Journal of Machine Learning Research 16 (2015) 47-75 Submitted 2/13; Revised 7/14; Published 1/15

Simultaneous Pursuit of Sparseness and Rank Structures for Matrix Decomposition

Qi Yan YANXX195@UMN.EDU
School of Statistics
University of Minnesota
Minneapolis, MN 55414, USA

Jieping Ye JIEPING.YE@ASU.EDU
Computer Science and Engineering
Arizona State University
Tempe, AZ 85287 USA

Xiaotong Shen XSHEN@UMN.EDU
School of Statistics
University of Minnesota
Minneapolis, MN 55414, USA

Editor: Aapo Hyvarinen

Abstract

In multi-response regression, pursuit of two different types of structures is essential to battle the curse of dimensionality. In this paper, we seek a sparsest decomposition representation of a parameter matrix in terms of a sum of sparse and low rank matrices, among many overcomplete decompositions. On this basis, we propose a constrained method subject to two nonconvex constraints, respectively for sparseness and low- rank properties. Computationally, obtaining an exact global optimizer is rather challenging. To overcome the difficulty, we use an alternating directions method solving a low-rank subproblem and a sparseness subproblem alternatively, where we derive an exact solution to the low-rank subproblem, as well as an exact solution in a special case and an approximated solution generally through a surrogate of the L_0-constraint and difference convex programming, for the sparse subproblem. Theoretically, we establish convergence rates of a global minimizer in the Hellinger-distance, providing an insight into why pursuit of two different types of decomposed structures is expected to deliver higher estimation accuracy than its counterparts based on either sparseness alone or low-rank approximation alone. Numerical examples are given to illustrate these aspects, in addition to an application to facial imagine recognition and multiple time series analysis.

Keywords: blockwise decent, nonconvex minimization, matrix decomposition, structure pursuit

1. Introduction

In multivariate analysis, data as well as parameters are usually expressed in terms of a matrix form, as opposed to a vector representation in univariate analysis. This occurs frequently in multi-class classification (Amit et al., 2007), matrix completion (Cai et al., 2010; Jain et al., 2010), collaborative filtering (Srebro et al., 2005), computer vision (Wright, 2009), among others. In situations as such, it essential to identify and employ certain lower-dimensional structures to battle the curse of dimensionality due to an increase in dimensionality from multivariate attributes. In this article, we explore rank and sparseness structures through matrix decomposition simultaneously in estimating large matrices through a novel notation of seeking a sparsest decomposition from a class of overcomplete decompositions.

Statistically, different structures have dramatically different interpretations. A low rank property of a matrix describes global information across different tasks, whereas sparseness concerns local information of specific task. For instance, for face images, the global information corresponds to the overall shape of a face, but the local information characterizes specific facial expression such as laugh and cry. In linear time-invariant (LTI) system, a low rank property corresponds to a low-order LTI system and a sparseness property captures an LTI system with a sparse impulse response (Porat, 1997). In a high-dimensional situation, betting on one type of structure may not be adequate to battle the curse of dimensionality. In this article, we seek a sparsest decomposition for the purpose of dimension reduction, from a class of overcomplete decompositions into simpler sparse and low-rank components. Specifically, a matrix $\boldsymbol{\Theta}$ is decomposed as $\boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2$, for a sparse $\boldsymbol{\Theta}_1$ and low-rank $\boldsymbol{\Theta}_2$ components, where $\boldsymbol{\Theta}_1$and $\boldsymbol{\Theta}_2$ are chosen from many such decompositions, with a smallest effective degrees of freedom, leading to high accuracy of parameter estimation. Our objective is to reconstruct the parameter matrix by identifying a sparsest decomposition consisting of simpler components. Such a decomposition can be used to provide a simpler and more efficient description of a complex system in terms of its simpler components. This results in more efficient structure representations leading to higher accuracy of parameter estimation in high-dimensional data analysis.

In this paper, we consider a multi-response linear regression problem in which a random sample $(\boldsymbol{a}_i, \boldsymbol{z}_i)_{i=1}^n$ is observed with a k-dimensional response vector $\boldsymbol{z}_i$ following

$$\boldsymbol{z}_i = \boldsymbol{a}_i^T \boldsymbol{\Theta} + \boldsymbol{\epsilon}_i, \quad E\boldsymbol{\epsilon}_i = 0,\ Cov(\boldsymbol{\epsilon}_i) = \sigma^2 \boldsymbol{I}; \quad i = 1, \dots, n, \tag{1}$$

where $\boldsymbol{a}_i$ is a p-dimensional design vector, is independent of random error ϵ_i, and $\boldsymbol{I}$ is the identity matrix. Model (1) reduces to the univariate case when $k = 1$, and becomes a multivariate autoregressive model when $\boldsymbol{a}_i = \boldsymbol{z}_{i-1}$. Through matrix decomposition, we decompose a $p \times k$ regression parameter matrix $\boldsymbol{\Theta}$ into a sum of a sparse matrix $\boldsymbol{\Theta}_1$ and a low rank matrix $\boldsymbol{\Theta}_2$ for structure exploration, that is, $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2$. Model (1) is expressible in a matrix form

$$\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{\Theta} + \boldsymbol{e}; \tag{2}$$

where $\boldsymbol{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)^T \in \mathbb{R}^{n \times k}$, $\boldsymbol{A} = (a_1, \cdots, a_n)^T$ is a $n \times p$ matrix, and $\boldsymbol{e} = \boldsymbol{(\epsilon_1, \cdots, \epsilon_n)}^T \in \mathbb{R}^{n \times k}$ are the data, design and error matrices. In (1), we estimate $\boldsymbol{\Theta}$ based on n paired observation vectors $(\boldsymbol{a}_i, \boldsymbol{z}_i)_{i=1}^n$, with prior knowledge that $\boldsymbol{\Theta}_1$ is sparse in the number of its nonzero entries, and rank $r(\boldsymbol{\Theta}_2)$ is low relative to $\min(n, k, p)$. Our goal is to recover the parameter $\boldsymbol{\Theta}$ by identifying $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$.

In the literature, the simultaneous exploration of rank and sparseness structures through matrix decomposition has received some attention, yet has not been well-studied. For robust principal component analysis (RPCA) where $\boldsymbol{A} = \boldsymbol{I}_{n \times p}$ is the $n \times p$ identity matrix with its diagonals and off-diagonals being one and zero, Yuan & Yang (2013) and Chandrasekaran et al. (2011) employed a linear combination of the L_1 sparsity regularization and the nuclear-norm regularization, and Zhou & Tao (2011) used a randomized projections based low rank approximations and thresholding for sparsity pursuit. Moreover, Wright et al. (2013) recovers the sparse and low-rank components by minimizing a linear combination of the L_1-norm for sparsity and the nuclear-norm for low rank pursuit, while Waters et al. (2011) develops a greedy algorithm to pursue the sparse and low rank structures. For multiple task learning, Chen et al. (2010) studies sparse and low rank structures separately through convex regularization. In essence, most the existing literature focuses exclusively on a unique matrix decomposition of $\boldsymbol{\Theta}$ with $\boldsymbol{A} = \boldsymbol{I}_{n\times p}$ or $\boldsymbol{A}$ to be a set of random linear measurements, and without noise or with small noise that is essentially ignorable. For instance, Chandrasekaran et al. (2011) provided sufficient conditions for exact recovery of a convex relaxation method without noise; Wright et al. (2013) proved that recovering a target matrix is possible from a small set of randomly selected linear measurements when the number of measurements is sufficiently large. Among these, Agarwal et al. (2012) considered a general $\boldsymbol{A}$ and derived a theorem that bounds the Frobenius-norm error obtained through regularized convex relaxation under a "spikiness" condition that the max-norm of the low rank component $\|\boldsymbol{\Theta}_2\|_{\max}$ is less than $\frac{\alpha}{\sqrt{pk}}$ for some fixed $\alpha > 0$.

In this paper, we consider a general design matrix $\boldsymbol{A}$ and parameter matrices $(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$, for regression analysis, where $\boldsymbol{A}$ represents features of observations which is deterministic, and can be any matrix with n rows and p columns. Of particular interest is reconstruction of $\boldsymbol{\Theta}$ in a high-dimensional situation in which (p, k) may exceed the sample size n. Computationally, we use an alternating direction method separating low-rank pursuit from sparsity pursuit alternatively, where an exact solution to the low-rank problem and that to the sparsity pursuit problem when $\boldsymbol{A} = \boldsymbol{I}_{n\times p}$ or an approximated solution for a general $\boldsymbol{A}$ is obtained. In either case, the final solution is shown to be stationary without and with maximum block improvement (Chen et al., 2012) for $\boldsymbol{A} = \boldsymbol{I}_{n\times p}$ and a general $\boldsymbol{A}$. Theoretically, we establish error bound for the proposed method in the Hellinger-distance for reconstruction of $\boldsymbol{\Theta}$, based on which rates of convergence are obtained. Numerically, the proposed method compares favorably against two strong competitors in simulations.

The paper is organized as follows. Section 2 develops a computational method through the alternating directions method and a closed-form solution for a rank problem. Section 3 investigates statistical properties of the proposed method, followed by simulation studies and a real data example in Section 4. Finally, technical proofs are contained in Section 5.

2. Proposed Method

In this section, we explore a structure decomposition of a parameter matrix in the form $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2$ under model (1), then develops computational methods in two situations and discuss their properties.

2.1 Structure Decomposition

Due to non-uniqueness of such a decomposition under model (1), we seek one decomposition, among many overcomplete decompositions, that minimizes the effective degrees of freedom of $\boldsymbol{\Theta}$ Efron (2004), defined as

$$\mathrm{Eff}(\boldsymbol{\Theta}) = \min_{\{\boldsymbol{\Theta}=\boldsymbol{\Theta}_1+\boldsymbol{\Theta}_2:\|\boldsymbol{\Theta}_1\|_0 \le \max(0,p+k-2r(\boldsymbol{\Theta}_2)-2)\}} \|\boldsymbol{\Theta}_1\|_0 + (p+k-r(\boldsymbol{\Theta}_2))r(\boldsymbol{\Theta}_2),$$

where $\|\cdot\|_0$ is the L_0-norm of a matrix, or the number of nonzero entries of the matrix, and $r(\cdot)$ denotes the rank of a matrix. In other words, we identify a decomposition minimizing the effective degrees of freedom $\mathrm{Eff}(\boldsymbol{\Theta})$, among all candidate decompositions. Lemma 1 below says that the minimal of $\mathrm{Eff}(\boldsymbol{\Theta})$ is unique in $(\|\boldsymbol{\Theta}_1\|_0, r(\boldsymbol{\Theta}_2))$ under the constraint that $\|\boldsymbol{\Theta}_1\|_0 \le \max(0, p+k-2r(\boldsymbol{\Theta}_2)-2) \le 2\max(p,k)$.

Lemma 1 *The minimizer of Eff$(\boldsymbol{\Theta})$ is unique with respect to* $(\|\boldsymbol{\Theta}_1\|_0, r(\boldsymbol{\Theta}_2))$ *if* $\|\boldsymbol{\Theta}_1\|_0 \le \max(0, p+k-2r(\boldsymbol{\Theta}_2)-2)$. *Moreover,*

$$\mathit{Eff}(\boldsymbol{\Theta}) \le \min((p+k-r(\boldsymbol{\Theta}))r(\boldsymbol{\Theta}), \|\boldsymbol{\Theta}\|_0)).$$

Model (1) is identifiable with respect to $\boldsymbol{\Theta}$ but may not be so in $(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ even when $\boldsymbol{A}$ is of full rank, due to non-uniqueness of a decomposition $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2$.

2.2 Estimation

To pursue structures of low-rank and sparsity through matrix decomposition simultaneously, we propose a constrained likelihood method subject to two nonconvex constraints:

$$\min_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \|\boldsymbol{A\Theta_1} + \boldsymbol{A\Theta_2} - \boldsymbol{Z}\|_F^2, \quad \text{subject to} \quad \|\boldsymbol{\Theta}_1\|_0 \le s_1, \quad r(\boldsymbol{\Theta}_2) \le s_2, \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius-norm defined as the L_2-norm of all entries of a matrix, and s_1 and s_2 are integer-valued tuning parameters with $0 \le s_1 \le \max(p,k)$ and $1 \le s_2 \le \min(n,k,p)$ based on the consideration that the rank function and the sparsity measure are integer-valued.

When $\boldsymbol{A} = \boldsymbol{I}_{n\times p}$, (3) is simplified as

$$\min_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \|\boldsymbol{Z} - \boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_F^2 \quad \text{subject to } \|\boldsymbol{\Theta}_1\|_0 \le s_1, \quad r(\boldsymbol{\Theta}_2) \le s_2, \tag{4}$$

where a special structure may be taken into account to solve this nonconvex minimization.

When $\boldsymbol{A} \ne \boldsymbol{I}_{n\times p}$ is any matrix of full rank, the two constraints in (3) are either defined by the L_0-function or the rank function, imposing computational challenges. To develop an efficient algorithm to solve (3), we approximate the $\|\boldsymbol{\Theta}_1\|_0 = \sum_{i,j} I(|\theta_{ij}| \ne 0)$ by its computational surrogate—the truncated L_1-function $\sum_{\theta_{ij}\in\boldsymbol{\Theta}_1} \frac{1}{\tau}\min(|\theta_{ij}|,\tau)$ Shen et al. (2012) as $\tau \to 0^+$. This leads to a computational surrogate of (3):

$$\min_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} f(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2), \text{ subject to } \frac{1}{\tau}\sum_{i,j} \min(|\theta_{ij}|,\tau) \le s_1, \; r(\boldsymbol{\Theta}_2) \le s_2, \tag{5}$$

where $f(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) = \|\boldsymbol{A}(\boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2) - \boldsymbol{Z}\|_F^2$ and τ is a nonnegative tuning parameter.

2.3 Method for Nonconvex Minimization

This section will develop computational strategies for (4) and (5) separately, based on blockwise coordinate decent as well as maximum block improvement (MBI, Chen et al., 2012). First, we separate the task of sparsity pursuit for $\boldsymbol{\Theta}_1$ from that of rank minimization for $\boldsymbol{\Theta}_2$, where $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ correspond to two blocks for decent. Second, we apply MBI to assure that blockwise coordinate decent yields a stationary solution for nonconvex minimization, which would be otherwise impossible. In addition, for (5), we develop a gradient project method to permit fast computation of a constrained problem through the means of unconstrained optimization.

The strategy of blockwise coordinate decent proceeds as follows. For (4) and (5), we solve it in $\boldsymbol{\Theta}_2$ given $\boldsymbol{\Theta}_1$ and solve them in $\boldsymbol{\Theta}_1$ given $\boldsymbol{\Theta}_2$, alternatively. In each step of alternating blocks, we proceed with the block giving the maximum block improvement.

2.3.1 Nonconvex minimization (4): a special case

For (4), when $\boldsymbol{\Theta}_2$ is held fixed, (4) has a global minimizer can be obtained through componentwise thresholding defined by the L_0-function as follows:

$$\hat{\boldsymbol{\Theta}}_1(\boldsymbol{Z}, \boldsymbol{\Theta}_2) = \left(I\left\{ |z_{ij} - \theta_{ij}^{(2)}| > \lambda \right\} \cdot (z_{ij} - \theta_{ij}^{(2)}) \right)_{p \times k}, \tag{6}$$

where $\theta_{ij}^{(2)}$ is the ijth entry of $\boldsymbol{\Theta}_2$ and λ is any number between the s_1th and (s_1+1)th largest entries of $|\boldsymbol{Z} - \boldsymbol{\Theta}_2|$.

When $\boldsymbol{\Theta}_1$ is held fixed, a global minimizer of (4) is

$$\hat{\boldsymbol{\Theta}}_2(\boldsymbol{Z}, \boldsymbol{\Theta}_1) = \boldsymbol{U}\boldsymbol{D}_{s_2}\boldsymbol{V}^T, \tag{7}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are given by singular value decomposition (SVD) of $\boldsymbol{Z} - \boldsymbol{\Theta}_1 = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ and $\boldsymbol{D}_{s_2}$ is a diagonal matrix retaining the largest s_2 singular values of $\boldsymbol{Z} - \boldsymbol{\Theta}_1$ and truncating other singular values at zero.

Our algorithm for computing (4) is summarized.

Step 1.(Initialization) Supply a good initial estimate $(\hat{\boldsymbol{\Theta}}_1^{(0)}, \hat{\boldsymbol{\Theta}}_2^{(0)})$ in (4). Specify precision $\delta > 0$.

Step 2.(Iteration) At iteration m, update $\hat{\boldsymbol{\Theta}}_2^{(m)}$ in (7) with $\boldsymbol{\Theta}_1 = \hat{\boldsymbol{\Theta}}_1^{(m-1)}$. Then update $\hat{\boldsymbol{\Theta}}_1^{(m)}$ in (6) with $\boldsymbol{\Theta}_2 = \hat{\boldsymbol{\Theta}}_2^{(m)}$.

Step 3.(Stopping rule) Terminate if $|f(\hat{\boldsymbol{\Theta}}_1^{(m)}, \hat{\boldsymbol{\Theta}}_2^{(m)}) - f(\hat{\boldsymbol{\Theta}}_1^{(m-1)}, \hat{\boldsymbol{\Theta}}_2^{(m-1)})| \le \delta$, where $f(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) = \|\boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2 - \boldsymbol{Z}\|_F^2$. Let m^* be the index at termination. The estimate is then $(\hat{\boldsymbol{\Theta}}_1^{(m^*)}, \hat{\boldsymbol{\Theta}}_2^{(m^*)})$.

2.3.2 Nonconvex minimization (5): A general case

The problem of solving for $\boldsymbol{\Theta}_2$ in (5) given $\boldsymbol{\Theta}_1$ reduces to that of constrained rank minimization

$$\min_{\boldsymbol{\Theta}_2} \|\boldsymbol{A}\boldsymbol{\Theta}_2 - (\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{\Theta}_1)\|_F^2 \quad \text{subject to} \quad r(\boldsymbol{\Theta}_2) \le s_2, \tag{8}$$

provided that $\mathbf{\Theta}_1$ satisfies the sparsity constraint in (5). Now write $\mathbf{\Theta}_2 \equiv \boldsymbol{CF}$, where $\boldsymbol{C}$ and $\boldsymbol{F}$ are $p \times r$ and $r \times k$ matrices with $r \leq s_2$, consisting of a basis of the column space and that of the row space of $\mathbf{\Theta}_2$, respectively. Note that $\{\mathbf{\Theta}_2 : r(\mathbf{\Theta}_2) \leq s_2\} = \{\mathbf{\Theta}_2 : \mathbf{\Theta}_2 = \boldsymbol{CF}, r \leq s_2\}$. Then solving (8) is equivalent to that

$$\min_{\boldsymbol{C},\boldsymbol{F}} \|\boldsymbol{A}(\boldsymbol{CF}) - (\boldsymbol{Z} - \boldsymbol{A}\mathbf{\Theta}_1)\|_F^2, \tag{9}$$

An application of an argument of (Xing et al., 2012) yields a global minimizer of (9), which has an analytic form

$$\hat{\mathbf{\Theta}}_2(\mathbf{\Theta}_1) = \hat{\boldsymbol{C}}\hat{\boldsymbol{F}}, \quad \hat{\boldsymbol{C}} = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}_w, \quad \hat{\boldsymbol{F}} = \boldsymbol{D}_w\boldsymbol{V}_w^T, \tag{10}$$

where $\boldsymbol{D}$ is a $r(\boldsymbol{A}) \times r(\boldsymbol{A})$ diagonal singular vector matrix based on SVD of $\boldsymbol{A} = \boldsymbol{UDV}^T$, $\boldsymbol{D}_w$ is also a diagonal matrix of s_2 leading singular values of $\boldsymbol{W} \equiv \boldsymbol{U}^T(\boldsymbol{Z} - \boldsymbol{A}\mathbf{\Theta}_1)$ and U_w, V_w are matrices consisting of the corresponding right and left singular vectors.

Note that computation involves only the first s_2 largest singular values. Therefore, we employ the randomized truncated SVD method (Halko et al., 2011), for efficient computation of a large problem. This amounts to a complexity of order $O(pk \log r)$, as compared to $O(\min(pk^2, p^2k))$ of a conventional SVD method (Golub & Van, 1996).

Solving for $\mathbf{\Theta}_1$ in (5) given $\mathbf{\Theta}_2$, on the other hand, becomes the problem of sparsity pursuit. In particular, we solve, assuming that $r(\mathbf{\Theta}_2) \leq s_2$,

$$\min_{\mathbf{\Theta}_1} \|\boldsymbol{A}\mathbf{\Theta}_1 - (\boldsymbol{Z} - \boldsymbol{A}\mathbf{\Theta}_2)\|_F^2, \quad \text{subject to} \quad \frac{1}{\tau}\sum_{\theta_{ij} \in \mathbf{\Theta}_1} \min(|\boldsymbol{\theta}_{ij}|, \tau) \leq s_1, \tag{11}$$

which is solved iteratively by a difference of convex (DC) programming, constructing a convex set containing the original constrained set. The constraint in (5) is defined by $J(\mathbf{\Theta}_1) = S_1(\mathbf{\Theta}_1) - S_2(\mathbf{\Theta}_1)$ with $S_1(\mathbf{\Theta}_1) = \frac{1}{\tau}\sum|\theta_{ij}|$ and $S_2(\mathbf{\Theta}_1) = \frac{1}{\tau}\sum\max(|\theta_{ij}| - \tau, 0)$ are convex in $\mathbf{\Theta}_1$. Then a sequence of upper approximations of $J(\mathbf{\Theta}_1)$ is constructed: At iteration step m by $J^{(m)}(\mathbf{\Theta}_1) = \sum_{\theta_{ij} \in \mathbf{\Theta}_1}\left(\frac{|\theta_{ij}|}{\tau} I(|\hat{\theta}_{ij}^{(m-1)}| \leq \tau) + I(|\hat{\theta}_{ij}^{(m-1)}| > \tau)\right)$. This yields a sequence of convex minimization subproblems with convex constraints: At iteration step m, we solve

$$\min\nolimits_{\mathbf{\Theta}_1} \|\boldsymbol{A}\mathbf{\Theta}_1 - (\boldsymbol{Z} - \boldsymbol{A}\mathbf{\Theta}_2)\|_F^2, \quad \text{subject to} \quad J^{(m)}(\mathbf{\Theta}_1) \leq s_1. \tag{12}$$

For (12), we develop a gradient projection method. First, we generalize an l_1-ball result of (Liu & Ye, 2009) to (12).

Lemma 2 *(Projection) For any set $K \subseteq \{1, 2, \cdots, n\}$,*

$$\boldsymbol{x}^* = \mathcal{T}_{K,z}(\boldsymbol{v}) = \underset{\boldsymbol{x} \in \mathbb{R}^n : \sum_{i \in K}|x_i| \leq z}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}\|_2^2,$$

where $\mathcal{T}_{K,z} : \mathbb{R}^n \to \mathbb{R}^n$ is a projection operator defined by

$$\mathcal{T}_{K,z}(\boldsymbol{v})_i = sign(v_i)\max(|v_i| - \lambda^*, 0)$$

where $\lambda^* = 0$ if $\sum_{i \in K}|v_i| \leq z$ or $i \notin K$ and $\lambda^* = \frac{\sum_{i \in K \setminus K_0}|v_i| - z}{|K| - |K_0|}$ otherwise, and $K_0 = \{j : \sum_{i \in K}\max(|v_i| - |v_j|, 0) - z > 0\}$.

Before solving (12), we simply extend the fast iterative shrinkage-thresholding (FISTA) algorithm (Beck & Teboulle, 2009) to solving (13).

Lemma 3 *For any set K defined in Lemma 2, a global minimizer of*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n:\sum_{i\in K}|x_i|\le z} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 \tag{13}$$

can be obtained by FISTA iteratively: At iteration step t:

$$\begin{aligned}
\boldsymbol{x}^{(t)} &= \mathcal{T}_{K,z}\Big(\boldsymbol{y}^{(t)} - \frac{1}{2L}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{y}^{(t)}-\boldsymbol{b})\Big),\\
\rho_{t+1} &= \frac{1+\sqrt{1+4\rho_t^2}}{2},\\
\boldsymbol{y}^{(t+1)} &= \boldsymbol{x}^{(t)} + \left(\frac{\rho_t-1}{\rho_{t+1}}\right)(\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(k-1)}),
\end{aligned}$$

where L is the largest singular value of $\boldsymbol{A}$.

Next we solve (12) using Lemma 3, which yields an analytic updating formula in a matrix form.

Then a global minimizer of (12) is computed using an iterative scheme with respect to t as follows:

$$\begin{aligned}
&\boldsymbol{v}^{(1)} = \hat{\boldsymbol{\Theta}}_1^{(m,0)} = \hat{\boldsymbol{\Theta}}_1^{(m-1)}, \quad \rho_1 = 1,\\
&\hat{\boldsymbol{\Theta}}_1^{(m,t)} = \mathcal{T}_{K^{(m)},z^{(m)}}\left(\boldsymbol{v}^{(t)} - \frac{1}{2\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})}\boldsymbol{A}^T[\boldsymbol{A}\boldsymbol{v}^{(t)} - (\boldsymbol{Z}-\boldsymbol{A}\boldsymbol{\Theta}_2)]\right),\\
&\rho_{t+1} = \frac{1+\sqrt{1+4\rho_t^2}}{2}, \quad \boldsymbol{v}^{(t+1)} = \hat{\boldsymbol{\Theta}}_1^{(m,t)} + \left(\frac{\rho_t-1}{\rho_{t+1}}\right)(\hat{\boldsymbol{\Theta}}_1^{(m,t)} - \hat{\boldsymbol{\Theta}}_1^{(m,t-1)}),
\end{aligned} \tag{14}$$

where $K^{(m)} = \{(i,j) : |\hat{\theta}_{ij}^{(m-1)}| \le \tau\}$, $z^{(m)} = \tau(s_1 - \sum_{\theta_{ij}\in\boldsymbol{\Theta}_1} I(|\hat{\theta}_{ij}^{(m-1)}| > \tau))$ and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix.

The algorithm is summarized as follows.

Algorithm 2:

Step 1.(Initialization) Supply a good initial estimate $(\hat{\boldsymbol{\Theta}}_1^{(0)}, \hat{\boldsymbol{\Theta}}_2^{(0)})$ in (5). Specify precision $\delta > 0$.

Step 2.(Iteration) At iteration m, compute candidate $\hat{\boldsymbol{\Theta}}_2$ in (10) with $\boldsymbol{\Theta}_1 = \hat{\boldsymbol{\Theta}}_1^{(m-1)}$ and candidate $\hat{\theta}_{ij} \in \hat{\boldsymbol{\Theta}}_1$ in (14) with $\boldsymbol{A}\boldsymbol{\Theta}_2 = \boldsymbol{A}\hat{\boldsymbol{\Theta}}_2^{(m-1)}$.

Step 3.(Maximum block improvement) At each iteration m, determine which of the two candidates $(\hat{\boldsymbol{\Theta}}_1, \hat{\boldsymbol{\Theta}}_2^{(m-1)})$ and $(\hat{\boldsymbol{\Theta}}_1^{(m-1)}, \hat{\boldsymbol{\Theta}}_2)$ for updating according to the amounts of improvement. That is, update $(\hat{\boldsymbol{\Theta}}_1^{(m)}, \hat{\boldsymbol{\Theta}}_2^{(m)}) = (\hat{\boldsymbol{\Theta}}_1, \hat{\boldsymbol{\Theta}}_2^{(m-1)})$ if $f(\hat{\boldsymbol{\Theta}}_1, \hat{\boldsymbol{\Theta}}_2^{(m-1)}) \le f(\hat{\boldsymbol{\Theta}}_1^{(m-1)}, \hat{\boldsymbol{\Theta}}_2)$; update $(\hat{\boldsymbol{\Theta}}_1^{(m)}, \hat{\boldsymbol{\Theta}}_2^{(m)}) = (\hat{\boldsymbol{\Theta}}_1^{(m-1)}, \hat{\boldsymbol{\Theta}}_2)$ otherwise.

Step 4.(Stopping rule) Terminate if $|f(\hat{\boldsymbol{\Theta}}_1^{(m)}, \hat{\boldsymbol{\Theta}}_2^{(m)}) - f(\hat{\boldsymbol{\Theta}}_1^{(m-1)}, \hat{\boldsymbol{\Theta}}_2^{(m-1)})| \le \delta$. Denote by m^* the index at termination. The final estimate is

$$\hat{\boldsymbol{\Theta}}_1 = \hat{\boldsymbol{\Theta}}_1^{(m^*)}, \quad \hat{\boldsymbol{\Theta}}_2 = \hat{\boldsymbol{C}}\hat{\boldsymbol{F}},$$

where $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{F}}$ are defined in (10) with $\boldsymbol{\Theta}_1 = \hat{\boldsymbol{\Theta}}_1$.

2.4 Computational Properties

This section discusses computational properties of Algorithms 1 and 2. For nonconvex minimization, our methods may not guarantee a global minimizer for (3). However, the following lemma says that our solution of Algorithms 1 and 2 yields a stationary point of the cost function. Note that the scheme of maximum block improvement is essential for the result of Lemma 5.

Lemma 4 *The minimal cost function $f(\hat{\mathbf{\Theta}}_1^{(m)}, \hat{\mathbf{\Theta}}_2^{(m)})$ in Algorithm 1 is strictly decreasing in m before termination. Moreover, the solution is a stationary point of $f(\mathbf{\Theta}_1, \mathbf{\Theta}_2)$ in that $\theta_{ij}^{(*)} = \operatorname{argmin}_{\theta_{ij} \in \mathbf{\Theta}_k; k=1,2} f((\mathbf{\Theta_1^*}, \mathbf{\Theta_2^*}) \setminus \theta_{ij})$, where $(\mathbf{\Theta}_1, \mathbf{\Theta}_2) \setminus \theta_{ij}$ is the set of parameters of $(\mathbf{\Theta}_1, \mathbf{\Theta}_2)$ without one component θ_{ij} in $\mathbf{\Theta}_1$ or $\mathbf{\Theta}_2$, and $(\mathbf{\Theta_1}, \mathbf{\Theta_2})$ satisfy the constraints in* (5).

Lemma 5 *If $\boldsymbol{A}$ is of full rank, then $\hat{\mathbf{\Theta}}_1$ computed from* **Algorithm 2** *satisfies the constraints in* (12). *Moreover, the minimal cost function $f(\hat{\mathbf{\Theta}}_1^{(m)}, \hat{\mathbf{\Theta}}_2^{(m)})$ is strictly decreasing in m before termination. Finally, if the solution $(\hat{\mathbf{\Theta}}_1, \hat{\mathbf{\Theta}}_2)$ satisfies* (5) *and it is a stationary point of $f(\mathbf{\Theta}_1, \mathbf{\Theta}_2)$ in that*

$$\theta_{ij}^{(*)} = \underset{\theta_{ij} \in \mathbf{\Theta}_k; k=1,2}{\operatorname{argmin}} f((\mathbf{\Theta_1^*}, \mathbf{\Theta_2^*}) \setminus \theta_{ij}),$$

where $(\mathbf{\Theta}_1, \mathbf{\Theta}_2) \setminus \theta_{ij}$ is the set of parameters of $(\mathbf{\Theta}_1, \mathbf{\Theta}_2)$ without one component θ_{ij} in $\mathbf{\Theta}_1$ or $\mathbf{\Theta}_2$, and $(\mathbf{\Theta_1}, \mathbf{\Theta_2})$ satisfy the constraints in (5).

With regard to the computational complexity of Algorithms 1 and 2, the method of truncated SVD yields an approximated SVD with a complexity of $O(pk \log r + (p+k)r^2)$ operations (Halko et al., 2011). Sorting requires a complexity of $O(pk \log(pk))$. For FISTA, the convergence rate is $O(1/t^2)$ (Beck & Teboulle, 2009), where t is the number of iterations. Overall, the computational complexity of Algorithm 1 is $O(pk \log(pk) + (p+k)r^2)I_2$, while that of Algorithm 2 is $O((pk \log r + (p+k)r^2 + I_1/\varepsilon^2)I_2$, where ε denotes the precision specified in Algorithm 2, and I_1 and I_2 is the number of DC iteration and blockwise iteration, respectively. Based on our experience, I_1 and I_2 are about between 3 and 20.

3. Theory

This section drives a finite-sample probability error bound for reconstruction of the true $\mathbf{\Theta}^0$ by $\hat{\mathbf{\Theta}}^{L_0}$, which is a global minimizer of (3) in that $\hat{\mathbf{\Theta}}^{L_0} = \hat{\mathbf{\Theta}}_1^{L_0} + \hat{\mathbf{\Theta}}_2^{L_0}$. Note that existence of a global minimizer is assured by the fact that the cost function (3) is bounded blow by zero. Moreover, we will provide an insight into simultaneous pursuit of the low rank and sparsity structures through matrix decomposition by contrasting the proposed method with (s_1, s_2) against low rank approximation alone with $(s_1 = 0, s_2)$ and sparsity pursuit alone with $(s_1, s_2 = 0)$.

Let $\|\mathbf{\Theta}\|_\infty = \max_i \sum_j |\theta_{ij}|$ and $\|\mathbf{\Theta}\|_{\max} = \max_{ij} |\theta_{ij}|$ are the L_∞-norm and max norm respectively. Before proceeding, we define a parameter space Λ as $\{\mathbf{\Theta} = \mathbf{\Theta}_1 + \mathbf{\Theta}_2 : \|\mathbf{\Theta}_1\|_0 \leq s_1, \|\mathbf{\Theta}_1\|_{\max} \leq l_1, \mathbf{\Theta}_2 = \boldsymbol{C}\boldsymbol{F}, \max(\|\boldsymbol{C}\|_\infty, \|\boldsymbol{F}^T\|_\infty) \leq l_2\}$, where $l_1, l_2 > 0$ are constant, $\boldsymbol{C}$ is a $p \times s_2$ matrix, $\boldsymbol{F}$ is a $s_2 \times k$ matrix, $\boldsymbol{F}^T$ is the transport of $\boldsymbol{F}$ and $s_2 > 0$ is an upper bound of $r(\mathbf{\Theta}_2)$. Let $g(\mathbf{\Theta}, \boldsymbol{Z})$ be the probability density of $\boldsymbol{Z}$ with respect to dominating measure ν on Λ. Define the Hellinger distance between two densities as

$$h(\mathbf{\Theta}, \mathbf{\Theta}') = \frac{1}{2}\left(\int (g^{1/2}(\mathbf{\Theta}, \boldsymbol{Z}) - g^{1/2}(\mathbf{\Theta}', \boldsymbol{Z}))^2 d\nu\right)^{1/2}, \tag{15}$$

which will be used to measure estimation accuracy.

The following technical assumptions are made.

Assumption A: (Norm-relation) For any $\mathbf{\Theta}, \mathbf{\Theta}' \in \Lambda$ and any $\delta > 0$,

$$\int \sup_{\|\mathbf{\Theta}-\mathbf{\Theta}'\|_{\max} \le \delta} (g^{1/2}(\mathbf{\Theta}, y) - g^{1/2}(\mathbf{\Theta}', y))^2 d\nu(y) \le M^2\delta^2,$$

where M might depend on $p,\ k,\ s_1,\ s_2$ and l_1, l_2.

Assumption A specifies a norm relation between the metric $\|\cdot\|_{\max}$ over parameters and the Hellinger distance over the corresponding densities. This can be verified given a specific form of g.

Theorem 1 gives a probability error bound for $\hat{\mathbf{\Theta}}^{L_0}$ under probability P under the true $\mathbf{\Theta}^0$. Let (s_1^0, s_2^0) be the degree of sparsity and rank, as defined in $\mathrm{Eff}(\mathbf{\Theta}^0)$ in Lemma 1.

Theorem 6 *Under* ***Assumptions A****, for any* $\epsilon \ge \epsilon_{n,p,k}$

$$P\left(h(\hat{\mathbf{\Theta}}^{L_0}, \mathbf{\Theta}^0) \ge \epsilon\right) \le 5\exp(-c_1 n\epsilon^2),$$

$\epsilon_{n,p,k} = \frac{C_{p,k}}{\sqrt{n}}\sqrt{\log(\frac{\sqrt{n}}{C_{p,k}})}$ *with*

$$C_{p,k} = c_2\sqrt{\log(2^9 M c_4(l_2^3 + l_1))}\sqrt{(p+k)s_2^0 + s_1^0} + c_2\sqrt{s_1^0 \log\left(e\frac{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{s_1^0}\right)}. \tag{16}$$

If $\log(r(\mathbf{\Theta}^0)) \le d s_2^0$ *for some* $d > 0$, *then it can be simplified:*

$$C_{p,k} = c_3\sqrt{\log(M)}\sqrt{(p+k-s_2^0)s_2^0},$$

where $c_1 - c_3$ *are positive constants and* M *is defined in Assumption A. Moreover, as* $n, p, k \to \infty$, $h^2(\hat{\mathbf{\Theta}}^{L_0}, \mathbf{\Theta}^0) = O_p(\epsilon_{n,p,k}^2)$, *and* $Eh^2(\hat{\mathbf{\Theta}}^{L_0}, \mathbf{\Theta}^0) = O(\epsilon_{n,p,k}^2)$, *where* $O_p(\cdot)$ *and* E *denote the stochastic order and the expectation under* P.

Corollary 1 gives an order of $\epsilon_{n,p,k}$ in three extreme situations with M held fixed.

Corollary 1 *Suppose* M *in Assumptions A is a constant independent of* (p, k, s_1, s_2).

(i) When $\mathbf{\Theta}^0$ *is extremely sparse, that is,* $\|\mathbf{\Theta}^0\|_0 \le p + k - 2$, $C_{p,k}$ *in* (16) *is no worse than* $O\left(\sqrt{\|\mathbf{\Theta}^0\|_0 \log((p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)/\|\mathbf{\Theta}^0\|_0)}\right)$.

(ii) When $\mathbf{\Theta}^0$ *is a low-rank matrix,* $C_{p,k}$ *in* (16) *is no worse than* $O\left(\sqrt{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}\right)$.

(iii) *When* $\mathbf{\Theta}^0$ *is dense, say* $\|\mathbf{\Theta}^0\|_0 \geq cpk$ *for a constant* $0 < c \leq 1$, *and of full rank,* $C_{p,k}$ *in* (16) *is* $O\left(\max\left(\sqrt{(p+k-s_2^0)s_2^0}, \sqrt{s_1^0 \log(\frac{pk}{s_1^0})}\right)\right)$.

Then $C^L_{p,k} = O\left(\sqrt{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}\right)$.

Corollary 2 and Theorem 2 give a similar result under the Hellinger distance and the Kullback-Leibler distance, respectively, assuming that $\boldsymbol{\epsilon}_i$ follows a normal distribution.

Corollary 2 *If* $\boldsymbol{\epsilon}_i$ *in* (1) *follows* $N(0, \sigma^2 I_{k\times k})$, $\|\boldsymbol{A}\|_\infty$ *is bounded, then the results in Corollary 1 continue to hold.*

Theorem 7 *Under the same assumptions in Corollary 2, we have, for any* $\epsilon \geq \epsilon_{n,p,k}$,

$$P\left(K(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0}) \geq 4\epsilon^2\right) \leq 5\exp(-c_1 n\epsilon^2).$$

where $K(\cdot,\cdot)$ *is Kullback-Leibler distance under normality and* $\epsilon_{n,p,k}$ *and* c_2 *remain to be the same as in Theorem 1. As* $n, p, k \to \infty$, $K(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0}) = O_p(\epsilon^2_{n,p,k})$ *and* $EK(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L^0}) = O(\epsilon^2_{n,p,k})$.

Theorem 3 gives an error bound for $\|\hat{\mathbf{\Theta}}^{L_0} - \mathbf{\Theta}^0\|_F^2$ under the normal assumption when $\boldsymbol{A} = \boldsymbol{I}_{n\times p}$.

Theorem 8 *Assume that* $\boldsymbol{A} = \boldsymbol{I}_{n\times p}$ *with* $n = \max(p, k)$. *Under the same assumptions in Corollary 2 with* $\sigma = O(\frac{1}{\sqrt{\max(p,k)}})$, *as* $n, p, k \to \infty$, $\|\hat{\mathbf{\Theta}}^{L_0} - \mathbf{\Theta}^0\|_F^2 = O_p(C'_{p,k}\log(\frac{1}{C'_{p,k}}))$, *where*

$$C'_{p,k} = \frac{\log(\max(p,k)) \cdot [(p+k)s_2^0 + s_1^0] + s_1^0 \log\left(e^{\frac{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{s_1^0}}\right)}{\max(p^2, k^2)}$$

4. Numerical Examples

This section examines operating characteristics of the proposed method through simulations, and demonstrates its effectiveness on applications in image reconstruction and in time series analysis. In the literature, it is known that the state-of-art methods are the low-rank approximation method subject to rank restriction as well as its regularized version, which outperforms the low-rank approximation method with the trace-norm (Xing et al., 2012; She, 2013; Zhou & Tao, 2011). In Section 4.1, we contrast our proposed method with pursuing low rank and sparsity structures through matrix decomposition simultaneously, with the former low rank approximation method subject to rank restriction (low-rank alone), as well as the method based on sparsity pursuit alone (sparsity alone). Here Algorithm 2 are used. Most importantly, in Section 4.2, we compare the proposed method using Algorithm 1 with two strong competitors the method of Go Decomposition (GoDec, Zhou & Tao, 2011) and the method augmented Lagrange multipliers (ALM, Lin et al., 2009) when $\boldsymbol{A} = \boldsymbol{I}_{n\times p}$ in (2). In simulations, codes for ALM and GoDec are used at the authors' website, and the initial values for Algorithms 1 and 2 are set to be the zero-matrix

4.1 Simulation I: Operating Characteristics

The simulated example is generated as follows. First, a $n \times p$ design matrix $\boldsymbol{A}$ is sampled with each entry being iid $N(0,1)$. Second, the true $\boldsymbol{\Theta}_1$ is a $p \times k$ matrix with all diagonals one and two more non-zeros (2 and 2) being randomly chosen with equal probability, and the true $\boldsymbol{\Theta}_2$ is generated by multiplying a $p \times r$ matrix with a $r \times k$ matrix with each entry following $N(1,1)$. Moreover, each entry of $\boldsymbol{E}$ is iid $N(0, 0.25)$. Throughout the simulations, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are held fixed with different values of (n, p, k).

The proposed method is trained with a training set, and the optimal tuning parameters, minimizing the prediction mean squares error over an independent tuning set, are obtained through a bisection search over integer values. Then a method's performance is examined over a test set. The training, tuning and testing data sizes are n, $4n$ and $2n$.

For parameter estimation, we employ the mean squares error to evaluate performance

$$\frac{1}{4n}\|\boldsymbol{A}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^0)\|_F^2. \tag{17}$$

For rank recovery, we calculate the absolute difference between an estimated rank $\hat{r}$ and the true rank r_0, that is $|\hat{r} - r_0|$. For sparsity pursuit, we define the true positive (TP) as a ratio of the true positive numbers of nonzero estimates over the number of nonzeros in the true model, and the false positive (FP) as a ratio of the false positive numbers of nonzero estimates over the number of zeros in the true model. Here "Low rank alone", "Sparsity alone" and "Ours" indicate the low rank method subject to rank restriction, the sparsity pursuit method, and the proposed method

As indicated in Table 1, the proposed method performs favorably against its counterpart—the low rank approximation method subject to rank restriction and sparsity pursuit alone, across all situations with different values of n, p and k. Moreover, the proposed method enables to identify two structures through matrix decomposition simultaneously. In particular, it recovers the true rank of the matrix with nearly zero $|\hat{r} - r_0|$-values as compared to relatively large $|\hat{r} - r_0|$-values, ranging from 6.7 to 29.6, for its low-rank counterpart. At the same time, the proposed method has high true positives ranging from .92 to 1.00 and low false positives between 0.00 and 0.01, as compared to true positives ranging 0.04 to .44 and false positives between 0.03 and 0.20 of its counterpart based on sparsity pursuit. This suggests that pursuit of two types of structures is indeed advantageous than that of either one structure individually. This is mainly because these two structures are complementary to each other. As a result, higher parameter estimation accuracy, as measured by the MSE values, can be realized. In fact, the amount of improvement is large, which ranges from 147% to 1185400%. To see how each method performs as (n, p) increases, we fix $k = 5$.

As suggested by Table 2, the proposed method yields more stable performance than its two counterparts whose performance deteriorates rapidly, as the level of difficulty of a problem escalates when p and k increase.

4.2 Simulation II: Comparison

To compare with ALM (Lin et al., 2009) and GoDec (Zhou & Tao, 2011) for RPCA, consider the case of $\boldsymbol{A} = \boldsymbol{I}_{n\times p}$ in (2) and $p = k$ as in these papers. GoDec minimizes

$$\min_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2} \|\boldsymbol{Z} - \boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_F^2 \quad \text{subject to } \mathrm{card}(\boldsymbol{\Theta}_1) \le s_1,\ \mathrm{rank}(\boldsymbol{\Theta}_2) \le s_2, \tag{18}$$

	Ours				Low-rank alone		Sparsity alone		
				$p = 20,\ k = 10$					
n	$\lvert \hat{r} - r_0 \rvert$	TP	FP	MSE	$\lvert \hat{r} - r_0 \rvert$	MSE	TP	FP	MSE
50	0.00	1.00	0.01	0.68	6.71	1.68	0.44	0.07	1367.03
	(0.00)	(0.00)	(0.03)	(0.15)	(0.52)	(0.28)	(0.29)	(0.01)	(173.50)
				$p = 30,\ k = 20$					
	Ours				Low-rank alone		Sparsity alone		
n	$\lvert \hat{r} - r_0 \rvert$	TP	FP	MSE	$\lvert \hat{r} - r_0 \rvert$	MSE	TP	FP	MSE
50	0.00	1.00	0.00	1.54	15.69	7.79	0.12	0.14	4650.35
	(0.00)	(0.00)	(0.00)	(0.30)	(2.41)	(1.03)	(0.21)	(0.02)	(511.55)
100	0.00	1.00	0.00	0.51	16.94	2.16	0.13	0.05	4399.38
	(0.00)	(0.00)	(0.00)	(0.08)	(0.24)	(0.18)	(0.22)	(0.01)	(429.41)
				$p = 20,\ k = 30$					
	Ours				Low-rank alone		Sparsity alone		
n	$\lvert \hat{r} - r_0 \rvert$	TP	FP	MSE	$\lvert \hat{r} - r_0 \rvert$	MSE	TP	FP	MSE
50	0.00	1.00	0.00	1.06	16.66	5.06	0.43	0.06	4276.25
	(0.00)	(0.00)	(0.00)	(0.17)	(0.76)	(0.62)	(0.28)	(0.01)	(508.06)
100	0.00	1.00	0.00	0.46	16.99	1.88	0.53	0.06	4087.58
	(0.00)	(0.00)	(0.00)	(0.05)	(0.10)	(0.16)	(0.20)	(0.01)	(406.97)
				$p = 40,\ k = 30$					
	Ours				Low-rank alone		Sparsity alone		
n	$\lvert \hat{r} - r_0 \rvert$	TP	FP	MSE	$\lvert \hat{r} - r_0 \rvert$	MSE	TP	FP	MSE
50	0.00	1.00	0.00	4.08	1.88	19.39	0.09	0.20	12018.68
	(0.00)	(0.00)	(0.00)	(1.21)	(0.59)	(1.57)	(0.20)	(0.04)	(1422.84)
				$p = 50,\ k = 20$					
	Ours				Low-rank alone		Sparsity alone		
n	$\lvert \hat{r} - r_0 \rvert$	TP	FP	MSE	$\lvert \hat{r} - r_0 \rvert$	MSE	TP	FP	MSE
100	0.00	1.00	0.00	0.95	16.86	5.05	0.04	0.03	11262.97
	(0.00)	(0.00)	(0.00)	(0.15)	(0.35)	(0.40)	(0.14)	(0.01)	(1003.69)
				$p = 200,\ k = 100$					
	Ours				Low-rank alone		Sparsity alone		
n	$\lvert \hat{r} - r_0 \rvert$	TP	FP	MSE	$\lvert \hat{r} - r_0 \rvert$	MSE			
300	3.76	0.92	0.00	8.26	29.56	54.24	–	–	–
	(1.24)	(0.23)	(0.00)	(0.86)	(7.84)	(0.81)	(–)	(–)	(–)

Table 1: Results of Simulation I. Algorithm 2 is used for computation.

where card$(\cdot)$ denotes the cardinality, and $s_j \geq 0$ are tuning parameters as in our case. Similarly, ALM that focuses on the non-noisy situation minimizes

$$\min_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2} \|\boldsymbol{\Theta}_2\|_* + \lambda \sum_{\theta_{ij} \in \boldsymbol{\Theta}_1} |\theta_{ij}|, \quad \text{subject to } \boldsymbol{Z} = \boldsymbol{\Theta}_1 + \boldsymbol{\Theta}_2, \tag{19}$$

where $\| \cdot \|_*$ is the nuclear-norm of a matrix.

Our simulation example remains the same as before except that the positions of nonzero elements in $\boldsymbol{\Theta}_2$ are randomly sampled with equal probability, in particular, $.1p$ and $.3p$ nonzeros are randomly chosen without replacement. For tuning, grid search is employed for GoDec in (18), with $1 \leq s_1 \leq (p + k)$ and $1 \leq s_2 \leq \min(p, k, 50)$; λ is fixed at $\frac{1}{\sqrt{p}}$ for (19).

		Ours				Low-rank alone		Sparsity alone		
n	p	$\lvert\hat{r}-r_0\rvert$	TP	FP	MSE	$\lvert\hat{r}-r_0\rvert$	MSE	TP	FP	MSE
50	20	0.00	1.00	0.002	0.58	2.00	0.84	0.433	0.08	570
		(0.00)	(0.00)	(0.006)	(0.14)	(0.00)	(0.18)	(0.30)	(0.02)	(73)
50	30	0.00	0.57	0.01	1.29	1.97	1.98	0.18	0.08	3772.33
		(0.00)	(0.17)	(0.01)	(0.32)	(0.17)	(0.42)	(0.27)	(0.01)	(542.38)
50	40	0.00	1.00	0.001	3.57	1.67	5.43	0.07	0.05	1998
		(0.00)	(0.00)	(0.003)	(1.58)	(0.60)	(1.73)	(0.18)	(0.01)	(257)
50	50	0.82	0.36	0.01	487.43	0.82	12255	0.05	0.03	3797
		(0.84)	(0.38)	(0.01)	(1081.68)	(0.81)	(79570)	(0.15)	(0.01)	(539)
100	20	0.00	1.00	0.01	0.23	2.00	0.32	0.53	0.08	541
		(0.00)	(0.00)	(0.03)	(0.05)	(0.00)	(0.05)	(0.21)	(0.02)	(58)
100	30	0.00	0.71	0.01	0.36	2.00	0.54	0.19	0.03	1461
		(0.00)	(0.25)	(0.01)	(0.05)	(0.00)	(0.08)	(0.21)	(0.01)	(147)
100	40	0.00	0.98	0.01	0.53	2.00	0.83	0.10	0.03	1929
		(0.00)	(0.10)	(0.02)	(0.08)	(0.00)	(0.11)	(0.20)	(0.01)	(179)

Table 2: Results for Simulation I with fixed $k = 5$. Algorithm 2 is used for computation.

From Tables 3, it is evidenced that the proposed method outperforms ALM uniformly in terms of the MSE while being comparable to GoDec, in all the situations with different values of (p, k, σ). Moreover, it always recovers the true rank of the matrix perfectly with $|\hat{r} - r_0| = 0$. Although ALM has comparable high TP values, its FP values are high as well in that they are at least 0.6488. As a result, ALM never captures the true rank.

4.3 AR Face Database 20pt Markup

For face image reconstruction, we use a subset of *AR Face Data* for this experiment. The original image is available at `http://www-prima.inrialpes.fr/FGnet/data/05-ARFace/markup_large.png`, which is a colored one with size of $186 \times 200 \times 3$. To enable detailed testing, the image has been labeled with 20 facial features on the face. We convert the image into black and white and reduce it to size 171×180. The target image is displayed in Figure 1.

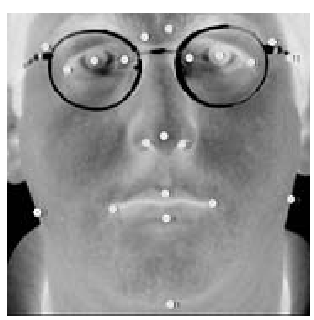

Figure 1: The converted AR face image with markup points.

Twenty one markup points around eyes, nose, mouth and cheeks, which are used to test face recognition or verification performance when the exact location of the face and features

nonzeros	p	k	σ	Method	$\lvert\hat{r}-r_0\rvert$	TP	FP	MSE
0.1p	50	30	0.1	Ours	0.0000	0.9940	0.0000	0.2366
					(0.0000)	(0.0343)	(0.0002)	(0.0251)
				ALM	13.0300	1.000	0.6488	1.5057
					(0.6735)	(0.0000)	(0.0082)	(0.0576)
				GoDec	0.0000	0.9940	0.0000	0.2363
					(0.0000)	(0.0342)	(0.0001)	(0.0245)
			1	Ours	0.0000	0.0320	0.0000	2.5308
					(0.0000)	(0.0839)	(0.0001)	(0.2418)
				ALM	13.3900	0.9280	0.6540	15.0569
					(0.6651)	(0.1223)	(0.0080)	(0.5758)
				GoDec	0.0000	0.0300	0.0001	2.5537
					(0.0000)	(0.0823)	(0.0003)	(0.2523)
	200	100	0.1	Ours	0.0000	0.9770	0.0000	0.2345
					(0.0000)	(0.0337)	(0.0000)	(0.0169)
				ALM	54.3100	1.0000	0.7034	4.9984
					(0.7745)	(0.0000)	(0.0022)	(0.0510)
				GoDec	0.0000	0.9755	0.0000	0.2330
					(0.0000)	(0.0344)	(0.0000)	(0.0160)
			1	Ours	0.0000	0.0075	0.0000	2.4469
					(0.0000)	(0.0206)	(0.0000)	(0.1387)
				ALM	54.2400	0.9456	0.7059	49.9838
					(0.7264)	(0.0456)	(0.0023)	(0.5095)
				GoDec	0.0000	0.0085	0.0000	2.4476
					(0.0000)	(0.0236)	(0.0000)	(0.1395)
0.3p	50	30	0.1	Ours	0.0000	0.9933	0.0002	0.2507
					(0.0000)	(0.0201)	(0.0003)	(0.0277)
				ALM	13.0000	1.0000	0.6472	1.5057
					(0.6195)	(0.0000)	(0.0079)	(0.0576)
				GoDec	0.0000	0.9953	0.0001	0.2489
					(0.0000)	(0.0171)	(0.0003)	(0.0271)
			1	Ours	0.0000	0.0373	0.0000	2.8870
					(0.0000)	(0.0624)	(0.0001)	(0.2410)
				ALM	13.37	0.9407	0.6531	15.0569
					(0.6301)	(0.0621)	(0.0080)	(0.5758)
				GoDec	0.0000	0.0327	0.0001	2.8983
					(0.0000)	(0.0653)	(0.0002)	(0.2504)
	200	100	0.1	Ours	0.0000	0.9867	0.0001	0.2495
					(0.0000)	(0.0164)	(0.0001)	(0.0198)
				ALM	54.3500	1.0000	0.7030	4.9984
					(0.6571)	(0.0000)	(0.0023)	(0.0510)
				GoDec	0.0000	0.9882	0.0000	0.2479
					(0.0000)	(0.0152)	(0.0001)	(0.0191)
			1	Ours	0.0000	0.0080	0.0000	2.8254
					(0.0000)	(0.0122)	(0.0000)	(0.1402)
				ALM	54.2200	0.9467	0.7054	49.9838
					(0.6289)	(0.0297)	(0.0022)	(0.5095)
				GoDec	0.0000	0.0075	0.0000	2.8237
					(0.0000)	(0.0135)	(0.0000)	(0.1409)

Table 3: Results for Simulation II . Algorithm 1 is used for computation.

are known. To identify the locations, we extract sparse ($\boldsymbol{\Theta}_1$) and low-rank ($\boldsymbol{\Theta}_2$) structures for the face images as described by the matrix decomposition into $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$. For this purpose, $\boldsymbol{A}$ in (3) is set to be the identity matrix of size 171×171. Figures 2 and 3 display two decomposed structures for the AR face images by the proposed method with different sparse and rank constraint parameters in (3).

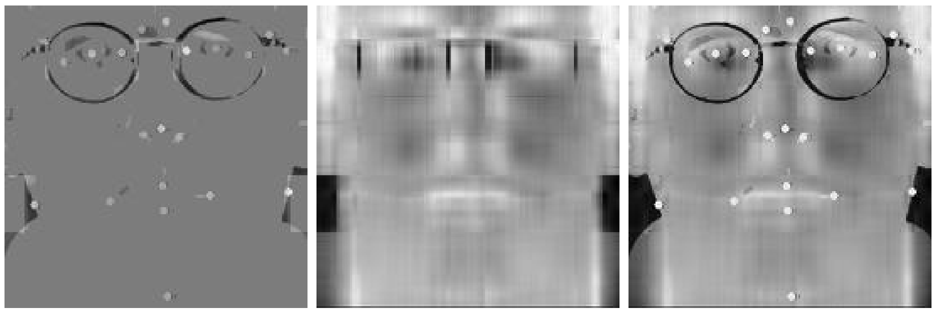

Figure 2: Extracted sparsity (first), low-rank (second) structures as well as the reconstructed image by the proposed method for AR face images; where the tuning parameters are set to $s_1 = 2500$, $s_2 = 5$.

Figure 3: Extracted sparsity (first), low-rank (second) structures as well as the reconstructed image by the proposed method for AR face images; where the tuning parameters are set to $s_1 = 2100$, $s_2 = 10$.

As indicated in Figures 2 and 3, the sparseness structure describes characteristics/detailed marks of the face, whereas the low-rank structure displays the rough outlook of the human face. This confirms our discussion regarding local and global features in the Introduction. Visually, both the first panels in Figures 2 and 3 preserve at least 60% markup points, especially the points around nose two sides of face and lip. In other words, the sparsity structure captures most of markup points. Similarly, the second panels retain the overall look of the face. Most interestingly, this decomposition tends to remove the glasses from the human face.

4.4 Greek Letters Image Reconstruction

Now consider a 26×31 black-white image of two Greek letters β and ϕ, where its noisy version is obtained by adding noise $N(0, 1)$ after dividing the original matrix values by 100. The ratio of the maximum value of the image to the noise standard deviation is about 2.5. The images are displayed in Figure 4.

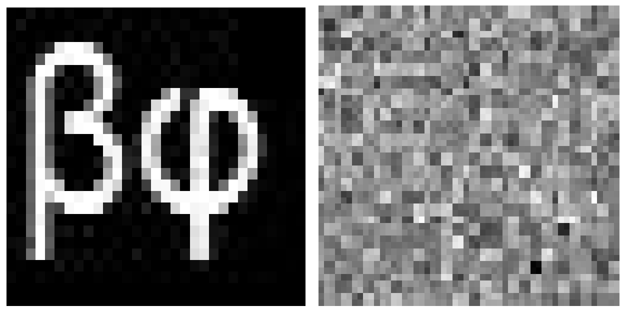

Figure 4: Original image (left) versus its noisy version (right).

Our goal is reconstruction of the original image from its noise version, with a focus on restoration of detailed structures of the letters. Towards this end, we apply the proposed method and contrast with its counterpart based on sparse pursuit alone and low-rank approximations. Specifically, let $\boldsymbol{A}$ to be the identity matrix of size 31×31 and $\boldsymbol{\Theta}$ be a 31×26 parameter matrix in (3). For each method, grid search is performed for tuning, with $s_1 = (10, 20, 30, 50)$, $1 \leq s_2 \leq \min(p, k) = 26$ and $\tau = (0.05, 0.1, 0.2)$. For each method, the 10-$fold$ cross-validation is employed. The reconstructed images are displayed in Figure 5.

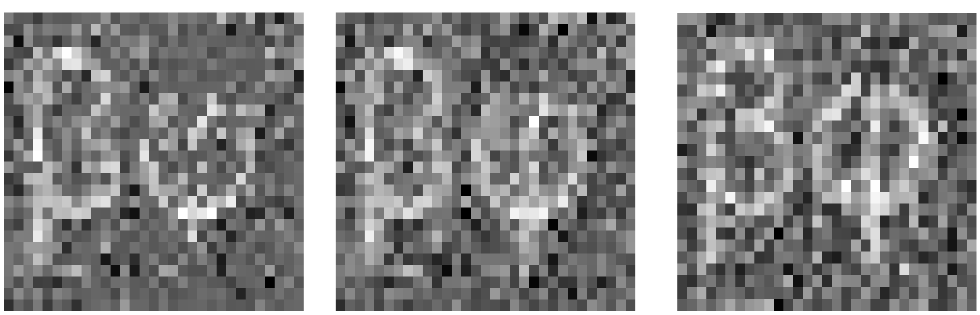

Figure 5: Reconstructed images based on sparsity alone (first), low-rank alone (second) and our method (third). Algorithm 2 is used for computation.

Visually, the first two reconstructed images by the low-rank method and the sparsity method give the rough shape of two letters, but the letters β and ϕ not distinguishable with blurred segments in places, especially the right middle of β and the top of ϕ. By comparison, the third reconstructed image by our method enables to reconstruct the complete shape of these two letters, and yield the best quality of reconstruction.

4.5 US Macroeconomic Time Series

This subsection examines multiple time series data described in (Stock & Watson, 2012). The data measures 143 US macroeconomic variables quarterly over a time span from February 1, 1959 to November 1, 2008. These variables are categorized into 13 groups and are summarized in Table 4.

Group	Description	Examples of series	# series
1	GDP component	GDP, consumption, investment	16
2	IP	IP, capacity utilization	14
3	Employment	Sectoral&total employment and hours	20
4	Unemployment rate	Unemployment rate, total and by duration	7
5	Housing	Housing starts, total and by region	6
6	Inventories	NAPM inventories, new orders	6
7	Prices	Price indexes, aggregate&disaggregate, commodity prices	37
8	Wages	Average hourly earning, unit labor cost	6
9	Interest rates	Treasuries, corporate, term spreads, public-private spreads	13
10	Money	M1, M2, business loans, consumer credit	7
11	Exchange rates	Average&selected trading partners	5
12	Stock prices	Various stock price indexes	5
13	Consumer expectations	Michigan consumer expectations	1

Table 4: Economic indicators collected for U.S. macroeconomic time series.

For data analysis, we consider time series starting from August 1, 1959 to November 1, 2008 due to incomplete initial observations. Our goal is one-step ahead forecasting, and contrast the proposed method with low-rank alone and sparsity alone in terms of forecasting accuracy. Using a multivariate autoregressive model, that is, $\boldsymbol{y}_t = \boldsymbol{y}_{t-1}^T\boldsymbol{\Theta} + \boldsymbol{\epsilon}_i$, we place it in the framework of (1), where $\boldsymbol{y}_t$ is a vector that records the values of various macroeconomic variables at time point t, and $\boldsymbol{\epsilon}_i$ follows normal distribution. In the presence of multiplicity and non-stationarity for economics data like this, we consider some transformations. For instance, log growth rates for quantity variables are differenced, nominal interest rates are differenced, as well as the logarithms of changes in rates of inflation for price series are differenced. See (Stock & Watson, 2012) for processing the data set. For this data set, $p = k = 143$ in (1) and the design matrix $\mathbf{A}$ is specified by the time series, which can written as $\boldsymbol{A} = (\boldsymbol{y}_{t_0}, \boldsymbol{y}_{t_0+1}, \ldots, \boldsymbol{y}_{t_0+d-1})^T$.

A one-step ahead K-fold cross validation (CV) criterion is used for tuning the time series (Arlot & Celisse, 2010). In particular, for design matrix $\boldsymbol{A}$, at each fold i, we use observations i to $n-K+i-1$ for training and the observation $n-K+i$ for tuning, where K is a pre-assigned integer and $K-1$ indicates the number of folds. Note that the values of p and k are close to the sample size n for this time series. We therefore choose $K \leq 20$ to maintain adequate training samples.

For tuning, the CV is optimized over a set of grids for $s_1 = (10, 20, 50, 100, 200)$, $1 \leq s_2 \leq \min(p, k)$ and $\tau = (0.02, 0.05, 0.1, 0.2)$. The results for $K = 11$ are reported in Table 5. The results for other K values are omitted due to similarity.

As suggested by Table 5, the proposed method outperforms its counterparts pursuing sparseness and low-rank alone. The amount of improvement over the low rank method and

	Ours	Low-rank alone	Sparsity alone
$K = 11$	301.22	348.02	3111.89

Table 5: Prediction errors of U.S. macroeconomic data for $K = 11$. Here "Low rank alone", "Sparsity alone" and "Ours" indicate our method for low rank pursuit only, for sparsity pursuit only and for simultaneous pursuit of low rank and sparsity. Algorithm 2 is used for computation.

the sparsity method is 15% and 933%, respectively. The Q-Q plots in Figure 6 indicate that the model assumption is adequate although some departure from normality has been detected. Overall, the proposed method performs reasonably well.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation Grants IIS-0953662 and DMS-0906616, and National Institute of Health Grants R01 LM010730, 2R01GM081535 and 1R01HL105397. The authors thank the editor, the action editor and three referees for their helpful comments and suggestions.

Appendix

Proof of Lemma 1: Let $df(s, r) = s + (p + k - r)r$. By definition of the effective degrees of freedom, we obtain that

$$\mathrm{Eff}(\boldsymbol{\Theta}) \leq \min(df(0, r(\boldsymbol{\Theta}^0)), df(\|\boldsymbol{\Theta}^0\|_0, 0)).$$

To prove uniqueness in terms of (s, r), suppose there exist $(\bar{s}, \bar{r}) \neq (\bar{s}', \bar{r}')$ such that $df(\bar{s}, \bar{r}) = df(\bar{s}', \bar{r}') = \min_{s,r} df(s, r)$. Without loss of generality, assume $\bar{r} = \bar{r}' - n_0 < \bar{r}'$, where $n_0 > 0$ is a positive integer. If $n_0 \leq \min(p, k) - \bar{r}$ and $\bar{r} < \min(p, k)$, then $\bar{s} + (p + k - \bar{r})\bar{r} = \bar{s}' + (p + k - \bar{r}')\bar{r}'$ implies that $\bar{s} = \bar{s}' + n_0(p + k - 2\bar{r} - n_0) \geq n_0(p + k - 2\bar{r} - n_0) > p + k - 2\bar{r} - 1$, which contradicts with the assumption that $s < p + k - 2r - 1$. Otherwise, if $\bar{r} = \min(p, k)$, $\bar{s}$ must be zero. This completes the proof. ■

Proof of Lemma 2: Let $x_i = v_i$ for $i \notin K$. Then the problem reduces to the standard l_1 ball problem.

$$\operatorname*{argmin}_{\sum_{i \in K} |x_i| \leq z} \frac{1}{2} \sum_{i \in K} (x_i - v_i)^2.$$

The results follows by the proof of Theorem 1 of (Liu & Ye, 2009). ■

Proof of Lemma 3: It suffices to derive the basic step of ISTA in (Amit et al., 2007) for (13). Consider the following quadratic approximation of problem (13) at a given point $\boldsymbol{y}$:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n : \sum_{i \in K} |x_i| \leq z} Q_L(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}\|_2^2 + \langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}) \rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \tag{20}$$

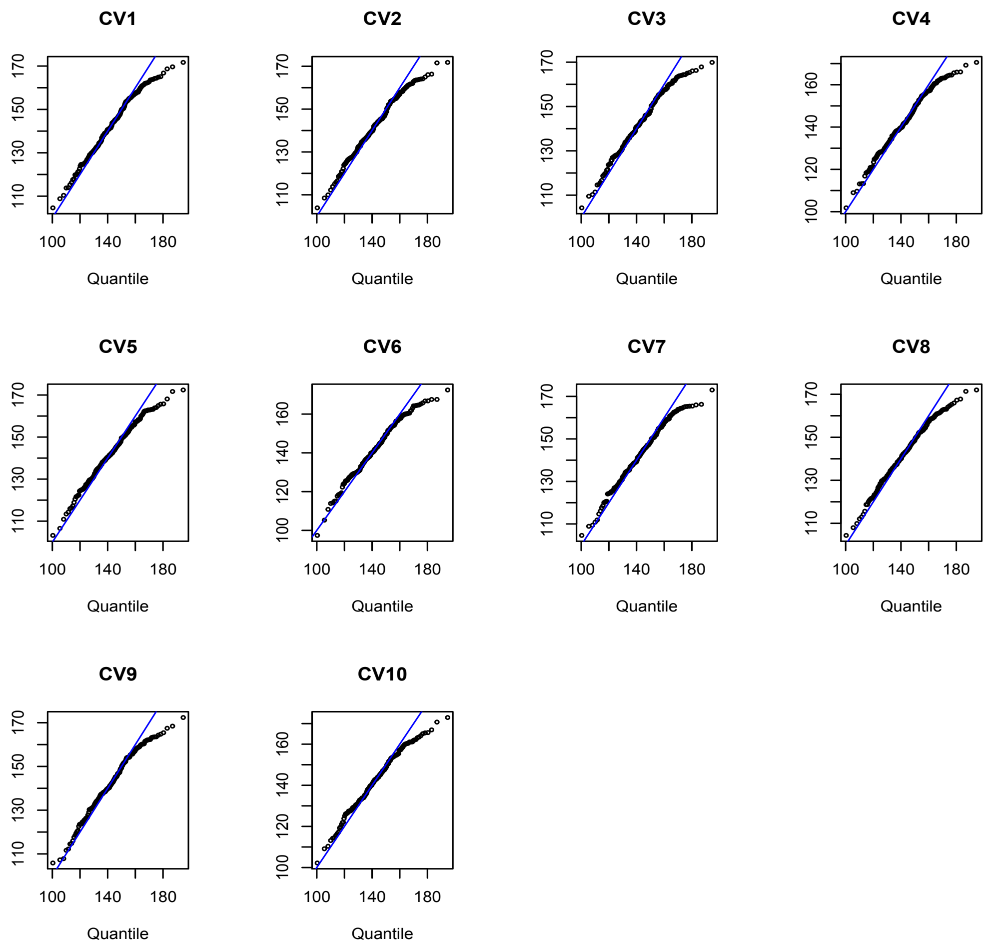

Figure 6: Q-Q plots for each-fold in U.S. macroeconomic time series data example, where points on a straight line indicates non-departure from normality.

where L is a Lipschitz constant of the function $\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})$ with respect to $\boldsymbol{x}$. Solving (20) is equivalent to that of

$$\min_{\boldsymbol{x}\in\mathbb{R}^n:\sum_{i\in K}|x_i|\leq z} \|\boldsymbol{x}-\big(\boldsymbol{y}-\frac{1}{L}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{b})\big)\|_2^2.$$

By Lemma 2, the solution is $\mathcal{T}_{K,z}\big(\boldsymbol{y}-\frac{1}{L}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{b})\big)$. The basic step of ISTA thus can be written as $\boldsymbol{x}^{(t)}=\mathcal{T}_{K,z}\big(\boldsymbol{x}^{(t-1)}-\frac{1}{L}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}^{(t-1)}-\boldsymbol{b})\big)$. Then, Lemma 3 follows by taking L to be $\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})$, where $\lambda_{\max}(\cdot)$ denotes the largest singular value. ■

Proof of Lemma 4: By (6) and (7), for any integer $m\geq 1$,

$$f(\hat{\boldsymbol{\Theta}}_1^{(m)},\hat{\boldsymbol{\Theta}}_2^{(m)})\geq f(\hat{\boldsymbol{\Theta}}_1^{(m)},\hat{\boldsymbol{\Theta}}_2^{(m+1)})\geq f(\hat{\boldsymbol{\Theta}}_1^{(m+1)},\hat{\boldsymbol{\Theta}}_2^{(m+1)}).$$

Meanwhile, it follows from (6) that

$$\begin{aligned}
f(\hat{\boldsymbol{\Theta}}_1^{(m)},\hat{\boldsymbol{\Theta}}_2^{(m)}) &= \|\boldsymbol{Z}-\hat{\boldsymbol{\Theta}}_2^{(m)}\|_F^2-\|\hat{\boldsymbol{\Theta}}_1^{(m)}\|_F^2\\
&\geq \|\boldsymbol{Z}-\hat{\boldsymbol{\Theta}}_2^{(m+1)}-\hat{\boldsymbol{\Theta}}_1^{(m)}\|_F^2\\
&\geq \|\boldsymbol{Z}-\hat{\boldsymbol{\Theta}}_2^{(m+1)}\|_F^2-\|\hat{\boldsymbol{\Theta}}_1^{(m)}\|_F^2.
\end{aligned}$$

Therefore $\|\boldsymbol{Z}-\hat{\boldsymbol{\Theta}}_2^{(m)}\|_F^2$ is lower bounded and decreasing in m. Moreover, by the monotone properties of $f(\hat{\boldsymbol{\Theta}}_1^{(m)},\hat{\boldsymbol{\Theta}}_2^{(m)})$, $\|\hat{\boldsymbol{\Theta}}_1^{(m)}\|_F^2$ converges as $m\to\infty$. Then there exists a subsequence $\{m_k\}$ such that $(\hat{\boldsymbol{\Theta}}_1^{(m_k)},\hat{\boldsymbol{\Theta}}_2^{(m_k)})\to(\hat{\boldsymbol{\Theta}}_1^{(m^*)},\hat{\boldsymbol{\Theta}}_2^{(m^*)})$.

Let $R_{ij}(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2)\in\operatorname{argmin}_{\theta_{ij}\in\boldsymbol{\Theta}_1 \text{ or } \theta_{ij}\in\boldsymbol{\Theta}_2} f((\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2)\setminus\theta_{ij})$. Let the cost function for θ_{ij} to be $f_m(\theta_{ij})=f((\hat{\boldsymbol{\Theta}}_1^{(m)},\hat{\boldsymbol{\Theta}}_2^{(m)})\setminus\theta_{ij})$, where other components of $(\hat{\boldsymbol{\Theta}}_1^{(m)},\hat{\boldsymbol{\Theta}}_2^{(m)})$ are held fixed. Then

$$\begin{aligned}
f_{m_k}(R_{ij}(\hat{\boldsymbol{\Theta}}_1^{(m^*)},\hat{\boldsymbol{\Theta}}_2^{(m^*)})) &\geq f_{m_k}(R_{ij}(\hat{\boldsymbol{\Theta}}_1^{(m_k)},\hat{\boldsymbol{\Theta}}_2^{(m_k)}))\\
&\geq \min\Big(f((\hat{\boldsymbol{\Theta}}_1^{(m_k)},\hat{\boldsymbol{\Theta}}_2^{(m_k+1)})),f((\hat{\boldsymbol{\Theta}}_1^{(m_k)},\hat{\boldsymbol{\Theta}}_2^{(m_k)}))\Big)\\
&\geq f((\hat{\boldsymbol{\Theta}}_1^{(m_k+1)},\hat{\boldsymbol{\Theta}}_2^{(m_k+1)})).
\end{aligned}$$

As $m\to\infty$, by continuity of $f(\cdot)$, $f_{(m^*)}(R_{ij}(\hat{\boldsymbol{\Theta}}_1^{(m^*)},\hat{\boldsymbol{\Theta}}_2^{(m^*)}))\geq f(\hat{\boldsymbol{\Theta}}_1^{(m^*)},\hat{\boldsymbol{\Theta}}_2^{(m^*)}))$, where the equality holds by the definition of R_{ij}. Hence, for each $\theta_{ij}\in\boldsymbol{\Theta}_l$; $l=1,\ 2$, $\hat{\theta}_{ij}^{(m^*)}=R_{ij}(\hat{\boldsymbol{\Theta}}_1^{(m^*)},\hat{\boldsymbol{\Theta}}_2^{(m^*)})$ is the optimal componentwise solution. The results of Lemma 4 then follow. ■

Proof of Lemma 5: First we prove that $\hat{\boldsymbol{\Theta}}_1^{(m)}$ satisfies

$$\sum_{\theta_{ij}\in\boldsymbol{\Theta}_1}\left(\frac{|\hat{\theta}_{ij}^{(m)}|}{\tau}I(|\hat{\theta}_{ij}^{(m)}|\leq\tau)+I(|\hat{\theta}_{ij}^{(m)}|>\tau)\right)\leq s_1. \tag{21}$$

Toward this end, we rewrite the left side of (21) as

$$\begin{aligned}
&\sum_{\theta_{ij}\in\mathbf{\Theta}_1}\left(\frac{|\hat{\theta}_{ij}^{(m)}|}{\tau}I(|\hat{\theta}_{ij}^{(m-1)}|\leq\tau)+I(|\hat{\theta}_{ij}^{(m-1)}|>\tau)\right)\\
&+\sum_{\theta_{ij}\in\mathbf{\Theta}_1}\left(\frac{|\hat{\theta}_{ij}^{(m)}|}{\tau}I(|\hat{\theta}_{ij}^{(m)}|\leq\tau)+I(|\hat{\theta}_{ij}^{(m)}|>\tau)-\frac{|\hat{\theta}_{ij}^{(m)}|}{\tau}I(|\hat{\theta}_{ij}^{(m-1)}|\leq\tau)-I(|\hat{\theta}_{ij}^{(m-1)}|>\tau)\right)\\
&=\sum_{\theta_{ij}\in\mathbf{\Theta}_1}\left(\frac{|\hat{\theta}_{ij}^{(m)}|}{\tau}I(|\hat{\theta}_{ij}^{(m-1)}|\leq\tau)+I(|\hat{\theta}_{ij}^{(m-1)}|>\tau)\right)+I_m, \qquad (22)
\end{aligned}$$

where $I_m=\sum_{\theta_{ij}\in\mathbf{\Theta}_1}\frac{|\hat{\theta}_{ij}^{(m)}|-\tau}{\tau}\left(I(|\hat{\theta}_{ij}^{(m)}|\leq\tau)-I(|\hat{\theta}_{ij}^{(m-1)}|\leq\tau)\right)$. Note that it follows from the DC construction that

$$\sum_{\theta_{ij}\in\mathbf{\Theta}_1}\left(\frac{|\hat{\theta}_{ij}^{(m)}|}{\tau}I(|\hat{\theta}_{ij}^{(m-1)}|\leq\tau)+I(|\hat{\theta}_{ij}^{(m-1)}|>\tau)\right)\leq s_1.$$

Thus, to establish (21), we only need to prove $I_m\leq 0$. Rewrite I as

$$I_m=\begin{cases}0 & \text{if } \min(|\hat{\theta}_{ij}^{(m)}|,|\hat{\theta}_{ij}^{(m-1)}|)>\tau \text{ or } \max(|\hat{\theta}_{ij}^{(m)}|,|\hat{\theta}_{ij}^{(m-1)}|)\leq\tau,\\ \sum_{\theta_{ij}\in\mathbf{\Theta}_1}(\frac{|\hat{\theta}_{ij}^{(m)}|}{\tau}-1) & \text{if } |\hat{\theta}_{ij}^{(m)}|\leq\tau \text{ and } |\hat{\theta}_{ij}^{(m-1)}|>\tau,\\ -\sum_{\theta_{ij}\in\mathbf{\Theta}_1}(\frac{|\hat{\theta}_{ij}^{(m)}|}{\tau}-1) & \text{if } |\hat{\theta}_{ij}^{(m)}|>\tau \text{ and } |\hat{\theta}_{ij}^{(m-1)}|\leq\tau,\end{cases}$$

implying that $I_m\leq 0$. Then, (21) follows.

For stationarity, note that it follows from (21) that

$$f(\hat{\mathbf{\Theta}}_1^{(m-1)},\hat{\mathbf{\Theta}}_2^{(m-1)})\geq f(\hat{\mathbf{\Theta}}_1^{(m-1)},\hat{\mathbf{\Theta}}_2)\geq f(\hat{\mathbf{\Theta}}_1^{(m)},\hat{\mathbf{\Theta}}_2^{(m)}),$$

where $\hat{\mathbf{\Theta}}_2$ is defined in Step 2 of **Algorithm 2**.

Suppose that termination index m^* is infinite. Then we will prove that $\hat{\mathbf{\Theta}}_1^{(m)}\to\hat{\mathbf{\Theta}}_1^{(m^*)}$ as $m\to m^*=\infty$. When $m^*=\infty$, $\hat{\mathbf{\Theta}}_1^{(m)}$ must be updated infinitely because $\hat{\mathbf{\Theta}}_2^{(m)}$ is analytically solved. First consider, at step m, $\mathbf{\Theta}_1$ is updated whereas $\mathbf{\Theta}_2=\hat{\mathbf{\Theta}}_2^{(m)}$. Denote by $\Lambda(\mathbf{\Theta}_1,\mathbf{\Theta}_2,\lambda^*)$ the dual problem of (12), where λ^* is the optimal Lagrange multiplier and $\mathbf{\Theta}_2=\hat{\mathbf{\Theta}}_2^m$. Then

$$\begin{aligned}
f(\hat{\mathbf{\Theta}}_1^{(m)},\hat{\mathbf{\Theta}}_2^{(m)})-f(\hat{\mathbf{\Theta}}_1^{(m+1)},\hat{\mathbf{\Theta}}_2^{(m+1)})&=\Lambda(\hat{\mathbf{\Theta}}_1^{(m)},\hat{\mathbf{\Theta}}_2^{(m)},\lambda^*)-\Lambda(\hat{\mathbf{\Theta}}_1^{(m+1)},\hat{\mathbf{\Theta}}_2^{(m)},\lambda^*)\\
&\quad-\lambda^*\sum_{\theta_{ij}\in\mathbf{\Theta}_1}\left(\frac{|\hat{\theta}_{ij}^{(m)}|}{\tau}I(|\hat{\theta}_{ij}^{(m)}|\leq\tau)+I(|\hat{\theta}_{ij}^{(m)}|>\tau)-s_1\right)
\end{aligned}$$

The equality holds because $\hat{\mathbf{\Theta}}_1^{(m+1)}$ is the global minimizer of a convex problem (12), attaining at constraint boundaries, i.e $\sum_{\theta_{ij}\in\mathbf{\Theta}_1}\left(\frac{|\hat{\theta}_{ij}^{(m+1)}|}{\tau}I(|\hat{\theta}_{ij}^{(m)}|\leq\tau)+I(|\hat{\theta}_{ij}^{(m)}|>\tau)-s_1\right)=0$.

An application of the Taylor expansion to $\Lambda(\mathbf{\Theta}_1, \hat{\mathbf{\Theta}}_2^{(m)}, \lambda^*)$ at $\mathbf{\Theta}_1 = \hat{\mathbf{\Theta}}_1^{(m+1)}$ yields that

$$\begin{aligned} & f(\hat{\mathbf{\Theta}}_1^{(m)}, \hat{\mathbf{\Theta}}_2^{(m)}) - f(\hat{\mathbf{\Theta}}_1^{(m+1)}, \hat{\mathbf{\Theta}}_2^{(m+1)}) \\ &= \langle \frac{\partial \Lambda}{\partial \mathbf{\Theta}_1}(\hat{\mathbf{\Theta}}_1^{(m+1)}, \hat{\mathbf{\Theta}}_2^{(m)}, \lambda^*), \hat{\mathbf{\Theta}}_1^{(m)} - \hat{\mathbf{\Theta}}_1^{(m+1)} \rangle + \frac{1}{2} \langle \boldsymbol{A}(\hat{\mathbf{\Theta}}_1^{(m)} - \hat{\mathbf{\Theta}}_1^{(m+1)}), \boldsymbol{A}(\hat{\mathbf{\Theta}}_1^{(m)} - \hat{\mathbf{\Theta}}_1^{(m+1)}) \rangle \\ &\quad - \lambda^* \sum_{\theta_{ij} \in \mathbf{\Theta}_1} \left(\frac{|\hat{\theta}_{ij}^{(m)}|}{\tau} I(|\hat{\theta}_{ij}^{(m)}| \leq \tau) + I(|\hat{\theta}_{ij}^{(m)}| > \tau) - s_1 \right), \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. The first term in the right side of the equality is zero, because $\hat{\mathbf{\Theta}}_1^{(m+1)}$ is the global minimizer and the third term is no less than zero by (21). Thus,

$$\begin{aligned} f(\hat{\mathbf{\Theta}}_1^{(m)}, \hat{\mathbf{\Theta}}_2^{(m)}) - f(\hat{\mathbf{\Theta}}_1^{(m+1)}, \hat{\mathbf{\Theta}}_2^{(m+1)}) &\geq \frac{1}{2} \langle \boldsymbol{A}(\hat{\mathbf{\Theta}}_1^{(m)} - \hat{\mathbf{\Theta}}_1^{(m+1)}), \boldsymbol{A}(\hat{\mathbf{\Theta}}_1^{(m)} - \hat{\mathbf{\Theta}}_1^{(m+1)}) \rangle \\ &\geq \frac{\lambda_{\min}(\boldsymbol{A}^T \boldsymbol{A})}{2} \|\hat{\mathbf{\Theta}}_1^{(m)} - \hat{\mathbf{\Theta}}_1^{(m+1)}\|_F^2, \end{aligned} \tag{23}$$

where $\lambda_{\min(\cdot)}$ is the smallest eigenvalue of a matrix. Therefore $f(\hat{\mathbf{\Theta}}_1^{(m)}, \hat{\mathbf{\Theta}}_2^{(m)})$ is lower bounded and decreasing in m, implying $f(\hat{\mathbf{\Theta}}_1^{(m)}, \hat{\mathbf{\Theta}}_2^{(m)})$ converges to some limit f^* as $m \to \infty$. By (23), convergence of $\hat{\mathbf{\Theta}}_1^{(m)} \to \hat{\mathbf{\Theta}}_1^{(m^*)}$ is established. Next consider the case in which $\mathbf{\Theta}_2$ is only updated finitely, say before step m_0, using the same notation with proof of Lemma 4, then for any $m > m_0$

$$f_m(R_{ij}(\hat{\mathbf{\Theta}}_1^{(m^*)}, \hat{\mathbf{\Theta}}_2^{(m^*)})) \geq f_m(R_{ij}(\hat{\mathbf{\Theta}}_1^{(m)}, \hat{\mathbf{\Theta}}_2^{(m)})) = f((\hat{\mathbf{\Theta}}_1^{(m+1)}, \hat{\mathbf{\Theta}}_2^{(m+1)})).$$

The second equality holds because the MBI is employed. As $m \to m^*$, by continuity of function f, $f_{(m^*)}(R_{ij}(\hat{\mathbf{\Theta}}_1^{(m^*)}, \hat{\mathbf{\Theta}}_2^{(m^*)})) \geq f(\hat{\mathbf{\Theta}}_1^{(m^*)}, \hat{\mathbf{\Theta}}_2^{(m^*)}))$, where the equality holds by the definition of R_{ij}. Finally, we consider the case in which $\mathbf{\Theta}_2$ is updated infinitely. Then there is a subsequence $\{m_k\}$ such that $\hat{\mathbf{\Theta}}_2^{(m_k)} \to \hat{\mathbf{\Theta}}_2^{(m^*)}$. Similarly, $f_{m^*}(R_{ij}(\hat{\mathbf{\Theta}}_1^{(m^*)}, \hat{\mathbf{\Theta}}_2^{(m^*)})) = f(\hat{\mathbf{\Theta}}_1^{(m^*)}, \hat{\mathbf{\Theta}}_2^{(m^*)}))$. Hence, for each $\theta_{ij} \in \mathbf{\Theta}_l$, $l = 1,\ 2$, $\hat{\theta}_{ij}^{(m^*)} = R_{ij}(\hat{\mathbf{\Theta}}_1^{(m^*)}, \hat{\mathbf{\Theta}}_2^{(m^*)})$ is the optimal componentwise solution. The results of Lemma 5 then follow. ∎

Let $\mathcal{B}_{S,r} = \{\mathbf{\Theta} = \mathbf{\Theta}_1^S + \mathbf{\Theta}_2 : r(\mathbf{\Theta}_2) = r\} \cap \Lambda$, a sub-parameter space with known sparsity structure S and rank r. Denote $H(\cdot, \Lambda)$ and $H^B(\cdot, \Lambda)$ to be the L_∞ entropy and bracketing Hellinger metric entropy for set Λ, respectively. The next two technical lemmas concern the size of the parameter space.

Lemma 9 *Suppose that Assumptions A is met.*

$$H^B(t, \mathcal{B}_{S,r}) \leq |S| \log(2Ml_1/t) + (p+k) r \log(2Ml_2^3/t),$$

where l_1, l_2 are constant and $M > 1$ is defined in Assumption A.

Lemma 10 *Suppose that Assumptions A is satisfied. If $s_1 = s_1^0$, $s_2 = s_2^0$, then*

$$\begin{aligned} H^B(t, \Lambda) \leq & 2(p+k) s_2^0 \log(2Ml_2^3 \epsilon/t) + s_1^0 \log((1 + 2Ml_1)/t) \\ & + 2 s_1^0 \log \left(e \frac{(p + k - r(\mathbf{\Theta}^0)) r(\mathbf{\Theta}^0)}{s_1^0} \right). \end{aligned}$$

Proof of Lemmas 6 and 7: For Lemma 6, note that $\Lambda = \cup_{|S| \leq s_1^0} \cup_{r \leq s_2^0} \mathcal{B}_{S,r}$. It suffices to calculate the entropy for each $\mathcal{B}_{S,r}$.

Let $\Lambda_2 = \{(\mathbf{\Theta}_1, \mathbf{\Theta}_2) : \mathbf{\Theta}_1, \mathbf{\Theta}_2$ satisfy conditions defined in $\Lambda\}$. For $\mathbf{\Theta} = \mathbf{\Theta}_1 + \mathbf{\Theta}_2$ and $(\mathbf{\Theta}_1, \mathbf{\Theta}_2) \in \Lambda_2$, define $\mathcal{B}_\delta(\mathbf{\Theta}_1, \mathbf{\Theta}_2) = \{(\mathbf{\Theta}_1', \mathbf{\Theta}_2') \in \Lambda_2 : \|\mathbf{\Theta}_1 - \mathbf{\Theta}_1'\|_{\max} + \|\mathbf{\Theta}_2 - \mathbf{\Theta}_2'\|_{\max} \leq \delta\}$ to be the neighborhood of $(\mathbf{\Theta}_1, \mathbf{\Theta}_2)$. For any $\mathbf{\Theta}' = \mathbf{\Theta}_1' + \mathbf{\Theta}_2'$ with $(\mathbf{\Theta}_1', \mathbf{\Theta}_2') \in \mathcal{B}_\delta(\mathbf{\Theta}_1, \mathbf{\Theta}_2)$, by Assumption A,

$$\int \sup_{\mathcal{B}_\delta(\mathbf{\Theta}_1, \mathbf{\Theta}_2)} (g^{1/2}(\mathbf{\Theta}, y) - g^{1/2}(\mathbf{\Theta}', y))^2 d\nu(y) \leq M^2 \delta^2.$$

Combined the above with Lemma 2.1 of (Ossiander, 1987), we have

$$H^B(t, \mathcal{B}_{S,r}) \leq H(M^{-1}t, \mathcal{B}_{S,r}). \tag{24}$$

Since $\|\mathbf{\Theta}_1\|_{\max}$ is bounded by l_1, by constructing a $2t$-net on $\mathcal{B}_{S,r}$ through the outer product of the t-nets on $\mathbf{\Theta}_1^S$ and $\mathbf{\Theta}_2$ defined in the parameter space Λ, we can show that

$$H(M^{-1}t, \mathcal{B}_{S,r}) \leq |S| \log(2Ml_1/t) + H_r(M^{-1}t) \tag{25}$$

where $|S|$ is the number of nonzeros in $\mathbf{\Theta}_1$ and $H_r(M^{-1}t)$ is the entropy for $\mathbf{\Theta}_2$ with rank r. Let $\boldsymbol{C}$ be a basis of column of $\mathbf{\Theta}_2$, then there exists an $k \times r$ matrix $\boldsymbol{F}$ such that $\mathbf{\Theta}_2 = \boldsymbol{CF}$. Hence

$$\|\mathbf{\Theta}_2 - \mathbf{\Theta}_2'\|_{\max} = \|\boldsymbol{CF} - \boldsymbol{C}'\boldsymbol{F}'\|_{\max} \leq \|\boldsymbol{C}\|_\infty \|\boldsymbol{F} - \boldsymbol{F}'\|_{\max} + \|\boldsymbol{F}^T\|_\infty \|\boldsymbol{C} - \boldsymbol{C}'\|_{\max}.$$

where $\|\mathbf{\Theta}_{p\times k}\|_\infty = \max_{1\leq i \leq p} \sum_{i=1}^k |\theta_{ij}|$ is the L_∞ matrix-norm and $\|\mathbf{\Theta}\|_{\max} = \max_{\theta_{ij} \in \mathbf{\Theta}} |\theta_{ij}|$ is the max norm. Note that $\|\boldsymbol{C}\|_\infty$ and $\|\boldsymbol{F}^T\|_\infty$ are bounded by l_2. This yields

$$H_r(M^{-1}t) \leq (p+k) r \log \frac{2l_2^3 M}{t}.$$

This, together with (24) and (25), implies Lemma 6.

For Lemma 7, note that

$$\begin{aligned}
&\exp(H^B(t, \Lambda)) \leq \exp(H(M^{-1}t, \Lambda)) \\
&= \sum_{r=0}^{s_2^0} \sum_{|S|=0}^{s_1^0} \sum_{i=0}^{|S|} \binom{s_1^0}{i} \binom{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0) - s_1^0}{|S| - i} \exp(H(M^{-1}t, \mathcal{B}_{S,r})) \\
&\leq \binom{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{s_1^0} \left(\sum_{|S|=0}^{s_1^0} \binom{s_1^0}{|S|} (2Ml_1/t)^{|S|} \right) \left(\sum_{r=0}^{s_2^0} (2Ml_2^3/t)^{(p+k)r} \right) \\
&\equiv \binom{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{s_1^0} \times I \times II.
\end{aligned}$$

Note that $\sum_{k=0}^n \binom{n}{k} a^k b^{n-k} = (a+b)^n$. Then $I = (1+\frac{2Ml_1}{t})^{s_1^0}$ and $II \leq (s_2^0+1) \left(\frac{2Ml_2^3 \epsilon}{t}\right)^{(p+k)s_2^0}$. Thus,

$$\begin{aligned}
H^B(t, \Lambda) &\leq \log \binom{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{s_1^0} + \log(s_2^0+1) + s_1^0 \log(1 + \frac{2Ml_1}{t}) + (p+k)s_2^0 \log(\frac{2Ml_2^3}{t}) \\
&\leq 2(p+k)s_2^0 \log(2Ml_2^3/t) + s_1^0 \log(\frac{1+2Ml_1}{t}) + 2s_1^0 \log\left(e \frac{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{s_1^0}\right),
\end{aligned}$$

where e is the natural number and $0 < t < 1$. The last inequality follows Theorem 2.6 of (Stanica & Montgomery, 2001) that $\binom{b}{a} \leq \frac{b^{b+1/2}}{\sqrt{2\pi}a^{a+1/2}(b-a)^{b-a+1/2}} \leq \exp((a+1/2)\log(b/a) + a) \leq \exp(2a\log(b/a)+a)$ for any integer $0 < a < b$. This completes the proof. ■

Proof of Theorem 1: We apply a large deviation inequality in Theorem 2 of (Wong & Shen, 1995). To this end, we verify (1.2) there. By Lemma 7,

$$\int_{\epsilon^2/2^8}^{2^{1/2}\epsilon} \left(H^B(t/c_4, \Lambda)\right)^{1/2} dt \leq \int_{\epsilon^2/2^8}^{2^{1/2}\epsilon} \sqrt{2(p+k)s_2^0 \log(2Ml_2^3 c_4/t) + s_1^0 \log((1+2Ml_1)c_4/t)} dt$$
$$+ \int_{\epsilon^2/2^8}^{2^{1/2}\epsilon} \sqrt{2s_1^0 \log\left(e\frac{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{s_1^0}\right)} dt \equiv I_1 + I_2,$$

for some constant $c_4 > 0$, say $c_4 = 10$. Then, for ϵ small,

$$\begin{aligned} I_1 &\leq \sqrt{2}\epsilon\sqrt{2(p+k)s_2^0 \log(2^9 Ml_2^3 c_4/\epsilon^2) + s_1^0 \log((1+2Ml_1)2^8 c_4/\epsilon^2)} \\ &\leq 2\epsilon\sqrt{(p+k)s_2^0 + s_1^0}\sqrt{\log(2^9 Mc_4(l_2^3 + l_1)) + 2\log\frac{1}{\epsilon}} \\ &\leq 2\sqrt{2}\epsilon\sqrt{\log(2^9 Mc_4(l_2^3+l_1))}\sqrt{(p+k)s_2^0 + s_1^0} \cdot \sqrt{\log\frac{1}{\epsilon}}. \end{aligned}$$

Similarly,

$$I_2 \leq 2\epsilon\sqrt{s_1^0 \log\left(e\frac{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{s_1^0}\right)}.$$

Let $\epsilon_{n,p,k} = \frac{C_{p,k}}{\sqrt{n}}\log(\frac{C_{p,k}}{\sqrt{n}})$ where

$$C_{p,k} = 2\sqrt{2}c_5^{-1}\sqrt{\log(2^9 Mc_4(l_2^3+l_1))}\sqrt{(p+k)s_2^0 + s_1^0} + 2c_5^{-1}\sqrt{s_1^0 \log\left(e\frac{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{s_1^0}\right)}.$$

Then, for any $\epsilon \geq \epsilon_{n,p,k}$ and $c_5 = \frac{512}{(2/3)^{5/12}}$

$$\int_{\epsilon^2/2^8}^{2^{1/2}\epsilon} \left(H^B(t/c_4, \Lambda)\right)^{1/2} dt \leq c_5^{-1}\sqrt{n}\epsilon^2.$$

By Theorem 2 of (Wong & Shen, 1995), $P\left(h(\hat{\mathbf{\Theta}}^{L_0}, \mathbf{\Theta}^0) \geq \epsilon\right) \leq 5\exp(-c_1 n\epsilon^2)$, which yields $Eh^2(\hat{\mathbf{\Theta}}^{L_0}, \mathbf{\Theta}^0) = O(\epsilon_{n,p,k}^2)$ by using the fact that $h(\hat{\mathbf{\Theta}}^{L_0}, \mathbf{\Theta}^0) \leq 1$.

Consider a special situation when $\log(r(\mathbf{\Theta}^0)) \leq ds_2^0$ for some constant $d > 0$ that is independent of p, k. Note that $s_1^0 < p+k-s_2^0$ and $p+k-r(\mathbf{\Theta}^0) \leq p+k-s_2^0$. Then

$$\begin{aligned} s_1^0 \log\left(e\frac{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{s_1^0}\right) &\leq (p+k-s_2^0)\log\left(e\frac{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{p+k-s_2^0}\right) \\ &\leq (p+k-s_2^0)\log(er(\mathbf{\Theta}^0)) \leq 2d(p+k-s_2^0)s_2^0. \end{aligned}$$

Thus, $I_1 + I_2$ is upper bounded by

$$2\sqrt{2}\epsilon\left(\sqrt{\log(2^9 Mc_4(l_2^3+l_1))}+\sqrt{d}\right)\sqrt{(p+k)s_2^0+s_1^0}\cdot\sqrt{\log\frac{1}{\epsilon}}.$$

Let $c_3 = 2\sqrt{2}c_5^{-1}\left(\sqrt{\log(2^9 c_4(l_2^3+l_1))}+\sqrt{d}\right)\sqrt{(p+k)s_2^0+s_1^0}$. The result then follows. This completes the proof. ■

Proof of Corollary 1: If $\mathbf{\Theta}^0$ is sparse and $\|\mathbf{\Theta}^0\|_0 \le p+k-2$, then by the definition of effective degrees of freedom $s_0 = s_1^0 + (p+k-s_2^0)s_2^0 \le \|\mathbf{\Theta}^0\|_0$. This implies that

$$\begin{aligned}
C_{p,k} &= O\left(\sqrt{\|\mathbf{\Theta}^0\|_0}\right) + O\left(\sqrt{\|\mathbf{\Theta}^0\|_0 \log\big(p+k-r(\mathbf{\Theta}^0)\big)r(\mathbf{\Theta}^0)/\|\mathbf{\Theta}^0\|_0\big)}\right)\\
&= O\left(\sqrt{\|\mathbf{\Theta}^0\|_0 \log\big(p+k-r(\mathbf{\Theta}^0)\big)r(\mathbf{\Theta}^0)/\|\mathbf{\Theta}^0\|_0\big)}\right).
\end{aligned}$$

The second inequality is because of nondecreasingness of $\sqrt{x}$ and $\sqrt{x\log(a/x)}$ in x for $x \le a/e$.

If $\mathbf{\Theta}^0$ is low-rank, we have

$$C_{p,k} = O\left(\sqrt{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)} + \sqrt{s_1^0\log\big((p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)/s_1^0\big)}\right).$$

Note that $s_1^0\log\big((p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)/s_1^0\big) \le \log\big((p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)/e\big)$. The result follows.

If $\mathbf{\Theta}^0$ is dense and of full rank, then $(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)$ is of order $O(pk)$. Hence $C_{p,k}$ can be written as $O\left(\sqrt{(p+k-s_2^0)s_2^0}+\sqrt{s_1^0\log(\frac{pk}{s_1^0})}\right)$. This completes the proof. ■

Proof of Corollary 2: It suffices to show the Assumption A is met. Let $f(\boldsymbol{\mu}_i, \boldsymbol{y}) = \frac{1}{(\sqrt{2\pi}\sigma)^k}\exp\big(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{\mu}_i)^T(\boldsymbol{y}-\boldsymbol{\mu}_i)\big)$ for $i=1,2$. $\boldsymbol{\mu}_1 = \boldsymbol{a}^T\mathbf{\Theta}$ and $\boldsymbol{\mu}_2 = \boldsymbol{a}^T\mathbf{\Theta}'$. Then

$$\begin{aligned}
&\int \sup_{\|\mathbf{\Theta}-\mathbf{\Theta}'\|_{\max}\le\delta}(f^{1/2}(\boldsymbol{\mu}_1,\boldsymbol{y})-f^{1/2}(\boldsymbol{\mu}_2,\boldsymbol{y}))^2 d\boldsymbol{y}\\
&\le 2-2\frac{1}{(\sqrt{2\pi}\sigma)^k}\int \inf_{\|\mathbf{\Theta}-\mathbf{\Theta}'\|_{\max}\le\delta}\exp\left(-\frac{\|\boldsymbol{y}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\|_2^2+\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_2^2/2}{2\sigma^2}\right)d\boldsymbol{y}\\
&\le 2-2\inf_{\|\mathbf{\Theta}-\mathbf{\Theta}'\|_{\max}\le\delta}\exp\left(-\frac{\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_2^2}{4\sigma^2}\right)\\
&\le \frac{(\|\boldsymbol{a}\|_1)^2\|\mathbf{\Theta}-\mathbf{\Theta}'\|_{\max}^2}{4\sigma^2} \le \frac{(\|\boldsymbol{a}\|_1)^2\delta^2}{4\sigma^2}.
\end{aligned}$$

The second inequality follows from the invariance property of the normal distribution. Corollary 2 follows when $\|\boldsymbol{a}\|_1$ is bounded. This completes the proof. ■

Proof of Theorem 2: After some calculations, we obtain that

$$\begin{aligned}
h^2(\mathbf{\Theta}, \mathbf{\Theta}^0) &= 2\left(1 - \prod_{i=1}^n \frac{1}{(\sqrt{2\pi}\sigma)^k}\int \exp\big[-\frac{1}{4\sigma^2}(\|\boldsymbol{y}_i - \boldsymbol{a}_i^T\mathbf{\Theta}\|^2 + \|\boldsymbol{y}_i - \boldsymbol{a}_i^T\mathbf{\Theta}^0\|^2)\big]d\boldsymbol{y}\right)\\
&= 2\left(1 - \prod_{i=1}^n \exp\big[-\frac{1}{8\sigma^2}\|\boldsymbol{a}_i^T(\mathbf{\Theta} - \mathbf{\Theta}^0)\|^2\right)\\
&= 2\left(1 - \exp(-\frac{1}{8\sigma^2}\|\boldsymbol{A}(\mathbf{\Theta} - \mathbf{\Theta}^0)\|_F^2)\right),\\
K(\mathbf{\Theta}^0, \mathbf{\Theta}) &= \frac{1}{2\sigma^2}\|\boldsymbol{A}(\mathbf{\Theta} - \mathbf{\Theta}^0)\|_F^2.
\end{aligned}$$

When $\epsilon < 1$,

$$\begin{aligned}
P(K(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0}) \geq 4\epsilon^2) &= P\left(\frac{1}{8\sigma^2}\|\boldsymbol{A}(\hat{\mathbf{\Theta}}^{L_0} - \mathbf{\Theta}^0)\|_F^2 \geq \epsilon^2\right)\\
&\leq P\left(\frac{1}{8\sigma^2}\|\boldsymbol{A}(\hat{\mathbf{\Theta}}^{L_0} - \mathbf{\Theta}^0)\|_F^2 \geq -\log(1 - \frac{\epsilon^2}{2})\right)\\
&= P\left(2\left(1 - \exp(-\frac{1}{8\sigma^2}\|\boldsymbol{A}(\hat{\mathbf{\Theta}}^{L_0} - \mathbf{\Theta}^0)\|_F^2)\right) \geq \epsilon^2\right)\\
&= P\left(h^2(\hat{\mathbf{\Theta}}^{L_0}, \mathbf{\Theta}^0) \geq \epsilon^2\right).
\end{aligned}$$

For any $\epsilon \geq \epsilon_{n,p,k}$, it follows from Theorem 1 and Corollary 2 that

$$\begin{aligned}
EK(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0}) &\leq EK(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0})I\{K(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0}) \leq 4\epsilon^2\} + EK(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0})I\{K(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0}) > 4\epsilon^2\}\\
&\leq 4\epsilon^2 + \left(EK^2(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0})\right)^{1/2}\left(P(K^2(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0}) > 4\epsilon^2)\right)^{1/2}.
\end{aligned}$$

By the triangle inequality, $\|\boldsymbol{A}\mathbf{\Theta}^0 - \boldsymbol{A}\hat{\mathbf{\Theta}}^{L_0}\|_F - \|\boldsymbol{\epsilon}\|_F \leq \|\boldsymbol{A}\mathbf{\Theta}^0 + \boldsymbol{\epsilon} - \boldsymbol{A}\hat{\mathbf{\Theta}}^{L_0}\|_F$. Note that $\hat{\mathbf{\Theta}}^{L_0}$ is a global minimizer of (3). Then $\|\boldsymbol{A}\mathbf{\Theta}^0 + \boldsymbol{\epsilon} - \boldsymbol{A}\hat{\mathbf{\Theta}}^{L_0}\|_F \leq \|\boldsymbol{\epsilon}\|_F$. Hence

$$K(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0}) = \frac{1}{2\sigma^2}\|\boldsymbol{A}(\mathbf{\Theta}^0 - \hat{\mathbf{\Theta}}^{L_0})\|_F^2 \leq \frac{2}{\sigma^2}\|\boldsymbol{\epsilon}\|_F^2.$$

Thus,

$$\begin{aligned}
EK(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0}) &\leq 4\epsilon^2 + \left(E\frac{4}{\sigma^4}\|\boldsymbol{\epsilon}\|_F^4\right)^{1/2}P\big(K^2(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0}) > 4\epsilon^2\big)\\
&\leq 4\epsilon^2 + 10\exp(-c_1 n\epsilon^2 + \log\sqrt{3nk}).
\end{aligned}$$

The results in Theorem 2 follow by letting $\epsilon = \epsilon_{n,p,k}$ and using the fact that $\log k \leq C_{p,k}^2$. This completes the proof. ■

Proof of Theorem 3: Without loss of generality, assume $p \geq k$ and $n = p$. When $\sigma = O(1/\sqrt{p})$, by Theorem 2, we have

$$\begin{aligned}
\|\hat{\mathbf{\Theta}}^{L_0} - \mathbf{\Theta}^0\|_F^2 = 2\sigma^2 K(\mathbf{\Theta}^0, \hat{\mathbf{\Theta}}^{L_0}) &= O_P(\frac{\epsilon_{n,p,k}^2}{p})\\
&= O_P\left(\frac{C_{p,k}^2}{p^2}\log(\frac{\sqrt{p}}{C_{p,k}})\right)\\
&= O_P\left(\frac{C_{p,k}^2}{p^2}\log(\frac{p^2}{C_{p,k}^2})\right), \qquad (26)
\end{aligned}$$

where

$$C_{p,k} = O\left(\sqrt{\log(p)}\sqrt{(p+k)s_2^0 + s_1^0} + \sqrt{s_1^0 \log\left(e\frac{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{s_1^0}\right)}\right). \quad (27)$$

(27) comes from the proof of Corollary 2 with M in **Assumption A** being $O(\sqrt{p})$. Thus,

$$\|\hat{\mathbf{\Theta}}^{L_0} - \mathbf{\Theta}^0\|_F^2 = O_P\left(C'_{p,k}\log(\frac{1}{C'_{p,k}})\right)$$

with

$$C'_{p,k} = \frac{\log(p)\cdot[(p+k)s_2^0 + s_1^0] + s_1^0 \log\left(e\frac{(p+k-r(\mathbf{\Theta}^0))r(\mathbf{\Theta}^0)}{s_1^0}\right)}{p^2}.$$

This completes the proof. ∎

References

Agarwal, A., Negahban, S. and Wainwright, M. (2012). Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. *The Annals of Statistics*, Vol. 40, No. 2, 1171–1197.

Amit, Y., Fink, M., Srebro, N., and Ullman, S. (2007). Uncovering shared structures in multiclass classification. *Proceedings of the 24th Annual International Conference on Machine learning*, 17–24.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4, 40–79.

Beck, Amir and Teboulle, Marc (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, Vol. 2, No. 1, 183-202.

Bunea, F., and She, Y. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics*, 39(2), 1282-1309.

Cai, J.F., Candès, E.J. and Shen, Z. (2010).A singular value thresholding algorithm for matrix completion. Arxiv preprint arXiv:0810.3286.

Cai, T. T., Liu, W. and Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106, 594-607.

Candes, E., Li. X., Ma, U., and Wright, J. (2009). Robust principal component analysis. *Journal of ACM*, 58(1), 1-37.

Candes, E.J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found of Comput. Math.*, 9, 717-772.

Chandrasekaran, V., Sanghavi, S., Parrilo, P.A., and Willsky, A.S. (2011). Rank-sparsity incoherence for matrix decomposition, *SIAM J. Optim.*, 21, 572-596.

Chen, B. , He, S., Li, Z. and Zhang, S. (2012). Maximum block improvement and polynomial optimization., *SIAM Journal on Optimization*, 22, 87-107.

Chen, J., Liu, J., and Ye, J. (2010). Learning incoherent sparse and low-rank patterns from multiple tasks. *The Sixteenth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining.* (SIGKDD 2010).

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99, 619-642. (with discussion).

Ganesh, A., Min, K., Wright, J. and Ma, Y. (2012). Principal component pursuit with reduced linear measurements. *International Symposium on Information Theory.*

Golub, G. and Van Loan, C. (1996). *Matrix Computations.* Third edition. London: The Johns Hopkins University Press.

Halko, N., Martinsson P. G. and Tropp, J. A.. (2011). Finding structure with randomness: stochastic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2), 217-288.

Jain, P., Meka, R. and Dhillon, I. (2010). Guaranteed rank minimization via singular value projection. *Advances in Neural Information Processing Systems*, 23, 937–945.

Kolmogorov, A.N. and Tihomirov, V.M. (1959). ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspekhi Mat. Nauk.* **14** *3-86.* [In Russian. English translation, *Ameri. Math. Soc. Transl.* **2, 17,** 277-364.(1961)].

Lin, Z., Chen, M., Wu, L. and Ma, Y. (2009). The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC Technical Report UILU-ENG-09-2215.*

Liu, J., and Ye, J. (2009). Efficient Euclidean projections in linear time. *The Twenty-Sixth International Conference on Machine Learning.*

Negahban, S. and Wainwright, M.J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2), 1069-1097.

Ossiander, M. (1987). A central limit theory under metric entropy with L_2 bracketing. *Ann. Probab.* **15** 897-919.

Porat, B. (1997). *A Course in Digital Signal Processing*, New York: John Wiley.

She, Y. (2013). Reduced rank vector generalized linear models for feature extraction. *Statistics and Its Interface*, Vol 6, 197-209.

Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107, 223-232.

Srebro, N., Rennie, J.D.M., and Jaakkola, T.S. (2005). Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems*, 17, 1329–1336.

Stanica, P., and Montgomery, A.P. (2001). Good lower and upper bounds on binomial coefficients. *J. Ineq. in Pure. Appl. Math.*, 2, art 30.

Stock, J. H. and Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business and Economic Statistics.*

Zhou, T. and Tao, D. (2011). GoDec: Randomized low-rank & sparse matrix decomposition in noisy case. *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA.

Waters, A.E., Sankaranarayanan, A.C. and Baraniuk, R.G.(2011). SpaRCS: Recovering low-rank and sparse matrices from compressive measurements. *Neural Information Processing Systems*, Granada, Spain.

Wong, W.H., and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.*, 23, 339-362.

Wright, J., Ganesh, A., Min, K., and Ma, Y. (2013). Compressive principal component pursuit. *Information and Inference*.

Wright, J., Ganesh, A., Rao, S., and Ma, Y. (2009). Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Advances in Neural Information Processing Systems*.

Xing, S., Zhu, Y., Shen, X., and Ye, J. (2012). Optimal exact rank minimization for noisy data. *Proceeding the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Beijing, China.

Yuan, X. and Yang, J. (2013). Sparse and low-rank matrix decomposition via alternating direction methods. *Pacific Journal of Optimization*, 9(1), 167-180.