman, Kristen, and Sha, Fei. Learning with whom to share in multi-task feature learning. In *Proc. ICML*, 2011.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Liu, Jun, Zhao, Zheng, Wang, Jie, and Ye, Jieping. Safe screening with variational inequalities and its application to lasso. In *Proc. ICML*, 2014.

Lu, Zhaosong, Monteiro, Renato, and Yuan, Ming. Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Mathematical Programming*, 131(1-2):163–194, 2012.

Mazumder, Rahul, Hastie, Trevor, and Tibshirani, Robert. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

Mishra, Bamdev, Meyer, Gilles, Bach, Francis, and Sepulchre, Rodolphe. Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4):2124–2149, 2013.

Ogawa, Kohei, Suzuki, Yoshiki, and Takeuchi, Ichiro. Safe screening of non-support vectors in pathwise svm computation. In *Proc. ICML*, 2013.

Shalev-Shwartz, Shai, Gonen, Alon, and Shamir, Ohad. Large-scale convex minimization with a low-rank constraint. In *Proc. ICML*, 2011.

Sim, Terence, Baker, Simon, and Bsat, Maan. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 25(12):1615–1618, 2003.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Tibshirani, Robert, Bien, Jacob, Friedman, Jerome, Hastie, Trevor, Simon, Noah, Taylor, Jonathan, and Tibshirani, Ryan J. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B*, 74(2):245–266, 2012.

Toh, Kim-Chuan and Yuan, Sangwoon. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(15):615–640, 2010.

Wang, Jie and Ye, Jieping. Two-layer feature reduction for sparse-group lasso via decomposition of convex sets. In *Proc. NIPS*, 2014.

Wang, Jie, Zhou, Jiayu, Wonka, Peter, and Ye, Jieping. Lasso screening rules via dual polytope projection. In *Proc. NIPS*, 2013.

Wang, Jie, Wonka, Peter, and Ye, Jieping. Scaling svm and least absolute deviations via exact data reduction. In *Proc. ICML*, 2014a.

Wang, Jie, Zhou, Jiayu, Liu, Jun, Wonka, Peter, and Ye, Jieping. A safe screening rule for sparse logistic regression. In *Proc. NIPS*, 2014b.

Watson, G. Alistair. Characterization of the subdifferential of some matrix norms. *Linear Algebra and Its Applications*, 170: 33–45, 1992.

Xiang, Zhen James, Xu, Hao, and Ramadge, Peter J. Learning sparse representations of high dimensional data on large scale dictionaries. In *Proc. NIPS*, 2011.

Yuan, Ming, Ekici, Ali, Lu, Zhaosong, and Monteiro, Renato. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B*, 69(3):329–346, 2007.