Sparse trace norm regularization

Jianhui Chen · Jieping Ye

Abstract We study the problem of estimating multiple predictive functions from a dictionary of basis functions in the nonparametric regression setting. Our estimation scheme assumes that each predictive function can be estimated in the form of a linear combination of the basis functions. By assuming that the coefficient matrix admits a sparse low-rank structure, we formulate the function estimation problem as a convex program regularized by the trace norm and the $\ell_1$-norm simultaneously. We propose to solve the convex program using the accelerated gradient (AG) method; we also develop efficient algorithms to solve the key components in AG. In addition, we conduct theoretical analysis on the proposed function estimation scheme: we derive a key property of the optimal solution to the convex program; based on an assumption on the basis functions, we establish a performance bound of the proposed function estimation scheme (via the composite regularization). Simulation studies demonstrate the effectiveness and efficiency of the proposed algorithms.

Keywords Low Rank · Regression · Gradient method · Performance bound

1 Introduction

We study the problem of estimating multiple predictive functions from noisy observations. Such a problem has received broad attention in many areas of statistics and
machine learning (Bunea et al. 2006; Huang et al. 2011; Lounici et al. 2009; Negahban and Wainwright 2011). This line of work can be roughly divided into two categories: parametric estimation and non-parametric estimation; a common and important theme for both categories is the appropriate assumption of the structure in the model parameters (parametric setting) or the coefficients of the dictionary (nonparametric setting).

There has been an enormous amount of literature on effective function estimation based on different sparsity constraints, including the estimation of the sparse linear regression via $\ell_1$-norm penalty (Bickel et al. 2009; Bunea et al. 2006; Tibshirani 1996; Zhang 2009), and the estimation of the linear regression functions using group lasso estimator (Huang et al. 2011; Lounici et al. 2009). More recently, trace norm regularization has become a popular tool for approximating a set of linear models and the associated low-rank matrices in the high-dimensional setting (Negahban and Wainwright 2011; Rohde and Tsybakov 2011); the trace norm is the tightest convex surrogate (Fazel et al. 2001) for the (non-convex) rank function under certain conditions, encouraging the sparsity in the singular values of the matrix of interest. One limitation of the use of trace norm regularization is that the resulting model is dense in general. However, in many real-world applications (Pati and Kailath 1994), the underlying structure of multiple predictive functions may be sparse as well as low-rank; the sparsity leads to explicitly interpretable prediction models and the low rank implies essential subspace structure information. Similarly, the $\ell_1$-norm is the tightest convex surrogate for the non-convex cardinality function (Boyd and Vandenberghe 2004), encouraging the sparsity in the entries of the matrix. This motivates us to explore the use of the combination of the trace norm and the $\ell_1$-norm as a composite regularization (called sparse trace norm regularization) to induce the desirable sparse low-rank structure.

Trace norm regularization (minimization) has been investigated extensively in recent years. Efficient algorithms have been developed for solving convex programs with trace norm regularization (Toh and Yun 2010; Fazel et al. 2001); sufficient conditions for exact recovery from trace norm minimization have been established in Recht et al. (2010); consistency of trace norm minimization has been studied in Bach (2008); trace norm minimization has been applied for matrix completion (Candès and Recht 2009) and collaborative filtering (Srebro et al. 2005; Rennie and Srebro 2005). Similarly, $\ell_1$-norm regularization has been well studied in the literature, such as the efficient algorithms for convex optimization (Efron et al. 2004; Friedman et al. 2007; Toh and Yun 2010), theoretical guarantee of the performance (Candès and Tao 2005; Zhang 2009), and model selection consistency (Zhao and Yu 2006).

In this paper, we focus on estimating multiple predictive functions simultaneously from a finite dictionary of basis functions in the nonparametric regression setting. Our function estimation scheme assumes that each predictive function can be approximated using a linear combination of those basis functions. By assuming that the coefficient matrix of the basis functions admits a sparse low-rank structure, we formulate the function estimation problem as a convex formulation, in which the combination of the trace norm and the $\ell_1$-norm is employed as a composite regularization to induce a sparse low-rank structure in the coefficient matrix. The simultaneous sparse and low-rank structure is different from the incoherent sparse and low-rank structures studied in Candès et al. (2009) and Chandrasekaran et al. (2011). We propose to solve the
function estimation problem using the accelerated gradient method; we also develop efficient algorithms to solve the key components involved in the proposed method. We conduct theoretical analysis on the proposed convex formulation: we first present some basic properties of the optimal solution to the convex formulation (Lemma 2); we then present an assumption associated with the geometric nature of the basis functions over the prescribed observations; based on such an assumption, we derive a performance bound for the combined regularization for function estimation (Theorem 2). We conduct simulations on benchmark data to demonstrate the effectiveness and efficiency of the proposed algorithms.

**Notation** Denote $\mathbb{N}_n = \{1, \ldots, n\}$. For any matrix $\Theta$, denote its trace norm by $\|\Theta\|_*$, i.e., the sum of the singular values; denote its operator norm by $\|\Theta\|_2$, i.e., the largest singular value; denote its $\ell_1$-norm by $\|\Theta\|_1$, i.e., the sum of absolute value of all entries.

### 2 Problem formulation

Let $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R}^k$ be a set of prescribed sample pairs (fixed design) associated with $k$ unknown functions $\{f_1, \ldots, f_k\}$ as

$$y_{ij} = f_j(x_i) + w_{ij}, \quad i \in \mathbb{N}_n, \ j \in \mathbb{N}_k, \quad (1)$$

where $f_j : \mathbb{R}^d \to \mathbb{R}$ is an unknown regression function, $y_{ij}$ denotes the $j$th entry of the response vector $y_i \in \mathbb{R}^k$, and $w_{ij} \sim \mathcal{N}(0, \sigma_w^2)$ is a stochastic noise variable. Let $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times d}$, $Y = [y_1, \ldots, y_n]^T \in \mathbb{R}^{n \times k}$, and $W = (w_{ij})_{i,j} \in \mathbb{R}^{n \times k}$. Denoting

$$\mathcal{F} = (f_j(x_i))_{i,j} \in \mathbb{R}^{n \times k}, \quad i \in \mathbb{N}_n, \ j \in \mathbb{N}_k, \quad (2)$$

we can rewrite Eq. (1) in a compact form as $Y = \mathcal{F} + W$.

Let $\{g_1, \ldots, g_h\}$ be a set of $h$ pre-specified basis functions as $g_i : \mathbb{R}^d \to \mathbb{R}$, and let $\Theta = [\theta_1, \ldots, \theta_k] \in \mathbb{R}^{h \times k}$ be the coefficient matrix. We define

$$\hat{g}_j(x) = \sum_{i=1}^h \theta_{ij} g_i(x), \quad j \in \mathbb{N}_k, \quad (3)$$

where $\theta_{ij}$ denotes the $i$th entry in the vector $\theta_j$. Note that in practice the basis functions $\{g_i\}$ can be estimators from different methods, or different values of the tuning parameters of the same method.

We consider the problem of estimating the unknown functions $\{f_1, \ldots, f_k\}$ using the composite functions $\{\hat{g}_1, \ldots, \hat{g}_k\}$ defined in Eq. (3), respectively. Denote

$$\mathcal{G}_X = (g_j(x_i))_{i,j} \in \mathbb{R}^{n \times h}, \quad i \in \mathbb{N}_n, \ j \in \mathbb{N}_h, \quad (4)$$

and define the empirical error as
\[ \hat{S}(\Theta) = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} (\hat{g}_j(x_i) - y_{ij})^2 = \frac{1}{N} \| G\Theta - Y \|_F^2, \tag{5} \]

where \( N = n \times k \). Our goal is to estimate the model parameter \( \Theta \) of a sparse low-rank structure from the given \( n \) sample pairs \( \{(x_i, y_i)\}_{i=1}^{n} \). Such a structure induces the sparsity and the low rank simultaneously in a single matrix of interest.

Given that the functions \( \{f_1, \ldots, f_k\} \) are coupled via \( \Theta \) in some coherent sparse and low-rank structure, we propose to estimate \( \Theta \) as

\[ \hat{\Theta} = \arg \min_{\Theta} \left( \hat{S}(\Theta) + \alpha \| \Theta \|_* + \beta \| \Theta \|_1 \right), \tag{6} \]

where \( \alpha \) and \( \beta \) are regularization parameters (estimated via cross-validation), and the linear combination of \( \| \Theta \|_* \) and \( \| \Theta \|_1 \) is used to induce the sparse low-rank structure in \( \Theta \). The optimization problem in Eq. (6) is non-smooth convex and hence admits a globally optimal solution; it can be solved using many sophisticated optimization techniques (Toh et al. 1999; Fazel et al. 2001); in Sect. 3, we propose to apply the accelerated gradient method (Nesterov 2004) to solve the optimization problem in Eq. (6).

### 3 Optimization algorithm

In this section, we consider to apply the accelerated gradient (AG) algorithm (Beck and Teboulle 2009; Nesterov 2004, 2013) to solve the (non-smooth and convex) optimization problem in Eq. (6). We also develop efficient algorithms to solve the key components involved in AG.

#### 3.1 Accelerated gradient algorithm

The AG algorithm has attracted extensive attention in the machine learning community due to its optimal convergence rate among all first order techniques and its ability of dealing with large scale data. The general scheme in AG is standard (Nesterov 2004, 2013) and its main steps for solving Eq. (6) can be described as below: at the \( k \)th iteration, the intermediate (feasible) solution \( \Theta_k \) can be obtained via

\[ \Theta_k = \arg \min_{\Theta} \left( \frac{\gamma_k}{2} \left\| \Theta - \left( \Phi_k - \frac{1}{\gamma_k} \nabla \hat{S}(\Phi_k) \right) \right\|_F^2 + \alpha \| \Theta \|_* + \beta \| \Theta \|_1 \right), \tag{7} \]

where \( \Phi_k \) denotes a searching point constructed on the intermediate solutions from previous iterations, \( \nabla \hat{S}(\Phi_k) \) denotes the derivative of the loss function in Eq. (5) at \( \Phi_k \), and \( \gamma_k \) specifies the step size which can be determined by iterative increment until the condition

\[ \hat{S}(\Theta_k) \leq \hat{S}(\Phi_k) + (\nabla \hat{S}(\Phi_k), \Theta_k - \Phi_k) + \frac{\gamma_k}{2} \| \Theta_k - \Phi_k \|_F^2, \]

 Springer
is satisfied. The operation in Eq. (7) is commonly referred to as proximal operator (Moreau 1965), and its efficient computation is critical for the practical convergence of the AG-type algorithm. Next we present an efficient alternating optimization procedure to solve Eq. (7) with a given $\gamma_k$.

3.1.1 Dual formulation

The problem in Eq. (7) is not easy to solve directly; next we show that this problem can be efficiently solved in its dual form. By reformulating $\|\Theta\|_*$ and $\|\Theta\|_1$ into the equivalent dual forms, we convert Eq. (7) into a max-min formulation as

$$
\begin{align*}
\max_{L, S} & \quad \min_{\Theta} \|\Theta - \hat{\Phi}\|_F^2 + \hat{\alpha} \langle L, \Theta \rangle + \hat{\beta} \langle S, \Theta \rangle \\
\text{subject to} & \quad \|L\|_2 \leq 1, \quad \|S\|_{\infty} \leq 1,
\end{align*}
$$

(8)

where $\hat{\Phi} = \Phi_k - \nabla \hat{S}(\Phi_k)/\gamma_k$, $\hat{\alpha} = 2\alpha/\gamma_k$, and $\hat{\beta} = 2\beta/\gamma_k$. It can be verified that in Eq. (8) the Slater condition is satisfied and strong duality holds (Boyd and Vandenberghe 2004). Also the optimal $\Theta$ can be expressed as a function of $L$ and $S$ given by

$$
\Theta = \hat{\Phi} - \frac{1}{2}(\hat{\alpha}L + \hat{\beta}S).
$$

(9)

By substituting Eqs. (9) into (8), we obtain the dual form of Eq. (7) as

$$
\begin{align*}
\min_{L, S} & \quad \|\hat{\alpha}L + \hat{\beta}S - 2\hat{\Phi}\|_F^2 \\
\text{subject to} & \quad \|L\|_2 \leq 1, \quad \|S\|_{\infty} \leq 1.
\end{align*}
$$

(10)

3.1.2 Alternating optimization

The optimization problem in Eq. (10) is smooth convex and it has two optimization variables. For such type of problems, coordinate descent (CD) method is routinely used to compute its globally optimal solution (Grippo and Sciandrone 2000). To solve Eq. (10), the CD method alternatively optimizes one of the two variables with the other variable fixed. Our analysis below shows that the variables $L$ and $S$ in Eq. (10) can be optimized efficiently. Note that the convergence rate of the CD method is not known; however, it converges very fast in practice ($<10$ iterations in our experiments).

**Optimization of $L$** For a given $S$, the variable $L$ can be optimized via solving the following problem:

$$
\begin{align*}
\min_{L} & \quad \|L - \hat{L}\|_F^2 \\
\text{subject to} & \quad \|L\|_2 \leq 1,
\end{align*}
$$

(11)

where $\hat{L} = (2\hat{\Phi} - \hat{\beta}S)/\hat{\alpha}$. The optimization on $L$ above can be interpreted as computing an optimal projection of a given matrix over a unit spectral norm ball. Our analysis
shows that the optimal solution to Eq. (11) can be expressed in an analytic form as summarized in the following theorem.

**Theorem 1** For arbitrary $\hat{L} \in \mathbb{R}^{h \times k}$ in Eq. (11), denote its SVD by $\hat{L} = U \Sigma V^T$, where $r = \text{rank}(\hat{L})$, $U \in \mathbb{R}^{h \times r}$, $V \in \mathbb{R}^{k \times r}$, and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r \times r}$. Let $\hat{\sigma}_i^* = \min(\sigma_i, 1)$, $i = 1, \ldots, r$. Then the optimal solution to Eq. (11) is given by

$$L^* = U \hat{\Sigma} V^T, \quad \hat{\Sigma} = \text{diag}(\hat{\sigma}_1^*, \ldots, \hat{\sigma}_r^*).$$

**Proof** Assume the existence of a set of left and right singular vector pairs shared by the optimal $L^*$ to Eq. (11) and the given $\hat{L}$ for their non-zero singular values. Under such an assumption, it can be verified that the singular values of $L^*$ can be obtained via

$$\min_{\hat{\sigma}_i} \ (\hat{\sigma}_i - \sigma_i)^2$$

subject to $0 \leq \hat{\sigma}_i \leq 1$, $i = 1, \ldots, r$,

from which the optimal solution is given by $\hat{\sigma}_i^* = \min(\sigma_i, 1)$ ($\forall i$); hence the expression of $L^*$ coincides with Eq. (12). Therefore, all that remains is to show that our assumption (on the left and right singular vector pairs of $L^*$ and $\hat{L}$) holds.

Denote the Lagrangian associated with the problem in Eq. (11) as

$$h(L, \lambda) = \|L - \hat{L}\|_F^2 + \lambda (\|L\|_2 - 1),$$

where $\lambda$ denotes the dual variable. Since $0$ is strictly feasible in Eq. (11), namely, $\|0\|_2 < 1$, strong duality holds for Eq. (11). Let $\lambda^*$ be the optimal dual variable to Eq. (11). Therefore we have $L^* = \arg \min_L h(L, \lambda^*)$. It is well known that $L^*$ minimizes $h(L, \lambda^*)$ if and only if $0$ is a subgradient of $h(L, \lambda^*)$ at $L^*$, i.e.,

$$0 \in 2(L^* - \hat{L}) + \lambda^* \partial \|L^*\|_2.$$

For any matrix $Z$, the subdifferential of $\|Z\|_2$ is given by (Watson 1992)

$$\partial \|Z\|_2 = \text{conv} \left\{u_z v_z^T : \|u_z\| = \|v_z\| = 1, Z v_z = \|Z\|_2 u_z \right\},$$

where $\text{conv}\{c\}$ denotes the convex hull of the set $c$. Specifically, any element of $\partial \|Z\|_2$ has the form

$$\sum_i \alpha_i u_{zi} v_{zi}^T, \quad \alpha_i \geq 0, \quad \sum_i \alpha_i = 1,$$

where $u_{zi}$ and $v_{zi}$ are any left and right singular vectors of $Z$ corresponding to its largest singular value (the top singular values may share a common value). From Eq. (13) and
Sparse trace norm regularization

the definition of \( \partial \| Z \|_2 \), there exist \( \{ \hat{\alpha}_i \} \) such that \( \hat{\alpha}_i > 0 \), \( \sum_i \hat{\alpha}_i = 1 \), \( \sum_i \hat{\alpha}_i u_{li} v_{li}^T \in \partial \| L^* \|_2 \), and

\[
\hat{L} = L^* + \frac{\lambda^*}{2} \sum_i \hat{\alpha}_i u_{li} v_{li}^T, \tag{14}
\]

where \( u_{li} \) and \( v_{li}^T \) correspond to any left and right singular vectors of \( L^* \) corresponding to its largest singular value. Since \( \hat{\alpha}_i > 0 \), Eq. (14) verifies the existence of a set of left and right singular vector pairs shared by \( L^* \) and \( \hat{L} \). This completes the proof. \( \square \)

The optimization variable \( L \) in Eq. (11) has the size of \( h \) by \( k \), where \( h \) denotes the number of pre-specified basis functions \( \{ g_i \} \) and \( k \) denotes the number of unknown functions \( \{ f_i \} \). It is worth noting that usually \( h \) and \( k \) are relatively small values and hence the SVD on \( L \) can be obtained efficiently.

Optimization of \( S \) For a given \( L \), the variable \( S \) can be optimized via solving the following problem:

\[
\min_S \| S - \hat{S} \|_F^2, \quad \text{subject to} \quad \| S \|_\infty \leq 1, \tag{15}
\]

where \( \hat{S} = (2\hat{\Phi} - \hat{\alpha}L)/\hat{\beta} \). Similarly, the optimization on \( S \) can be interpreted as computing a projection of a given matrix over an infinity norm ball. It also admits an analytic solution as summarized in the following theorem.

Lemma 1 For any matrix \( \hat{S} \), the optimal solution to Eq. (15) is given by

\[
S^* = \text{sgn}(\hat{S}) \circ \min(|\hat{S}|, 1), \tag{16}
\]

where \( \circ \) denotes the component-wise multiplication operator, and \( 1 \) denotes the matrix with entries 1 of appropriate size.

4 Theoretical analysis

In this section, we present a performance bound for the function estimation scheme in Eq. (3). Such a performance bound measures how well the estimation scheme can approximate the regression functions \( \{ f_j \} \) in Eq. (2) via the sparse low-rank coefficient \( \Theta \).

4.1 Basic properties of the optimal solution

We first present some basic properties of the optimal solution defined in Eq. (6); these properties are important building blocks of our following theoretical analysis.

Lemma 2 Consider the optimization problem in Eq. (6) for \( h, k \geq 2 \) and \( n \geq 1 \). Given \( n \) sample pairs as \( X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times d} \) and \( Y = [y_1, \ldots, y_n]^T \in \mathbb{R}^{n \times k} \). Let \( \mathcal{F} \) and \( \mathcal{G}_X \) be defined in Eqs. (2) and (4), respectively; let \( \sigma_{X(l)} \) be the largest singular
values of $G_X$. Assume that $W \in \mathbb{R}^{n \times k}$ has independent and identically distributed (i.i.d.) entries as $w_{ij} \sim \mathcal{N}(0, \sigma_w^2)$. Take

$$\alpha + \beta = \frac{2\sigma_X(i)\sigma_w\sqrt{n}}{N} \left(1 + \sqrt{\frac{k}{n} + t}\right), \quad (17)$$

where $N = n \times k$ and $t$ is a universal constant. Then with probability of at least $1 - \exp(-nt^2/2)$, for the minimizer $\hat{\Theta}$ in Eq. (6) and any $\Theta \in \mathbb{R}^{h \times k}$, we have

$$\frac{1}{N} \|G_X \hat{\Theta} - \mathcal{F}\|_F^2 \leq \frac{1}{N} \|G_X \Theta - \mathcal{F}\|_F^2 + 2\alpha \|S_0(\hat{\Theta} - \Theta)\|_* + 2\beta \|\hat{\Theta} - \Theta\|_J(\Theta)\|_1, \quad (18)$$

where $S_0$ is an operator defined in Lemma 1 of the Appendix.

**Proof** From the definition of $\hat{\Theta}$ in Eq. (6), we have

$$\hat{\Theta} = \Theta + \alpha \|\hat{\Theta}\|_* + \beta \|\hat{\Theta}\|_1 \leq \hat{\Theta} + \alpha \|\Theta\|_* + \beta \|\Theta\|_1. \quad (22)$$

By substituting $Y = \mathcal{F} + W$ and Eq. (5) into the previous inequality, we have

$$\frac{1}{N} \|G_X \hat{\Theta} - \mathcal{F}\|_F^2 \leq \frac{1}{N} \|G_X \Theta - \mathcal{F}\|_F^2 + \frac{2}{N} \langle W, G_X(\hat{\Theta} - \Theta) \rangle + \alpha (\|\Theta\|_* - \|\hat{\Theta}\|_*) + \beta (\|\Theta\|_1 - \|\hat{\Theta}\|_1). \quad (23)$$

Define the random event

$$\mathcal{A} = \left\{ \frac{1}{N} \|G_X^T W\|_2 \leq \frac{\alpha + \beta}{2} \right\}. \quad (19)$$

Taking $\alpha + \beta$ as the value in Eq. (17), it follows from Lemma 3 of the Appendix that $\mathcal{A}$ holds with probability of at least $1 - \exp(-nt^2/2)$. Therefore, we have

$$\langle W, G_X(\hat{\Theta} - \Theta) \rangle = \frac{\alpha + \beta}{\alpha + \beta} \langle W, G_X(\hat{\Theta} - \Theta) \rangle \leq \frac{\alpha}{\alpha + \beta} \|G_X^T W\|_2 \|\hat{\Theta} - \Theta\|_* + \frac{\beta}{\alpha + \beta} \|G_X^T W\|_\infty \|\hat{\Theta} - \Theta\|_1 \leq \frac{N}{2} \left( \alpha \|\hat{\Theta} - \Theta\|_* + \beta \|\hat{\Theta} - \Theta\|_1 \right),$$

where the second inequality follows from $\|G_X^T W\|_2 \geq \|G_X^T W\|_\infty$. Therefore, under $\mathcal{A}$, we have

$$\frac{1}{N} \|G_X \hat{\Theta} - \mathcal{F}\|_F^2 \leq \frac{1}{N} \|G_X \Theta - \mathcal{F}\|_F^2 + \alpha \|\hat{\Theta} - \Theta\|_* + \beta \|\hat{\Theta} - \Theta\|_1 + \alpha (\|\Theta\|_* - \|\hat{\Theta}\|_*) + \beta (\|\Theta\|_1 - \|\hat{\Theta}\|_1). \quad (25)$$

From Corollary 1 and Lemma 2 of the Appendix, we complete the proof.
4.2 Main assumption

We introduce a key assumption on the dictionary of basis functions $G_X$. Based on such an assumption, we derive a performance bound for the sparse trace norm regularization formulation in Eq. (6).

**Assumption 1** For a matrix pair $\Theta$ and $\Delta$ of size $h \times k$, let $s \leq \min(h, k)$ and $q \leq h \times k$. We assume that there exist constants $\kappa_1(s)$ and $\kappa_2(q)$ such that

$$
\kappa_1(s) \triangleq \min_{\Delta \in R(s, q)} \frac{\|G_X \Delta\|_F}{\sqrt{N} \|S_0(\Delta)\|_*} > 0, \quad \kappa_2(q) \triangleq \min_{\Delta \in R(s, q)} \frac{\|G_X \Delta\|_F}{\sqrt{N} \|\Delta J(\Theta)\|_1} > 0,
$$

where the restricted set $R(s, q)$ is defined as

$$
R(s, q) = \left\{ \Delta \in \mathbb{R}^{h \times k}, \Theta \in \mathbb{R}^{h \times k} \mid \Delta \neq 0, \text{ rank}(S_0(\Delta)) \leq s, \ |J(\Theta)| \leq q \right\},
$$

and $|J(\Theta)|$ denotes the number of nonzero entries in the matrix $\Theta$.

Our assumption on $\kappa_1(s)$ in Eq. (20) is closely related to but less restrictive than the RSC condition used in Negahban and Wainwright (2011); its denominator is only a part of the one in RSC and in a different matrix norm as well. Our assumption on $\kappa_2(q)$ is similar to the RE condition used in Bickel et al. (2009) except that its denominator is in a different matrix norm; our assumption can also be implied by sufficient conditions similar to the ones in Bickel et al. (2009).

4.3 Performance bound

We derive a performance bound for the sparse trace norm structure obtained by solving Eq. (6). This bound measures how well the optimal $\hat{\Theta}$ can be used to approximate $F$ by evaluating the averaged estimation error, i.e., $\|G_X \hat{\Theta} - F\|_F^2/N$.

**Theorem 2** Consider the optimization problem in Eq. (6) for $h, k \geq 2$ and $n \geq 1$. Given $n$ sample pairs as $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times d}$ and $Y = [y_1, \ldots, y_n]^T \in \mathbb{R}^{n \times k}$, let $F$ and $G_X$ be defined in Eqs. (2) and (4), respectively; let $\sigma_{X(l)}$ be the largest singular value of $G_X$. Assume that $W \in \mathbb{R}^{n \times k}$ has i.i.d. entries as $w_{ij} \sim \mathcal{N}(0, \sigma_w^2)$. Take $\alpha + \beta$ as the value in Eq. (17). Then with probability of at least $1 - \exp(-nt^2/2)$, for the minimizer $\hat{\Theta}$ in Eq. (6), we have

$$
\frac{1}{N} \|G_X \hat{\Theta} - F\|_F^2 \leq (1 + \epsilon) \inf_{\Theta} \left\{ \frac{1}{N} \|G_X \Theta - F\|_F^2 \right\} + \mathcal{E}(\epsilon) \left( \frac{\alpha^2}{\kappa_1^2(2r)} + \frac{\beta^2}{\kappa_2^2(c)} \right),
$$

where $\inf$ is taken over all $\Theta \in \mathbb{R}^{h \times k}$ with $\text{rank}(\Theta) \leq r$ and $|J(\Theta)| \leq c$, and $\mathcal{E}(\epsilon) > 0$ is a constant depending only on $\epsilon$. 

© Springer
Proof. Denote $\Delta = \hat{\Theta} - \Theta$ in Eq. (18). We have

$$
\frac{1}{N} \|G_X \hat{\Theta} - F\|_F^2 \leq \frac{1}{N} \|G_X \Theta - F\|_F^2 + 2\alpha \|S_0(\Delta)\|_* + 2\beta \|\Delta J(\Theta)\|_1.
$$

(22)

Given rank$(S_0(\Delta)) \leq 2r$ (from Lemma 1 of the Appendix) and $|J(\Theta)| \leq c$, we derive upper bounds on the components $2\alpha \|S_0(\Delta)\|_*$ and $2\beta \|\Delta J(\Theta)\|_1$ over the restricted set $R(2r, c)$ based on Assumption 1, respectively. It follows that

$$
2\alpha \|S_0(\Delta)\|_* \leq \frac{2\alpha}{\kappa_2(2r)\sqrt{N}} \|G_X (\hat{\Theta} - \Theta)\|_F \\
\leq \frac{2\alpha}{\kappa_2(2r)\sqrt{N}} (\|G_X \hat{\Theta} - F\|_F + \|G_X \Theta - F\|_F) \\
\leq \frac{\alpha^2\tau}{\kappa_1(2r)} + \frac{1}{N\tau} \|G_X \hat{\Theta} - F\|_F^2 + \frac{\alpha^2\tau}{\kappa_2(2r)} + \frac{1}{N\tau} \|G_X \Theta - F\|_F^2,
$$

(23)

where the last inequality above follows from $2ab \leq a^2\tau + b^2/\tau$ for $\tau > 0$. Similarly, we have

$$
2\beta \|\Delta J(\Theta)\|_1 \leq \frac{\beta^2\tau}{\kappa_2(c)} + \frac{1}{N\tau} \|G_X \hat{\Theta} - F\|_F^2 + \frac{\beta^2\tau}{\kappa_2(c)} + \frac{1}{N\tau} \|G_X \Theta - F\|_F^2.
$$

(24)

Substituting Eqs. (23) and (24) into Eq. (22), we have

$$
\frac{1}{N} \|G_X \hat{\Theta} - F\|_F^2 \leq \frac{\tau + 2}{(\tau - 2)N} \|G_X \Theta - F\|_F^2 + \frac{2\tau^2}{\tau - 2} \left(\frac{\alpha^2}{\kappa_1^2(2r)} + \frac{\beta^2}{\kappa_2^2(c)}\right).
$$

Setting $\tau = 2 + 4/\epsilon$ and $E(\epsilon) = 2(\epsilon + 2)^2/\epsilon$ in the inequality above, we complete the proof.

By choosing specific values for $\alpha$ and $\beta$, we can refine the performance bound described in Eq. (21). It follows from Eq. (17) we have

$$
\min_{\alpha, \beta, \alpha + \beta = \gamma} \left(\frac{\alpha^2}{\kappa_1^2(2r)} + \frac{\beta^2}{\kappa_2^2(c)}\right) = \frac{\gamma^2}{\kappa_1^2(2r) + \kappa_2^2(c)}
$$

$$
\gamma = \frac{2\sigma_X(l)w}{N} \sqrt{n} \left(1 + \sqrt{\frac{k}{n} + t}\right),
$$

where the equality of the first equation is achieved by setting $\alpha$ and $\beta$ proportional to $\kappa_1^2(2r)$ and $\kappa_2^2(q)$, i.e., $\alpha = \gamma \kappa_1^2(2r)/(\kappa_1^2(2r) + \kappa_2^2(c))$ and $\beta = \gamma \kappa_2^2(c)/(\kappa_1^2(2r) + \kappa_2^2(c))$. © Springer
\[
\frac{1}{N} \| G_X \hat{\Theta} - F \|_F^2 \leq (1 + \epsilon) \inf_{\Theta} \left\{ \frac{1}{N} \| G_X \Theta - F \|_F^2 + \frac{4\mathcal{E}(\epsilon)\sigma_{X(l)}^2 \sigma_n^2}{N^2 (\kappa_1^2(2r) + \kappa_2^2(c))} \left( 1 + \sqrt{\frac{k}{n}} + \epsilon \right)^2 \right\}.
\]

Note that the performance bound above is independent of the value of \(\alpha\) and \(\beta\), and it is tighter than the one described in Eq. (21).

5 Experiments

In this section, we conduct empirical study on the sparse trace norm regularization formulation in Eq. (6). Specifically, we choose the pre-specified basis function \(g_j(x_i) = x_{ij} (\forall i, j)\), and hence Eq. (6) can be explicitly expressed as

\[
\hat{\Theta} = \arg \min_{\Theta} \left( \| X \Theta - Y \|_F^2 + \alpha \| \Theta \|_* + \beta \| \Theta \|_1 \right).
\]

For demonstration, we evaluate the effectiveness of Eq. (25) on benchmark data sets; we also conduct numerical studies on the convergence of AG for solving Eq. (25) and the convergence of the alternating optimization algorithm for solve Eq. (10).

Performance evaluation We apply the sparse trace norm regularization formulation (S.TrNorm) on multi-label classification problems, in comparison with the trace norm regularization formulation (TrNorm) and the \(\ell_1\)-norm regularization formulation (OneNorm). AUC, Macro F1 and Micro F1 (Lewis et al. 2004) are used as the classification performance measures. Note that Macro F1 is obtained via computing the respective F1 measure separately for each label, and then computing the mean of the resulting F1 measures, while Micro F1 is obtained via computing the F1 measure across the involved training data for all labels as a single group.

Four benchmark data sets, including Business, Arts, and Health from Yahoo webpage data sets (Ueda and Saito 2002) and Scene from LIBSVM multi-label data sets,\(^1\) are employed in this experiment. The reported experimental results are averaged over 10 random repetitions of the data sets into training and test sets of the ratio 1:9. We use the AG method to solve the S.TrNorm formulation, and stop the iterative procedure of AG if the change of the objective values in two successive iterations is smaller than \(10^{-8}\) or the iteration numbers larger than \(10^5\). The regularization parameters \(\alpha\) and \(\beta\) are determined via double cross-validation from the set \(\{10^{-2} \times i\}_{i=1}^{10} \cup \{10^{-1} \times i\}_{i=2}^{10} \cup \{2 \times i\}_{i=1}^{10}\).

We present the averaged performance of the competing algorithms in Table 1. The main observations are summarized as follows: (1) S.TrNorm achieves the best

\(^1\) http://www.csie.ntu.edu.tw/~cjlin.
Table 1  Averaged performance (with standard derivation) comparison in terms of AUC, Macro F1, and Micro F1

| Data set (n, d, m) | Business (9,968, 16,621, 17) | Arts (7,441, 17,973, 19) | Health (9,109, 18,430, 14) | Scene (2,407, 294, 6) |
|-------------------|-----------------------------|-------------------------|-----------------------------|----------------------|
| AUC               |                             |                         |                             |                      |
| S.TrNorm          | 85.42 ± 0.31                | 76.31 ± 0.15            | 86.18 ± 0.56                | 91.54 ± 0.18         |
| TrNorm            | 83.43 ± 0.41                | 75.90 ± 0.27            | 85.24 ± 0.42                | 90.33 ± 0.24         |
| OneNorm           | 81.95 ± 0.26                | 70.47 ± 0.18            | 83.60 ± 0.32                | 88.42 ± 0.31         |
| Macro F1          |                             |                         |                             |                      |
| S.TrNorm          | 48.83 ± 0.13                | 32.83 ± 0.25            | 60.05 ± 0.36                | 51.65 ± 0.33         |
| TrNorm            | 47.24 ± 0.15                | 31.90 ± 0.31            | 58.91 ± 0.24                | 50.59 ± 0.08         |
| OneNorm           | 46.28 ± 0.25                | 31.03 ± 0.46            | 58.01 ± 0.18                | 46.57 ± 1.10         |
| Micro F1          |                             |                         |                             |                      |
| S.TrNorm          | 78.26 ± 0.71                | 42.91 ± 0.27            | 67.22 ± 0.47                | 52.83 ± 0.35         |
| TrNorm            | 78.84 ± 0.11                | 42.08 ± 0.11            | 66.92 ± 0.42                | 52.06 ± 0.49         |
| OneNorm           | 78.16 ± 0.17                | 40.64 ± 0.52            | 66.37 ± 0.19                | 47.32 ± 0.13         |

Note that n, d, and m denote the sample size, dimensionality, and label number, respectively.

Fig. 1  Convergence plot of AG for solving Eq. (25) (left plot) and the alternating optimization algorithm for solving the dual formulation of the proximal operator in Eq. (10) (right plot).

Performance on all benchmark data sets (except on Business data) in this experiment; this result demonstrates the effectiveness of the induced sparse low-rank structure for multi-label classification tasks; (2) TrNorm outperforms OneNorm on all benchmark data sets; this result demonstrates the effectiveness of modeling a shared low-rank structure for high-dimensional text and image data analysis.

Numerical study  We study the practical convergence of AG by solving Eq. (25) on Scene data. Note that we set $\alpha = 1$, $\beta = 1$ in Eq. (25) for demonstration; for other parameter settings, we observe similar trends.

In the first experiment, we study the convergence of AG. We stop AG when the change of the objective values in two successive iterations smaller than $10^{-8}$. The convergence curves is presented in the left plot of Fig. 1. We observe that AG converges very fast, and its convergence speed is consistent with the theoretical convergence analysis in Nesterov (2004).
In the second experiment, we conduct numerical study on the alternating optimization algorithm (in Sect. 3.1.2) for solving the dual formulation of the proximal operator in Eq. (10). Similarly, the alternating optimization algorithm is stopped when the change of the objective values in two successive iterations smaller than $10^{-8}$. For illustration, in Eq. (10) we randomly generate the matrix $\hat{\Phi}$ of size 10,000 by 5,000 from $\mathcal{N}(0, 1)$; we then apply the alternating optimization algorithm to solve Eq. (10) and plot its convergence curve in the right plot of Fig. 1. Our experimental results show that the alternating optimization algorithm generally converges within 10 iterations and our results demonstrate the practical efficiency of this algorithm.

6 Conclusion

We study the problem of estimating multiple predictive functions simultaneously in the nonparametric regression setting. In our estimation scheme, each predictive function is estimated using a linear combination of a dictionary of pre-specified basis functions. By assuming that the coefficient matrix admits a sparse low-rank structure, we formulate the function estimation problem as a convex program with the trace norm and the $\ell_1$-norm regularization. We propose to employ the AG algorithm to solve the function estimation problem and also develop efficient algorithms for the key components involved in AG. We derive a key property of the optimal solution to the convex program; moreover, based on an assumption associated with the basis functions, we establish a performance bound of the proposed function estimation scheme using the composite regularization. Our simulation studies demonstrate the effectiveness and the efficiency of the proposed formulation. In the future, we plan to derive a formal sparse oracle inequality for the convex problem in Eq. (6) as in Bickel et al. (2009); we also plan to apply the proposed function estimation formulation to other real world applications.

7 Appendix

7.1 Operators $S_0$ and $S_1$

We define two operators, namely $S_0$ and $S_1$, on an arbitrary matrix pair (of the same size) based on Lemma 3 in Recht et al. (2010), as summarized in the following lemma.

Lemma 1 Given any $\Theta$ and $\Delta$ of size $h \times k$, let $\text{rank}(\Theta) = r$ and denote the SVD of $\Theta$ as

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

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

$$\hat{\Delta} = U^T \Delta V = \begin{bmatrix} \hat{\Delta}_{11} & \hat{\Delta}_{12} \\ \hat{\Delta}_{21} & \hat{\Delta}_{22} \end{bmatrix},$$
where $\hat{\Delta}_{11} \in \mathbb{R}^{r \times r}$, $\hat{\Delta}_{12} \in \mathbb{R}^{r \times (k-r)}$, $\hat{\Delta}_{21} \in \mathbb{R}^{(h-r) \times r}$, and $\hat{\Delta}_{22} \in \mathbb{R}^{(h-r) \times (k-r)}$. Define $S_0$ and $S_1$ as

$$S_0(\Theta, \Delta) = U \begin{bmatrix} \hat{\Delta}_{11} & \hat{\Delta}_{12} \\ \hat{\Delta}_{21} & 0 \end{bmatrix} V^T, \quad S_1(\Theta, \Delta) = U \begin{bmatrix} 0 & 0 \\ 0 & \hat{\Delta}_{22} \end{bmatrix} V^T.$$

Then the following conditions hold: $\text{rank}(S_0(\Theta, \Delta)) \leq 2r$, $\Theta S_1(\Theta, \Delta)^T = 0$, $\Theta^T S_1(\Theta, \Delta) = 0$.

The result presented in Lemma 1 implies a condition under which the trace norm on a matrix pair is additive. From Lemma 1 we can easily verify that

$$\|\Theta + S_1(\Theta, \Delta)\|_* = \|\Theta\|_* + \|S_1(\Theta, \Delta)\|_*,$$

for arbitrary $\Theta$ and $\Delta$ of the same size. To avoid clutter notation, we denote $S_0(\Theta, \Delta)$ by $S_0(\Delta)$, and $S_1(\Theta, \Delta)$ by $S_1(\Delta)$ throughout this paper, as the appropriate $\Theta$ can be easily determined from the context.

7.2 Bound on trace norm

As a consequence of Lemma 1, we derive a bound on the trace norm of the matrices of interest as summarized below.

**Corollary 1** Given an arbitrary matrix pair $\hat{\Theta}$ and $\Theta$, let $\Delta = \hat{\Theta} - \Theta$. Then

$$\|\hat{\Theta} - \Theta\|_* + \|\Theta\|_* - \|\hat{\Theta}\|_* \leq 2\|S_0(\Delta)\|_*.$$

**Proof** From Lemma 1 we have $\Delta = S_0(\Delta) + S_1(\Delta)$ for the matrix pair $\Theta$ and $\Delta$. Moreover,

$$\|\hat{\Theta}\|_* = \|\Theta + S_0(\Delta) + S_1(\Delta)\|_* \geq \|\Theta + S_1(\Delta)\|_* - \|S_0(\Delta)\|_*$$

$$= \|\Theta\|_* + \|S_1(\Delta)\|_* - \|S_0(\Delta)\|_*,$$

where the inequality above follows from the triangle inequality and the last equality above follows from Eq. (28). Using the result in Eq. (29), we have

$$\|\hat{\Theta} - \Theta\|_* + \|\Theta\|_* - \|\hat{\Theta}\|_* \leq \|\Delta\|_* + \|\Theta\|_* - \|\Theta\|_* - \|S_1(\Delta)\|_* + \|S_0(\Delta)\|_*$$

$$\leq 2\|S_0(\Delta)\|_*.$$

We complete the proof of this corollary.

7.3 Bound on $\ell_1$-norm

Analogous to the bound on the trace norm in Corollary 1, we also derive a bound on the $\ell_1$-norm of the matrices of interest in the following lemma. For arbitrary matrices
Sparse trace norm regularization

\(\Theta\) and \(\Delta\), we denote by \(J(\Theta) = \{(i, j)\}\) the coordinate set (the location set of nonzero entries) of \(\Theta\), and by \(J(\Theta)_{\perp}\) the associated complement (the location set of zero entries); we denote by \(\Delta J(\Theta)\) the matrix of the same entries as \(\Delta\) on the set \(J(\Theta)\) and of zero entries on the set \(J(\Theta)_{\perp}\). We now present a result associated with \(J(\Theta)\) and \(J(\Theta)_{\perp}\) in the following lemma. Note that a similar result for the vector case is presented in Bickel et al. (2009).

**Lemma 2** Given a matrix pair \(\hat{\Theta}\) and \(\Theta\) of the same size, the inequality below always holds

\[
\|\hat{\Theta} - \Theta\|_1 + \|\Theta\|_1 - \|\hat{\Theta}\|_1 \leq 2\|\hat{\Theta}_{J(\Theta)} - \Theta_{J(\Theta)}\|_1.
\]

**Proof** It can be verified that the inequality

\[
\|\Theta_{J(\Theta)}\|_1 - \|\hat{\Theta}_{J(\Theta)}\|_1 \leq \|\hat{\Theta} - \Theta\|_{J(\Theta)_{\perp}}\|_1,
\]

and the equalities

\[
\Theta_{J(\Theta)_{\perp}} = 0, \quad \|\hat{\Theta} - \Theta\|_{J(\Theta)_{\perp}}\|_1 - \|\hat{\Theta}_{J(\Theta)_{\perp}}\|_1 = 0.
\]

hold. Therefore we can derive

\[
\|\hat{\Theta} - \Theta\|_1 + \|\Theta\|_1 - \|\hat{\Theta}\|_1 = \|\hat{\Theta} - \Theta\|_{J(\Theta)}\|_1 + \|\hat{\Theta} - \Theta\|_{J(\Theta)_{\perp}}\|_1
\]

\[
+ \|\Theta_{J(\Theta)}\|_1 + \|\Theta_{J(\Theta)_{\perp}}\|_1 - \|\hat{\Theta}_{J(\Theta)}\|_1 - \|\hat{\Theta}_{J(\Theta)_{\perp}}\|_1
\]

\[
\leq 2\|\hat{\Theta} - \Theta\|_{J(\Theta)}\|_1.
\]

This completes the proof of this lemma.

7.4 Concentration inequality

**Lemma 3** Let \(\sigma_X(l)\) be the maximum singular value of the matrix \(G_X \in \mathbb{R}^{n \times h}\); let \(W \in \mathbb{R}^{n \times k}\) be the matrix of i.i.d entries as \(w_{ij} \sim \mathcal{N}(0, \sigma_w^2)\). Let \(\lambda = 2\sigma_X(l)\sigma_w\sqrt{n} \left(1 + \sqrt{\frac{k}{n}} + t\right) / N\). Then

\[
\Pr \left(\frac{1}{N}\|WG_X\|_2 \leq \frac{\lambda}{2}\right) \geq 1 - \exp \left(-\frac{1}{2}nt^2\right).
\]

**Proof** It is known (Szarek 1991) that a Gaussian matrix \(\hat{W} \in \mathbb{R}^{n \times k}\) with \(n \geq k\) and \(\hat{w}_{ij} \sim \mathcal{N}(0, 1/n)\) satisfies

\[
\Pr \left(\|\hat{W}\|_2 > 1 + \sqrt{\frac{k}{n}} + t\right) \leq \exp \left(-\frac{1}{2}nt^2\right), \quad (30)
\]

\(\odot\) Springer
where $t$ is a universal constant. From the definition of the largest singular value, there exists a vector $b \in \mathbb{R}^h$ of length 1, i.e., $\|b\|_2 = 1$, such that $\|W^T G_X b\|_2 = \|W^T G_X b\|_2 \leq \|W\|_2 \|G_X b\|_2 \leq \sigma_X(l) \|W\|_2$. Since $w_{ij} / (\sigma_w \sqrt{n}) \sim N(0, 1/n)$, we have
\[
\Pr \left( \frac{1}{N} \|W^T G_X\|_2 > \frac{\lambda}{2} \right) \leq \Pr \left( \frac{1}{N} \sigma_X(l) \|W\|_2 > \frac{\lambda}{2} \right).
\]
Applying the result in Eq. (30) into the inequality above, we complete the proof of this lemma.

References

Bach FR (2008) Consistency of trace norm minimization. J Mach Learn Res 9:1019–1048
Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J Imag Sci 2(1):183–202
Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of Lasso and dantzig selector. Ann Stat 37(4):1705–1732
Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, Cambridge
Bunea F, Tsybakov AB, Wegkamp MH (2006) Aggregation and sparsity via $\ell_1$ penalized least squares. In: Proceedings of the 19th annual conference on learning theory, pp 379–391
Candès EJ, Tao T (2005) Decoding by linear programming. IEEE Trans Inf Theory 51(12):4203–4215
Candès EJ, Recht B (2009) Exact matrix completion via convex optimization. Found Comput Math 9:717–772
Candès EJ, Li X, Ma Y, Wright J (2009) Robust principal component analysis? J ACM 58(1):1–37
Chandrasekaran V, Sanghavi S, Parrilo PA, Willsky AS (2011) Rank-sparsity incoherence for matrix decomposition. SIAM J Optim 21(2):572–596
Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32(2):407–499
Fazel M, Hindi H, Boyd SP (2001) A rank minimization heuristic with application to minimum order system approximation. In: Proceedings of the American control conference, pp 4734–4739
Friedman J, Hastie T, Höfling H, Tibshirani R (2007) Pathwise coordinate optimization. Ann Appl Stat 1(2):302–332
Grippo L, Sciandrone M (2000) On the convergence of the block nonlinear gauss-seidel method under convex constraints. Oper Res Lett 26(3):127–136
Huang J, Zhang T, Metaxas DN (2011) Learning with structured sparsity. J Mach Learn Res 12:3371–3412
Lewis DD, Yang Y, Rose TG, Li F, Dietterich G, Li F (2004) RCV1: a new benchmark collection for text categorization research. J Mach Learn Res 5:361–397
Lounici K, Pontil M, Tsybakov AB, van de Geer SA (2009) Taking advantage of sparsity in multi-task learning. In: The 22nd international conference on learning theory
Moreau JJ (1965) Proximité et dualité dans un espace hilbertien. Bull de la Société Mathématique de France 93:273–299
Negahban S, Wainwright MJ (2011) Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Ann Stat 39(2):1069–1097
Nesterov Y (2004) Introductory lectures on convex optimization: a basic course. Springer, Berlin
Nesterov Y (2013) Gradient methods for minimizing composite objective function. Math Program 140(1):125–161
Pati YC, Kailath T (1994) Phase-shifting masks for microlithography: automated design and mask requirements. J Opt Soc Am A 11(9):2438–2452
Recht B, Fazel M, Parrilo PA (2010) Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. SIAM Rev 52(3):471–501
Rennie JDM, Srebro N (2005) Fast maximum margin matrix factorization for collaborative prediction. In: Proceedings of the 22nd international conference on, machine learning, pp 713–719
Rohde A, Tsybakov AB (2011) Estimation of high-dimensional low rank matrices. Ann Stat 39(2):887–930
Srebro N, Rennie JDM, Jaakola TS (2005) Maximum-margin matrix factorization. In: Advances in neural information processing systems, vol 17
Szarek SJ (1991) Condition numbers of random matrices. J Complex 7(2):131–149
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288
Toh KC, Todd MJ, Tutuncu R (1999) SDPT3: a MATLAB software package for semidefinite programming. Optim Methods Soft 11:545–581
Toh KC, Yun S (2010) An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. Pac J Optim 6:615–640
Ueda N, Saito K (2002) Single-shot detection of multiple categories of text using parametric mixture models. In: Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, pp 626–631
Watson GA (1992) Characterization of the subdifferential of some matrix norms. Linear Algebra Appl 170:33–45
Zhang T (2009) Some sharp performance bounds for least squares regression with $l_1$ regularization. Ann Stat 37(5A):2109–2144
Zhao P, Yu B (2006) On model selection consistency of Lasso. J Mach Learn Res 7:2541–2563