THRESHOLDING MULTIVARIATE REGRESSION AND GENERALIZED PRINCIPAL COMPONENTS

A Dissertation

by

RANYE SUN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Mohsen Pourahmadi
Co-Chair of Committee,	Raymond J. Carroll
Committee Members,	Michael T. Longnecker
	Vijay P. Singh
Head of Department,	Simon Sheather

May 2014

Major Subject: Statistics

ABSTRACT

As high-dimensional data arises from various fields in science and technology, traditional multivariate methods need to be updated. Principal component analysis and reduced rank regression are two of the most important multivariate statistical techniques that have seen major changes in recent years. To improving the statistical performance and achieve fast computational efficiency, recent approaches aim at regularizing both the row and column factors of the low-rank matrix approximation by adopting the Lasso-type penalties. Thresholding is another powerful technique for regularizing the row and column factors without solving an optimization problem. This dissertation research covers two novel applications of the idea of thresholding: the thresholding reduced rank multivariate regression and the generalized principal component analysis/singular value decomposition (SVD). The following two paragraphs give brief introductions to each of the two topics, respectively.

Uncovering a meaningful relationship between the responses and the predictors is a fundamental goal in multivariate regression problems, which can be very challenging when data are high-dimensional. Dimension reduction and regularization techniques are applied extensively to alleviate the curse of dimensionality. It is desirable to estimate the regression coefficient matrix by low-rank matrices constructed from its SVD. We reduce such regression problems to sparse SVD problems for correlated data matrices and generalize the fast iterative thresholding for sparse SVDs algorithm to this situation. This generalization inherits the computational and statistical advantages of the original algorithm including its sparse initialization, novel ways of estimating the thresholding levels and the thresholded subspace iterations. It guarantees the orthogonality of the singular vectors and computes them simultaneously and not sequentially as in the existing methods. We also place this algorithm in an optimization framework by introducing a specific bi-convex objective function. An iterative algorithm that minimizes the objective function, via closed form iterates, is proposed and its convergence is established. This enables us to study the large sample properties of the solution of the multivariate regression problem and establishes consistency of the estimators as the sample size tends to infinity. The methodology and the potential adverse impact of dependence on the earlier algorithms are illustrated using simulation and real data.

The second part of this dissertation considers transposable data matrices where both their rows and columns are correlated. Such datasets are routinely encountered in fields such as econometrics, bio-informatics, chemometrics, network data and so on. While methods to approximate the high-dimensional data matrices have been extensively researched for uncorrelated and independent situations, they are much less so for the transposable data matrices. A generalization of principal component analysis and the related weighted least squares matrix decomposition with respect to a transposable quadratic norm for such data matrices along with their regularized counterparts have been proposed recently. We replace this optimization framework by thresholding the factors in the decompositions and propose a fast iterative thresholding for sparse generalized matrix decomposition algorithm to find sparse factors of the data matrix and account for the two-way dependencies simultaneously. We show that our algorithm is suitable for the reduced rank regression and canonical correlation analysis for two-way dependent data, which is done by connecting them with the generalized matrix decomposition. These connections enable us to improve predictive accuracy in regression and to facilitate interpretation of our proposed algorithm. The effectiveness of the method is tested and illustrated through simulation and real data examples.

DEDICATION

To my mom, dad and my wife for their continuous support and priceless love.

ACKNOWLEDGEMENTS

First and foremost I would like to express my greatest gratitude to my advisor, mentor and friend, Dr. Mohsen Pourahmadi, for his excellent guidance, constant inspiration and systematic mentoring throughout my entire doctoral study. I cannot thank him enough for all his contributions of time, ideas, and supports to make my Ph.D. experience productive and stimulating. I respect him for his excitement for work, his attitude towards life, and his generosity to his students. Being a graduate student in a foreign country can be very difficult with lots of issues other than school work to deal with. Whenever I needed help, Dr. Pourahmadi was always there with his patience, kindness and faith in me to keep me motivated through the difficult times. I want to thank him for everything he did for me. He is the best and the kindest professor I have ever met.

Special thanks to the members of my dissertation committee: Dr. Raymond Carroll, Dr. Michael Longnecker and Dr. Vijay Singh, for generously giving their time and valuable expertise to improve my research plan. Dr. Carroll is a world leading statistician. Dr. Longnecker is the Associate Department Head. Dr. Singh is a world class researcher in environmental engineering. It is an honor to have them on my committee. Their insights in research helped me to improve my work.

I am thankful to Dr. Michael Longnecker who is a great teacher and a good friend. As the Associate Department Head, he is extremely caring and helpful to all the students. His constant advice, help and support makes my life go so much more smoothly, specially when I was working as a graduate teaching assistant. I feel very lucky to have him around in my journey to my doctoral degree in Statistics.

I thank Dr. Zhonggen Su and Dr. Zhengyan Lin at Zhejiang University who

provided much support in my being here. Their excitement about statistics and compassion for students have inspired me to study statistics when I was an undergraduate student and have opened the door for me to this wonderful statistical world.

I want to thank all my colleagues in the Department and all my friends. You have made my life in College Station so colorful and exciting.

Finally, I would like to dedicate this dissertation to my family especially my parents Jianming and Xu and my wife Yanru, who have always had my back and are truly special to me. Without their endless love and support, I would never go this far.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

As high-dimensional data arises from various fields in science and technology, traditional multivariate methods need to be updated. Dimension reduction is important in analyzing the high-dimensional data matrices, where low-rank matrix approximation and regularization techniques are of great interest. Principal component analysis (PCA) and reduced rank regression (RRR) are two of the most important multivariate statistical techniques that have seen major changes in recent years. In the modern approaches, particular attention is paid to improving the statistical performance and achieving fast computational efficiency. Given these goals, recent approaches aim at regularizing both the row and column factors of the low-rank matrix approximation by adopting the Lasso-type penalties. Thresholding is another powerful technique for regularizing the row and column factors without solving an optimization problem. This dissertation research covers two novel applications of the idea of thresholding: the thresholding reduced rank multivariate regression and the generalized PCA/singular value decomposition (SVD).

High-dimensional data matrices usually have structural dependencies where both the rows and columns are dependent. Ignoring these dependencies can lead to poor statistical performances. In this section, we first introduce the transposable data matrix with two-way dependencies and two real-data applications as the motivation for this research. We next review the classical dimension reduction tools for the low-rank matrix approximation in multivariate analysis: the singular value decomposition and the principal component analysis in Sections 1.2 and 1.3. These two approaches are central to the regularized low-rank matrix approximation methods. After introducing the Lasso-type regularization in Section 1.4, the literature review for various regularization methods in the low-rank model and in the multivariate linear regressions are presented in Sections 1.5 and 1.6, respectively.

1.1 Transposable Data

Transposable data matrices are routinely encountered in fields such as econometrics, bio-informatics, chemometrics, network data, and so on. Such data matrices, where rows and columns are both dependent, have drawn much attention in recent statistical analyses. Recovering the true subspace or low-rank signal from the transposable data through low-rank matrix approximation is crucial in the statistical analysis.

For example, the macroeconomic data analyzed in Stock and Watson (2012) consisting of 144 U.S. macroeconomic time series for a total of 195 quarterly observations is a transposable dataset, because there are strong dependencies among feature variables (columns) and temporal dependencies among observations (rows). To make accurate forecasting it is ideal to extract a parsimonious set of common factors from the data matrix. This idea is important and useful especially for datasets where the series are correlated and the number of observations are close to or less than the number of variables.

Another important example is the functional MRI (fMRI) data (Lindquist, 2008; Allen et al., 2013) that consists of measurements of brain images (columns) over time (rows) and exhibits spatial and temporal dependencies. Each pixel in the fMRI brain images corresponds to a measure of activation in the brain (Lazar, 2008). Finding major brain activation patterns is a primary analysis goal in the fMRI studies. It requires the algorithms to extract the important information from a mix of noise and signal, or a low-rank matrix approximation is desired.

In Sections 1.2 and 1.3, we first review classical dimension reduction tools like the singular value decomposition and the principal component analysis, which are the fundamental methods for the low-rank matrix approximation.

1.2 The Singular Value Decomposition

The singular value decomposition (SVD), a fundamental conceptual and computational technique in linear algebra, has been used widely in the recent high-dimensional data situations for dimension reduction, data visualization, data compression, and information extraction by relying on its first few singular vectors (Golub and Van Loan, 1996).

Let $Y \in R^{n\times q}$ be a matrix, its SVD is of the form:

$$Y = UDV' = \sum_{i=1}^{q} d_i \mathbf{u}_i \mathbf{v}_i, \tag{1.1}$$

where the columns of $U = [\mathbf{u}_1, ..., \mathbf{u}_q] \in R^{n\times q}$ and $V = [\mathbf{v}_1, ..., \mathbf{v}_q] \in R^{q\times q}$ are the left and right singular vectors, respectively. The matrices U and V are orthonormal with $U'U = I_q, V'V = I_q$ and the matrix $D \in R^{q\times q}$ is a diagonal matrix with non-negative entries, called the singular values of Y. According to the Eckart-Young Theorem, for a given r, truncating the SVD gives the best rank-r approximation to a matrix Y. Let $||Y||_F$ denote the Frobenius norm of the matrix Y where $||Y||_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{q} y_{ij}^2 = tr(Y'Y)$.

Theorem 1.2.1 *(Eckart and Young, 1936) For any r, the matrix $Y^{(r)} = \sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i$ is the closest rank-r approximation to Y in the Frobenius norm,*

$$Y^{(r)} = \arg \min_{rank(B)=r} ||Y - B||_F^2,$$

where $d_i, \mathbf{u}_i, \mathbf{v}_i, i = 1, ..., r$ are the first r singular values and singular vectors of the matrix Y.

The Power method (Golub and Van Loan, 1996) is the most popular method for computing the SVD of a matrix, where a pair of left and right singular vectors is computed iteratively. For a matrix Y, it starts with an initial unit q-vector $\mathbf{v}^{(0)}$ and generates a sequence of unit vectors $\mathbf{u}^{(k)}$ and $\mathbf{v}^{(k)}, k = 1, 2, ...$ through the following two steps until convergence:

(1). Updating $\mathbf{u}$: $\mathbf{u}^{(k)} = Y\mathbf{v}^{(k-1)}/||Y\mathbf{v}^{(k-1)}||$,

(2). Updating $\mathbf{v}$: $\mathbf{v}^{(k)} = Y'\mathbf{u}^{(k)}/||Y'\mathbf{u}^{(k)}||$.

For $\mathbf{u}$ and $\mathbf{v}$ the unit vectors at convergence, the corresponding singular value is given by $d = \mathbf{u}'Y\mathbf{v}$. Then, the first rank-1 layer $d\mathbf{u}\mathbf{v}'$ is subtracted from Y and the power method is applied to the residual matrix $Y - d\mathbf{u}\mathbf{v}'$ to obtain the second pair of singular vectors, and so on.

An alternative to the power method is the idea of orthogonal iteration, which computes several leading left and right singular vectors 'at once' rather than a pair at a time (Golub and Van Loan, 1996). The orthogonal iteration method generalizes the power method from a vector to a subspace setup and achieves subspace orthogonalization through the QR decomposition. The QR decomposition is a decomposition of the matrix into an orthogonal matrix and a triangular matrix (Golub and Van Loan, 1996). For a matrix $Y \in R^{n \times q}$, the orthogonal iteration is a standard method for computing the subspaces spanned by its leading r singular vectors. Of course, for $r = 1$ the method reduces to the power method. It starts with an initial orthonormal matrix $V^{(0)} \in R^{q \times r}$ and generates sequences of orthonormal matrices $U^{(k)} \in R^{p \times r}$ and $V^{(k)} \in R^{q \times r}, k = 1, 2, ...$ through the four steps in Figure 1.1 until convergence. For U and V the orthonormal matrices at convergence, their columns are the leading left and right singular vectors of Y, which are orthogonal by construction using the QR decomposition.

1. Multiplication: $T_u^{(k)} = YV^{(k-1)}$,
2. QR decomposition: $U^{(k)}R_u^{(k)} = T_u^{(k)}$,
3. Multiplication: $T_v^{(k)} = Y'U^{(k-1)}$,
4. QR decomposition: $V^{(k)}R_v^{(k)} = T_v^{(k)}$.

Figure 1.1: The four key steps of the orthogonal iteration algorithm.

1.3 The Principal Component Analysis

The principal component analysis (PCA) is one of the most popular techniques in multivariate analysis, which is closely related to the SVD (Golub and Van Loan, 1996).

For a q-vector $\mathbf{y} = (y_1, ..., y_q)'$, the PCA is to explain the population covariance matrix Σ of $\mathbf{y}$ through r linear combinations of $\mathbf{y}$. Let $V = [\mathbf{v}_1, ..., \mathbf{v}_r] \in R^{q \times r}$, then the linear combinations of $\mathbf{y}$ given by $z_1 = \mathbf{v}_1'\mathbf{y}, ..., z_r = \mathbf{v}_r'\mathbf{y}$ are called the principal components (PC) if the z_i's have the maximum variance and are uncorrelated with each other. More precisely, the PCA computes V by solving the following optimization problems:

$$\begin{aligned}
&\mathbf{v}_1 = \arg\max_{\mathbf{v}} \mathbf{v}'\Sigma\mathbf{v} \text{ subject to } \mathbf{v}'\mathbf{v} = 1,\\
&\mathbf{v}_2 = \arg\max_{\mathbf{v}} \mathbf{v}'\Sigma\mathbf{v} \text{ subject to } \mathbf{v}'\mathbf{v} = 1, \mathbf{v}'\mathbf{v}_1 = 0,\\
&\vdots\\
&\mathbf{v}_r = \arg\max_{\mathbf{v}} \mathbf{v}'\Sigma\mathbf{v} \text{ subject to } \mathbf{v}'\mathbf{v} = 1, \mathbf{v}'\mathbf{v}_j = 0, \text{ for } j = 1, 2, .., r-1,
\end{aligned} \tag{1.2}$$

which leads to the orthogonal matrix V with $V'V = I_r$. The vectors $\mathbf{v}_1, ..., \mathbf{v}_r$ are called the loadings of the PCs, where the $\mathbf{v}_i$ is the ith eigenvector of Σ and

$var(z_i) = var(\mathbf{v}_i'\mathbf{y}) = d_i^2$ is its ith largest eigenvalue (Johnson and Wichern, 2007).

For the sample data, the PCA of the centered data matrix $Y \in R^{n\times q}$ finds the eigen-decomposition of its sample covariance matrix $S = \frac{1}{n}Y'Y$, an estimator of the population covariance Σ. The PCA provides the dimension reduction by forming the first r PCs of the original q variables, where the PCs are possibly easier to interpret and visualize.

Since the SVD is usually used to solve the eigen decomposition problem, there is a direct relation between them where the loading matrix V of the PCs can be found by the SVD of $Y = \tilde{U}\tilde{D}\tilde{V}'$ (Golub and Van Loan, 1996). More precisely,

$$Y'Y = (\tilde{U}\tilde{D}\tilde{V}')'(\tilde{U}\tilde{D}\tilde{V}') = \tilde{V}\tilde{D}^2\tilde{V}',$$

by using the orthogonality of the singular vectors $\tilde{U}$ and $\tilde{V}$. By (1.1), the columns of $\tilde{V}$ corresponding to the first r largest singular values are the eigenvectors of $Y'Y$, which are the solutions for (1.2). Thus, the right singular vectors $\tilde{V}$ of Y are the same as the loading matrix of PCs of Y. This connection is crucial in introducing the sparse PCA and SVD algorithms in Sections 1.5 and 1.6.

For transposable matrices, the generalized principal component analysis (GPCA), first proposed by Escoufier (1977), is a natural generalization of the PCA. To highlight the key idea of the GPCA, we give a heuristic account on how to compute its loading matrix before introducing it more formally in Section 3.3.2. Given two symmetric positive-definite matrices Ω and Σ, the GPCA finds the loading matrix $V = [\mathbf{v}_1, ..., \mathbf{v}_r]$ by maximizing the following criterion incorporating the matrices Ω

and Σ:

$$\begin{aligned}
\mathbf{v}_1 &= \arg\max_{\mathbf{v}} \mathbf{v}'\Sigma Y'\Omega Y\Sigma\mathbf{v} \text{ subject to } \mathbf{v}'\Sigma\mathbf{v} = 1,\\
\mathbf{v}_2 &= \arg\max_{\mathbf{v}} \mathbf{v}'\Sigma Y'\Omega Y\Sigma\mathbf{v} \text{ subject to } \mathbf{v}'\Sigma\mathbf{v} = 1, \mathbf{v}'\Sigma\mathbf{v}_1 = 0,\\
&\vdots\\
\mathbf{v}_r &= \arg\max_{\mathbf{v}} \mathbf{v}'\Sigma Y'\Omega Y\Sigma\mathbf{v} \text{ subject to } \mathbf{v}'\Sigma\mathbf{v} = 1, \mathbf{v}'\Sigma\mathbf{v}_j = 0, \text{ for } j < r.
\end{aligned} \tag{1.3}$$

Different from the PCA, the loading matrix V of GPCA observes the generalized orthogonality constraint $V'\Sigma V = I_r$. The generalized principal components (GPC) are given by $Y\Sigma\mathbf{v}_1, ..., Y\Sigma\mathbf{v}_r$. The matrices Ω and Σ are closely related to the dependency structures of the transposable data discussed in details in the later sections.

Unfortunately, classical PCA and SVD encounter major problems in the high-dimensional data situations. The sample eigenvectors and singular vectors of PCA and SVD are not consistent estimators for their population counterparts (Johnstone and Lu, 2009; Fan and Lv, 2010). In addition, when eigenvectors and singular vectors of PCA and SVD have too many non-zero entries, they are hard to interpret and their use in practice could lead to misleading conclusions. Methods imposing sparsity or smoothness on the singular vectors and values have been shown to lead to consistency in high-dimensional settings (Johnstone and Lu, 2009). When the irrelevant entries in the eigenvectors and singular vectors are forced to zero, the statistical efficiency of the model and its interpretability are improved.

We review the Lasso-type regularization for the linear regression models in Section 1.4. We then illustrate its usage in the low-rank matrix approximation and the multivariate regression model in Sections 1.5 and 1.6, respectively.

1.4 Lasso-Type Regularizations

In this subsection, we review the method of least-squares estimation of the regression parameters with the Lasso-type (l_1) penalties on the coefficients (Tibshirani, 1996).

Consider the linear model for the response $\mathbf{y} \in R^{n\times 1}$ and covariate $X \in R^{n\times p}$:

$$\mathbf{y} = X\beta + e, \tag{1.4}$$

where $\beta \in R^{p\times 1}$ is the coefficient vector and e is the noise. The ordinary least-squares estimator of (1.4) $\hat{\beta} = (X'X)^{-1}X'\mathbf{y}$ does not perform well for the high-dimensional data situation when $n \leq p$ as shown in Hastie et al. (2009).

The least absolute shrinkage and selection operator (Lasso) is one of the most popular approaches for selecting the most significant variables and estimating regression coefficients simultaneously. It penalizes the least-squares regression using the l_1 penalty on the coefficients and finds the Lasso solution $\hat{\beta}$ by minimizing the objective function

$$\frac{1}{2}||\mathbf{y} - X\beta||^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{1.5}$$

where λ denotes the tuning parameter controlling the sparsity of the coefficients. Note that $\lambda = 0$ corresponds to the least-squares estimator.

A number of innovative approaches are available to compute the Lasso solution. Two representative examples are the least angle regression (LARS) algorithm (Efron et al., 2004) and the coordinate descent algorithm (Friedman et al., 2008). The LARS algorithm starts with the variable which is most correlated with the response, and computes the whole solution path of the Lasso as the tuning parameter λ changes.

The coordinate descent algorithm solves (1.5) by minimizing over one β_i at a time while the other β's are kept fixed and cycles through the parameters $\beta_i, i = 1, ..., p$, until convergence.

The Lasso regression is very popular, but has some drawbacks such as the bias problem (Fan and Li, 2001). There are several alternative Lasso-type penalty functions designed to fix these drawbacks. Zou (2006) proposed the adaptive Lasso with the weighted l_1 penalty leading to the objective function

$$\frac{1}{2}||\mathbf{y} - X\beta||^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

where the weights are data-driven with $w_j = 1/|\hat{\beta}_j|^\gamma$, $\hat{\beta}_j$ the ordinary least-squares estimator and $\gamma > 0$. Zou and Hastie (2005) proposed the elastic net where the penalty is a linear combination of the l_1 and l_2 penalties. Given nonnegative λ_1 and λ_2, the objective function for elastic net is given by

$$(1 + \lambda_2)||Y - X\beta||^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2,$$

which generalizes both the Lasso ($\lambda_2 = 0$) and the ridge regression ($\lambda_1 = 0$).

There are extensive research in this area where other penalized least-squares or likelihood methods with various types of regularizations have been developed. A representative but incomplete list of references is Fan and Li (2001), Yuan and Lin (2007), Zhao and Yu (2007) and Meier et al. (2008).

1.5 The Low-Rank Matrix Model

In this subsection, we review the regularization methods in low-rank matrix approximation. To find low-rank signal from the data matrix $Y \in R^{n \times q}$, the following setup (Yang et al., 2013; Allen et al., 2013) is considered

$$Y = B + E, \tag{1.6}$$

where Y and $B = UDV'$ denote the data and the signal matrices, D the singular values, U and V the left and right factors, respectively.

The goal is to find a low-rank structure for the signal in the data matrix. If Y is the spatial-temporal fMRI data set described in Section 1.1, the rows would represent locations in the brain image and the columns point to the time effect. We present an overview of various methods for regularizing and computing the sparse singular vectors in U and V. There are two categories of algorithms: (i) the optimization-based sequential algorithms and (ii) the subspace iteration algorithms.

1.5.1 The Sequential Algorithms

Various regularization methods for computing the singular vectors have been proposed where the solutions are found sequentially through rank-one approximations. More precisely, the first rank-1 approximation is computed by imposing penalties on the vectors $\mathbf{u}$ and $\mathbf{v}$. Then, the first computed layer $d\mathbf{u}\mathbf{v}'$ is subtracted from Y and the procedure is repeated on the residual matrix. The first pair $\mathbf{u}$ and $\mathbf{v}$, of $U = [\mathbf{u}_1, ..., \mathbf{u}_r]$ and $V = [\mathbf{v}_1, ..., \mathbf{v}_r]$, is found by minimizing the following objective function:

$$\frac{1}{2}||Y - d\mathbf{u}\mathbf{v}'||_F^2 + P_\lambda(\mathbf{u}, \mathbf{v}), \tag{1.7}$$

with respect to the triplet $(d, \mathbf{u}, \mathbf{v})$, where λ is the tuning parameter and $P_\lambda(\mathbf{u}, \mathbf{v})$ is a penalty function on $\mathbf{u}$ and $\mathbf{v}$. Some penalty functions introduced in recent years are listed below.

1. The sparse PCA via regularized SVD algorithm in Shen and Huang (2008) and the sparse SVD (SSVD) algorithm in Lee et al. (2010) use the additive penalty function:

   $$P_\lambda(\mathbf{u}, \mathbf{v}) = \lambda_u ||\mathbf{u}||_1 + \lambda_v ||\mathbf{v}||_1,$$

   where λ_u and λ_v are the tuning parameters for the left and right singular vectors, respectively. Using additive penalties with two penalty parameters allows different levels of sparsity on $\mathbf{u}$ and $\mathbf{v}$.

2. The penalized matrix decomposition (PMD) algorithm proposed by Witten et al. (2009) relies on the following constraints/penalties:

   $$||\mathbf{u}||_2 \leq 1, ||\mathbf{v}||_2 \leq 1, P_u(\mathbf{u}) \leq c_u, P_v(\mathbf{v}) \leq c_v,$$

   where $P_u(\cdot), P_v(\cdot)$ are the Lasso or fused Lasso penalty, c_u and c_v are the corresponding tuning parameters.

3. The sparse reduced rank regression algorithm (SRRR) in Chen et al. (2012a) applies multiplicative penalty on the singular vectors:

   $$P_\lambda(\mathbf{u}, \mathbf{v}) = \lambda \sum_{i=1}^{n} \sum_{j=1}^{p} w_{ij} |d u_i v_j|,$$

   where u_i and v_j are the ith and jth entries of the vectors $\mathbf{u}$ and $\mathbf{v}$, respectively, and w_{ij}'s are data-driven weights as in the adaptive Lasso.

4. The penalized SVDs approach in Huang et al. (2009) uses a more general form of the multiplicative penalty function to regularize the singular vectors in the context of two-way functional data.

Unfortunately, the above regularization methods assume the entries of Y are i.i.d and ignore the row and column dependencies present in the transposable data. Ignoring the two-way dependencies in transposable data is known to lead to poor statistical performance (Efron, 2009; Allen and Tibshirani, 2010; Allen et al., 2013).

1.5.2 The Sequential Algorithm For Transposable Data

In Section 1.5.1, the low-rank matrix approximation problem was solved by minimizing the Frobenius norm: $||Y - d\mathbf{u}\mathbf{v}'||_F^2$. This loss function treats errors with equal weight and the covariances in the transposable data are ignored. To permit unequal weights according to the dependence structure of the data, Allen et al. (2013) proposed a generalized least-squares matrix decomposition (GMD) framework to directly accounts for the known covariance matrices Ω and Σ. Define the transposable quadratic norm ((Ω, Σ)-norm) of a matrix A as $||A||_{\Omega,\Sigma}^2 = tr(A'\Omega A\Sigma)$ to replace the Frobenius norm in finding the best low-rank approximation by minimizing the (Ω, Σ)-norm

$$||Y - d\mathbf{u}\mathbf{v}'||_{\Omega,\Sigma}^2 \ , \tag{1.8}$$

subject to the generalized orthogonality conditions

$$\mathbf{u}'\Omega\mathbf{u} = \mathbf{v}'\Sigma\mathbf{v} = 1. \tag{1.9}$$

It turns out that for normally distributed data matrix defined next, the transposable quadratic norm (1.8) is proportional to the log-likelihood of the transposable data matrix $Y \in R^{n\times q}$.

As a generalization of normal random vectors, a matrix-variate normal distribution (Gupta and Nagar, 1999) is defined and denoted by

$$Y \sim MN_{n,q}(M, \Omega^{-1}, \Sigma^{-1}), \tag{1.10}$$

where Ω^{-1}, Σ^{-1} denote the rows and columns covariance matrices and M denotes the mean matrix of the data. The definition in (1.10) means that the vectorized Y is distributed as

$$vec(Y) \sim N_{nq}(vec(M), \Omega^{-1} \otimes \Sigma^{-1}),$$

with a separable covariance structure where $\otimes$ denotes the Kronecker product and the *vec* operator forms a vector by stacking up the columns of a matrix.

It is easy to show that the log-likelihood function of Y can be written as:

$$l(Y|\Omega^{-1}, \Sigma^{-1}) \propto tr\{(Y - d\mathbf{u}\mathbf{v}')'\Omega(Y - d\mathbf{u}\mathbf{v}')\Sigma\} = ||Y - d\mathbf{u}\mathbf{v}'||^2_{\Omega,\Sigma},$$

where the right hand side is (1.8).

To find the sparse GMD factors, Allen et al. (2013) proposed to regularize (1.8) using the Lasso penalty $P(\mathbf{u}, \mathbf{v}) = \lambda_u|\mathbf{u}|_1 + \lambda_v|\mathbf{v}|_1$ on $\mathbf{u}$ and $\mathbf{v}$. However, their algorithms are still sequential as in Shen and Huang (2008), Witten et al. (2009) and Lee et al. (2010), which lack orthogonality of the columns of U and V when $r > 1$.

1.5.3 Subspace Iterations and FIT-SSVD

It is known that the sequential algorithms have expensive computation costs and cannot guarantee the orthogonality of the regularized singular vectors (Yang et al., 2013). Hence, a novel approach for low-rank approximation using the orthogonal iteration in (1.6) was proposed by Yang et al. (2013). Their fast iterative thresholding for sparse SVDs (FIT-SSVD) algorithm computes the two subspaces spanned by the leading left and right singular vectors simultaneously which guarantees the orthogonality of the singular vectors. More precisely, the key ideas that distinguish the FIT-SSVD method from the earlier sequential methods are listed below:

1. the use of orthogonal iteration to compute the subspaces spanned by the first r singular vectors in U and V,
2. the use of thresholding to replace the smaller entries of U and V by zeros, and novel, inexpensive ways of estimating the threshold levels,
3. sparse initialization by deleting the rows and columns of the data matrix with low signal.

Rather than solving the optimization problems in Section 1.5.1, the FIT-SSVD algorithm is based on thresholding. It adopts a thresholding function, like the familiar soft-thresholding

$$S(y, \gamma) = sgn(y)(|y| - \gamma)_+,$$

the hard-thresholding

$$H(y, \gamma) = y1_{|y|>\gamma}$$

or SCAD (Fan and Li, 2001), where γ is the threshold level. The threshold level γ is selected to be $\sqrt{2\log n}$ motivated by the asymptotic results from the Gaussian sequence models (Johnstone, 2011) or using the idea of “m out of n” bootstrapping the data. These novel ways of estimating the threshold level avoid choosing tuning parameters by the computationally expensive cross-validation methods and hence lead to fast computational performance. A detailed introduction of the FIT-SSVD is given in Section 2.2.1.

Selecting the threshold level in the FIT-SSVD algorithm relies heavily on the independence of the entries of a data matrix. Generalizations of the FIT-SSVD to account for the row-dependence are proposed in Section 2, where the correlations are incorporated in selecting the threshold levels.

1.6 Multivariate Linear Regression

In this subsection, we review the reduced rank regression (RRR), its connection to various multivariate methods and various ways to regularize the regression coefficient matrix.

Given n observations on the q-vector of responses $\mathbf{y}$ and the p-vector of predictors $\mathbf{x}$, the multivariate linear regression model is

$$Y = XB + E, \tag{1.11}$$

where $Y = (\mathbf{y}_1, ..., \mathbf{y}_n)' \in R^{n\times q}, X = (\mathbf{x}_1, ..., \mathbf{x}_n)' \in R^{n\times p}$ and $B \in R^{p\times q}$ denote the responses, covariates and regression coefficients matrices, and E the noise matrix consists of iid normal random variables. Note that at least for a full-rank design matrix, X can be removed from (1.11) by left multiplying both sides by $(X'X)^{-1}X'$ leading to

$$\tilde{Y} = (X'X)^{-1}X'Y \quad = \quad (X'X)^{-1}X'XB + (X'X)^{-1}X'E.$$

It appears to be of the form (1.6), but with Y and E replaced by $\tilde{Y} = \hat{B}_{OLS} = (X'X)^{-1}X'Y$ and $\tilde{E} = (X'X)^{-1}\ X'E$. This situation with the $\tilde{E}$ will be extensively studied in this dissertation.

The number of parameters in B can be quite large when both dimensions p and q are large, and its regularization is advisable when either dimension exceeds the sample size n. An early approach to regularization of B is the RRR (Anderson, 1951; Izenman, 1975) where one finds the least-squares estimate of B subject to the rank constraint rank$(B) = r$, for a given integer r. Hence, the coefficient matrix $B = \Theta_1\Theta_2$ can be written as a product of two lower dimensional matrices $\Theta_1 \in R^{q\times r}$ and $\Theta_2 \in R^{r\times p}$ of rank r. Thus, the RRR reduces the potentially large number of parameters in B from pq to $r(p + q)$ which is linear in p and q. Let the matrices X and Y be centered, given any positive-definite matrix $W \in R^{q\times q}$, the solution Θ_1 and Θ_2 of the RRR problem is computed by minimizing a weighted sum-of-squares criterion

$$tr\{(Y - X\Theta_1\Theta_2)'W(Y - X\Theta_1\Theta_2)\}, \tag{1.12}$$

which is of the form of (Ω, Σ)-norm, see Figure 1.2. Then, the solution is given by (Reinsel and Velu, 1998; Izenman, 2008)

$$\begin{aligned}\Theta_1 &= \Sigma_{XX}^{-1}\Sigma_{XY}W^{1/2}P = \hat{B}_{OLS}W^{1/2}P,\\ \Theta_2 &= P'W^{-1/2},\end{aligned}$$

where $P = [p_1, ..., p_r] \in R^{q\times r}$ and p_i is the eigenvector corresponding to the ith largest eigenvalue of the matrix

$$W^{1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}W^{1/2},$$

where

$$\Sigma_{XX} = Cov(X, X), \Sigma_{XY} = \Sigma'_{YX} = Cov(X, Y), \Sigma_{YY} = Cov(Y, Y),$$

which are proportional to $X'X, X'Y$ and $Y'Y$, since X and Y are centered.

The RRR problem provides a very general framework subsuming various widely used techniques in multivariate statistics, see Figure 1.2.

1. Setting $W = I$ and $Y = X$, the (1.12) reduces to a PCA problem and have solutions $\Theta_1 = P$ and $\Theta_2 = P'$, where $P = V$ is the eigenvectors of Σ_{YY} as in Section 1.3. Thus, the RRR solves the PCA problem for data Y.

2. In the canonical correlation analysis (CCA) (Hotelling, 1935, 1936), the goal is to find vectors g and h such that the correlation $corr(g'\mathbf{x}, h'\mathbf{y})$ is maximized. Let $G = (g_1, ..., g_r) \in R^{p\times r}$ and $H = (h_1, ..., h_r) \in R^{q\times r}$, the CCA is to find matrices G and H that minimize all the eigenvalues of $(YH-XG)'(YH-XG)$. It has been shown in Reinsel and Velu (1998) that by letting $W = \Sigma_{YY}^{-1}$ in (1.12), the solutions of the CCA are $G = \Theta_1$ and $H = \Theta_2'$ as in the RRR.

3. The CCA setup can be reduced to the Fisher's linear discriminant analysis (Fisher, 1936) if the response is a vector of binary variables.

4. The CCA setup can also be reduced to the correspondence analysis (Hirschfeld, 1935) if both the responses and the predictors are binary variables as shown in (Izenman, 2008).

There is a host of regression estimators that either regularize the weighted least-squares estimators (WLS) or regularize the likelihood function. In the literature, particular attention is paid to using the penalty on the singular values and singular

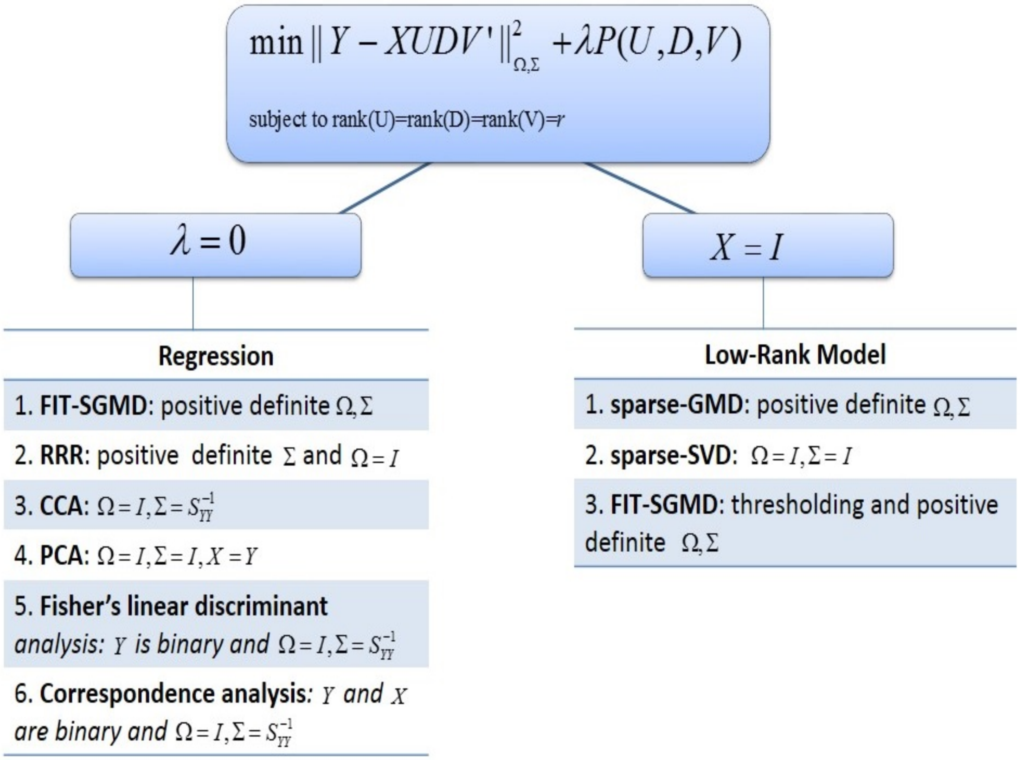

$\min \| Y - XUDV' \|^2_{\Omega,\Sigma} + \lambda P(U,D,V)$
subject to rank(U)=rank(D)=rank(V)=r
$\lambda = 0$
$X = I$
Regression
1. **FIT-SGMD**: positive definite Ω,Σ
2. **RRR**: positive definite Σ and $\Omega = I$
3. **CCA**: $\Omega = I, \Sigma = S_{YY}^{-1}$
4. **PCA**: $\Omega = I, \Sigma = I, X = Y$
5. **Fisher's linear discriminant** *analysis: Y is binary and* $\Omega = I, \Sigma = S_{YY}^{-1}$
6. **Correspondence analysis***: Y and X are binary and* $\Omega = I, \Sigma = S_{YY}^{-1}$
Low-Rank Model
1. **sparse-GMD**: positive definite Ω,Σ
2. **sparse-SVD**: $\Omega = I, \Sigma = I$
3. **FIT-SGMD**: thresholding and positive definite Ω,Σ

Figure 1.2: A diagram of RRR-related methods.

vectors of the coefficient matrix. The rank-r approximation of the coefficient matrix B is found using the SVD of $B = \sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i'$ (Chen et al., 2012a). We show the details of the connection between the solution of (1.12) and the SVD in Section 3. Using the weighted Frobenius norm or the (Ω, Σ)-norm, the regularized RRR finds the minimizer of

$$||(Y - XB)W^{1/2}||_F^2 + P_\lambda(B) = ||Y - XB||_{I,W}^2 + P_\lambda(B),$$

where W denotes a weight matrix and $P_\lambda(B)$ is a penalty function on B or its SVD factors U, D and V. When $W = \Sigma_{YY}^{-1}$, where Σ_Y is the population covariance matrix of the response Y and $Y \sim N(XB, \Sigma_{YY})$, the above objective function is proportional to the regularized log-likelihood function:

$$l(Y|\Sigma_{YY}) \propto tr\{(Y - XB)'\Sigma_{YY}^{-1}(Y - XB)\} = ||Y - XB||^2_{I,\Sigma_{YY}^{-1}}.$$

We list first those methods where penalties are imposed on the singular values of the coefficient matrix:

1. Yuan et al. (2007) proposed a nuclear norm penalized (NNP) least-squares estimator by setting

$$P_\lambda(B) = \lambda||B||_* = \sum_{i=1}^{r} d_i,$$

   where $d_i(B)$ denotes the ith singular value of B. The NNP encourages sparse singular values, hence it performs dimension reduction and coefficient estimation at the same time. However, computing the estimates is very challenging in practice due to the nuclear norm constraint. Cai et al. (2010), Toh and Yun (2010) and Lu et al. (2012) have conducted extensive research to solve this optimization problem.

2. Bunea et al. (2011) proposed the rank selection criterion (RSC) by setting

$$P_\lambda(B) = \lambda \sum_{i=1}^{r} I(d_i \neq 0),$$

   where $I(\cdot)$ is an indicator function. It has low computational complexity compared to the NNP and has the explicit solution $\hat{B} = (X'X)^g X'YVD^{-1}H(D)V'$, where UDV' is the SVD of the predictor $X(X'X)^g X'Y$, $H(D) = diag\{d_i I(d_i > \lambda), i = 1, ..., r\}$. The RSC provides consistent estimators of the rank of the coefficient matrix when both n and p go to infinite.

3. Dobrev and Schaumburg (2013) proposed the Tikhonov regularization of reduced rank regression (TRRR) by setting

$$P_\lambda(B) = \lambda||R(\Theta_1\Theta_2')W^{1/2}||_F^2 = \lambda||\Theta_1\Theta_2'||^2_{R'R,W},$$

   subject to $\Theta_1'\Sigma_{XX}\Theta_1 = I_r$. Here, $B = \Theta_1\Theta_2'$ and R is a pre-determined matrix which may be chosen to differentially penalize certain directions in the parameter space. Its solution can be obtained by solving a generalized eigenvalue problem $|\Sigma_{XY}W\Sigma_{YX} - \rho(\Sigma_{XX} + \lambda R'R)| = 0$, where ρ is the generalized eigenvalue.

4. Chen et al. (2012b) proposed the adaptive nuclear norm penalization (ANN) by penalizing the mean (predictor) XB instead of B,

$$P_\lambda(B) = \lambda\sum_{i=1}^{r} w_i d_i(XB),$$

   where $d_i(XB)$ denotes the ith singular value of XB. It has the explicit solution $\hat{B} = (X'X)^g X'YVD^{-1}S_{\lambda w}(D)V'$, where UDV' is the SVD of the predictor $X(X'X)^gX'Y$, $S_\lambda(D) = diag\{(d_i - \lambda)_+, i = 1, ..., r\}$, where the soft-thresholding operator acts on the singular values of the matrix XB. The ANN method directly tackles the prediction matrix approximation and imposes the sparsity on XB rather than B. The selection of the tuning parameters $\{\lambda, w\}$ in the ANN are data-driven.

These regularized RRR approaches penalize the singular values of the coefficient matrix and encourage sparsity in the singular values and hence restricts the rank.

In order to account for the possible sparsity of the singular vectors, we list two representative methods below that enforce sparsity penalties on the singular vectors of the coefficient matrix.

1. Chen et al. (2012a) proposed the sparse reduced rank regression by setting

   $$P_\lambda(B) = \sum_{k=1}^{r} \lambda_k \sum_{i=1}^{p} \sum_{j=1}^{q} w_{ijk} |d_k u_{ik} v_{jk}|,$$

   where u_{ik}, v_{jk} are entries of $\mathbf{u}_k$ and $\mathbf{v}_k$ and w_{ijk}'s are the data-driven weights as used in the adaptive lasso. The iterative exclusive extraction algorithm is proposed to estimate B with sparse SVD structure starting from some initial consistent estimator of B, e.g. the reduced rank least-squares estimator $\hat{B}_{OLS} = \sum_{l=1}^{r} \hat{d}_l \hat{\mathbf{u}}_l \hat{\mathbf{v}}_l'$. They reduce the task of regularizing B into r parallel sparse unit rank regressions by decomposing the response matrix Y into r layers $Y_l = Y - X(\hat{B}_{OLS} - \hat{B}_{OLS,l})$, where $\hat{B}_{OLS,l} = \hat{d}_l \hat{\mathbf{u}}_l \hat{\mathbf{v}}_l'$ and $l = 1, ..., r$, and solve the sparse regression of Y_l's on X with unit rank coefficient matrix.

2. Chen and Huang (2012) proposed a group lasso penalty of sparse reduced rank regression by setting

   $$P_\lambda(B) = \lambda \sum_{i=1}^{p} |\Theta_{i.,1}|,$$

   where $B = \Theta_1 \Theta_2'$, $\Theta_{i.,1}$ is the ith row of Θ_1, subject to condition $\Theta_2' \Theta_2 = I_r$. It uses the idea of group lasso in Yuan and Lin (2007) and the numerical solution can be obtained through the subgradient or variational method.

Unfortunately, these regularization methods either simply assume the entries of noise E are iid distributed or assume E consists of independent columns, where the dependencies among rows and columns as in the transposable data are ignored.

The rest of the dissertation is organized as follows. We present our proposed novel regularization approach for the reduced-rank regression in Section 2, where its optimization problem is extensively discussed. We then extend our approach to the transposable data situation and incorporate the row and column dependencies in Section 3. Analyses of a micro-array data and a macroeconomic data follow the development of methodologies in each section. We conclude and discuss some future research topics in Section 4, and present the proofs and additional simulations in the Appendix.

2. REDUCED RANK MULTIVARIATE REGRESSION: THRESHOLDING AND OPTIMIZATION

Uncovering the meaningful relationship between the responses and the predictors is a fundamental goal in multivariate regression problems, which can be very challenging when data are high-dimensional. Dimension reduction and regularization techniques are applied extensively to alleviate the curse of dimensionality. It is desirable to estimate the regression coefficient matrix by low-rank matrices constructed from its SVD. In this section, we integrate the reduced-rank regression approach with the regularization techniques and reduce such regression problems to sparse SVD problems for correlated data matrices and generalize the FIT-SSVD algorithm in Yang et al. (2013) to this situation. We also place Yang et al.'s algorithm in an optimization framework by introducing a specific bi-convex objective function. This enables us to study the large sample properties of the solution of the multivariate regression problem and establish consistency of the estimators as the sample size tends to infinity.

2.1 Background

There are very close and synergistic connections between the reduced rank multivariate regression (Anderson, 1951; Izenman, 1975) and the SVD of its coefficient matrix (Reinsel and Velu, 1998; Yuan et al., 2007; Chen et al., 2012a). Consider the multivariate linear regression model

$$Y = XB + E, \tag{2.1}$$

where Y, X and B denote the $n \times q$, $n \times p$ and $p \times q$ matrices of the responses, covariates and regression coefficients, and E the noise matrix which consists of iid normal random variables.

To reduce the potentially large number of parameters in B, the reduced rank regression (RRR) finds the least-square estimate of B subject to the rank constraint $\text{rank}(B) = r \leq \min(p, q)$. Its solution is known (Reinsel and Velu, 1998; Chen et al., 2012a) to relate to the low-rank approximation property of the SVD of $B = \sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i'$ as discussed in Section 1.6. Note that the special case of $p = n$ and $X = I_p$ leads to the model

$$Y = B + E, \tag{2.2}$$

where the low-rank approximation of B has been studied as a free-standing low-rank model in the recent literature of high-dimensional data analysis, see Section 1.5. However, it has been noted (Johnstone and Lu, 2009) that for high-dimensional data the classical SVD lacks good computational and statistical properties.

A regularized least-squares approach to the RRR proposed by Yuan et al. (2007) penalizes the sum of the singular values of the coefficient matrix, it encourages sparsity in the singular values and hence restricts the rank of B. Unfortunately, this approach and its variants (Bunea et al., 2011) do not take into account the possible sparsity of the singular vectors. Regularization of the singular vectors has been proposed by Shen and Huang (2008, p.123), Witten et al. (2009), Lee et al. (2010) and Allen et al. (2013) where the solution is found sequentially through rank-one approximations of the data matrix Y in (2.2). More generally, Chen et al. (2012a) have introduced a regularized reduced rank regression method by considering a low-rank approximation to the ordinary least square estimate (OLS) of the coefficient matrix in the multivariate regression (2.1) using an adaptive lasso penalty on the singular vectors $\mathbf{u}_l, \mathbf{v}_l, l = 1, \cdots, m$. A notable drawback of this sequential approach is that the orthogonality of the singular vectors cannot be guaranteed.

A common feature of these sequential algorithms is that they provide solutions of certain penalized optimization problems. A novel non-optimization based iterative approach for low-rank approximation of high dimensional data in (2.2) is the fast iterative thresholding for sparse SVDs (FIT-SSVD) algorithm proposed by Yang et al. (2013). Unlike Shen and Huang (2008), Witten et al. (2009) and Lee et al. (2010) which compute singular vectors sequentially one at a time, the FIT-SSVD algorithm computes the two subspaces spanned by the leading left and right singular vectors using the idea of orthogonal iteration (Golub and Van Loan, 1996, Chapter 8) which guarantees the orthogonality of the singular vectors. More precisely, the key ideas that distinguish the FIT-SSVD method from the earlier sequential methods are the use of orthogonal iteration, thresholding and its novel sparse initialization, see Section 1.5.3. The theory and simulation studies in Yang et al. (2013) confirm that the FIT-SSVD algorithm is computationally much faster than the earlier sequential algorithms.

Unlike the sequential algorithms, the FIT-SSVD algorithm is neither motivated by nor based on solving optimization problems. In this section, our first contribution is to place the FIT-SSVD algorithm in an optimization framework. We introduce a suitable bi-convex objective function and an iterative algorithm to minimize it via closed form iterates. This setup enables us to study the large sample properties of the FIT-SSVD solution and establish consistency of the estimators as the sample size n tends to infinity. Our second contribution is to reduce the more general regression problem (2.1) to the low-rank model (2.2), recognize and deal with its correlated error by generalizing the FIT-SSVD algorithm to the correlated data situation. We propose a fast iteratively thresholded sparse reduced rank regression (FIT-SRRR) algorithm which addresses the lack of orthogonality of the singular vectors in Chen et al. (2012a) and accounts for the correlation. The FIT-SRRR methodology allows the use of covariates and guarantees that the SVD layers are orthogonal. It makes effective use of the distribution of errors in finding the threshold levels and inherits all the good computational and statistical properties of the FIT-SSVD algorithm. Our simulation study and data analysis reveal the considerable gain when the dependence in the data is accounted for.

The rest of the Section 2 is organized as follows. In Section 2.2, after briefly reviewing the FIT-SSVD algorithm in Yang et al. (2013), we place it in an optimization framework. We develop in Section 2.3 the FIT-SRRR algorithm for the sparse reduced rank regression. We illustrate our methodology using simulations and real data in Sections 2.4 and 2.5, respectively. Section 2.6 concludes this section.

2.2 Thresholding for Sparse SVDs

In this subsection, we briefly review the FIT-SSVD algorithm and place it in an optimization setup by proposing a bi-convex objective function.

2.2.1 Overview

Recalling Section 1.2, the power method (Golub and Van Loan, 1996) is the most basic tool for computing the singular vectors of a matrix. Given an initial vector, it iteratively computes a pair of left and right singular vectors at a time. The alternative technique of orthogonal iteration computes several left and right singular vectors 'at once' and generalizes the power method from a vector to a subspace setup. It achieves subspace orthogonalization through the QR decomposition. A key step of the FIT-SSVD algorithm is based on the idea of orthogonal iteration.

In the low-rank model (2.2), given an initial matrix $V^{(0)}$ of the right singular vectors of B with r orthonormal columns, the FIT-SSVD algorithm at the kth iteration applies multiplication to B and B' by $V^{(k-1)}$ and $U^{(k)}$ followed by thresholding. Then, the QR decomposition of the thresholded matrices gives the orthonormal matrices U and V. It iterates according to the four steps in the Figure 2.1 until convergence.

1. Right-to-left Multiplication and Thresholding: $U^{(k),thr} = \eta(BV^{(k-1)}, \gamma_u)$,
2. Orthonormalization with QR Decomposition: $U^{(k)}R_u^{(k)} = U^{(k),thr}$,
3. Left-to-right Multiplication and Thresholding: $V^{(k),thr} = \eta(B'U^{(k)}, \gamma_v)$,
4. Orthonormalization with QR Decomposition: $V^{(k)}R_v^{(k)} = V^{(k),thr}$.

Figure 2.1: The four key steps of the FIT-SSVD algorithm.

In Figure 2.1, the function $\eta(.)$ is a pre-selected threshold function, like the familiar soft-thresholding with $\eta(y, \gamma) = S(y, \gamma) = sgn(y)(|y| - \gamma)_+$, hard-thresholding $\eta(y, \gamma) = H(y, \gamma) = y1_{|y|>\gamma}$ or SCAD (Fan and Li, 2001), where γ is the threshold level. We work with the soft-thresholding function in the theoretical development here and discuss the other cases in the Appendix A.

Choosing the proper threshold level γ requires the knowledge of the distribution of the noise E. In fact, selection of γ in Yang et al. (2013) relies heavily on the assumption of independence of the entries of the noise matrix. When this assumption is violated, we present a generalized FIT-SSVD algorithm in Section 2.3, which incorporates the dependence in the data.

2.2.2 An Objective Function Framework

Although the FIT-SSVD algorithm for model (2.2) as developed in Yang et al. (2013) is not optimizing-based, we introduce a suitable objective function in this subsection and place it in an optimization-based framework.

Consider minimizing the objective function,

$$\Psi(U, D, V) = ||Y - UDV'||_F^2 + \lambda_u \sum_{i=1}^{p} \sum_{k=1}^{r} |u_{ik} d_k| + \lambda_v \sum_{j=1}^{q} \sum_{k=1}^{r} |v_{jk} d_k|, \qquad (2.3)$$

over (U, D, V) subject to

$$U'U = I_r, V'V = I_r, \qquad (2.4)$$

where Y, U and V are $p \times q$, $p \times r$ and $q \times r$ matrices, $r \leq m = min(p, q)$, and λ_u, λ_v are the regularization parameters. Note that (2.3) as an optimization problem with respect to matrices U, D, V is not bi-convex due to the two equality conditions on U and V in (2.4). We reduce it to a bi-convex optimization problem, following Witten et al. (2009) and Wittstock (1984, Definition 1.1) and finesse the equality conditions (2.4) and replace them by

$$(U'U - I) \preceq 0, (V'V - I) \preceq 0, \qquad (2.5)$$

where $A \preceq 0$ means that the matrix A is negative semi-definite. Now, it is evident that for V fixed, (2.3) subject to (2.5) is convex in UD, and similarly it is convex in VD for U fixed. Thus, the function $\Psi(\cdot)$ in (2.3) is bi-convex in UD and VD subject to (2.5). It can be minimized (Gorski et al., 2007) by iteratively minimizing the convex functions (2.6) and (2.7) below with respect to $\tilde{U} = UD$ and $\tilde{V} = VD$

while keeping the other fixed:

$$||Y - \tilde{U}V'||_F^2 + \lambda_u \sum_{i=1}^{p}\sum_{k=1}^{r} |\tilde{u}_{ik}|, \tag{2.6}$$

$$||Y - U\tilde{V}'||_F^2 + \lambda_v \sum_{j=1}^{q}\sum_{k=1}^{r} |\tilde{v}_{jk}|. \tag{2.7}$$

Fortunately, the minimizers of (2.6) and (2.7) have closed forms as component-wise soft-thresholding operators acting on Y. Moreover, the minimizers $\tilde{U}^{(k)}$ and $\tilde{V}^{(k)}$ in the kth iteration of (2.3) have the same form as those in the kth updating steps $U^{(k),thr}$, $V^{(k),thr}$ in the FIT-SSVD algorithm for certain threhsolding levels, see Figure 2.1. These observations are summarized in the following proposition and its proof is given in the Appendix A.

Proposition 2.2.1 **(a)** *Given the data matrix Y and model (2.2), the objective function $\Psi(.)$ in (2.3) subject to conditions (2.5) is bi-convex and can be minimized by alternatively minimizing (2.6)-(2.7) with respect to $\tilde{U}, \tilde{V}$.*
(b) *The solution $\tilde{U}$ of (2.6) for V fixed is the component-wise soft-thresholding of YV, i.e. $S(YV, \frac{1}{2}\lambda_u) = [S(YV)_{ij}, \frac{1}{2}\lambda_u]_{i=1,\dots,n;j=1,\dots,r}$. Similarly, the solution $\tilde{V}$ of (2.7) for U fixed is $S(Y'U, \frac{1}{2}\lambda_v)$, where $S(A, \lambda)$ for a matrix A denotes soft-thresholding every entry of A with threshold level λ.*
(c) *The FIT-SSVD algorithm is equivalent to iteratively minimizing (2.6)-(2.7) for fixed λ_u, λ_v, and then obtaining the orthonormal matrices U and V through the QR decompositions of their solutions.*

Proposition 2.2.1 shows that the FIT-SSVD algorithm provides the solution for the estimator (U, V) of (2.3) subject to (2.5) through iteratively solving for the matrices U and V. Compared to the sequential algorithms in Lee et al. (2010),Witten et al. (2009) and Chen et al. (2012a), it is a matrix-based algorithm. It solves for the matrices U and V directly as opposed to solving for their column vectors sequentially. Specifically, if the threshold levels γ_u, γ_v in the FIT-SSVD algorithm are set to be $\frac{1}{2}\lambda_u, \frac{1}{2}\lambda_v$ as in (2.3), then the soft-thresholding estimators of the kth iteration $\tilde{U}^{(k)}, \tilde{V}^{(k)}$ in (2.6) and (2.7) are identical to the estimators $U^{(k),thr}, V^{(k),thr}$ in the FIT-SSVD (Figure 2.1, Step 1 and 3).

An advantage of placing the FIT-SSVD algorithm in the optimization-based framework is that the asymptotic properties of its solution can be studied using the techniques developed for the Lasso-type objective functions (Knight and Fu, 2000; Zou, 2006; Chen et al., 2012a; Chen and Huang, 2012). We discuss the existence of a local minimum of (2.3) and the selection consistency of its solutions in the Appendix A.

2.3 The Generalized Thresholding for Sparse Reduced Rank Regression

In this subsection, we reduce the multivariate regression model (2.1) to a low-rank model for a correlated data matrix, and then generalize the FIT-SSVD to the correlated data situation. The transformed model (2.9) below is the bridge connecting the regression problem to the low-rank model and the standard SVD problems. Compared to the FIT-SSVD algorithm the key changes in the FIT-SRRR are in selecting the threshold levels for correlated data and the separate updating of U and V, which no longer is symmetric in U and V due to the dependence in the rows of the data matrix.

The close connection between the FIT-SSVD and FIT-SRRR methodologies is explained by noting that at least for a full-rank design matrix, X can be removed from (2.1) by left multiplying both sides by $(X'X)^{-1}X'$ leading to

$$(X'X)^{-1}X'Y \quad = \quad (X'X)^{-1}X'XB + (X'X)^{-1}X'E. \tag{2.8}$$

It appears to be of the form (2.2), but with Y and E replaced by $\tilde{Y} = \hat{B} = (X'X)^{-1}X'Y$ the least-square estimate of B, and the transformed noise $\tilde{E} = (X'X)^{-1} X'E$. However, the model (2.8) is more general than (2.2), since with

$$\tilde{Y} = B + \tilde{E}, \tag{2.9}$$

the columns of $\tilde{E}$ are iid $N_p(0, \sigma^2\Sigma)$ where $\Sigma = (X'X)^{-1}$ is known. Depending on the rank of the design matrix X, the following three cases are of interest:

I: X is orthonormal and $X'X = I$: Then, the proposed FIT-SRRR algorithm reduces to the FIT-SSVD algorithm of Yang et al. (2013). The latter will be used verbatim in computing the sparse SVD of the coefficient matrix B where $\tilde{Y} = X'Y$ is used as the data matrix in the algorithm.

II: X has full column rank: Then, the entries of transformed noise $\tilde{E}$ in (2.9) are no longer iid, but row-wise dependent, so that the FIT-SSVD algorithm is not directly applicable, and it needs to be modified to account for the correlation.

III: X is less than full-rank: In this case, there is no unique least square estimator of B in (2.8) because the Gram matrix $X'X$ is singular. Some alternative estimators are the Moore-Penrose inverse (Bunea et al., 2011), and the ridge estimator. Throughout the Section 2, following Chen et al. (2012a, p. 8), the ridge estimator is used where a small positive constant $\epsilon = 10^{-4}$ is added to the diagonal elements of the Gram matrix to make it invertible.

In the rest of this subsection, we describe the details of the FIT-SRRR algorithm, especially in updating $U^{(k)}$, $V^{(k)}$ and choosing the corresponding threshold levels.

1. Right-to-left Multiplication and Threshold $U^{(k),thr} = \eta(\tilde{Y}V^{(k-1)}, \gamma_u)$, where γ_u is selected by Algorithm 1,
2. Orthonormalization with QR Decomposition: $U^{(k)}R_u^{(k)} = U^{(k),thr}$,
3. Left-to-right Multiplication and Threshold $V^{(k),thr} = \eta(\tilde{Y}'U^{(k)}, \gamma_v)$, where γ_v is selected by Algorithm 2,
4. Orthonormalization with QR Decomposition: $V^{(k)}R_v^{(k)} = V^{(k),thr}$.

Figure 2.2: The four key steps of the FIT-SRRR algorithm.

2.3.1 Threshold Levels and Updating the $U^{(k)}, V^{(k)}$

The goal of thresholding is to retain the entries of $\tilde{Y}$ with high signal and replace the others with zero. Due to row-dependence in the data matrix in (2.9), unlike the FIT-SSVD, finding $U^{(k)}, V^{(k)}$ in Step 1 and 3 in Figure 2.2 are no longer symmetric, although they both require thresholding. We highlight this difference using some properties of matrix normal distributions (Gupta and Nagar, 1999). Recall that a random matrix Y $(m \times n)$ is said to have a matrix normal distribution and denoted as $Y \sim MN_{m,n}(M, \Sigma, \Omega)$, where M is the mean matrix and Σ, Ω denote the row and column covariance matrices. The following properties of linear transformations of matrix normal distributions are needed in the sequel.

Proposition 2.3.1 **(a)** *Let $Y \sim MN_{m,n}(M, \Sigma, \Omega)$ and $\mathbf{a}$ be a suitable vector, then $Y\mathbf{a} \sim N_m(M\mathbf{a}, (\mathbf{a}'\Omega\mathbf{a})\Sigma)$, and $Y'\mathbf{a} \sim N_n(M'\mathbf{a}, (\mathbf{a}'\Sigma\mathbf{a})\Omega)$.*

(b) *If $\tilde{E} \sim MN_{p,q}(0, \sigma^2\Sigma, I_q)$ as in (2.9), and $\mathbf{u}, \mathbf{v}$ are suitable vectors of unit norm, then the entries of $\tilde{E}\mathbf{v} \sim N_p(0, \sigma^2\Sigma)$ are dependent, while those of $\tilde{E}'\mathbf{u} \sim N_q(0, \sigma^2(\mathbf{u}'\Sigma\mathbf{u})I_q)$ are independent.*

We discuss the details of selecting the threshold levels and the updating procedures for $U^{(k)}, V^{(k)}$ through their lth column $\mathbf{u}_l^{(k)}, \mathbf{v}_l^{(k)}, l = 1 \cdots, r$. To update $U^{(k)}$, let $V^{(k-1)}$ be the update of V at the $(k-1)$st iteration and $\mathbf{v}_l^{(k-1)}$ be its lth column. Right multiplying both sides of (2.9) by $\mathbf{v}_l^{(k-1)}$ and using the SVD of B leads to the mean model,

$$\tilde{Y}\mathbf{v}_l^{(k-1)} = B\mathbf{v}_l^{(k-1)} + \tilde{E}\mathbf{v}_l^{(k-1)}. \tag{2.10}$$

For the moment, let γ_{ul} be a threshold level used in (2.10). Then, $[\tilde{Y}\mathbf{v}_l^{(k-1)}]^{thr}$, a thresholded version of the response vector, would serve as an estimator of the mean vector in (2.10). Repeating this procedure for all $l = 1, \cdots, r$ leads to the matrix $[\tilde{Y}V^{(k-1)}]^{thr}$, and the orthonormal matrix $U^{(k)}$ is obtained from its QR decomposition.

From Proposition 2.3.1(b), since the noise $\tilde{E}\mathbf{v}_l^{(k-1)}$ in (2.10) is dependent with covariance matrix Σ, the universal threshold level (Donoho and Johnstone, 1994) used in Yang et al. (2013) is not applicable. Then, we need to modify the FIT-SSVD to account for the dependence and heterogeniety in (2.10). Johnstone and Silverman (1997) and Kovac and Silverman (2000) suggest more general thresholding methods for correlated and heteroscedastic noise. In brief, their methods view the entries of the noise vector $\tilde{E}\mathbf{v}_l^{(k-1)}$ as independent heterogeneous random variables and ignore the correlation structure.

Berkner and Wells (1998, 2001) and Delouille et al. (2004) generalize the above thresholding methods by incorporating the correlations. More precisely, the threshold level γ_{ujl} for the jth entry of the vector $\tilde{Y}\mathbf{v}_l^{(k-1)}$ is selected as

$$\gamma_{ujl} \quad = \quad \hat{\sigma}_{ujl}\gamma_{ul}, \tag{2.11}$$

where from Proposition 2.3.1(b), $\hat{\sigma}^2_{ujl} = \hat{\sigma}^2\sigma_{jj}$ and σ_{jj} is the j^{th} diagonal entry of Σ. The γ_{ul}, following Berkner and Wells (2001), is given by

$$\gamma_{ul} = \sqrt{2(1+\delta)\log(p)}, \tag{2.12}$$

where

$$\delta = \max_{j \neq j'}(|corr(\{\tilde{Y}\mathbf{v}_l^{(k-1)}\}_j, \{\tilde{Y}\mathbf{v}_l^{(k-1)}\}_{j'})|),$$

is the maximum of the magnitudes of the correlations in $\tilde{Y}\mathbf{v}_l^{(k-1)}$. We adopt this thresholding procedure for the correlated case in updating $U^{(k)}$, and summarize the details in Algorithm 1.

```
**Algorithm 1:** Selection of the threshold level $\gamma_{ul} = g_u(\tilde{Y}, U^{(k-1)}, V^{(k-1)}, \hat{\sigma})$.
  **Input**:
  1. Data matrix $\tilde{Y}$, covariance matrix $\Sigma$ for the rows of the noise matrix;
  2. Previous estimators of the singular vectors $U^{(k-1)}, V^{(k-1)}$;
  3. An estimate of $\hat{\sigma}$.
  **Output**:
  Threshold level vectors $\gamma_{ul}$ for $l = 1, ..., r$.
  $\hat{\sigma}^2_{ujl} \leftarrow \hat{\sigma}^2\sigma_{jj}$, where $\sigma_{jj}$ is the $j^{th}$ diagonal entry of $\Sigma$;
  $\gamma_{ul} \leftarrow \sqrt{2(1+\delta)\log(p)}$ where $\delta = \max_{j \neq j'}(|corr(\{\tilde{Y}\mathbf{v}_l^{(k-1)}\}_j, \{\tilde{Y}\mathbf{v}_l^{(k-1)}\}_{j'})|)$;
  $\gamma_{ujl} \leftarrow \hat{\sigma}_{ujl}\gamma_{ul}$, $j = 1, ..., p$;
  **return** $\gamma_{ul} = (\gamma_{u1l}, \ldots, \gamma_{upl})'$, $l = 1, ..., r$.
```

We update $V^{(k)}$ through its lth column $\mathbf{v}_l^{(k)}, l = 1, ..., r$. Right multiplying the transpose of both side of (2.9) by $\mathbf{u}_l^{(k)}$, it follows that

$$\tilde{Y}'\mathbf{u}_l^{(k)} = B'\mathbf{u}_l^{(k)} + \tilde{E}'\mathbf{u}_l^{(k)}, \tag{2.13}$$

where the entries of the vector $\tilde{E}'\mathbf{u}_l^{(k)}$ are iid $N(0, \sigma^2 \mathbf{u}_l^{(k)\prime} \Sigma \mathbf{u}_l^{(k)})$, see Proposition 2.3.1(b). Since the noises in (2.13) are uncorrelated, a theoretically sensible (though not actionable) threshold level for $\tilde{Y}'\mathbf{u}_l^{(k)}$ would be $\gamma_{vl} = \mathbf{E}\{||(\tilde{E}'\mathbf{u}_l^{(k)})||_\infty\}$ (Yang et al., 2013, Section 2.4). We adjust their Algorithm 3 to obtain the threshold levels $\gamma_{vl}, l = 1, ..., r$ for the regression setup. Since the SVD of the matrix B is assumed to be sparse, let L_u, L_v denote the index sets for U and V in which every element in a row is zero, and H_u, H_v be the complimentary sets of L_u and L_v. Then, after a reordering of the rows and columns of B (still denoted as B), it can be partitioned as,

$$B = UDV' = \begin{bmatrix} B_{H_uH_v} & B_{H_uL_v} \\ B_{L_uH_v} & B_{L_uL_v} \end{bmatrix} = \begin{bmatrix} \underbrace{\begin{matrix} B_{H_uH_v} \\ 0_{L_uH_v} \end{matrix}}_{p\times|H_v|} & \underbrace{\begin{matrix} 0_{H_uL_v} \\ 0_{L_uL_v} \end{matrix}}_{p\times|L_v|} \end{bmatrix}.$$

Using a compatible partitioning of X, (2.1) can be rewritten as,

$$Y = \begin{bmatrix} \underbrace{\begin{matrix} X_{11}B_{H_uH_v} \\ X_{21}B_{H_uH_v} \end{matrix}}_{n\times|H_v|} & \underbrace{\begin{matrix} 0 \\ 0 \end{matrix}}_{n\times|L_v|} \end{bmatrix} + E, \tag{2.14}$$

where X_{11}, X_{21} are the corresponding submatrices of X. Note that in our setting, the (2,1)-block or the submatrix $X_{21}B_{H_uH_v}$ in (2.14) is not a zero matrix as in Yang et al. (2013, Section 2.4) and the pure noise part is a $n \times |L_v|$ matrix. Following Yang et al. (2013, Section 2.4), if the dimension of the pure noise is large enough, say, $n|L_v| > pq\log(pq)$, we find the threshold level by using the rule of "m out n" bootstrap (Bickel *et al.,* 1997). Otherwise, we use the universal threshold level

$$\gamma_{vl} = \hat{\sigma}_{vl}\sqrt{2\log(q)}, \tag{2.15}$$

where $\hat{\sigma}^2_{vl} = \hat{\sigma}^2\mathbf{u}_l^{(k)\prime}\Sigma\mathbf{u}_l^{(k)}$, for $l \in \{1, ..., r\}$. The details of choosing the threshold level is presented in the Algorithm 2. Then, $[\tilde{Y}'\mathbf{u}_l^{(k)}]^{thr}$, a thresholded version of the response vector in (2.13), would serve as an estimator of the mean vector. Repeating this procedure for all $l = 1, \cdots, r$ leads to the matrix estimator $[\tilde{Y}'U^{(k)}]^{thr}$, where the orthonormal matrix $V^{(k)}$ is obtained from its QR decomposition.

```
Algorithm 2: Selection of the threshold level $\gamma_{vl} = g_v(\tilde{Y}, U^{(k)}, V^{(k-1)}, \hat{\sigma})$.
Input:
1. Data matrix $\tilde{Y}$, covariance matrix $\Sigma$ for the columns of the noise matrix;
2. Previous estimators of the singular vectors $U^{(k)}, V^{(k-1)}$;
3. Pre-specified number M of bootstraps;
4. An estimate of $\hat{\sigma}$.
Output:
Threshold level $\gamma_{vl}$ for $l = 1, ..., r$.
Subset selection: $L_u = \{i : u_{i1}^{(k)} = ... = u_{ir}^{(k)} = 0\}$,
$L_v = \{j : v_{j1}^{(k-1)} = ... = v_{jr}^{(k-1)} = 0\}$, $H_u = L_u^c$, $H_v = L_v^c$;
if $n|L_v| > pq\log(pq)$ then
    for $t$ in $1, \cdots, M$ do
        Sample $pq$ entries from the pure noise part in (2.14) and reshape them
        into a matrix $Z \in R^{q\times p}$;
        $C = [C_{:1}, \dots, C_{:r}] \leftarrow Z[\Sigma^{1/2}U^{(k)}]_{H_u} \in^{q\times r}$;
        $D_{t:} \leftarrow (||C_{:1}||_\infty, \cdots, ||C_{:r}||_\infty) \in R^{1\times r}$ ;
        $\gamma_{vl} \leftarrow \text{median}(D_{:l})$, $l = 1, \dots, r$;
else
    $\gamma_{vl} \leftarrow \hat{\sigma}_{vl}\sqrt{2\log(q)}$, where $\hat{\sigma}^2_{vl} = \hat{\sigma}^2\mathbf{u}_l^{(k)\prime}\Sigma\mathbf{u}_l^{(k)}$;
return $\gamma_{vl}$, $l = 1, ..., r$.
```

2.3.2 Implementing the Algorithm

The FIT-SRRR algorithm is designed to inherit the statistical and computational properties of the FIT-SSVD algorithm. Its main steps are described in Algorithm 3, where the sub-algorithms for selecting the threshold levels for U and V are given in Section 2.3.1.

The initial orthonormal matrices $U^{(0)} \in R^{p\times r}$ and $V^{(0)} \in R^{q\times r}$ are chosen as in Yang et al. (2013) by first reducing the dimensionality of data matrix as in Johnstone and Lu (2009) and then computing its ordinary SVD. We stop the FIT-SRRR Algorithm after the kth iteration if the maximum distance between the successive iterates is small, i.e. $max\{||P_{U^{(k)}} - P_{U^{(k-1)}}||_2^2, ||P_{V^{(k-1)}} - P_{V^{(k-1)}}||_2^2\}$ is less than or equal to a preselected $\epsilon = 10^{-8}$, where $P_A = AA'$ is a projection matrix for an orthonormal matrix A, and $||A||_2$ denotes the spectral norm of the matrix A.

2.4 Simulations

In this subsection, we use simulations to assess and compare the performance of the FIT-SRRR with the existing methods, and report the results in next two subsections. The first subsection compares the FIT-SRRR with the FIT-SSVD for correlated data matrices in model (2.2). The second compares the FIT-SRRR in the regression model (2.1) with the iterative exclusive extraction algorithm (IEEA) in Chen et al. (2012a) described in Subsection 2.4.2.

Throughout, the rank of the true underlying matrix B is assumed to be known. The parameters and setups are as the same as in the simulations in Yang et al. (2013) and Chen et al. (2012a), i.e. we keep their setups for the threshold function, Huberization, bootstrap, cross-validation and initial values. In particular, for the FIT-SRRR, the bootstrap parameter $M = 100$ in Algorithm 2, the threshold function is the hard thresholding $\eta(\cdot) = H(\cdot)$ in Algorithm 3, and $\hat{\sigma} = 1.4826$

Algorithm 3: The FIT-SRRR Algorithm.

```
Input:
1. Data matrix $\tilde{Y}$, covariance matrix $\Sigma$ for the rows of the noise matrix;
2. Target rank r, and an estimate of $\hat{\sigma}$;
3. Thresholding function $\eta$;
4. Algorithms $g_u$ and $g_v$ to calculate the threshold levels $\gamma_u, \gamma_v$;
5. Initial orthonormal matrices $U^{(0)} \in R^{p \times r}$, $V^{(0)} \in R^{q \times r}$.
Output:
Estimators $\hat{U} = U^{(\infty)}, \hat{V} = V^{(\infty)}$.
repeat
    Obtain the matrix $U^{(k),thr} \leftarrow [\mathbf{u}_1^{(k),thr}, ..., \mathbf{u}_r^{(k),thr}]$, where
    $\mathbf{u}_l^{(k),thr} = \eta(\tilde{Y}\mathbf{v}_l^{(k-1)}, \gamma_{ul})$ and $\gamma_{ul} = g_u(\tilde{Y}, U^{(k)}, V^{(k-1)}, \hat{\sigma})$ for $l = 1, \cdots, r$;
    Orthonormalization with QR decomposition for $U$: $U^{(k)} R_u^{(k)} \leftarrow U^{(k),thr}$;
    Obtain the matrix $V^{(k),thr} \leftarrow [\mathbf{v}_1^{(k),thr}, ..., \mathbf{v}_r^{(k),thr}]$ where
    $\mathbf{v}_l^{(k),thr} = \eta(\tilde{Y}'\mathbf{u}_l^{(k)}, \gamma_{vl})$ and $\gamma_{vl} = g_v(\tilde{Y}, U^{(k)}, V^{(k-1)}, \hat{\sigma})$ for $l = 1, \cdots, r$;
    Orthonormalization with QR decomposition for $V$: $V^{(k)} R_v^{(k)} \leftarrow V^{(k),thr}$;
until Convergence;
return Estimators $\hat{U} = U^{(\infty)}, \hat{V} = V^{(\infty)}$.
```

$MAD(as.vector(\tilde{Y}))$ is a multiple of the median absolute deviation (MAD) of the data. The repetitions for each simulation are $N = 1000$ times.

2.4.1 The Low-Rank Model with Correlated Data

In this subsection, we consider the low-rank model (2.2) with correlated data. Yang et al. (2013) have compared the FIT-SSVD algorithm with several sequential sparse SVD methods, such as SSVD in Lee et al. (2010), PMD-SVD in Witten et al. (2009), and found it to outperform them in terms of estimation accuracy and computation cost. Hence, it suffices here to compare the FIT-SRRR with only the FIT-SSVD algorithm.

The following three covariance structures for $\Sigma = (\sigma_{ij})$ of the correlated noise in model (2.2) are used to illustrate the effects of correlated error on the FIT-SSVD algorithm and the proposed FIT-SRRR algorithm which is designed to account for the correlation.

1. **Compound symmetry,** CS.

$$\sigma_{ij} = \begin{cases} \sigma^2 & \text{if } i = j, \\ \sigma^2\gamma & \text{if } i \neq j. \end{cases}$$

2. **Auto Regression,** $AR(1)$.

$$\sigma_{ij} = \sigma^2 \rho^{|i-j|}, i, j \in \{1, \cdots, p\},$$

3. **Moving Average,** $MA(1)$.

$$\sigma_{ij} = \begin{cases} \sigma^2 & \text{if } i = j, \\ \sigma^2 \frac{\theta}{1+\theta^2} & \text{if } |i-j| = 1, \end{cases}$$

where γ, ρ and θ are the parameters in CS, $AR(1)$ and $MA(1)$ structures, respectively.

We generate data matrices according to model (2.2) with covariance structures from the above list and set parameters γ, ρ and θ to the three levels: 0.1, 0.5, 0.9. We choose σ^2 so as to have four different levels of signal to noise ratio (SNR): 1, 0.5, 0.25 and 0.125, where the SNR is calculated following Chen et al. (2012a, Section 4) and Yang et al. (2013).

For the unit rank B, following the example in Lee et al. (2010) and Chen et al. (2012a), we let the signal $B = d\mathbf{u}\mathbf{v}'$ be a 50×100 matrix ($p = 50$ and $q = 100$), with

$d = 50$ and

$$
\begin{aligned}
\tilde{\mathbf{u}} &= (10, -10, 8, -8, 5, -5, rep(3,5), rep(-3,5), rep(0,34))', \\
\tilde{\mathbf{v}} &= (10, 9, 8, 7, 6, 5, 4, 3, rep(2,17), rep(0,75))', \\
\mathbf{u} &= \frac{\tilde{\mathbf{u}}}{||\tilde{\mathbf{u}}||}, \mathbf{v} = \frac{\tilde{\mathbf{v}}}{||\tilde{\mathbf{v}}||},
\end{aligned}
$$

where $rep(m, n)$ denotes a vector of length n, whose entries are all equal to m.

For the rank-2 B, we set $B = \sum_{l=1}^{2} d_l \mathbf{u}_l \mathbf{v}_l'$ where

$$
\begin{aligned}
\tilde{\mathbf{u}}_1 &= (10, -10, 8, -8, 5, -5, rep(3,4), rep(0,40))', \\
\tilde{\mathbf{u}}_2 &= (rep(0,40), rep(2,5), rep(-2,5))', \\
\tilde{\mathbf{v}}_1 &= (10, 9, 8, 7, 6, 5, 4, 3, rep(2,17), rep(0,75))', \\
\tilde{\mathbf{v}}_2 &= (rep(0,80), 5, 4, 3, rep(2,17))', \\
d_1 &= 40, d_2 = 30,\ \mathbf{u}_l = \frac{\tilde{\mathbf{u}}_l}{||\tilde{\mathbf{u}}_l||}, \mathbf{v}_l = \frac{\tilde{\mathbf{v}}_l}{||\tilde{\mathbf{v}}_l||}, l = 1, 2.
\end{aligned}
$$

We measure the estimation accuracy of a method using the average mean-squared error ratio from the 1000 simulation repetitions, i.e. $MSE_i = ||B - \hat{B}_i||_F^2$, and the average MSE-ratio is calculated by $\frac{1}{1000}\sum_{i=1}^{1000} \frac{MSE_{i,FIT-SSVD}}{MSE_{i,FIT-SRRR}}$, where a value greater than 1 indicates better performance of our proposed method.

Table 2.1 summarizes the results for the three correlation structures, four levels of SNRs and two ranks $r = 1, 2$. It is evident that for the CS structure, the FIT-SRRR universally outperforms the FIT-SSVD. It enjoys a significantly lower level of mean-square errors than its counterpart, especially when the correlation parameter γ is large. It holds true for all levels of SNRs, and both ranks $r = 1, 2$. For instance,

Rank	SNR	CS (γ)	Ratio	$AR1$ (ρ)	Ratio	$MA1$ (θ)	Ratio
1	1.00	0.1	1.01	0.1	1.00	0.1	1.00
		0.5	1.05	0.5	1.02	0.5	1.01
		0.9	2.92	0.9	1.04	0.9	1.03
	0.50	0.1	1.12	0.1	1.12	0.1	1.12
		0.5	1.23	0.5	1.27	0.5	1.20
		0.9	22.34	0.9	1.09	0.9	1.17
	0.25	0.1	1.09	0.1	1.11	0.1	1.12
		0.5	2.87	0.5	1.50	0.5	1.96
		0.9	19.15	0.9	2.05	0.9	1.96
	0.125	0.1	1.00	0.1	1.00	0.1	1.00
		0.5	6.85	0.5	1.01	0.5	1.12
		0.9	19.18	0.9	3.06	0.9	1.12
2	1.00	0.1	1.04	0.1	1.06	0.1	1.06
		0.5	2.14	0.5	1.18	0.5	1.11
		0.9	5.46	0.9	1.10	0.9	1.14
	0.50	0.1	1.11	0.1	1.09	0.1	1.10
		0.5	4.54	0.5	1.01	0.5	1.15
		0.9	4.67	1.74	1.69	0.9	0.85
	0.25	0.1	1.03	0.1	0.96	0.1	0.94
		0.5	4.38	0.5	1.03	0.5	0.87
		0.9	6.71	0.9	2.73	0.9	0.78
	0.125	0.1	1.05	0.1	0.95	0.1	0.95
		0.5	6.38	0.5	0.99	0.5	0.95
		0.9	11.07	0.9	4.47	0.9	0.94

Table 2.1: Average MSE-ratios of FIT-SSVD to FIT-SRRR for various correlation structures, SNRs and ranks.

the MSE-ratio in the CS situation reaches as high as 22. For the $AR(1)$ structure, the FIT-SRRR outperforms the FIT-SSVD in most situations, in other cases its performance is very close to that of the FIT-SSVD. The situation for the $MA(1)$ structure is quite different in the sense that for higher-rank B and lower SNR the FIT-SRRR underperforms FIT-SSVD. In terms of the strength of dependence in these three covariance structures, the CS allows the strongest dependence, followed by $AR(1)$ and $MA(1)$. Note that the parameter δ used in (2.12) does not fully characterize a correlation matrix. It only reflects the maximum of magnitudes of the correlation coefficients. For example, $\theta = 0.5$ in $MA(1)$ and $\gamma = 0.4$ in CS, both lead to a $\delta = 0.4$. However, the nature of the correlations in the two situations are very different.

We have also selected the threshold levels using the method in Johnstone and Silverman (1997) and Kovac and Silverman (2000). This amounts to using (2.12) with $\delta = 0$. Our simulations results show that the FIT-SRRR outperformed the FIT-SSVD in most of the $MA(1)$ situations, but the performances for the CS structure was not as strong as those reported in Table 2.1.

As a whole, in the presence of higher dependence and correlations, the FIT-SRRR outperforms the FIT-SSVD, but the latter performs well when the correlation is light and negligible. A better measure of dependence than (2.12) which incorporates the whole correlation structure in determining the threshold level is expected to improve the performance of the FIT-SRRR for correlated data.

2.4.2 Multivariate Reduced Rank Regression Model

In this subsection, we compare the FIT-SRRR with the IEEA algorithm by simulating data matrices using the regression model (2.1) where the rank of B is set to be 3.

The iterative exclusive extraction algorithm (IEEA) proposed by Chen et al. (2012a) estimates B with sparse SVD structure starting from some initial consistent estimator of B, e.g. the reduced rank least square estimator $\hat{B}_{OLS} = \sum_{l=1}^{r} \hat{d}_l \hat{\mathbf{u}}_l \hat{\mathbf{v}}_l'$. They reduce the task of regularizing B into r parallel sparse unit rank regressions by decomposing the response matrix Y into r layers $Y_l = Y - X(\hat{B}_{OLS} - \hat{B}_{OLS,l})$, where $\hat{B}_{OLS,l} = \hat{d}_l \hat{\mathbf{u}}_l \hat{\mathbf{v}}_l'$ and $l = 1, ..., r$, and solve the sparse regression of Y_l's on X with unit rank coefficient matrix. Chen et al. (2012a) show that the IEEA outperforms the ordinary least square (OLS), RRR and the nuclear norm penalized (NNP) least square in Yuan et al. (2007) in terms of estimation and prediction accuracies. Hence it suffices here to compare our FIT-SRRR only with the IEEA algorithm.

Two scenarios in terms of moderate ($p, q < n$) versus high model dimensions ($p, q > n$) are considered (Chen et al., 2012a), i.e. the Models I and II below. (The special case of the identity design matrix X has also been studied, and the results are presented in the Appendix A.) The response matrix Y is generated from (2.1) using the X and B described below. We construct the design matrix X by generating its rows from an iid $N_p(0, \Omega)$ distribution, where Ω is the covariance matrix of an $AR(1)$ process with the fixed parameter $\rho = 0.5$. Once the design matrix is generated, it remains fixed in all the replications of the simulation. For the coefficient matrix $B = \sum_{l=1}^{r=3} d_l \mathbf{u}_l \mathbf{v}_l'$, we consider two different configurations for the singular values corresponding to the well-separated case $D_1 = diag(20, 15, 10)$ and the less well-separated case $D_2 = diag(20, 18, 16)$ (Ma, 2011). The elements of E are generated from iid $N(0, \sigma^2)$.

Model I: $n = 50$, $p = q = 25$, the design matrix X has full column rank, and

$$\begin{aligned}
\tilde{\mathbf{u}}_1 &= (unif(5, J), rep(0, 20))', \\
\tilde{\mathbf{u}}_2 &= (rep(0, 5), unif(5, J))', \\
\tilde{\mathbf{u}}_3 &= (rep(0, 10), runif(5, J))', \\
\tilde{\mathbf{v}}_1 &= (rep(1, 5), rep(-1, 5), rep(0, 15))', \\
\tilde{\mathbf{v}}_2 &= (rep(0, 12), rep(1, 5), rep(-1, 5), rep(0, 3))', \\
\tilde{\mathbf{v}}_3 &= (rep(0, 6), \tilde{\mathbf{v}}_1[7:8], -\tilde{\mathbf{v}}_1[9:10], 1, -1, -\tilde{\mathbf{v}}_2[13:14], \tilde{\mathbf{v}}_2[15:16], rep(0, 9))', \\
\mathbf{u}_l &= \frac{\tilde{\mathbf{u}}_l}{||\tilde{\mathbf{u}}_l||}, \mathbf{v}_l = \frac{\tilde{\mathbf{v}}_l}{||\tilde{\mathbf{v}}_l||}, l = 1, 2, 3,
\end{aligned}$$

where $\tilde{\mathbf{v}}_l[a:b]$ denotes a vector whose entries are the corresponding entries of $\tilde{\mathbf{v}}_l$ from a to b, and $unif(m, J)$ denotes a vector of length m whose entries are iid uniformly distributed on the set of $J = [-1, -.3] \cup [.3, 1]$.

Model II: $n = 50$, $p = q = 60$, since $n < p = q$, the design matrix X is singular. For $\tilde{\mathbf{u}}_l$ and $\tilde{\mathbf{v}}_l$ we use the same setting in the Model I, except that we add 35 0's to each $\mathbf{u}_l$'s and $\mathbf{v}_l$'s to make them 60×1 vectors.

As before σ^2 is chosen to make the signal-to-noise ratio to be approximately equal to 1, 0.5, 0.25 and 0.125. The performance is measured in terms of estimation accuracy using $\text{MSE}_i = ||B - \hat{B}_i||_F^2$ and prediction accuracy using the $\text{PMSE}_i = ||XB - X\hat{B}_i||_F^2$, $i = 1, ..., 1000$. We report the average error ratios, i.e. the average MSE-ratio $= \frac{1}{1000}\sum_{i=1}^{1000} \frac{MSE_{i,IEEA}}{MSE_{i,FIT-SRRR}}$, the average PMSE-ratio $= \frac{1}{1000}\sum_{i=1}^{1000} \frac{PMSE_{i,IEEA}}{PMSE_{i,FIT-SRRR}}$, where a ratio greater than 1 indicates better performance of our proposed method.

Singular values	Model	Ratio	SNR			
			1	0.5	0.25	0.125
D_1=(20,15,10)	I	MSE	1.35	1.35	1.41	1.39
		PMSE	1.35	1.34	1.32	1.39
	II	MSE	2.11	2.10	2.08	1.37
		PMSE	2.36	2.12	2.14	1.44
D_2=(20,18,16)	I	MSE	2.74	2.52	2.66	2.71
		PMSE	2.73	2.51	2.64	2.67
	II	MSE	2.04	2.20	2.22	1.28
		PMSE	2.22	2.22	2.27	1.86

Table 2.2: Average ratios of estimation error (MSE) and prediction error (PMSE) of the IEEA and FIT-SRRR methods.

Table 2.2 reports the simulation results of ratios of the estimation error (MSE) and prediction error (PMSE). As a whole, we see that the FIT-SRRR universally outperforms the IEEA since every entry is greater than 1. It enjoys a lower level of MSEs and PMSEs for all levels of SNRs, singular values $D1$ and $D2$ and both Models I and II. To be specific, when the singular values are well-separated and the design matrix is of full-rank (upper panel, Model I), the MSE-ratio and PMSE-ratio reach as high as 1.41 and 1.39, respectively. For the lower panel, where the singular values are not well-separated, the FIT-SRRR performs even better. The ratios are greater than 2 in most situations. For instance, the MSE-ratio and PMSE-ratio reach as high as 2.74 and 2.73, respectively. This good performance of the FIT-SRRR algorithm in the lower panel is probably due to its capability to capture the whole subspace of the singular vectors of the coefficient matrix instead of estimating one layer at a time.

The FIT-SRRR algorithm inherits the fast computational properties of the FIT-SSVD. The whole simulation exercise only took a few seconds. For example, it takes about 0.06 system seconds on average for one run for the moderate dimension, well-separated singular values, using our R program running on a Windows 7 desktop with Intel Core i5 Duo CPU of a clock speed of 7.19 Gigahertz.

2.5 Example: Lung Cancer Data

The lung cancer data consists of expression levels of 12,625 genes, measured from 56 subjects divided into 4 groups: one group of normal subjects (Normal) plus patients with one of the three following types of lung cancer: pulmonary carcinoid tumors (Carcinoid), colon metastases (Colon), and Small cell carcinoma (SmallCell). Hence, the observed data matrix Y is $56 \times 12,625$. A detailed description of the data can be found in Bhattacharjee et al. (2001).

Since for each subject the cancer type information is available, it could be considered as a covariate in the multivariate regression model (2.1). Let X denote the 56×4 orthonormal design matrix indicating the subjects cancer category:

$$X = \begin{pmatrix} \frac{1}{\sqrt{20}}\mathbf{1_{20}} & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{13}}\mathbf{1_{13}} & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{17}}\mathbf{1_{17}} & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{6}}\mathbf{1_{6}} \end{pmatrix}, \tag{2.16}$$

where $\mathbf{1_{20}}$ is a vector of length 20, whose entries are all equal to 1. The length of the vector reflects the sample size of each category.

This data set has been analyzed using the SSVD in Lee et al. (2010), the penalized matrix decomposition (PMD) and the FIT-SSVD in Yang et al. (2013), and the IEEA algorithm in Chen et al. (2012a). The SSVD algorithm in Lee et al. (2010) applies penalized least square on rank-one matrix approximations with additive penalties on singular vectors, and the PMD in Witten et al. (2009) is a low-rank SVD decomposition with constraints on the singular vectors. Note that the PMD, SSVD and FIT-SSVD directly work with the data matrix without using the covariate information in (2.16), however, only the IEEA uses the covariate information. Unfortunately, all these methods, with the exception of the FIT-SSVD, lack the orthogonality of the singular vectors.

We apply the FIT-SRRR method using (2.16) as the design matrix and consider the first three layers or use $r = 3$ as in Lee et al. (2010), Yang et al. (2013) and Chen et al. (2012a), which allows comparison of the results of different methods.

Table 2.3 summarizes the sparsity of the first three singular vectors $\mathbf{v}_i, i = 1, 2, 3,$ for six different methods. Since the left singular vectors in U corresponds to subject categories and is relatively stable, we only focus on the right singular vectors in V

which corresponds to different genes. Not surprisingly, the singular vectors of the standard SVD is dense with no zeros; the PMD solution is equally dense; the SSVD, IEEA and FIT-SSVD give similar levels of sparsity in $\mathbf{v}_1$, while the FIT-SRRR generates approximately about 25% more zeros in $\mathbf{v}_1$. As for $\mathbf{v}_2$, the FIT-SRRR and IEEA give the highest level of sparsity, which is around 10% more than the SSVD and FIT-SSVD. For $\mathbf{v}_3$, the solutions of the FIT-SRRR, IEEA and SSVD give similar levels of sparsity, the number of zero entries are round 11000. However, the FIT-SSVD only gives 7937 zero entries in $\mathbf{v}_3$. It should be noted that by applying the FIT-SRRR, the number of genes zeroed out in each layer is huge, which means the number of genes selected are much less than 12,625: there are 1348, 1636, and 1710 genes involved in the first three layers.

	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$
SVD	0	0	0
PMD	0	0	0
SSVD	8666	9370	11297
IEEA	9118	10394	11536
FIT-SSVD	9173	8177	7937
FIT-SRRR	11277	10989	10915

Table 2.3: Number of zeros in the first three singular vectors in V for the lung cancer data.

Heat maps of the first three estimated layers using the FIT-SRRR are plotted in Figure 2.3. To better visualize the gene clustering, all entries of the plotted matrices are divided by the maximum absolute value of the entries, so that the range of all entries are between -1 and 1. The genes in Figure 2.3 are ordered from the smallest to the largest. In each panel of Figure 2.3, around 7000 genes in the middle white area (sparse entries) are excluded when plotting. The four blocks in each panel reveal the four groups of subjects. These from the top to the bottom are: SmallCell, Normal, Colon and Carcinoid.

In Figure 2.3, we can see a clear checkerboard structure which indicates biclustering and gene grouping. The first estimated layer suggests a significant contrast between the Carcinoid group and the Normal group. The second estimated layer shows a contrast between the Normal group and Colon group. The third estimated layer zeros out the Normal, Carcinoid and Colon groups and singles out the SmallCell group. Compared to the heat maps in Lee et al. (2010) and Yang et al. (2013), some weaker contrasts are also zeroed out. The reason might be due to accounting for the within-group effects using multivariate regression with the design matrix (2.16). Note that the relevant results in Chen et al. (2012a) are similar to ours, however, we have eliminated more than 2000 genes with more sparse structures especially in the first two layers. Unlike the results in Chen et al. (2012a), the FIT-SRRR method gives orthogonal sparse singular vectors of the first three layers.

2.6 Discussion and Future Research

We have placed the FIT-SSVD algorithm in an optimization framework by introducing a specific bi-convex objective function and presented the FIT-SRRR method for low-rank approximation of the regression coefficient matrix. The latter generalizes the FIT-SSVD algorithm in Yang et al. (2013) to the correlated data which requires finding the threshold level in this new setup. The FIT-SRRR algorithm inherits the good properties of the FIT-SSVD and is more efficient than the FIT-SSVD for correlated data situations as demonstrated through simulation experiments.

There are several potential directions for future research:

1. It should be noted that when selecting the threshold level in the FIT-SRRR algorithm, the parameter δ used in (2.12) does not fully characterize a correlation

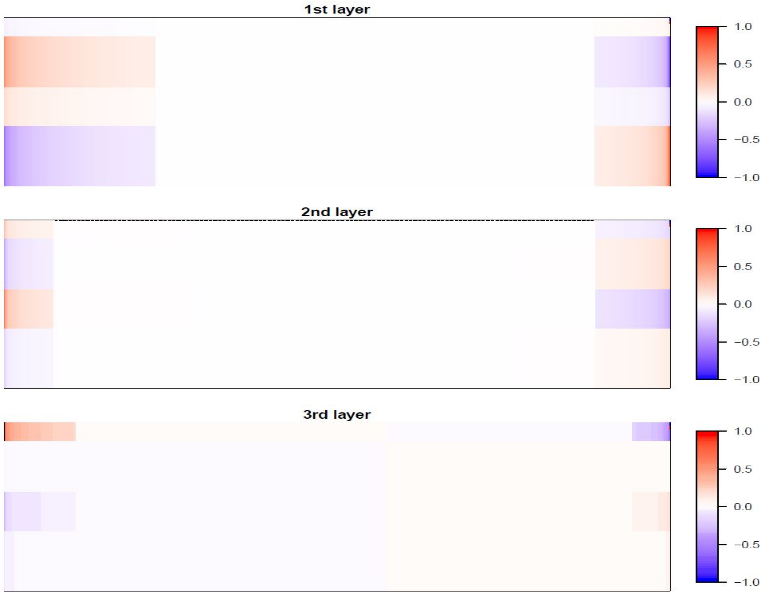


Figure 2.3: Heat maps of the first three layers by FIT-SRRR of lung cancer data.

matrix. It only reflects the maximum of magnitudes of the correlation coefficients. A method that uses the correlation matrix more effectively is the empirical Bayes block shrinkage approach in Wang and Wood (2006, 2010). Assuming data $\mathbf{y} \sim N(\mu, \Sigma)$, this method estimates μ in $\mathbf{y} = \mu + e$ through incorporating Σ to select the threshold level. It finds the posterior mean/median of the quadratic form $\mathbf{y}'\Sigma^{-1}\mathbf{y}$ as the threshold level through a pre-determined non-central chi-square priors on μ. If $\mathbf{y}$ is a vector with block-diagonal covariance structure, the shrinkage method partitions $\mathbf{y}$ into m non-overlapping sub-blocks $\mathbf{y}_i$'s, where $\mathbf{y}_i \sim N(\mu_i, \Sigma_i)$ for $i = 1, .., m$. It performs the thresholding on each block, allowing for different threshold levels for each block. It is of interest to incorporate this idea in our setup with properly selected priors on the means and block sizes to fully account for the correlation matrix, and for transposable data matrices which are correlated both in the rows and the columns (Allen et al., 2013). These generalizations of the FIT-SRRR remain open.

2. The FIT-SRRR algorithm assumes that the rank of B is known which is usually not the case in practice. Various methods of rank estimation are studied in Yuan et al. (2007) and Owen and Perry (2009). Yang et al. (2013) used the bi-cross-validation method in Owen and Perry (2009), while Chen et al. (2012a) assumed the rank is known. Developing proper rank estimation methods for B in the regression model (2.1) is of interest.

3. GENERALIZED PRINCIPAL COMPONENT ANALYSIS AND SINGULAR VALUE DECOMPOSITION

Transposable data are routinely encountered in fields such as econometrics, bioinformatics, chemometrics, network data and so on. The macroeconomic data and the fMRI data discussed in Section 1.2 are representative examples with strong dependencies among feature variables (columns) and temporal dependencies among observations (rows). The primary goal of many statistical analyses is to find the signal from the observed data. In this section, a generalization of SVD and PCA along with their regularized counterparts are considered for such data matrices. We propose thresholding-based algorithms to find sparse matrix decompositions of the transposable data matrix and to account for the two-way dependencies simultaneously.

3.1 Background

Recall the low-rank model in Section 1.5,

$$Y = UDV' + E, \tag{3.1}$$

where Y and $B = UDV' \in R^{n \times q}$ denote the data and the signal matrices, D is the diagonal matrix of the singular values of B, and U and V are the left and right factors, respectively. The noise E is assumed as a matrix-variate normal distribution (Gupta and Nagar, 1999):

$$E \sim MN_{n,q}(0, \Omega^{-1}, \Sigma^{-1}), \tag{3.2}$$

where $\Omega^{-1} \in R^{n\times n}, \Sigma^{-1} \in R^{q\times q}$ denote the rows and columns covariance matrices of the data. Hence, Y is a transposable data matrix with both rows and columns are dependent.

Most existing methods in the literature assume the entries of noise E in (3.1) are i.i.d. random variables and do not take into consideration the row-column correlations. They are focused on recovering B through regularizing its SVD $B = UDV' = \sum_{i=1}^{q} d_i \mathbf{u}_i \mathbf{v}_i'$. Regularization methods on the singular vectors of B have been proposed by Shen and Huang (2008), Witten et al. (2009) and Lee et al. (2010), where the U and V are found sequentially through rank-one approximation of the data matrix Y, see Section 1.5. However, it has been noted that ignoring the two-way dependencies in transposable data can lead to poor statistical performances (Efron, 2009; Allen and Tibshirani, 2010, 2012; Allen et al., 2013).

To incorporate two-way dependencies in (3.1)-(3.2), Allen et al. (2013) proposed the GMD which finds the best low-rank approximation of Y by minimizing the (Ω, Σ)-norm of the errors:

$$||Y - B||^2_{\Omega,\Sigma} = ||Y - UDV'||^2_{\Omega,\Sigma} = tr\{(Y - UDV')'\Omega(Y - UDV')\Sigma\}, \tag{3.3}$$

subject to the (generalized) orthogonality conditions:

$$U'\Omega U = I_r, V'\Sigma V = I_r. \tag{3.4}$$

The matrices U, V and the diagonal entries of D are referred to as the left and right GMD factors and the GMD values. The matrices Ω and Σ are the inverse row and column covariance matrices. They also can be interpreted as weighting matrices that weight data Y based upon its heteroscedasticity among rows and columns.

To solve (3.3)-(3.4), Allen et al. (2013) connects the mathematical solution of the GMD problem to that of the SVD solution of the de-correlated or sphered data matrix. Their computational GMD algorithm designed for massive datasets avoids the computationally expensive idea of sphering and relies on an iterative weighted power method. They also propose a framework for regularizing the GMD factors using the l_1 penalty to yield sparse U and V. However, their algorithms are still sequential and lack orthogonality of the columns of U and V when rank is greater than one.

Rather than computing the singular vectors sequentially one at a time as in Allen et al. (2013), we propose a fast iterative thresholding algorithm which is a generalization of the FIT-SSVD in Yang et al. (2013). Our fast iterative thresholding for sparse generalized matrix decomposition (FIT-SGMD) algorithm is designed to find a low-rank B for transposable data Y. Two major challenges in applying our algorithm to the transposable data matrix are:

(i) the U and V are required to satisfy the generalized orthogonality conditions (3.4),

(ii) selection of the threshold levels requires incorporating the row and column dependencies of the data matrix.

The FIT-SGMD algorithm is applicable to the multivariate regression setup, where the noise matrix E has dependencies as in (3.2) (Srivastava, 2009; Viroli, 2012). Recall that (Section 1.6), given n observations on the q-vector of responses $\mathbf{y}$ and p-vector of predictors $\mathbf{x}$, the multivariate linear regression model is written as

$$Y = XB + E, \tag{3.5}$$

where $Y = (\mathbf{y}_1, ..., \mathbf{y}_n)' \in R^{n \times q}, X = (\mathbf{x}_1, ..., \mathbf{x}_n)' \in R^{n \times p}$ and $B \in R^{p \times q}$ denote the responses, covariates, and regression coefficients matrices. When p and q are large, the RRR is widely used and known to be related to several techniques of multivariate analyses (Reinsel and Velu, 1998; Izenman, 2008), e.g. the principal component analysis (PCA) and the canonical correlation analysis (CCA), see Figure 1.2. We make the FIT-SGMD algorithm suitable for the RRR and CCA problems with two-way dependent data. This is done by connecting the RRR and CCA with the GMD problem through selecting suitable matrices Ω and Σ. These connections facilitate the interpretation of the estimates of the factors of $B = UDV'$ in (3.5).

Conceptually, the role of U and V in the RRR is to first transform the predictors $\mathbf{x}$ into a r-vector $\eta = U'\mathbf{x}$ and next transform η into the q-vector $DV'\eta$ to approximate the response $\mathbf{y}$. In CCA, U and V are used to form linear combinations $U'\mathbf{x}$ and $V'\mathbf{y}$ of predictors and responses with maximum correlation. When U and V are sparse, some irrelevant variables in $\mathbf{x}$ and $\mathbf{y}$ will be eliminated from the linear combinations, so that a sort of variable selection of predictors and responses is performed. In addition, for a given r, selecting the first r important factors of $\mathbf{x}$ and $\mathbf{y}$ leads to a sort of factor selection (Stock and Watson, 2012; Dobrev and Schaumburg, 2013). Thus, as consequence of applying the FIT-SGMD to (3.5), the coefficient matrix B is estimated from two-way dependent data situation by performing variable selection and factor selection simultaneously.

The rest of this section is organized as follows. We develop the FIT-SGMD approach to find the sparse GMD factors U and V for a transposable data matrix and prove the convergence of our generalized orthogonal iteration algorithm in Section 3.2. In Section 3.3, we connect the RRR and the CCA to the GMD problem and solve them using the FIT-SGMD algorithm. We illustrate the forecasting performance of the FIT-SGMD algorithm in regression (3.5) using a macroeconomic dataset and simulations in Section 3.4. Additional simulations are employed to evaluate our algorithm, followed by an analysis of an fMRI dataset in the Appendix B.

3.2 The Sparse Generalized Matrix Decomposition

We consider the model (3.1) with the noise matrix E as in (3.2) and known Ω, Σ. We first generalize the orthogonal iteration algorithm to compute the GMD factors U and V in (3.3)-(3.4). Then illustrate its ability to induce sparsity on U and V through thresholding and propose a fast iterative thresholding algorithm for sparse generalized matrix decomposition (FIT-SGMD).

3.2.1 The Generalized Orthogonal Iteration Algorithm

In this subsection, we compute the subspaces spanned by the GMD factors U and V (Algorithm 4). The mathematical solution of the GMD problem (3.3)-(3.4) is presented in Allen et al. (2013, Theorem 1) by sphering or de-correlating the data matrix as described next.

Let $\tilde{\Omega}, \tilde{\Sigma}$ be the square roots of the matrices Ω, Σ and $\tilde{\Omega}^{-1}, \tilde{\Sigma}^{-1}$ be their left matrix inverses. The solution U^*, V^* are computed based on the SVD of the sphered data $\tilde{Y} = \tilde{\Omega} Y \tilde{\Sigma}$, and then multiplying the covariances back. To be specific, $U^* = \tilde{\Omega}^{-1}\tilde{U}, V^* = \tilde{\Sigma}^{-1}\tilde{V}$, where $\tilde{U}, \tilde{V}$ are the singular vectors of the SVD of the decorrelated data matrix $\tilde{Y}$. They also showed that the solutions are the same when the positive-definiteness of Ω, Σ is relaxed to positive semi-definiteness.

Compared to Allen et al. (2013), where an iterative weighted power method is proposed for computing the GMD factors sequentially, our generalization inherits the advantages of the orthogonal iteration algorithm in computing the subspaces simultaneously, avoiding sequential computation and guaranteeing the orthogonality of the singular vectors. To incorporate the two-way dependencies in the data matrix, modifications to the orthogonal iteration are made as follows: Step 2 of the Algorithm

4 is now to right multiply Y by $\Sigma V^{(k-1)}$ instead of $V^{(k-1)}$, while Step 4 is adjusted to right multiply Y' by $\Omega U^{(k)}$ instead of $U^{(k)}$. In Step 3 (or 5), the so-called "Ω-QR decomposition (or Σ-QR decomposition)" is applied instead of the conventional QR decomposition to observe the conditions (3.4). The Ω-QR decomposition is discussed in the Appendix B, where it could be regarded intuitively as replacing the Frobenius norm by the Ω-norm for all the inner products in the Gram-Schmidt process for the conventional QR decomposition (Golub and Van Loan, 1996).

Algorithm 4: Generalized Orthogonal Iteration Algorithm.

```
  Input:
  1. Data matrix $Y$, matrices $\Omega, \Sigma$;
  2. Initial matrix $V^{(0)} \in R^{q \times r}$.
1 repeat
2     Right-to-Left multiplication: $Y_u^{(k)} \leftarrow Y \Sigma V^{(k-1)}$;
3     $\Omega$-QR decomposition: $U^{(k)} R_u \leftarrow Y_u^{(k)}$;
4     Left-to-Right multiplication: $Y_v^{(k)} = Y' \Omega U^{(k)}$;
5     $\Sigma$-QR decomposition: $V^{(k)} R_v \leftarrow Y_v^{(k)}$;
6 until Convergence;
7 return $\hat{U} = U^{(\infty)}, \hat{V} = V^{(\infty)}$.
```

Theorem 3.2.1 shows that the Algorithm 4 converges to the mathematical solution given in Allen et al. (2013, Theorem 1), which is the global solution of the GMD problem. Algorithm 4 avoids the computationally expensive idea of sphering data by taking square roots and inverses of Ω, Σ, and it computes the subspaces generated by U and V simultaneously rather than sequentially. Consequently, it guarantees that the solutions U and V satisfy conditions (3.4) or the generalized orthogonality with respect to Ω and Σ.

Theorem 3.2.1 *Let the mathematical solutions of the GMD factors be $U^* = \tilde{\Omega}^{-1}\tilde{U}$, $V^* = \tilde{\Sigma}^{-1}\tilde{V}$, where $\tilde{U}, \tilde{V}$ are the singular vectors of the SVD of the sphered data matrix $\tilde{Y} = \tilde{\Omega}Y\tilde{\Sigma}$, $\tilde{\Omega}, \tilde{\Sigma}$ denote the square roots of Ω, Σ and $\tilde{\Omega}^{-1}, \tilde{\Sigma}^{-1}$ denote their left matrix inverses, respectively. Then, the $U^{(\infty)}$ and $V^{(\infty)}$ in Algorithm 4 are equal to the GMD solutions U^* and V^*.*

Proof We outline the key steps in proving the theorem and relegate the complete version to the Appendix B. We show that the updates of U and V in the Algorithm 4 are equivalent to the updates of $\tilde{U}$ and $\tilde{V}$ in the orthogonal iteration algorithm for computing the SVD of $\tilde{Y}$. When writing the updating steps of U and V in terms of $\tilde{Y}$, $\tilde{U}$ and $\tilde{V}$ in Algorithm 4, we have:

$$\begin{aligned} UR_1 &= \tilde{\Omega}^{-1}\tilde{U}R_1 = Y\Sigma V = \tilde{\Omega}^{-1}\tilde{Y}\tilde{V}, \\ VR_2 &= \tilde{\Sigma}^{-1}\tilde{V}R_2 = Y'\Omega U = \tilde{\Sigma}^{-1}\tilde{Y}'\tilde{U}, \end{aligned}$$

where R_1, R_2 are the corresponding R matrices for QR decompositions of $Y\Sigma V$ and $Y'\Omega U$. By the proof of convergence of orthogonal iteration algorithm in Golub and Van Loan (1996, Chapter 8), the Algorithm 4 is equivalent to the orthogonal iteration for $\tilde{Y}$, which converges to the SVD of $\tilde{Y}$.

3.2.2 Thresholding for Sparse GMD

In this subsection, we describe the FIT-SGMD (Algorithm 5) by first presenting the overall procedures for updating U and V, then illustrating how to apply thresholding techniques, and finally selecting the proper threshold levels. The FIT-SGMD computes the sparse factors U and V in the model (3.1) subject to the conditions (3.4). It is based on the Algorithm 4 and incorporates thresholding when updating U and V. In this subsection, we assume that Ω and Σ are positive definite matrices.

Algorithm 5: The FIT-SGMD algorithm.

```
Input:
1. Data matrix $Y$ and matrices $\Omega, \Sigma$;
2. Thresholding function $\eta$, $\gamma_u, \gamma_v$ from the Algorithm 6 ;
3. Initial matrices $U^{(0)} \in R^{n \times r}, V^{(0)} \in R^{q \times r}$.
Output:
Estimators $\hat{U} = U^{(\infty)}, \hat{V} = V^{(\infty)}$.
repeat
    Right-to-Left multiplication: $Y_u^{(k)} \leftarrow Y \Sigma V^{(k-1)}$ ;
    Thresholding: $(Y_u)^{(k),thr} \leftarrow \eta(Y_u^{(k)}, \gamma_u)$;
    $\Omega$-QR decomposition: $U^{(k)} R_u^{(k)} \leftarrow (Y_u)^{(k),thr}$;
    Left-to-Right multiplication: $Y_v^{(k)} \leftarrow Y' \Omega U^{(k)}$;
    Thresholding: $(Y_v)^{(k),thr} \leftarrow \eta(Y_v^{(k)}, \gamma_v)$;
    $\Sigma$-QR decomposition: $V^{(k)} R_v^{(k)} \leftarrow (Y_v)^{(k),thr}$;
until Convergence;
return Estimators $\hat{U} = U^{(\infty)}, \hat{V} = V^{(\infty)}$.
```

Since the procedures for updating U and V are symmetric, we describe the details only for U and illustrate the details of thresholding and updating procedure through its lth column $\mathbf{u}_l, l = 1 \cdots, r$. The goal of thresholding is to retain the entries of Y with high signal and replace the others with zero.

Let $U^{(k-1)}, V^{(k-1)}$ be the updates of U and V at $(k-1)$st iteration, $\mathbf{v}_l^{(k-1)}$ be the lth column of $V^{(k-1)}$, and let $D^{(k)} = diag(d_1^{(k)}, ..., d_r^{(k)})$. Right multiplying both side of (3.5) by $\Sigma \mathbf{v}_l^{(k-1)}$ and using conditions (3.4), it follows that

$$Y_{Left,l}^{(k)} = Y \Sigma \mathbf{v}_l^{(k-1)} = B \Sigma \mathbf{v}_l^{(k-1)} + E \Sigma \mathbf{v}_l^{(k-1)} \approx \mathbf{u}_l^{(k)} d_l^{(k)} + E \Sigma \mathbf{v}_l^{(k-1)}. \tag{3.6}$$

Let γ_{ul} be a given threshold level in (3.6), then $(Y_{Left,l})^{(k),thr}$ a thresholded version of $Y_{Left,l}^{(k)}$ serves as an estimator of the mean vector $\mathbf{u}_l^{(k)} d_l^{(k)}$. Repeating the previous procedure for $l = 1, \cdots, r$, leads to $(Y_{Left})^{(k),thr}$.

Next, we discuss how to obtain the optimal threshold level γ_{ul} in (3.6). The following properties of linear transformations of matrix normal distributions are needed.

Proposition 3.2.1 **(a)** *Let $Y \sim MN_{m,n}(M, \Omega, \Sigma)$ and A be a suitable matrix, then $YA \sim N_{m,n}(MA, \Omega, A\Sigma A')$, and $AY \sim N_{m,n}(AM, A\Omega A', \Sigma)$.*
(b) *If $E \sim MN_{n,q}(0, \Omega^{-1}, \Sigma^{-1})$ as in (3.2), and $\mathbf{u}, \mathbf{v}$ are suitable vectors satisfying $\mathbf{u}'\Omega\mathbf{u} = \mathbf{v}'\Sigma\mathbf{v} = 1$, then $E\Sigma\mathbf{v} \sim N_{n,1}(0, \Omega^{-1}, 1) = N_n(0, \Omega^{-1})$ and $E'\Omega\mathbf{u} \sim N_q(0, \Sigma^{-1})$.*

From Proposition 3.2.1(b) the noise vector $E\Sigma\mathbf{v}_l^{(k-1)}$ in (3.6) is dependent with covariance matrix Ω^{-1}, and as a result the universal threshold level (Donoho and Johnstone, 1994) used in Yang et al. (2013) is not applicable. We need to account for the dependence and heterogeniety in (3.6). Johnstone and Silverman (1997); Kovac and Silverman (2000) suggest more general thresholding methods which can be used for correlated and heteroscedastic noise. In brief, their methods view the entries of the vector $E\Sigma\mathbf{v}_l^{(k-1)}$ as independent heterogeneous random variables and ignore the correlation structure.

Berkner and Wells (1998, 2001); Delouille et al. (2004) generalize the universal thresholding by incorporating the correlations. More precisely, the thresholding level γ_{ujl} for the jth entry of the vector $Y\Sigma\mathbf{v}_l^{(k-1)}$ is selected as

$$\gamma_{ujl} \quad = \quad \hat{\sigma}_{ujl}\gamma_{ul}, \tag{3.7}$$

where σ_{ujl}^2 is the variance of the $(Y\Sigma\mathbf{v}_l^{(k-1)})_j$ entry and from Proposition 3.2.1(b), $\hat{\sigma}_{ujl}^2 = \sigma_{jj}$, where σ_{jj} is the j^{th} diagonal entry of Ω^{-1}. The value γ_{ul} (Berkner and Wells, 2001) is

$$\gamma_{ul} = \sqrt{2(1+\delta)\log(n)}, \tag{3.8}$$

where

$$\delta = \max_{j \neq j'}(|corr(\{Y\Sigma\mathbf{v}_l^{(k-1)}\}_j, \{Y\Sigma\mathbf{v}_l^{(k-1)}\}_{j'})|).$$

We adopt the same thresholding procedure for the correlated case in updating $U^{(k)}$, and illustrate its details in Algorithm 6.

```
Algorithm 6: Selection of the threshold level $\gamma_{ul} = g_u(\tilde{Y}, U^{(k-1)}, V^{(k-1)}, \hat{\sigma})$.
    Input:
    1. Data matrix $Y$ and matrices $\Omega, \Sigma$;
    2. Previous estimators of the singular vectors $U^{(k-1)}, V^{(k-1)}$.
    Output:
    Threshold level vectors $\gamma_{ul}$ for $l = 1, ..., r$.
    $\hat{\sigma}^2_{ujl} \leftarrow \sigma_{jj}$, where $\sigma_{jj}$ is the $j^{th}$ diagonal entry of $\Sigma$;
    $\gamma_{ul} \leftarrow \sqrt{2(1+\delta)\log(p)}$ where $\delta = \max_{j \neq j'}(|corr(\{\tilde{Y}\mathbf{v}_l^{(k-1)}\}_j, \{\tilde{Y}\mathbf{v}_l^{(k-1)}\}_{j'})|)$;
    $\gamma_{ujl} \leftarrow \hat{\sigma}_{ujl}\gamma_{ul}$, $j = 1, ..., n$;
    return $\gamma_{ul} = (\gamma_{u1l}, \dots, \gamma_{upl})'$, $l = 1, ..., r$.
```

The FIT-SGMD is displayed in the Algorithm 5, where γ_u and γ_v are obtained through the Algorithm 6. We use Algorithm 4 to find the initial matrices $U^{(0)}, V^{(0)}$ and stop after the kth iteration if the maximum distance between the successive iterates is small (Yang et al., 2013): i.e. $max\{||P_{U^{(k)}} - P_{U^{(k-1)}}||_2^2, ||P_{V^{(k-1)}} - P_{V^{(k-1)}}||_2^2\}$ is less than or equal to a preselected $\epsilon = 10^{-8}$. Here $P_A = AA'$ is a projection for a matrix A and $||A||_2$ denotes the spectral norm of the matrix A. The thresholding function is the hard-thresholding $\eta(y, \gamma) = H(y, \gamma) = y1_{\{|y| > \gamma\}}$, and the Ω, Σ are assumed known as in the Algorithm 5.

3.3 The FIT-SGMD for Supervised Learning

In this subsection, we apply the FIT-SGMD algorithm to the multivariate regression problems by reducing the model (3.5) to (3.1) and demonstrating that the FIT-SGMD is suitable for the RRR and CCA problems through connecting to the GMD problem (3.3)- (3.4). Finally, we discuss the connection between the FIT-SGMD and the generalized principal component analysis (GPCA) and propose an algorithm to compute the sparse eigenvectors in the GPCA.

For a full-rank design matrix X, left multiplying both sides of (3.5) by $(X'X)^{-1}X'$ leads to

$$\tilde{Y} = B + \tilde{E}, \tag{3.9}$$

where $\tilde{Y} = \hat{B} = (X'X)^{-1}X'Y$ is the least-square estimate of B, and $\tilde{E} = (X'X)^{-1}$ $X'E \sim MN_{p,q}(0, T\Omega^{-1}T', \Sigma^{-1})$ with $T = (X'X)^{-1}X'$. When X is less than full-rank, some alternatives to the least-square estimators are the Moore-Penrose inverse (Bunea et al., 2011) and the ridge estimator (Chen et al., 2012a). Throughout this section, the Moore-Penrose inverse G of $X'X$ is used, where the transformed data has noise $\tilde{E} \sim N_{p,q}(0, T\Omega^{-1}T', \Sigma^{-1})$ with $T = GX'$.

The matrices Ω and Σ are usually unknown and there is a wide range of choices for them in various area of statistics. They could be chosen as the well-known time series autoregressive (Shaman, 1969) and moving average processes (Galbraith and Galbraith, 1974), spatial models like random fields (Rue and Held, 2005), or graphic models like graphic Laplacian (Merris, 1994) and reverse distance structures, see Allen et al. (2013). When Ω, Σ are unknown, it is very challenging to estimate them and the GMD factors from the data at the same time. The transposable regularized covariance model (TRCM) in Allen and Tibshirani (2010, 2012) allows us to estimate the Ω, Σ. Unfortunately, as Allen et al. (2013) point out, the TRCM is computationally expensive, and can easily get stuck in one of its many local optima, and is very sensitive to the choice of its initial starting values. Hence, estimating these matrices is still an open problem. In the next Section 3.3.1, we show that selecting Ω, Σ as the sample covariances of Y and X reveals the close connection among the RRR, CCA and GMD.

3.3.1 Connections among RRR, CCA and GMD.

The regression model (3.5) is a general framework for many techniques of multivariate analysis. It is common to estimate B by minimizing a weighted sum-of-squares as an objective function:

$$tr\{(Y - BX)'W(Y - BX)\} = ||Y - XB||_{I,W},$$

where W is a positive-definite symmetric matrix of weights. It is known that (Reinsel and Velu, 1998; Izenman, 2008) the RRR is closely related to the PCA, CCA, the Fisher's linear discriminant analysis (Fisher, 1936) and the correspondence analysis (Hirschfeld, 1935), see Figure 1.2.

In this subsection, we show that the RRR and CCA can be reduced to the GMD problem (3.3)-(3.4) using the transposable quadratic (Ω, Σ)-norm. Theorem 3.3.1 indicates that when the matrices Ω and Σ are replaced by the sample covariance matrix of the predictors and inverse sample covariance matrix of the responses, the RRR,CCA and GMD problem are closely related and have the same objective function. More precisely, the RRR and CCA problems are a GMD problem for the matrix $GX'Y$, and the solution triplet $(\hat{U}, \hat{V}, \hat{D})$ for the three optimization problems below is recognized as $(\hat{U}, \hat{V}, \hat{D}) = \arg\max_{U,V,D} tr\{VDU'X'Y\Sigma\}$ (see the Appendix B).

Theorem 3.3.1 *Let G be the Moore-Penrose generalized inverse of $X'X$ and $GX'Y = UDV'$ with U and V satisfying the condition (3.4). Set $\Omega = S_{XX}$, $\Sigma = S_{YY}^{-1}$, where S_{XX}, S_{YY} are the sample covariances of the predictors and the responses, respectively, and $\tilde{V} = \Sigma V D$. Then, the following three optimization problems:*

$$\begin{aligned} \textit{Regression}: &\qquad min_{U,D,V}||Y - XUDV'||^2_{I,\Sigma} \\ \textit{CCA}: &\qquad min_{U,D,V}||Y\tilde{V} - XU||^2_{I,I} \\ \textit{GMD}: &\qquad min_{U,D,V}||GX'Y - UDV'||^2_{\Omega,\Sigma} \end{aligned}$$

are identical with the same solution triplet (U, V, D). That is, if $(\hat{U}, \hat{V}, \hat{D})$ solves any problem, then it will also solve the remaining two problems. In fact, $(\hat{U}, \hat{V}, \hat{D}) = \arg\max_{U,D,V} tr\{VDU'X'Y\Sigma - D^2/2\}$

The proof of Theorem 3.3.1 is given in the Appendix B. The theorem connects the GMD with RRR and CCA, which enables us to interpret the U and V estimated by the FIT-SGMD. Recall that the matrices U and V consist of the r columns $U = [\mathbf{u}_1, ..., \mathbf{u}_r]$ and $V = [\mathbf{v}_1, ..., \mathbf{v}_r]$, where X and Y are the $n \times p$ and $n \times q$ matrices of stacked n observations of the p-vector of predictors $\mathbf{x}$ and q-vector of responses $\mathbf{y}$. In RRR, the role of U and V is to first transform the predictors $\mathbf{x}$ into a r-vector η through $\eta_i = \mathbf{u}_i'\mathbf{x}$, then to transform η into a q-vector $DV'\eta$ to approximate the responses $\mathbf{y}$. In CCA, $\mathbf{u}_i'\mathbf{x}$ and $(\Sigma\mathbf{v}_i)'\mathbf{y}$ are the linear combinations of predictors and responses for $i = 1, ..., r$, which have maximum correlation among all possible linear combinations. Zero entries in $\mathbf{u}_i, \mathbf{v}_i$ indicate that the corresponding responses and predictors are eliminated from the newly formed linear combinations, consequently variable selection of predictors and responses is performed. In other words, sparse $\mathbf{u}_i$ and $\mathbf{v}_i$ indicate that the connections between the predictors and responses only involve a subset of predictors and responses. In addition, the rank r in RRR and CCA is the rank of U and V, which determines the number of pathways relating the responses to the predictors: it is the number of factors of the responses and predictors in RRR and indicates the number of pairs of linear combinations of responses and predictors in CCA. Hence, when applying the FIT-SGMD to (3.5) it performs variable selection and factor selection simultaneously.

3.3.2 Generalized Principal Component Analysis

In this subsection, we first introduce the GPCA (see Section 1.3) for transposable data in details, then show that the Algorithm 4 can be used to perform GPCA, and finally adjust our FIT-SGMD algorithm to perform the sparse GPCA problem, where sparse GPCs are found by thresholding.

As discussed in Section 1.3, for a q-vector variable $\mathbf{y} = (y_1, ..., y_q)'$, the PCA is to explain the covariance matrix of $\mathbf{y}$ through the r PCs $z_1 = \mathbf{v}_1'\mathbf{y}, ..., z_r = \mathbf{v}_r'\mathbf{y}$. For a centered data matrix $Y \in R^{n\times q}$, the PCA finds the eigen-decomposition of its sample covariance matrix $S = \frac{1}{n}Y'Y$. More precisely, the PCA computes the loading matrix of PCs $V = (\mathbf{v}_1, ..., \mathbf{v}_r)$ by solving the following optimization problems:

$$\begin{aligned}
&\mathbf{v}_1 = \arg\max_{\mathbf{v}} \mathbf{v}'S\mathbf{v} \text{ subject to } \mathbf{v}'\mathbf{v} = 1,\\
&\mathbf{v}_2 = \arg\max_{\mathbf{v}} \mathbf{v}'S\mathbf{v} \text{ subject to } \mathbf{v}'\mathbf{v} = 1, \mathbf{v}'\mathbf{v}_1 = 0,\\
&\vdots\\
&\mathbf{v}_r = \arg\max_{\mathbf{v}} \mathbf{v}'S\mathbf{v} \text{ subject to } \mathbf{v}'\mathbf{v} = 1, \mathbf{v}'\mathbf{v}_j = 0, \text{ for } j = 1, 2, .., r-1,
\end{aligned}$$

which leads to orthogonal V:

$$V'V = I_r.$$

When the matrix Y is transposable, GPCA finds the loading matrix $V = (\mathbf{v}_1, ..., \mathbf{v}_r)$ by maximizing the sample variance after incorporating the two-way dependencies in the data (Escoufier, 2006; Allen et al., 2013). To be specific, GPCA projects the data into a space induced by the (Ω, Σ)-norm where all the inner products in calculation of PCA is replaced by the (Ω, Σ)-norm. In GPCA,

$$\begin{aligned}
&\mathbf{v}_1 = \arg\max_{\mathbf{v}} \mathbf{v}'\Sigma Y'\Omega Y\Sigma\mathbf{v} \text{ subject to } \mathbf{v}'\Sigma\mathbf{v} = 1,\\
&\mathbf{v}_2 = \arg\max_{\mathbf{v}} \mathbf{v}'\Sigma Y'\Omega Y\Sigma\mathbf{v} \text{ subject to } \mathbf{v}'\Sigma\mathbf{v} = 1, \mathbf{v}'\Sigma\mathbf{v}_1 = 0,\\
&\vdots\\
&\mathbf{v}_r = \arg\max_{\mathbf{v}} \mathbf{v}'\Sigma Y'\Omega Y\Sigma\mathbf{v} \text{ subject to } \mathbf{v}'\Sigma\mathbf{v} = 1, \mathbf{v}'\Sigma\mathbf{v}_j = 0, \text{ for } j < r,
\end{aligned} \tag{3.10}$$

where the loading matrix V in GPCA satisfy the generalized orthogonality constraint in the Σ-norm:

$$V'\Sigma V = I_r, \tag{3.11}$$

and the generalized principal components (GPC) are given by $Y\Sigma\mathbf{v}_1, ..., Y\Sigma\mathbf{v}_r$. Set $d_i^2 = \mathbf{v}_i'\Sigma Y'\Omega Y\Sigma\mathbf{v}_i$, the proportion of variance explained by the ith GPC after taking consideration of the matrices Ω, Σ is given by $d_i^2/||Y||_{\Omega,\Sigma}^2$ for $i = 1, ..., r$. While the PCA uses linear projection to explain the variance, the GPCA uses (Ω, Σ)-norm projection, which gives an alternative way of explaining the variance-covariance in the data (Allen et al., 2013, Proposition 2 and Corollary 5).

Just as the SVD can be used to find the PCs as shown in Section 1, the Algorithm 4 can be used to find the GPCs. We can shown that the $\mathbf{v}_i$ in (3.10) are given by the ith right GMD factor $\mathbf{v}_i$ in Algorithm 4. As using the Algorithm 1 for GMD problem, this algorithm has the desirable computation properties especially for high-dimensional data, such as avoids the computationally expensive idea of sphering the data and computation of the reverses and square roots of Ω, Σ. In addition, the Algorithm 1 finds the subspace of GPC loading matrix V spanned by its leading vectors and guarantees the (generalized) orthogonality of V, which is an important feature of the sparse GPCA.

Considering the sparse PCA methodologies, where its goal is to find a set of sparse PC loadings that explains most of the variance to enhance the model interpretation. The sparse GPCA is a natural generalization of sparse PCA for transposable data matrix. Ma (2013) proposed an iterative thresholding sparse PCA algorithm to recover the sparse loading matrix of PCs, which is conceptually a predecessor of the FIT-SSVD algorithm. Given a transposable data matrix Y, we generalize and adjust the algorithms in Ma (2013) and in our FIT-SGMD and propose Algorithm 7 to solve for the sparse loading matrix in the GPCA.

Algorithm 7: Algorithm for sparse GPCA.

```
Input:
1. Data matrix $Y$ and matrices $\Omega, \Sigma$;
2. Thresholding function $\eta$, threshold level $\gamma$;
3. Initial matrix $V^{(0)} \in R^{q \times r}$.
Output:
Estimator $\hat{V} = V^{(\infty)}$.
$T \leftarrow Y'\Omega Y$;
repeat
    Multiplication: $T^{(k)} \leftarrow T\Sigma V^{(k-1)}$ ;
    Thresholding: $T^{(k),thr} \leftarrow \eta(T^{(k)}, \gamma)$;
    $\Omega$-QR decomposition: $V^{(k)}R^{(k)} \leftarrow T^{(k),thr}$;
until Convergence;
return Estimator $\hat{V} = V^{(\infty)}$.
```

3.4 Macroeconomic Data Analysis and Simulations

In this subsection, we evaluate the forecasting performance of our FIT-SGMD algorithm in regression model (3.5) using the macroeconomic dataset in Stock and Watson (2012), a common benchmark for high-dimensional regression forecasting. We next use simulation to further study some of the findings from the real data analysis.

3.4.1 A Macroeconomic Dataset

The dataset consists of 144 U.S. macroeconomic time series for a total of 195 quarterly observations from 1960:II through 2008:IV. There are 35 high-level aggregate series that are related by an identity to the remaining 109 lower-level disaggregate series in the dataset. For example, the (aggregate) gross domestic product (GDP) variable consists of the sum of GDP indices such as fixed investment, goods, services and so on. The 35 aggregate macroeconomic variables are used as the responses $\mathbf{y}$, while the 109 disaggregated series $\mathbf{x}$ are used as the predictors. Rather than directly studying the original series, Stock and Watson (2012, Supplement materials) describe stationary-inducing transformations of the data, where all data series are transformed by one or more of the following transformations: first- and second-order differences, logarithm transformation or first- and second-order difference of the logarithm.

The predictors based on the principal component regression (PCR) are shown in Stock and Watson (2012) and Dobrev and Schaumburg (2013) to have superior performance in forecasting compared to many other regularization methods. They use the first 5 principal components of $\mathbf{x}$ as factors in the PCR and the coefficients are estimated through least-squares. Such a regression model is denoted as PCR-5, and we use it as the benchmark to assess the performance of our FIT-SGMD algorithm.

We evaluate the forecast performance for the original and the transformed data, respectively, and perform out-of-sample one-step-ahead forecasts with rolling window size 100 (quarterly observations), i.e. we estimate the model using the time segment running from $t-99, t-99+1, ..., t$ and forecast the variables at time $t+1$, where $t = 100, ..., 194$. The forecast accuracy is measured using the RMSE, defined as $RMSE_j = \sqrt{\sum_t (y_{jt} - \hat{y}_{jt})^2}$, where $\hat{y}_{jt}$ is the prediction and y_{jt} is the observed value for the variable $j = 1, ..., 35$ at time $t = 101, .., 195$. We use the RMSE ratios of the FIT-SGMD to the PCR-5 benchmark. A ratio less than 1 indicates that our alternative method has a better forecast accuracy, otherwise the PCR-5 is better. The notation FIT-SGMD-r and PCR-r indicate selecting r layers or components. The sample covariances of $\mathbf{x}, \mathbf{y}$ are used as estimates of Ω and Σ in the FIT-SGMD.

3.4.1.1 Forecast results for the transformed data

We summarize the results in percentiles of relative RMSE ratios for various forecasting methods as in Stock and Watson (2012). To further highlight the performances, we also report the empirical distribution of the RMSE ratios for chosen intervals (< 0.9, $(0.9, 0.97)$, $(0.97, 1.03)$, $(1.03, 1.1)$, > 1.1). Any downward or upward deviations from 1, indicates better or worse forecasting accuracy relative to the PCR-5 benchmark (Dobrev and Schaumburg, 2013). The empirical distribution reveals any improvement for proposed methods compared to the PCR-5 if the left tail of the distribution is heavier than the right tail and the median is less than or equal to 1. More precisely, if the 50-,75- and 95-percentiles are less or equal to 1, and 5-, 25-percentiles are close to zero, it indicates that the compared method is better than the PCR-5 benchmark in forecasting. In addition, if the probabilities $P(ratio > 1.1), P(1.03 < ratio < 1.1)$ are very small and the probabilities $P(ratio < .9), P(.09 < ratio < .97)$ are relatively large, it indicates that our method is more accurate than the PCR-5 in terms of the RMSE.

FIT-SGMD-r	Percentiles					Empirical Distrituion				
r	5%	25%	50%	75%	95%	<.9	.9-.97	.97-1.03	1.03-1.1	>1.1
1	0.67	0.96	1.04	1.21	1.45	0.09	0.29	0.11	0.17	0.34
2	0.53	0.94	1.00	1.09	1.28	0.14	0.29	0.14	0.17	0.26
3	0.51	0.95	1.03	1.13	1.29	0.14	0.14	0.23	0.17	0.31
4	0.44	0.97	1.05	1.23	1.38	0.14	0.11	0.17	0.20	0.37
5	0.44	1.00	1.08	1.19	1.39	0.14	0.06	0.14	0.26	0.40
6	0.44	0.97	1.04	1.10	1.15	0.20	0.03	0.29	0.23	0.26
7	0.45	0.98	1.05	1.08	1.18	0.20	0.06	0.29	0.20	0.26
8	0.47	0.91	1.03	1.16	1.49	0.26	0.06	0.20	0.14	0.26
9	0.32	0.41	1.01	1.09	1.25	0.40	0.06	0.09	0.26	0.20

Table 3.1: Distributions of ratios of RMSE to PCR-5 for the transformed data.

Table 3.1 summarizes the percentiles and empirical distributions of the 35 ratios of RMSE of the FIT-SGMD to the PCR-5 benchmark and shows the results of the FIT-SGMD algorithm for ranks $r = 1, ..., 9$ in the upper panel and the results of the PCR with corresponding ranks in the lower panel. In complete agreement with the main conclusions in Stock and Watson (2012) and Dobrev and Schaumburg (2013), the PCR-5 is very competitive consistently across all the 35 series. We claim that a possible reason as to why the FIT-SGMD does not outperform the PCR-5 is due to the transformations of the original data. If the aggregate series are obtained from the disaggregate series linearly, the log-transform, for example, would possibly sabotage their original linear relationship. To examine this claim, in the next section we look at the forecasting performance of our procedure applied to the original data without any transformations aimed at reducing the data stationarity.

3.4.1.2 Forecast results for the original data

We perform rolling-window out-of-sample one-step-ahead and two-step-ahead forecasts with rolling window size 100 quarterly observations using the original series. Table 3.2 reports the one-step-ahead forecasts in the upper panel and the two-step-ahead forecasts in the lower panel. Results in the upper panel demonstrate that the FIT-SGMD with $r > 4$ clearly outperform the PCR-5, since all their 50 percentiles are below 1, and the $P(ratio > 1.1)$'s are close to zeros. For instance, the 95-percentiles of the FIT-SGMD-8 is 0.82, and its empirical $P(ratio > 0.9) = 0$, which means that all the RMSE ratios of the FIT-SGMD-8 to the PCR-5 are less than 0.9. Throughout the results, the FIT-SGMD-8 has the best performance in all the one-step-ahead forecasts where not only all the ratios are uniformly less than 0.9, but also it achieves the lowest median ratio at 0.53.

The pattern in the lower panel of Table 3.2 for the two-step-ahead forecasts is similar and consistent with the corresponding results in the one-step-ahead forecasts. The FIT-SGMD with $r > 4$ all have less than one 50-percentiles and have relatively low probabilities for large ratios. For instance, the 95-percentile of the FIT-SGMD-8 is only 0.96, and the probability $P(ratio < .9) = 0.89$.

In summary, the FIT-SGMD outperforms the PCR-5 when applied to the original data. Thus, it is remarkable that the FIT-SGMD is capable of producing accurate forecasts for the original and nonstationary data.

3.4.2 Simulations

It is remarkable that our analysis of the macroeconomic time series dataset in Stock and Watson (2012) revealed that the FIT-SGMD outperforms the PCR-5 for the original nonstationary data. Through, its performance was not as accurate as the PCR-5 for the transformed stationary data. If this phenomenon can be shown to hold widely, then it may obviate the need to transform the data to stationarity which can be a huge advantage for the FIT-SGMD method over the PCR in high-dimensional data situations. Deciding what transformations to use to reduce data to stationarity is a difficult task even for univariate time series data.

FIT-SGMD			Percentiles					Empirical Distrituion		
r	5%	25%	50%	75%	95%	<.9	.9-.97	.97-1.03	1.03-1.1	>1.1
					One-step-ahead					
3	0.44	0.70	0.96	1.28	1.76	0.40	0.11	0.03	0.09	0.37
4	0.49	0.62	0.85	1.07	1.50	0.54	0.00	0.17	0.06	0.23
5	0.51	0.64	0.78	0.95	1.38	0.68	0.09	0.06	0.03	0.14
6	0.56	0.68	0.77	0.98	1.16	0.71	0.03	0.03	0.11	0.11
7	0.50	0.56	0.62	0.78	0.92	0.83	0.11	0.03	0.03	0.00
8	0.38	0.45	0.53	0.73	0.82	1.00	0.00	0.00	0.00	0.00
9	0.42	0.51	0.64	0.84	0.94	0.86	0.11	0.00	0.03	0.00
					Two-step-ahead					
3	0.69	0.81	1.12	1.28	1.56	0.37	0.00	0.06	0.03	0.54
4	0.62	0.78	0.95	1.09	1.25	0.34	0.26	0.09	0.06	0.25
5	0.54	0.60	0.90	1.01	1.16	0.49	0.20	0.11	0.11	0.09
6	0.56	0.68	0.77	0.98	1.16	0.71	0.03	0.03	0.11	0.11
7	0.41	0.53	0.65	0.83	1.10	0.83	0.06	0.03	0.03	0.05
8	0.37	0.46	0.64	0.81	0.96	0.89	0.06	0.03	0.00	0.02
9	0.42	0.54	0.74	0.99	1.27	0.69	0.03	0.06	0.14	0.08

Table 3.2: Distributions of ratios of RMSE to PCR-5 for the original data.

In this subsection, to further understand the above phenomenon we rely on simulation to assess the forecast performance of both methods applied to the simulated nonstationary series and the corresponding transformed stationary series. We divide this section into two parts, (1) the data generating procedure and transformations and (2) analysis of the simulation results.

3.4.2.1 Data generating procedure and transformations

The dimensions of the simulated data are $n = 195, p = 90$ and $q = 35$ as in Stock and Watson (2012). We generate the matrix X with the following three different nonstationary features:

- Case I: Random walk. Let $X_{j,t} = X_{j,t-1} + \epsilon_{jt}$, $j = 1, ..., p$ and $t = 1, ..., n$, where ϵ_{jt} is a $N(0, 1)$ white noise.

- Case II: AR(2) with unit roots plus drift. Let $X_{j,t} = 1.03X_{j,t-1} - 0.03X_{j,t-2} + \epsilon_{jt} + c_j$ (with roots 1 and 0.03), where c_j is a constant for $j = 1, ..., p$. To be specific, c_j is the negative of the minimum value of the $AR(2)$ series without drift plus one. This guarantees that all entries generated are positive and feasible for taking logarithm.

- Case III: AR(3) with unit roots plus seasonality. Let $X_{j,t} = 1.2X_{j,t-1} - 0.21X_{j,t-2} + 0.01X_{j,t-3} + \epsilon_{jt} + c_j + sin(\pi * t/16) * 5$ (with roots 1,0.1,0.1), where $j = 1, ..., p$ and $t = 1, ..., n$.

After generating the data matrix X, the aggregate matrix Y is obtained by the linear transformation $Y = XB$ with the matrix $B = \sum_{i=1}^{m} d_i \mathbf{u}_i \mathbf{v}_i'$, where the first five singular values are $(177, 32, 30, 26, 22)$, while the others are less than 5. This indicates that a reduced rank regression model with rank $r = 5$ is appropriate.

In order to choose proper transformations of the simulated data to mimic the features of the macroeconomics data, we examine the original and transformed series in Stock and Watson (2012), and note that there are four worthy characteristics of the data:

(1) the original series are all positive or above zero, and many of them have large ranges;

(2) the original series often have trends or seasonality;

(3) the transformed series have smaller ranges, often between -1 to 1;

(4) the transformed series are smooth and have small variance along the x-axis.

Note that when X is generated by a random walk (Case I), its first order difference is a stationary white noise. For the other two cases, we consider the following different transformations: (1) first- and second-order differences, (2) first-order difference of the logarithm. Compared with the characteristics of the real data, the first-order difference of log transformation is the closest one to the transformation in Stock and Watson (2012). Hence, we use it as the transformation for Cases II and III, and use the first-order difference for Case I.

3.4.2.2 Simulation results

We produce rolling out-of-sample one-step-ahead forecasts with rolling window size 100. The forecast accuracy is measured using the RMSE. The replications for the simulation is 100 for each of the Cases I, II and III.

FIT-SGMD-r	Percentiles					Empirical Distrituion				
	5%	25%	50%	75%	95%	<.9	.9-.97	.97-1.03	1.03-1.1	>1.1
Original Data										
3	0.35	1.21	1.52	1.87	2.55	0.10	0.02	0.02	0.04	0.83
4	0.17	1.09	1.31	1.56	2.04	0.11	0.04	0.04	0.07	0.74
5	0.10	0.98	1.14	1.33	1.66	0.16	0.08	0.09	0.11	0.57
6	0.07	0.88	1.01	1.15	1.39	0.28	0.13	0.12	0.13	0.34
7	0.05	0.80	0.92	1.03	1.22	0.46	0.16	0.12	0.10	0.15
8	0.04	0.72	0.83	0.94	1.10	0.66	0.14	0.09	0.06	0.05
9	0.03	0.65	0.76	0.87	1.02	0.80	0.10	0.05	0.03	0.02
Transformed Data										
3	0.02	0.99	1.02	1.04	1.09	0.09	0.09	0.46	0.34	0.03
4	0.02	0.95	0.99	1.02	1.07	0.14	0.23	0.43	0.19	0.02
5	0.02	0.90	0.96	1.00	1.05	0.23	0.32	0.33	0.11	0.01
6	0.02	0.86	0.93	0.98	1.04	0.37	0.33	0.23	0.06	0.00
7	0.02	0.83	0.90	0.95	1.02	0.50	0.31	0.15	0.03	0.00
8	0.02	0.79	0.87	0.93	1.00	0.63	0.25	0.10	0.02	0.00
9	0.02	0.76	0.84	0.90	0.98	0.73	0.20	0.06	0.01	0.00

Table 3.3: Case I: Distributions of ratios of RMSE to PCR-5 for the original and transformed simulated data.

FIT-SGMD-r	Percentiles					Empirical Distrituion				
	5%	25%	50%	75%	95%	<.9	.9-.97	.97-1.03	1.03-1.1	>1.1
Original Data										
3	0.25	0.59	0.80	1.03	1.46	0.62	0.07	0.06	0.06	0.19
4	0.12	0.56	0.75	0.96	1.34	0.69	0.07	0.05	0.05	0.14
5	0.03	0.54	0.72	0.91	1.28	0.74	0.06	0.04	0.04	0.11
6	0.01	0.47	0.61	0.76	0.99	0.91	0.04	0.02	0.01	0.03
7	0.01	0.42	0.54	0.64	0.82	0.98	0.01	0.00	0.00	0.00
8	0.01	0.38	0.48	0.57	0.72	1.00	0.00	0.00	0.00	0.00
9	0.01	0.35	0.43	0.51	0.63	1.00	0.00	0.00	0.00	0.00
Transformed Data										
3	1.50	1.70	1.85	2.01	2.27	0.00	0.00	0.00	0.00	1.00
4	1.51	1.71	1.86	2.02	2.28	0.00	0.00	0.00	0.00	1.00
5	1.52	1.72	1.87	2.03	2.29	0.00	0.00	0.00	0.00	1.00
6	1.53	1.72	1.88	2.04	2.30	0.00	0.00	0.00	0.00	1.00
7	1.55	1.73	1.89	2.06	2.32	0.00	0.00	0.00	0.00	1.00
8	1.55	1.74	1.90	2.07	2.33	0.00	0.00	0.00	0.00	1.00
9	1.56	1.75	1.91	2.08	2.34	0.00	0.00	0.00	0.00	1.00

Table 3.4: Case II: Distributions of ratios of RMSE to PCR-5 for the original and transformed simulated data.

Tables 3.3-3.5 report percentiles and empirical distributions of ratios of RMSE for our forecast method relative to the PCR-5 benchmark, for the Cases I, II and III, respectively. The performances of the FIT-SGMD for the rank $r = 3, ..., 9$ are compared to the PCR-5 for both the original data and the transformed data. The pattern is similar to that of the results in Section 3.4.1, where the PCR-5 has superior performance across all the 35 aggregate series. We find the same pattern for the transformed data in Tables 3.4 and 3.5. However, the FIT-SGMD outperforms the benchmark for the original data panels among all three cases in Tables 3.3-3.4. The FIT-SGMD also outperforms the benchmark in the transformed panel in Table 3.3. In Table 3.4 for example, from the distributions of ratios of RMSE of our forecast method to the PCR-5, for the original data panel, it is evident that FIT-SGMD-6,7,8

FIT-SGMD-r	Percentiles					Empirical Distrituion				
	5%	25%	50%	75%	95%	<.9	.9-.97	.97-1.03	1.03-1.1	>1.1
Original Data										
3	0.41	0.89	1.21	1.60	2.36	0.26	0.05	0.04	0.06	0.58
4	0.22	0.82	1.12	1.45	2.11	0.31	0.06	0.05	0.06	0.52
5	0.05	0.79	1.05	1.37	1.96	0.35	0.07	0.06	0.07	0.45
6	0.03	0.69	0.89	1.10	1.51	0.52	0.09	0.07	0.07	0.25
7	0.02	0.60	0.76	0.91	1.19	0.74	0.07	0.05	0.05	0.09
8	0.01	0.54	0.67	0.79	1.02	0.87	0.05	0.03	0.02	0.03
9	0.01	0.47	0.59	0.71	0.91	0.94	0.03	0.01	0.01	0.01
Transformed Data										
3	1.50	1.74	1.90	2.07	2.36	0.00	0.00	0.00	0.00	1.00
4	1.51	1.75	1.91	2.08	2.37	0.00	0.00	0.00	0.00	1.00
5	1.51	1.75	1.92	2.09	2.38	0.00	0.00	0.00	0.00	1.00
6	1.52	1.76	1.93	2.09	2.39	0.00	0.00	0.00	0.00	1.00
7	1.52	1.76	1.93	2.10	2.41	0.00	0.00	0.00	0.00	1.00
8	1.53	1.77	1.94	2.11	2.42	0.00	0.00	0.00	0.00	1.00
9	1.53	1.77	1.94	2.12	2.43	0.00	0.00	0.00	0.00	1.00

Table 3.5: Case III: Distributions of ratios of RMSE to PCR-5 for the original and transformed simulated data.

and 9 clearly outperform the PCR-5. They have a lower RMSE in forecasting than the PCR-5, since all their 50-percentiles are below 1 and the $P(ratio > 1.03)$ is very close or equal to zero. Note that the FIT-SGMD are not as accurate as the PCR-5 for the transformed data since all the percentiles of the RMSE ratios are greater than 1.

Since the aggregate series are linear transformations of the disaggregate series in the original data, we illustrate their relationships after transformation. Consider a simple example where the original response y is linearly related to predictors x_1, x_2 as $y_t = b_1 x_{t,1} + b_2 x_{t,2}$. Let the first-order differences of y, x_1 and x_2 to be Δy, Δx_1, and Δx_2. Then the linear relationship remains as $\Delta y_t = b_1 \Delta x_{t,1} + b_2 \Delta x_{t,2}$, so that the linear relationship between responses and predictors does not change after the first-order differencing. However, the log transformation sabotages the linear relationship between X and Y, and causes the FIT-SGMD to fail in capturing the linear relationship for transformed data. Hence, the reason for the FIT-SGMD not outperforming the PCR-5 in the lower panels of Tables 3.4-3.5 is, perhaps, that the log transformation and the ensuing nonlinearity.

3.5 Conclusion

We considered transposable data matrices where both the rows and columns are correlated and solved the generalized matrix decomposition (GMD) problem by developing the FIT-SGMD algorithm to compute the sparse factors of the data matrix. The algorithm generalizes the FIT-SSVD algorithm in Yang et al. (2013) by accounting for the two-way dependencies in the rows and columns of the transposable data matrix. It can be used to compute the sparse components of RRR and CCA, and to yield factor and variable selections simultaneously. In applications to forecasting the macroeconomic time series data in Stock and Watson (2012), we find that the FIT-SGMD algorithm outperforms the benchmark PCR-5 for the original nonstationary data, while the PCR-5 is better for the transformed stationary data. Our simulation experiments confirm this curious phenomenon observed in the real data, and suggest that the FIT-SGMD algorithm outperforms the PCR-5 for the original nonstationary data when the response variables are linearly related to the predictors. Thus, the use of the FIT-SGMD in these situations may obviate the need to follow the traditional and subjective steps: (1) transforming a (high-dimensional) nonstationary data to (marginal) stationarity, (2) model-building and forecasting and (3) back-transforming the forecasts to the original scale.

4. CONCLUSIONS

This dissertation research consists of two novel applications of the idea of thresholding: the thresholding reduced rank multivariate regression and the generalized PCA/SVD. We have developed the methodologies, related theoretical results and applied them in the areas of macro-arrary gene expression, macroeconomics and brain image fMRI datasets. We showed that our proposed methods are flexible and can be applied to a wide variety of statistical analysis, such as the low-rank model, the reduced rank regression and the canonical correlation analysis.

The two thresholding methodologies have been developed in Sections 2-3. The developments in Section 2 considered large multivariate linear regression models and developed a method to estimate the regression coefficient matrix by low-rank matrices constructed from its sparse SVD. We presented the FIT-SRRR method for low-rank approximation of the regression coefficient matrix. It is a generalization of the FIT-SSVD algorithm for the correlated data which requires finding the threshold level in this new setup. The FIT-SRRR algorithm inherits the good properties of the FIT-SSVD and is more efficient than the FIT-SSVD for correlated data situations as demonstrated through simulation experiments. The developments in Section 3 considered the low-rank approximation of transposable data matrices where both their rows and columns are correlated. Rather than using the weighted least squares matrix decomposition with respect to a transposable quadratic norm as in Allen et al. (2013), we replace their optimization framework by thresholding the GMD factors and propose the FIT-SGMD algorithm while accounting for the two-way dependencies. The FIT-SGMD algorithm guarantees the orthogonality of the GMD factor, which is a desirable property.

The related theoretical results for the methodologies are developed in Sections 2 and 3. In Section 2, we have placed the FIT-SSVD algorithm in Yang et al. (2013) in an optimization framework by introducing a specific bi-convex objective function. This enables us to study the large sample properties of the solution and establish consistency of the estimators as the sample size tends to infinity. In Section 3, we have shown that our FIT-SGMD algorithm is suitable for a general framework, where the reduced rank regression and canonical correlation analysis are two important special cases. These connections enable us to improve the predictive accuracy in regression and to facilitate the interpretation of our proposed algorithm.

The computational results consist of the extensive simulation and real data applications. In Section 2, the simulation study of the FIT-SRRR algorithm produced superior performance compared to the existing counterparts. Using the real data, we showed the promise of applying this method in producing interpretable results, and improving estimation and forecast accuracy. In Section 3, our analysis of using the FIT-SGMD to the macroeconomic time series data in Stock and Watson (2012) revealed that it outperforms the benchmark PCR-5 for the original nonstationary data. Our simulation experiments confirm this curious phenomenon observed in the real data and suggest that the FIT-SGMD algorithm outperformed the PCR-5 for the original nonstationary data when the response variables are linearly related with the predictors. Since it is a difficult task to decide what transformations to use to reduce data to stationarity, even for univariate time series data, such findings are desirable in high-dimensional data situations. Thus, using the FIT-SGMD in these situations may obviate the following traditional subjective steps: (i) transforming a (high-dimensional) nonstationary data to stationarity, (ii) model-building and forecasting and (iii) back-transforming the forecasts to the original scale.

REFERENCES

Allen, G. I., Grosenick, L., and Taylor, J. (2013), "A Generalized Least Squares Matrix Decomposition," *Journal of the American Statistical Association*, To appear.

Allen, G. I. and Tibshirani, R. (2010), "Transposable Regularized Covariance Models with an Application to Missing Data Imputation," *The Annals of Applied Statistics*, 4(2), 764–790.

— (2012), "Inference with Transposable Data: Modeling the Effects of Row and Column Correlations," *Journal of the Royal Statistical Society: Series B*, 74(4), 721–743.

Anderson, T. W. (1951), "Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions," *The Annals of Mathematical Statistics*, 22, 327–351.

Asgarian, N. and Greiner, R. (2007), "Using Rank-1 Biclusters to Classify Microarray Data," *Bioinformatics*, 0, 1–10.

Berkner, K. and Wells, R. O. (1998), "A Correlation-Dependent Model for Denoising via Nonorthogonal Wavelet Transforms," *Computational Mathematics Laboratory, Rice University, Technical Reports*, 98–107.

— (2001), *Denoising via Nonorthogonal Wavelet Transforms*, Springer, New York, NY.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., and Meyerson, M. (2001), "Classification of Human Lung Carcinomas by MRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses," *Proceedings of the National*

Academy of Sciences, 13790–13795.

Bunea, F., She, Y., and Wegkamp, M. (2011), “Optimal Selection of Reduced Rank Estimators of High-dimensional Matrices,” *The Annals of Statistics*, 39(2), 1282–1309.

Cai, J., Candes, E. J., and Shen, Z. (2010), “A Singular Value Thresholding Algorithm for Matrix Completion,” *SIAM Journal on Optimization*, 20(4), 1956–1982.

Chen, K., Chan, K., and Chr.Stenseth, N. (2012a), “Reduced Rank Stochastic Regression with a Sparse Singular Value Decomposition,” *Journal of the Royal Statistical Society, Series B*, 74(2), 203–221.

Chen, K., Dong, H., and Chan, K. (2012b), “Reduced Rank Regression via Adaptive Nuclear Norm Penalization,” *Biometrika*, 100(4), 901–920.

Chen, L. and Huang, Z. H. (2012), “Sparse Reduced-Rank Regression for Simultaneous Dimension Reduction and Variable Selection in Multivariate Regression,” *Journal of the American Statistical Association*, 107(500), 1533–1545.

Delouille, V., Simoens, J., and von Sachs, R. (2004), “Smooth Design-Adapted Wavelets for Nonparametric Stochastic Regression,” *Journal of the American Statistical Association*, 99(467), 643–658.

Dobrev, D. and Schaumburg, E. (2013), “Robust Forecasting by Regularization,” *Unpublished Manuscript.*

Donoho, D. L. and Johnstone, I. M. (1994), “Ideal Spatial Adaptation by Wavelet Shrinkage,” *Biometrika*, 81(3), 425–455.

Eckart, C. and Young, G. (1936), “The Approximation of One Matrix by Another of Lower Rank,” *Psychometrika*, 1(3), 211–218.

Efron, B. (2009), “Are A Set of Microarrays Independent of Each Other?” *The Annals of Applied Statistics*, 3(3), 922–942.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), “Least Angle Regression," *The Annals of Statistics*, 32(2), 407–499.

Escoufier, Y. (1977), "Operator Related to A Data Matrix," *Recent Developments in Statistics*, 125–131.

— (2006), "Operator Related to A Data Matrix: A Survey," *Compstat 2006-Proceedings in Computational Statistics*, 285–297.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96(456), 1348–1360.

Fan, J. and Lv, J. (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, 20(1), 101–148.

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7(2), 179–188.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 9(3), 432–441.

Galbraith, R. and Galbraith, J. (1974), "On the Inverses of Some Patterned Matrices Arising in The Theory of Stationary Time Series," *Journal of Applied Probability*, 11(4), 63–71.

Golub, G. H. and Van Loan, C. F. (1996), *Matrix Computations (3rd ed.)*, Johns Hopkins University Press, Baltimore, MD.

Gorski, J., Pfeuffer, F., and Klamroth, K. (2007), "Biconvex Sets and Optimization with Biconvex Functions: A Survey and Extensions." *Mathematical Methods of Operations Research*, 66(3), 373–407.

Gupta, A. K. and Nagar, D. K. (1999), *Matrix Variate Distributions*, CRC Press, Boca Raton, FL.

Hastie, T., Tibshirani, R., and J., F. (2009), *The Elements of Statistical Learning*, Springer, New York, NY.

Hirschfeld, H. (1935), "A Connection Between Correlation and Contingency," *Proc. Cambridge Philosophical Society*, 31, 520–524.

Hotelling, H. (1935), "The Most Predictable Criterion," *Journal of Education Psychology*, 26(2), 139–142.

— (1936), "Relations Between Two Sets of Variates," *Biometrika*, 2(3-4), 321–377.

Huang, Z., Shen, H., and Buja, A. (2009), "The Analysis of Two-Way Functional Data Using Two-Way Regularized Singular Value Decompositions," *Journal of the American Statistical Association*, 104, 1609–1620.

Izenman, A. J. (1975), "Reduced-rank Regression for the Multivariate Linear Model," *Journal of Multivariate Analysis*, 5(2), 248–264.

— (2008), *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, Springer, New York, NY.

Johnson, R. A. and Wichern, D. W. (2007), *Applied Multivariate Statistical Analysis*, vol. 6, Prentice Hall, Upper Saddle River, NJ.

Johnstone, I. M. (2011), *Gaussian Estimation: Sequence and Multiresolution Models*, Unpublished Manuscript.

Johnstone, I. M. and Lu, A. Y. (2009), "On Consistency and Sparsity for Principal Components Analysis in High Dimensions," *Journal of the American Statistical Association*, 104(486), 682–693.

Johnstone, I. M. and Silverman, B. (1997), "Wavelet Threshold Estimators for Data with Correlated Noise," *Journal of the Royal Statistical Society, Series B*, 59(2), 319–351.

Knight, K. and Fu, W. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356–1378.

Kovac, A. and Silverman, B. (2000), "Extending the Scope of Wavelet Regression Methods by Coefficient-Dependent Thresholding," *Journal of the American Sta-*

tistical Association, 95(449), 172–183.

Lazar, N. (2008), *The Statistical Analysis of Functional MRI Data*, Springer, New York, NY.

Lazzeroni, L. and Owen, A. (2002), “Plaid Models for Gene Expression Data,” *Statistica Sinica*, 12(1), 61–86.

Lee, M., Shen, H., Huang, J., and Marron, J. S. (2010), “Biclustering via Sparse Singular Value Decomposition,” *Biometrics*, 66(4), 1087–1095.

Lindquist, M. (2008), “The Statistical Analysis of fMRI Data,” *Statistical Science*, 23(4), 439–464.

Lu, Z., Monteiro, R., and Yuan, M. (2012), “Convex Optimization Methods for Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression,” *Mathematical Programming*, 131(1-2), 163–194.

Ma, Z. (2013), “Sparse Principal Component Analysis and Iterative Thresholding,” *The Annals of Statistics*, 41(2), 772–801.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008), “The Group Lasso for Logistic Regression,” *Journal of the Royal Statistical Society: Series B*, 70(1), 53–71.

Merris, R. (1994), “Laplacian Matrices of Graphs: A Survey,” *Linear Algebra and Its Applications*, 197, 143–176.

Mitchell, T., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., Just, M., and Newman, S. (2004), “Learning to Decode Cognitive States From Brain Images,” *Machine Learning*, 57(1), 145–175.

Owen, A. B. and Perry, P. (2009), “Bi-cross-validation of the SVD and the Nonnegative Matrix Factorization,” *The Annals of Applied Statistics*, 3, 564–594.

Reinsel, G. C. and Velu, P. (1998), *Multivariate Reduced-rank Regression: Theory and Applications*, Springer, New York, NY.

Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Appli-*

cations, CRC Press, Boca Raton, FL.

Shaman, P. (1969), "On the Inverse of the Covariance Matrix of a First Order Moving Average," *Biometrika*, 56(3), 595–600.

Shen, H. and Huang, J. Z. (2008), "Sparse Principal Component Analysis via Regularized Low Rank Matrix Approximation," *Journal of Multivariate Analysis*, 99(6), 1015–1034.

Srivastava, M. (2009), "Estimation and Testing in General Multivariate Linear Models with Kronecker Product Covariance Structure," *Sankhyā: The Indian Journal of Statistics, Series A*, 71, 137–163.

Stock, J. and Watson, M. (2012), "Generalized Shrinkage Methods for Forecasting Using Many Predictors," *Journal of Business and Economic Statistics*, 30(4), 481–493.

Tibshirani, R. (1996), "Regrssion Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Toh, K. and Yun, S. (2010), "An Accelerated Proximal Gradient Algorithm for Nuclear Norm Regularized Linear Least Squares Problems," *Pacific Journal of Optimization*, 6, 615–640.

Viroli, C. (2012), "On Matrix-variate Regression Analysis," *Journal of Multivariate Analysis*, 111, 296–309.

Wang, X. and Wood, A. (2006), "Empirical Bayes Block Shrinkage of Wavelet Coefficients via the Non-central Chi-square Distribution," *Biometrika*, 66(4), 705–722.

— (2010), "Wavelet Estimation of an Unknown Function Observed with Correlated Noise," *Communications in Statistics - Simulation and Computation*, 39(2), 287–304.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009), "A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical

Correlation Analysis," *Biostatistics*, 10(3), 515–534.

Wittstock, G. (1984), "On Matrix Order and Convexity," *North-Holland Mathematics Studies*, 90, 175–188.

Yang, D., Ma, Z., and Buja, A. (2013), "A Sparse SVD Method for High-dimensional Data," *Journal of Computational and Graphical Statistics*, To appear.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007), "Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression," *Journal of the Royal Statistical Society, Series B*, 69(3), 329–346.

Yuan, M. and Lin, Y. (2007), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society, Series B*, 68(1), 49–67.

Zhao, P. and Yu, B. (2007), "Stagewise Lasso," *The Journal of Machine Learning Research*, 8, 2701–2726.

Zou, H. (2006), "The Adaptive Lasso And Its Oracle Properties." *Journal of the American Statistical Association*, 101(476), 1418–1429.

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, 67(2), 301–320.

APPENDIX A

SUPPLEMENTARY MATERIALS FOR SECTION 2

A.1 Additional Objective Functions for Hard-thresholding and SCAD

In Section 2.2.2, we placed the FIT-SSVD algorithm with the soft-thresholding function in an optimization problem. We briefly discuss here the corresponding choices of objective functions when the hard-thresholding or the SCAD (Fan and Li, 2001) are used in the FIT-SSVD algorithm.

When the hard-thresholding function $H(y, \lambda) = yI_{\{|y|>\lambda\}}$ is used in the FIT-SSVD, then its objective function $\Psi_H(U, D, V)$ has the form

$$\Psi_H(U, D, V) = ||Y - UDV'||_F^2 + \lambda_u^2 \sum_{i=1}^{p} \sum_{k=1}^{r} I_{\{|u_{ik}d_k| \neq 0\}} + \lambda_v^2 \sum_{j=1}^{q} \sum_{k=1}^{r} I_{\{|v_{jk}d_k| \neq 0\}}. \text{(A.1)}$$

For $\tilde{U} = UD, \tilde{V} = VD$, it can be shown that the solution $\tilde{U}$ of (A.1) for V fixed is found by component-wise hard-thresholding of YV, i.e. $H(YV, \lambda_u) = [H((YV)_{ij}, \lambda_u)]$ for $i = 1, ..., p$, $j = 1, ..., r$. Similarly, the solution $\tilde{V}$ of (A.1) for U fixed is $H(Y'U, \lambda_v)$.

For the SCAD function, let $h^{SCAD}(y, \lambda)$ denote the SCAD operator

$$h^{SCAD}(y, \lambda) = \begin{cases} sign(y)(|y| - \lambda)_+ & for\ |y| \leq 2\lambda; \\ \{(a-1)y - sign(y)a\}/(a-2) & for\ 2\lambda < |y| \leq a\lambda; \\ y & for\ |y| > a\lambda, \end{cases}$$

where $a > 2$ is another regularization parameter. The corresponding objective function has the form

$$\begin{aligned}\Psi_{SCAD}(U,D,V) = \quad & ||Y - UDV'||_F^2 + \sum_{i=1}^{p}\sum_{k=1}^{r}\{2\lambda_u|u_{ik}d_k|I_{\{|u_{ik}d_k|\leq\lambda_u\}} \\ & -\frac{(u_{ik}d_k)^2 - 2a_u\lambda_u|u_{ik}d_k| + \lambda_u^2}{a_u - 1}I_{\{\lambda_u<|u_{ik}d_k|\leq a_u\lambda_u\}} \\ & +(a_u+1)\lambda_u^2 I_{\{|u_{ik}d_k|>a_u\lambda_u\}}\} \\ & +\sum_{j=1}^{q}\sum_{k=1}^{r}\{2\lambda_v|v_{jk}d_k|I_{\{|v_{jk}d_k|\leq\lambda_v\}} \\ & -\frac{(v_{jk}d_k)^2 - 2a_v\lambda_v|v_{jk}d_k| + \lambda_v^2}{a_v - 1}I_{\{\lambda_v<|v_{ik}d_k|\leq a_v\lambda_v\}} \\ & +(a_v+1)\lambda_v^2 I_{\{|v_{ik}d_k|>a_v\lambda_v\}}\}\end{aligned}$$

For $\tilde{U} = UD, \tilde{V} = VD$, it can be shown that the solution $\tilde{U}$ for V fixed is found by component-wise SCAD, i.e. $h^{SCAD}(YV, \lambda_u) = [h^{SCAD}((YV)_{ij}, \lambda_u)]_{i=1,\ldots,p;j=1,\ldots,r}$, and the solution $\tilde{V}$ for U fixed is $h^{SCAD}(Y'U, \lambda_v)$.

A.2 Additional Simulations

In this subsection, we consider a special case of the multivariate reduced rank regression in Section 2.4.2, where the design matrix X is the identity matrix, and provide additional simulation results to further compare the FIT-SRRR with the IEEA and SSVD algorithms.

Let $B = d\mathbf{u}\mathbf{v}'$ in (1) have the same $d, \mathbf{u}, \mathbf{v}$ as in Section 2.4.1 setup and the design matrix X be a 50×50 identity matrix, then model (1) reduces into $Y = B + E$, where the entries of E are samples from $N(0, 1)$. This is also the same setup used in Lee et al. (2010) and Chen et al. (2012a), where they found out both the IEEA and SSVD methods significantly outperformed other existing methods, e.g. SVD, Plaid in Lazzeroni and Owen (2002), RoBiC in Asgarian and Greiner (2007) and SPCA in Shen and Huang (2008), in terms of misclassification rate and the mean-squared error (MSE). Hence, it suffices to only compare the FIT-SRRR with the IEEA and SSVD.

The results summarized in Table A.1 show that in terms of the MSE, the FIT-SRRR enjoys a lower level than those in the IEEA and SSVD. In terms of the misclassification rates, the FIT-SRRR has extremely low misclassification rates in the performance of both correctly identifying zero and non-zero entries. The overall misclassification rate of the FIT-SRRR is 0.27%, which is only about one fourth of the rates of its counterparts.

			Corrected identified		Total
Method	MSE		0s(%)	non-0s(%)	Error %
FIT-SRRR	**6.20**	**u**	33.96(99.88)	16(100)	0.08
		v	74.93(99.91)	24.70(98.8)	0.37
		Overall	108.89(99.90)	40.7(99.27)	**0.27**
IEEA	**6.65**	**u**	33.90(99.7)	16(100)	0.21
		v	73.69(98.25)	24.78(99.13)	1.53
		Overall	107.58(98.70)	40.78(99.46)	**1.09**
SSVD	**6.35**	**u**	33.77(99.32)	16(100)	0.46
		v	73.95(98.60)	24.73(98.93)	1.32
		Overall	107.71(98.82)	40.73(99.35)	**1.04**

Table A.1: Comparison of FIT-SRRR method with SRRR and SSVD for unit rank models.

A.3 Additional Heat Maps for the Lung Cancer Data

In this subsection, we perform the SSVD in Lee et al. (2010), the FIT-SSVD in Yang et al. (2013) and the IEEA in Chen et al. (2012a) on the lung cancer data, and plot their corresponding heat maps of the first three estimated layers as a supplementary for Section 2.5. The four blocks in each panel reveal the four groups of subjects. These from top to the bottom are: SmallCell, Normal, Colon and Carcinoid.

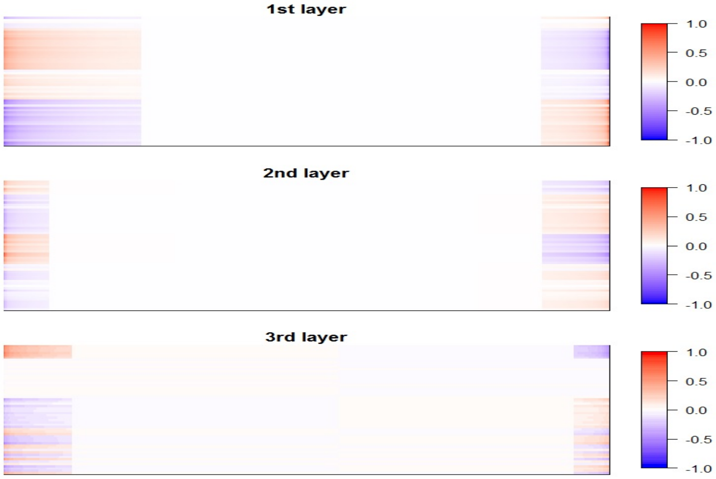


Figure A.1: Heat maps of the first three layers using SSVD of lung cancer data.

A.4 Proof of Proposition 2.2.1

(a) To show that (3) subject to (2.5) is biconvex in $\tilde{U}, \tilde{V}$, we need to show that (2.6) is convex in $\tilde{U}$ for V fixed and (2.7) is convex in $\tilde{V}$ for U fixed subject to (2.5). For (2.6), since the set of matrices U in (2.5) is convex (Wittstock, 1984), and both $||Y - \tilde{U}V'||_F^2$ and the L_1 norm of $\tilde{U}$ are convex functions, hence their sum is also convex. The convexity of (2.7) is shown similarly.

Using the bi-convexity of $\Psi(\cdot)$ in (2.3), this objective function can be minimized by iteratively minimizing the objective functions (2.6) and (2.7) (Gorski et al., 2007), where it is known that the objective function decreases monotonically at each iteration and the sequence generated by the iterative algorithm converges monotonically to its partial minimum (Gorski et al., 2007, Section 4).

(b) Here we only prove that the solution $\tilde{V}$ of (2.7) is found by componentwise soft-thresholding of $Y'U$. The the proof for $\tilde{U}$ in (2.6) is similar. For a fixed orthonormal matrix U, let $U^{\perp}$ be its complement so that the partitioned $\mathbf{U} = [U|U^{\perp}]$ is an orthogonal matrix. From the definition of the Frobenius norm $||A||_F^2 = tr(A'A)$,

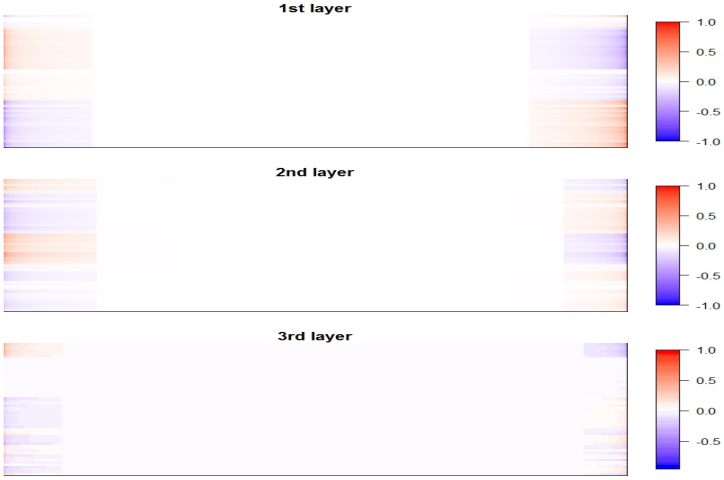


Figure A.2: Heat maps of the first three layers by FIT-SSVD of lung cancer data.

it follows that

$$\begin{aligned} & ||Y - U\tilde{V}'||_F^2 = ||Y' - \tilde{V}U'||_F^2 = tr\{(Y' - \tilde{V}U')'(Y' - \tilde{V}U')\mathbf{U}\mathbf{U}'\} \\ = \quad & ||(Y' - \tilde{V}U')\mathbf{U}||_F^2 = ||(Y' - \tilde{V}U')[U|U^{\perp}]||_F^2 = ||Y'U - \tilde{V}||_F^2 + ||Y'U^{\perp}||_F^2, \end{aligned}$$

where the last term $||Y'U^{\perp}||_F^2$ is free of $\tilde{V}$. Hence, minimizing (2.7) with respect to $\tilde{V}$ is equivalent to minimizing

$$\begin{aligned} & ||Y'U - \tilde{V}||_F^2 + \lambda_v \sum_{l=1}^{r}\sum_{i=1}^{q} |\tilde{v}_{il}| \\ = \quad & \sum_{l=1}^{r} ||Y'\mathbf{u}_l - \tilde{\mathbf{v}}_l||_F^2 + \lambda_v \sum_{l=1}^{r} |\tilde{\mathbf{v}}_l| = \sum_{l=1}^{r}\sum_{i=1}^{q}(Y_i'\mathbf{u}_l - \tilde{v}_{il})^2 + \lambda_v \sum_{l=1}^{r}\sum_{i=1}^{q} |\tilde{v}_{il}|, \end{aligned} \tag{A.2}$$

where $Y_i, \mathbf{u}_l, \tilde{\mathbf{v}}_l$ are the ith, lth columns of Y, U and $\tilde{V}$, respectively, and $\tilde{v}_{il}$ is the (i, l)th entry of $\tilde{V}$. Expanding the right hand side of (A.2), it follows that

$$RHS = \sum_{l=1}^{r}\sum_{i=1}^{q}(\tilde{v}_{il}^2 - 2\tilde{v}_{il}Y_i'\mathbf{u}_l + \lambda_v|\tilde{v}_{il}|) + \sum_{l=1}^{r}\sum_{i=1}^{q} \mathbf{u}_l'Y_iY_i'\mathbf{u}_l.$$

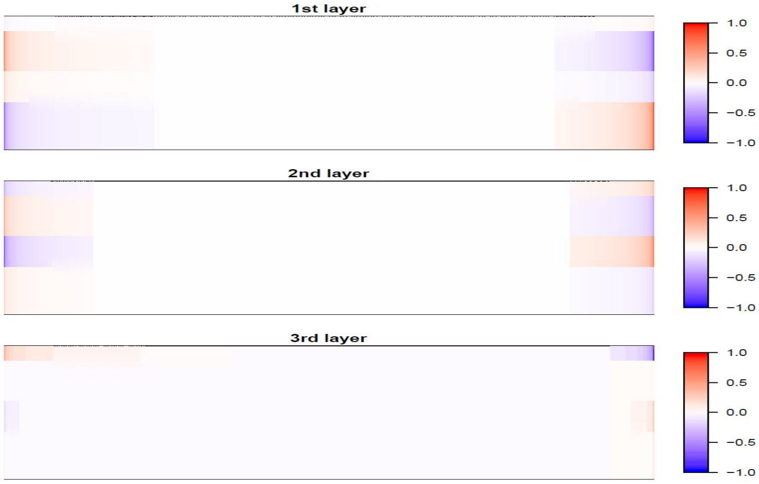


Figure A.3: Heat maps of the first three layers by IEEA of lung cancer data.

Thus, the minimizers $\tilde{v}_{il}$'s for all $i = 1, ..., q; l = 1, ..., r$ are given by the soft-thresholding of $Y_i'\mathbf{u}_l$, i.e. $S(Y_i'\mathbf{u}_l, \frac{1}{2}\lambda_v) = sgn(Y_i'\mathbf{u}_l)(|Y_i'\mathbf{u}_l| - \frac{1}{2}\lambda_v)_+$, hence, the solution of $\tilde{V}$ is $S(Y'U, \frac{1}{2}\lambda_v)$. Similarly, the solution of $\tilde{U}$ in (2.6) is $S(YV, \frac{1}{2}\lambda_v)$.

(c) Recall that in the FIT-SSVD algorithm the kth updating step for the estimators $U^{(k),thr}$,$V^{(k),thr}$ of UD, VD are $S(YV^{(k-1)}, \gamma_u)$ and $S(Y'U^{(k)}, \gamma_v)$, where γ_u, γ_v are the threshold levels. Fortunately from **(b)**, $\tilde{U}^{(k)}, \tilde{V}^{(k)}$ in the kth iteration of (2.3) are of the forms $S(YV^{(k-1)}, \frac{1}{2}\lambda_v)$, $S(Y'U^{(k)}, \frac{1}{2}\lambda_v)$, respectively. Hence, with $\gamma_u = \frac{1}{2}\lambda_u, \gamma_v = \frac{1}{2}\lambda_v$, the updates in the kth iteration in the FIT-SSVD algorithm have the same forms as those in minimizing the objective function in (2.3).

A.5 Existence of Local Minimum and Selection Consistency

In this part, for the objective function (2.3) we discuss existence of a local minimum and the selection consistency of the solutions. For simplicity, the singular value matrix D of the SVD of B is absorbed into the singular vectors and let $\lambda_u = \{\lambda_{il}^u\}$ and $\lambda_v = \{\lambda_{jl}^v\}$ for $i = 1, ..., p; j = 1, ..., q; l = 1, ..., r$. Then, the objective function

(2.3) can be written as

$$\Psi_{u,v}(U,V) = ||Y - UV'||_F^2 + \sum_{i=1}^{p}\sum_{l=1}^{r}\lambda_{il}^u|u_{il}| + \sum_{j=1}^{q}\sum_{l=1}^{r}\lambda_{jl}^v|v_{jl}|. \tag{A.3}$$

A major difference between (A.3) and the objective function $||Y - XUV'||_F^2 + \sum_{l=1}^{r} \sum_{i=1}^{p}\sum_{j=1}^{q} w_{ijl}|u_{il}v_{jl}|$ in Chen et al. (2012a), is the additivity of the penalties on U and V instead of being multiplicative. Imposing a multiplicative penalty directly on $B = UV'$ but not on its factors U and V leads to an identifiability problem since the decomposition of B is not unique. A secondary difference is that (A.3) is for the signal plus noise model where X is an identity matrix. Here, we state the main theoretical results and outline the key steps of the proof. Fore more details see the proof of Theorem 2.1 in Chen et al. (2012a).

Let the true $B = U^*V^{*\prime}$, where U^*, V^* are $p \times r, q \times r$ orthogonal matrices with rank $r \leq min(p,q)$. Let L_u, L_v denote the index sets for U^* and V^* in which every element in a row is zero, H_u, H_v be the complimentary sets of L_u and L_v, and $|H_u|, |H_v|$ denote the cardinality of the sets H_u and H_v. The following conditions are needed in the theoretical development.

1. The errors e_{ij} in E are iid with $E(e_{ij}) = 0$ and $Var(e_{ij}) = \sigma^2$.
2. Suppose $\lambda_{il}^u/\sqrt{p} \to 0$ for $i \in H_u$, $\lambda_{il}^u/\sqrt{p} \to \infty$ for $i \in L_u$, and $\lambda_{jl}^v/\sqrt{p} \to 0$ for $j \in H_v$, $\lambda_{jl}^v/\sqrt{p} \to \infty$ for $j \in L_v$, as $p \to \infty$.

Suppose $B \in \Delta^{(r)} = \cup_{L\in\Pi}\Delta_L^{(r)}$ as in Chen et al. (2012a), where $\Delta^{(r)}$ denotes the manifold structure of all $p \times q$ matrices with rank smaller than or equal to r, and Π denotes the set of all size-r subsets of $\{1, ..., q\}$. Let $\Delta_L^{(r)} = \{UV' : rank(U) = rank(V) = r, V_{11} = I_r\}$, where the upper-left block V_{11} of V is an identity matrix after rearranging and partitioning V. Consider a ball centered at B with radius h is defined by

$$
\begin{aligned}
N(B,h) &= \{(\check{U} + \frac{1}{\sqrt{p}}\check{\mathbf{A}})(\check{V} + \frac{1}{\sqrt{p}}\check{\mathbf{B}})'; ||\check{\mathbf{A}}||_F \le h, ||\check{\mathbf{B}}||_F \le h, \check{\mathbf{B}}_{11} = 0\} \\
&= \{(U^* + \frac{1}{\sqrt{p}}\mathbf{A})(V^* + \frac{1}{\sqrt{p}}\mathbf{B})'; \mathbf{A} = \check{\mathbf{A}}Q^{-1}, \mathbf{B} = \check{\mathbf{B}}Q, \mathbf{B}_{11} = 0\} \in \Delta^{(r)},
\end{aligned}
$$

where $\check{U} = U^*Q', \check{V} = V^*Q^{-1}$, Q is an $r \times r$ invertible matrix. We have the following theorems.

Theorem A.5.1 *(existence of a local minimum) Given a data matrix Y, suppose condition 1 are satisfied and $\lambda^u_{il}/\sqrt{p} \to 0$ for $i \in H_u$, $\lambda^v_{jl}/\sqrt{p} \to 0$ for $j \in H_v$, as $p \to \infty$. Then there exists a local minimizer $(\hat{U}, \hat{V})$ of $\Psi_{u,v}(U,V)$ in (A.3) which is $\sqrt{p}-$consistent in estimating U^*, V^*, i.e. $||\hat{U}^{(n)} - U^*|| = O_P(p^{-1/2})$ and $||\hat{V}^{(n)} - V^*|| = O_P(p^{-1/2})$.*

Proof Following Chen et al. (2012a), we show that for any given $\epsilon > 0$, there is a large enough h such that

$$
\liminf_n Pr\{\inf_{||\check{\mathbf{A}}||_F = ||\check{\mathbf{B}}||_F = h} \Psi_{u,v}(U^* + \frac{1}{\sqrt{p}}\mathbf{A}, V^* + \frac{1}{\sqrt{p}}\mathbf{B}) > \Psi_{u,v}(U^*, V^*)\} > 1 - \epsilon. \quad (A.4)
$$

Hence, with probability converging to 1, there exists a local minimum $\hat{B} = \hat{U}\hat{V}'$ inside the ball $N(B,h)$, and thus the corresponding $\hat{U}, \hat{V}$ satisfying $||\hat{U} - U^*||_F = O_P(p^{-1/2})$ and $||\hat{V} - V^*||_F = O_P(p^{-1/2})$.

To show (A.4) holds, let $\tilde{\Psi}(\mathbf{A}, \mathbf{B}) = \Psi_{u,v}(U^* + \frac{1}{\sqrt{p}}\mathbf{A}, V^* + \frac{1}{\sqrt{p}}\mathbf{B}) - \Psi_{u,v}(U^*, V^*)$,

$\mathbf{z} = vec(U^*\mathbf{B}' + \mathbf{A}V^{*\prime} + \frac{1}{\sqrt{p}}\mathbf{A}\mathbf{B}')$, and denote $\tilde{\mathbf{A}} = U^* + \frac{1}{\sqrt{p}}\mathbf{A}, \tilde{\mathbf{B}} = V^* + \frac{1}{\sqrt{p}}\mathbf{B}$, then

$$\begin{aligned} & \tilde{\Psi}(\mathbf{A}, \mathbf{B}) \\ = \quad & -2\mathbf{z}'vec(\frac{1}{\sqrt{p}}E) + \frac{1}{p}\mathbf{z}'\mathbf{z} + \sum_{l=1}^{r}\{\lambda_u \sum_{i=1}^{p}(|u_{il}^* + \frac{1}{\sqrt{p}}a_{il}| - |u_{il}^*|) \\ = \quad & +\lambda_v \sum_{j=1}^{q}(|v_{jl}^* + \frac{1}{\sqrt{p}}b_{jl}| - |v_{jl}^*|)\}. \end{aligned}$$

Since $\sqrt{p}(|u_{il}^* + \frac{1}{\sqrt{p}}a_{il}| - |u_{il}^*|) \to sgn(u_{il}^*)a_{il}$ and $\sqrt{p}(|v_{jl}^* + \frac{1}{\sqrt{p}}b_{jl}| - |v_{jl}^*|) \to sgn(v_{jl}^*)b_{jl}$, then,

$$\begin{aligned} & \Psi(\mathbf{A}, \mathbf{B}) \geq \\ & -2\tilde{\mathbf{z}}'vec(\frac{1}{\sqrt{p}}E) + \frac{1}{p}\tilde{\mathbf{z}}'\tilde{\mathbf{z}} + \sum_{l=1}^{r}\frac{1}{\sqrt{p}}\{\lambda_u \sum_{i=1}^{p} sgn(u_{il}^*)a_{il} + \lambda_v \sum_{j=1}^{q} sgn(v_{jl}^*)b_{jl}\} \end{aligned} \tag{A.5}$$

where $\tilde{\mathbf{z}} = vec(U^*\mathbf{B}' + \mathbf{A}V^{*\prime})$. From the definition of $N(B, h)$, $V_{11}^* = Q$, $(V^*Q^{-1})_{11}$ $= I_r$, and $\mathbf{B}_{11} = 0$, we have $(U^*\mathbf{B}' + \mathbf{A}V^{*\prime})_{11} = (U^*\mathbf{B}')_{11} + (\mathbf{A}V^{*\prime})_{11} = \mathbf{A}Q' = \check{\mathbf{A}}$. Then it follows that $\tilde{\mathbf{z}}'\tilde{\mathbf{z}}$ dominates the other two terms $\tilde{\mathbf{z}}'vec(\frac{1}{\sqrt{p}}E)$ and $\sum_{l=1}^{r}\frac{1}{\sqrt{p}}$ $\{\lambda_u \sum_{i=1}^{p} sgn(u_{il}^*)a_{il} + \lambda_v \sum_{j=1}^{q} sgn(v_{jl}^*)b_{jl}\}$ on the right side of (A.5).

Theorem A.5.2 *(selection consistency) Suppose Conditions 1-2 are satisfied. Then* $P(\hat{u}_{il} = 0) \to 1$ *for* $i \in L_u$, *and* $P(\hat{v}_{jl} = 0) \to 1$ *for* $j \in L_v$, *as* $p \to \infty$.

Proof Expanding (A.3), it follows that

$$\Psi_{u,v}(U, V) = \sum_{i=1}^{p}\sum_{j=1}^{q}(Y_{ij} - \sum_{l=1}^{r} u_{il}v_{jl})^2 + \sum_{i=1}^{p}\sum_{l=1}^{r}\lambda_{il}^{u}|u_{il}| + \sum_{j=1}^{q}\sum_{l=1}^{r}\lambda_{jl}^{v}|v_{jl}|.$$

To show that $P(\hat{u}_{il} = 0) \to 1$ for $i \in L_u$, we need to show that $P(\hat{u}_{il} \neq 0) \to 0$ for $i \in L_u$. Suppose $\hat{u}_{il} \neq 0$, by the KKT conditions, the first order condition for

(A.3) with respect to $\hat{u}_{il}$ is

$$\frac{2}{\sqrt{p}}\sum_{j=1}^{q}\{(Y_{ij}-\sum_{l=1}^{r}\hat{u}_{il}\hat{v}_{jl})\hat{v}_{jl}\}=\frac{\lambda_{il}^{u}}{\sqrt{p}}\frac{\hat{u}_{il}}{|\hat{u}_{il}|}. \tag{A.6}$$

The left hand side of (A.6) equals

$$\begin{aligned} LHS &= \frac{2}{\sqrt{p}}\sum_{j=1}^{q}\{[(Y_{ij}-\sum_{l=1}^{r}u_{il}^{*}v_{jl}^{*})+(\sum_{l=1}^{r}u_{il}^{*}v_{jl}^{*}-\sum_{l=1}^{r}\hat{u}_{il}\hat{v}_{jl})]\hat{v}_{jl}\} \\ &= \frac{2}{\sqrt{p}}\sum_{j=1}^{q}\{e_{ij}+O_P(\frac{1}{\sqrt{p}})\}\hat{v}_{jl}=O_P(1). \end{aligned}$$

On the other hand, from Condition 2 the right hand side of (A.6) equals

$$RHS=\frac{\lambda_{il}^{u}}{\sqrt{p}}\frac{\hat{u}_{il}}{|\hat{u}_{il}|}\to\infty \text{ as } p\to\infty, \text{ for } i\in L_u.$$

Therefore,

$$\begin{aligned} &P(\hat{u}_{il}\neq 0)\leq \\ &P\left(\frac{2}{\sqrt{p}}\sum_{j=1}^{q}\{(Y_{ij}-\sum_{l=1}^{r}\hat{u}_{il}\hat{v}_{jl})\hat{v}_{jl}\}=\frac{\lambda_{il}^{u}}{\sqrt{p}}\frac{\hat{u}_{il}}{|\hat{u}_{il}|}\right)\to 0 \text{ as } p\to\infty, \text{ for } i\in L_u. \end{aligned}$$

Similarly, we can show that $P(\hat{v}_{jl}\neq 0)\to 0$ as $p\to\infty$, for $j\in L_v$.

APPENDIX B

SUPPLEMENTARY MATERIALS FOR SECTION 3

B.1 Additional Simulations and the fMRI Data Analysis

In this subsections, we first use simulation to evaluate the performance of the FIT-SGMD, and then apply it to an fMRI dataset in Allen et al. (2013). In our study, we would like to compare our FIT-SGMD to the generalized penalized matrix factorization (GPMF) in Allen et al. (2013), which is designed to recover sparse or smooth GMD factors under conditions (3.4) by adding penalties on the (Ω, Σ)-norm of U and V. Allen et al. (2013) have compared the GPMF with SVD, two-way functional PCA (Huang et al., 2009) and sparse PCA (Shen and Huang, 2008), and have found that the GPMF outperformed others. Hence, it suffices to just compare our FIT-SGMD algorithms with the GPMF.

B.1.1 Simulations

We consider two scenarios, the first is a setup similar to Allen et al. (2013, Section 4.1), where the signal B consists of sparse U and smooth V (spatial and temporal data). The second is concerned with the signal B where both U and V are sparse.

All datasets $Y(n \times q)$ are generated from:

$$Y = B + E = UDV' + \Omega^{-1/2} Z \Sigma^{-1/2} = \sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i' + \Omega^{-1/2} Z \Sigma^{-1/2}, \tag{B.1}$$

where the entries of $Z = (Z_{ij})$'s are generated as iid $N(0, \sigma^2)$. We choose σ so as to have different levels of signal to noise ratio (SNR) where the SNR is computed following Allen et al. (2013, Section 4). We use the median absolute deviation (MAD) to estimate σ in the FIT-SGMD, $\hat{\sigma} = 1.4826MAD(as.vector(Y))$ following Yang et al. (2013). Throughout this part, the rank r of the true underlying matrix B is assumed to be known, the thresholding function in the FIT-SGMD is the hard thresholding $\eta(\cdot) = H(\cdot)$ and the number of replications for each simulation is $N = 100$ times.

The performances of the algorithms are measured by the root mean-squared error (RMSE) of the 100 simulation replications, where for replication i, the $RMSE_i = ||B - \hat{B}_i||_F, i = 1, ..., 100$. We use the RMSE ratios of comparing algorithms relative to the GPMF, i.e. $\frac{RMSE_{FIT-SGMD}}{RMSE_{GPMF}}$ and a value less than 1 indicates that our proposed algorithm has better performance than the GPMF.

B.1.1.1 Spatio-Temporal Simulated Data

We investigate the performance of the FIT-SGMD algorithm to the GPMF. We notice that in the early version of Allen et al. (2013, Section 2.4), they compared their generalized power method to the GPMF and found that although the former algorithm which is a mathematical algorithm does not enforce any sparse structure on the factor U and V, it has comparable results to that of the GPMF in terms of the estimating accuracy. Hence, in this subsection we also investigate the performance of our Algorithm 4 and denoted it as the GMD algorithm.

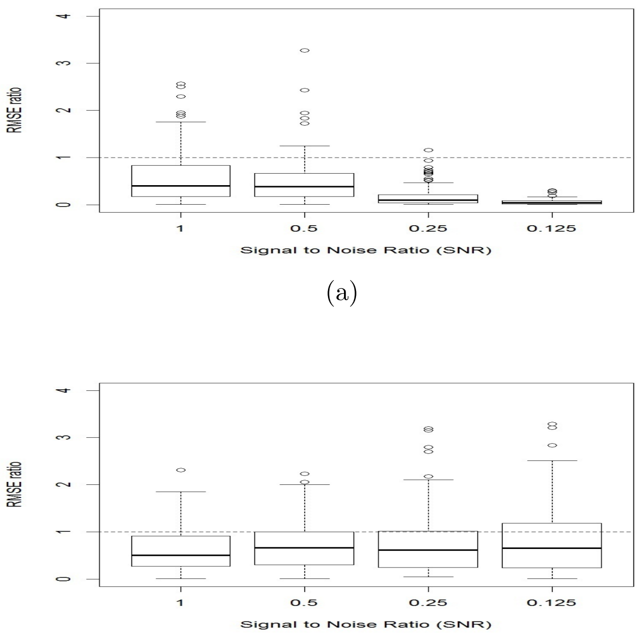
We use slight modifications of the setup in Allen et al. (2013, Section 4.1). Let the rank $r = 2$, $D = diag(1, 0.5)$, and two spatial factors $U = [\mathbf{u}_1|\mathbf{u}_2] \in R^{256\times 2}$ and two temporal factors $V = [\mathbf{v}_1|\mathbf{v}_2] \in R^{200\times 2}$. Specifically, the spatial factors are structured from two 16×16 images each with three non-overlapping non-zero signal blocks as displayed in the first column of Figure B.2. The vectors $\mathbf{u}_i, i = 1, 2$ are formed by stacking up the columns of the corresponding 16×16 image. The temporal factors are constructed as $\mathbf{v}_1 = sin(10\pi x)$ and $\mathbf{v}_2 = sin(2\pi x)$ for 200 equally spaced values

$x \in [0,1]$ as shown in the first column of Figure B.3. Rather than forming Ω, Σ as autoregressive covariance matrices as in Allen et al. (2013), we construct them to directly satisfy the conditions (3.4), i.e. $U'\Omega U = V'\Sigma V = I_r$. Due to the smoothness feature of V, we adjust our FIT-SGMD algorithm by skipping the thresholding of V (Step 6 in Algorithm 5). We first scale $\hat{Y}$ by $Y/\hat{\sigma}$ and then apply to it the GMD and FIT-SGMD algorithm.

	Panel I: FIT-SGMD		Panel II: GMD	
SNR	Median	Mean	Median	Mean
1	3.93E-01	7.00E-01	4.98E-01	6.14E-01
0.5	3.79E-01	5.15E-01	6.60E-01	7.22E-01
0.25	9.26E-02	1.80E-01	6.06E-01	7.78E-01
0.125	3.22E-02	5.61E-02	6.48E-01	8.48E-01

Table B.1: Median and mean summary for RMSE raitos. Panel I: Ratios of FIT-SGMD relative to GPMF. Panel II: Ratios of GMD relative to GPMF.

Table B.1 summarizes the two ratios ($\frac{RMSE_{GMD}}{RMSE_{GPMF}}, \frac{RMSE_{FIT-SGMD}}{RMSE_{GPMF}}$) of their mean and median for four levels of SNRs, while Figure B.1 presents the distributions of ratios using boxplots. Results demonstrate that the FIT-SGMD and GMD algorithm outperform the GPMF. From the ratios in Table B.1 (Panel II) and Figure B.1 (b), we see that the RMSE of GMD is about 50% lower than that of the GPMF for the four levels of SNR. From the ratios in Table B.1 (Panel I) and Figure B.1 (a), we see that our FIT-SGMD performs uniformly better than the GPMF, especially when the SNR is low. This occurs when SNR=0.25 and 0.125, resulting in the medians of ratios equal 0.093 and 0.032, respectively. In particular, we see that the boxplot of SNR=0.125 in Figure B.1 (a) is centered at the median with a very low level of dispersion, which indicates the RMSE of the FIT-SGMD is significantly lower than



(a)

(b)

Figure B.1: Boxplots of RMSE ratios. (a): Ratios of FIT-SGMD to GPMF. (b): Ratios of GMD to GPMF.

its counterpart. Among these three algorithms, the FIT-SGMD performs the best in terms of the low level of RMSE. A closer look of the RMSE values of these three algorithms (which is unreported here) reveals that all of them tend to have larger RMSE values as the SNR becomes lower. However, this effect is more significant and evident for the GPMF and GMD.

An example of these three methods' feature recovering performance is shown in Figures B.2-B.3, where the true and recovered Us and Vs are plotted when SNR$= 1$. The true spatial signals $\mathbf{u}_1, \mathbf{u}_2$ each consist of three non-overlapping non-zero blocks on a 16×16 grid, while the rest is zero. All three algorithms are able to recover the blocks in $\mathbf{u}_1, \mathbf{u}_2$, but their abilities to reveal the zeros are different. The FIT-SGMD outperforms the other two in terms of its ability to highlight the non-zero blocks

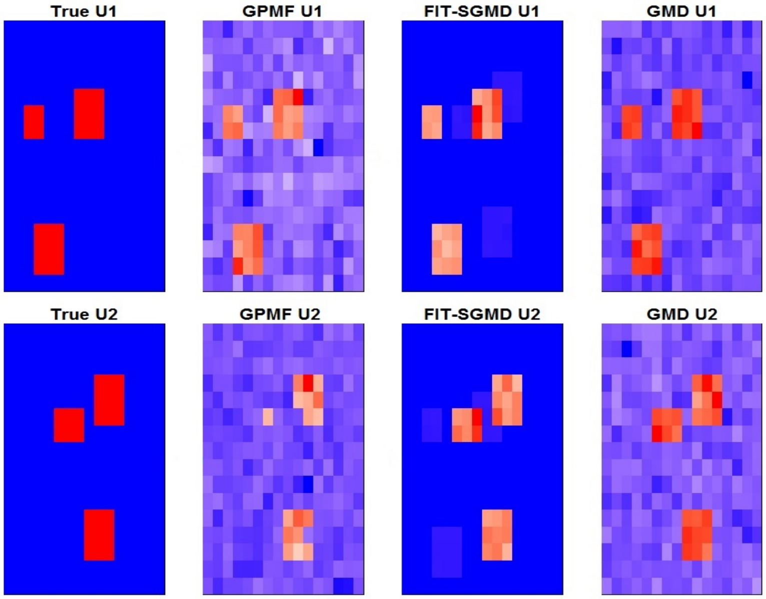


Figure B.2: Results for spatial signal $U = [\mathbf{u}_1, \mathbf{u}_2]$ from spatio-temporal simulation.

and force the rest to be zero. On the other hand, although the GPMF successfully recovers the majority of the blocks, it does not reveal the sparse structures of the two images. Its performance is similar to that of the GMD algorithm. Since the GMD algorithm does not impose any sparsity penalties on U, it is not surprising that almost all of the entries in its output are non-zeros. In Figure B.3, the three algorithms perform equally well in recovering V, except the GPMF for $\mathbf{v}_2$ which is overwhelmed by noise. Among them, the GMD has the smoothest outcomes.

B.1.1.2 Sparse U and V

In this subsection, model (3.1) with sparse U and V are considered. Since the GMD algorithm is not designed to recover sparse signals, we only compare the performances of the FIT-SGMD and GPMF. We set the rank $r = 3$, $n = 200$, $q = 200$

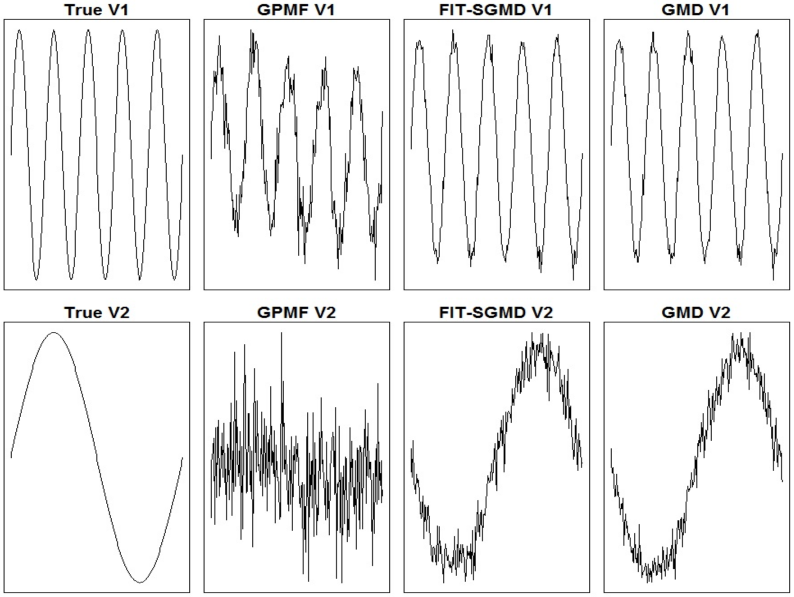


Figure B.3: Results for temporal signal $V = [\mathbf{v}_1, \mathbf{v}_2]$ from spatio-temporal simulation.

and $D = diag(100, 70, 50)$. Let $U = [\mathbf{u}_1|\mathbf{u}_2|\mathbf{u}_3], V = [\mathbf{v}_1|\mathbf{v}_2|\mathbf{v}_3]$:

$$\begin{aligned}
\mathbf{u}_1 &= (1, -1, -1, 1, 1, rep(0, 20), rep(0, 175))', \\
\mathbf{u}_2 &= (rep(0, 5), 1, -1, -1, 1, 1, rep(0, 15), rep(0, 175))', \\
\mathbf{u}_3 &= (rep(0, 10), 1, -1, -1, 1, 1, rep(0, 10), rep(0, 175))', \\
\mathbf{v}_1 &= (1, -1, -1, 1, 1, 1, unif(4, J), rep(0, 15), rep(0, 175))', \\
\mathbf{v}_2 &= (rep(0, 12), unif(4, J), 1, 1, -1, -1, 1, 1, rep(0, 3), rep(0, 175))', \\
\mathbf{v}_3 &= (rep(0, 6), \mathbf{v}_1[7:8], -\mathbf{v}_1[9:10], 1, -1, -\mathbf{v}_2[13:14], \\
&\quad \mathbf{v}_2[15:16], rep(0, 9), rep(0, 175))'.
\end{aligned}$$

where $\mathbf{v}_l[a:b]$ denotes a vector whose entries are the corresponding entries of $\mathbf{v}_l$ from a to b, and $unif(m, J)$ denotes a vector of length m whose entries are iid uniformly distributed on the set of $J = [-1, -0.3] \cup [0.3, 1]$. We choose matrices Ω, Σ satisfy the generalized orthogonal condition $U'\Omega U = V'\Sigma V = I$. As before the SNR is set to be the four levels: 1, 0.5, 0.25 and 0.125.

SNR	Median	Mean
1	9.85E-01	8.86E-01
0.5	6.41E-01	6.42E-01
0.25	4.47E-01	4.97E-01
0.125	2.91E-01	2.56E-01

Table B.2: Median and mean summary for RMSE ratios of FIT-SGMD to GPMF.

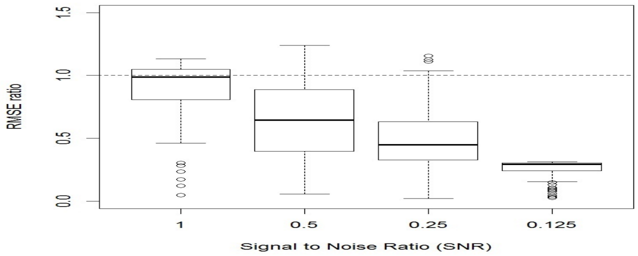


Figure B.4: Boxplots of RMSE ratios of FIT-SGMD to GPMF.

Table B.2 summarizes the mean and median of the RMSE ratios of the FIT-SGMD to the GPMF, and Figure B.4 presents their boxplots. The patterns in Table B.2 and Figure B.4 are similar and consistent with the corresponding results in Section B.1.1.1. As a whole, the FIT-SGMD universally outperforms the GPMF, where the FIT-SGMD enjoys a lower RMSEs for all levels of SNRs. We notice that the RMSE-ratio decreases as the SNR decreases, which indicates that the FIT-SGMD is more accurate in estimating the signal in (3.1) especially when the SNR is low.

In addition, the FIT-SGMD enjoys fast computational speed. The computation costs for $N = 100$ replications are 0.2 system second for the FIT-SGMD and 23.2 system second for the GPMF (Intel(R) Core (TM) i7-3770 CPU @3.40GHz).

B.1.2 The "StarPlus" fMRI Dataset

The StarPlus dataset in Mitchell et al. (2004) is a functional MRI dataset which has the typical transposable structure. In this dataset, images of a brain in the spatial domain are measured over time, where spatial dependence and temporal dependence are often considered (Lindquist, 2008). Following Allen et al. (2013), we select the data for subject number 04847, which consists of 4,698 voxels ($64 \times 64 \times 8$ images) measured for 20 tasks over 54 - 55 time points leading to a total 1098 observations and rearrange the data into a $4,698 \times 1,098$ matrix. We use their choices of Ω, Σ as the unweighted Laplacian structure for the row dependence and a kernel smoother with a window size of ten time points for the column dependence.

Using the FIT-SGMD, the first three $\mathbf{u}_i, i = 1, 2, 3$ of the brain images are presented in Figure B.5 where each $\mathbf{u}_i$ contains eight 64×64 images. Allen et al. (2013) has presented the first three $\mathbf{u}_i$'s by PCA, sparse PCA, GPCA, and found out that the noise overwhelms the PCA and sparse PCA methods but not the GPCA. In Figure B.5, sparse structures are presented where most of the noise is eliminated. Hence, the FIT-SGMD is suitable for this two-way dependent fMRI data, which have comparable results to those given in Allen et al. (2013).

B.2 Proof of Theorem 3.2.1

We divide the proof of Theorem 3.2.1 into two stages. In the first stage, we show that using Algorithm 4 for data matrix Y is equivalent to solving the SVD for the sphered data matrix $\tilde{Y} = \tilde{\Omega} Y \tilde{\Sigma}$ using the orthogonal iteration. In stage 2, we show that the orthogonal iteration for SVD has a general form of the orthogonal iteration for eigen decomposition, where the output of the latter algorithm is proved to converge to the mathematical solution in Golub and Van Loan (1996). Hence, output from Algorithm 4 converges to the mathematical solution of the GMD problem: $U^* = \tilde{\Omega}^{-1}\tilde{U}$, and $V^* = \tilde{\Sigma}^{-1}\tilde{V}$.

Stage 1: Let $\tilde{Y} = \tilde{U}\tilde{D}\tilde{V}'$ be the SVD of $\tilde{Y}$, $\tilde{\Omega}$ and $\tilde{\Sigma}$ denote the square roots of Ω, Σ and $\tilde{\Omega}^{-1}, \tilde{\Sigma}^{-1}$ denote their left matrix inverses, respectively. Since $\tilde{U} = \tilde{\Omega} U$ and $\tilde{V} = \tilde{\Sigma} V$, once the updating steps of U and V in Algorithm 4 are written with respect to $\tilde{Y}$, $\tilde{U}$ and $\tilde{V}$, it follows that:

$$\begin{aligned}
\text{Steps 2-3: } Y_u^{(k)} &= Y\Sigma V^{(k-1)} = (\tilde{\Omega}^{-1}\tilde{Y}\tilde{\Sigma}^{-1})\Sigma(\tilde{\Sigma}^{-1}\tilde{V}^{(k-1)}) = \tilde{\Omega}^{-1}\tilde{Y}\tilde{V}^{(k-1)} \\
&= U^{(k)}R_u^{(k)} = \tilde{\Omega}^{-1}\tilde{U}^{(k)}R_u^{(k)} \\
\text{Steps 4-5: } Y_v^{(k)} &= Y'\Omega U^{(k)} = (\tilde{\Sigma}^{-1}\tilde{Y}'\tilde{\Omega}^{-1})\Omega(\tilde{\Omega}^{-1}\tilde{U}^{(k)}) = \tilde{\Sigma}^{-1}\tilde{Y}'\tilde{U}^{(k)} \\
&= V^{(k)}R_v^{(k)} = \tilde{\Sigma}^{-1}\tilde{V}^{(k)}R_v^{(k)}.
\end{aligned}$$

Hence, $\tilde{\Omega}^{-1}$ and $\tilde{\Sigma}^{-1}$ are canceled from both sides of equations and it follows

$$\begin{aligned}
\tilde{U}^{(k)}R_u^{(k)} &= \tilde{Y}\tilde{V}^{(k-1)}, \\
\tilde{V}^{(k)}R_v^{(k)} &= \tilde{Y}'\tilde{U}^{(k)}.
\end{aligned}$$

From Section 1.2, the above are the two key steps of the orthogonal iteration for matrix $\tilde{Y}$. Hence, Algorithm 4 is equivalent to solving the SVD of $\tilde{Y}$ using the orthogonal iteration.

Stage 2: Suppose $B = UDV$, we show that this SVD can be converted into an eigenvalue problem of a larger matrix H

$$H = \begin{bmatrix} 0 & B' \\ B & 0 \end{bmatrix}.$$

Define

$$Q = \begin{bmatrix} V & V \\ U & -U \end{bmatrix},$$

then it is clear that Q is orthogonal and satisfies $HQ = QD$, i.e.

$$\begin{bmatrix} 0 & B' \\ B & 0 \end{bmatrix} \begin{bmatrix} V & V \\ U & -U \end{bmatrix} = \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & -D \end{bmatrix},$$

namely, its columns are the eigenvectors of H. Expanding the above matrix equation leads to

$$B'U = VD, BV = UD,$$

which are the key step of the orthogonal iteration algorithm. Hence, finding the eigen decomposition of matrix H is equivalent to finding the SVD of matrix B using the orthogonal iteration. Golub and Van Loan (1996, Theorems 7.3.1 and 8.2.2) shows that the orthogonal iteration for eigen decomposition converges to its mathematical solution. Thus, the orthogonal iteration for the SVD also converges to the mathematical solution of the SVD. Combining the above two stages, the convergence of Algorithm 4 is proved.

B.3 Proof of Theorem 3.3.1

Proof We prove the theorem by expanding the (Ω, Σ)-norms, rearranging terms and showing that their right-hand side are identical.

$$\begin{aligned}
\text{Regression:} \quad & ||Y - XUDV'||^2_{I,\Sigma} \\
= \quad & tr\{(Y - XUDV')'(Y - XUDV')\Sigma\} \\
= \quad & tr(Y'Y\Sigma) - tr(Y'XUDV'\Sigma) - tr(\Sigma VDU'X'Y) \\
& +tr(\Sigma VDU'X'XUDV') \\
= \quad & tr(I_m) - 2tr(VDU'X'Y\Sigma) + tr(D^2). \\
\text{GMD:} \quad & ||GX'Y - UDV'||^2_{\Omega,\Sigma} \\
= \quad & tr(Y'XG\Omega GX'Y\Sigma) - tr(Y'XG\Omega UDV'\Sigma) \\
& -tr(\Sigma VDU'\Omega GX'Y) + tr(\Sigma VDU'\Omega UDV').
\end{aligned}$$

Since G is the Moore-Penrose generalized inverse of $\Omega = X'X$, we have $G\Omega G = G$, $(G\Omega)' = \Omega G$ and $(\Omega G)' = G\Omega$. Then it follows,

$$\begin{aligned}
& ||GX'Y - UDV'||^2_{\Omega,\Sigma} \\
= \quad & tr(Y'XG\Omega GX'Y\Sigma) - 2tr(\Sigma VDU'\Omega GX'Y) + tr(\Sigma VDU'\Omega UDV') \\
= \quad & tr(Y'XG\Omega GX'Y\Sigma) - 2tr(VDU'X'Y\Sigma) + tr(D^2).
\end{aligned}$$

It should be noted that if X has full column rank, replacing G by $(X'X)^{-1}$ in (II) returns the same right hand side.

$$\begin{aligned}
\text{CCA:} \quad & ||Y\tilde{V} - XU||^2_{I,I} \\
= \quad & tr(U'X'XU) - 2tr(U'X'Y\tilde{V}) + tr(\tilde{V}'Y'Y\tilde{V}) \\
= \quad & tr(I_m) - 2tr(U'X'Y\Sigma VD) + tr(DV'\Sigma Y'Y\Sigma VD) \\
= \quad & tr(I_m) - 2tr(VDU'X'Y\Sigma) + tr(D^2),
\end{aligned}$$

given $\tilde{V} = \Sigma VD$.

Thus, the solution triplet $(\hat{U}, \hat{V}, \hat{D})$ for the three optimization problems is recognized as $(\hat{U}, \hat{V}, \hat{D}) = \arg\max_{U,V,D} tr\{VDU'X'Y\Sigma - D^2/2\}$. That is, if $(\hat{U}, \hat{V}, \hat{D})$ solves any problem, then it will also solve the remaining two problems.

B.4 The Ω-QR Decomposition

In this section, we introduce the so-called Ω-QR decomposition which is utilized in the generalized orthogonal iteration in Section 3.2.1. Intuitively, the Ω-QR decomposition uses the Ω-norm in place of the Frobenius norm for all the inner products in the original Gram-Schmidt process.

Given a positive definite matrix Ω, the Ω-QR decomposition of a matrix A is a decomposition of A into a general orthogonal matrix Q and a triangular matrix R, i.e. it finds the decomposition of A as

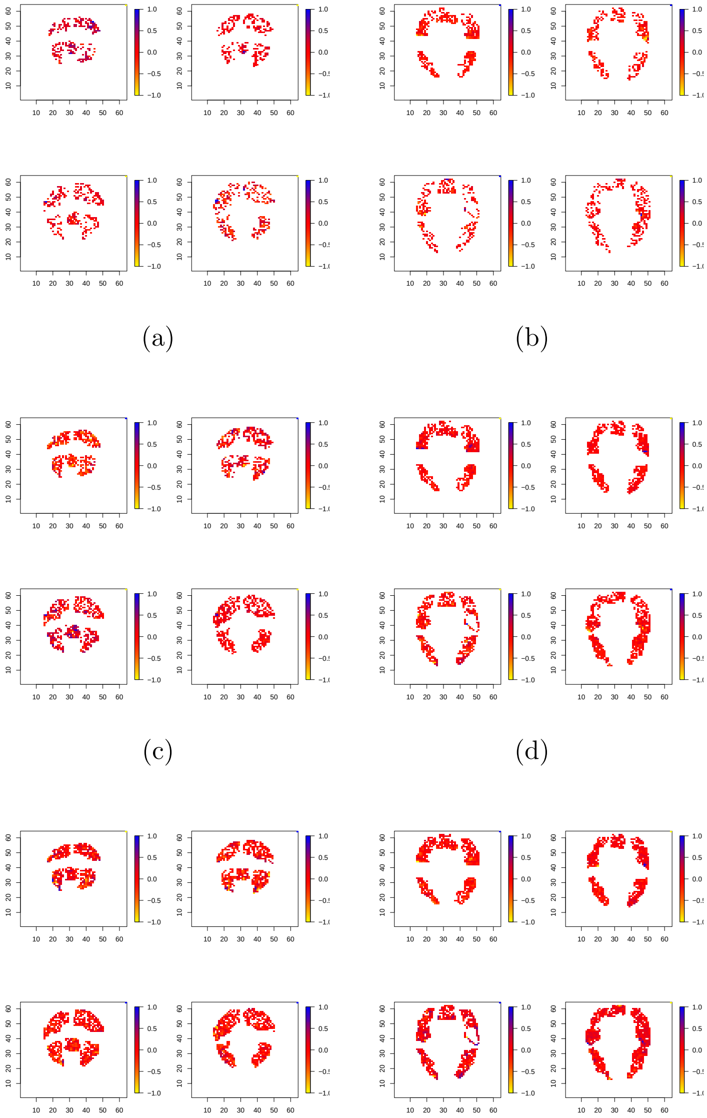
$$A = QR,$$

where $Q'\Omega Q = I$ and R is an upper triangular matrix. Next, we introduce the generalized Gram-Schmidt process, which incorporates the matrix Ω and observes the condition $Q'\Omega Q = I$.

Define $||u||_{\Omega}^2 = u'\Omega u$, $\mathbf{proj}_{\Omega,q} a = \frac{q'\Omega a}{q'\Omega q} q$, $A = [a_1, ..., a_n]$, then,

$$\begin{aligned}
u_1 &= a_1, \quad q_1 = u_1/||u_1||_{\Omega}; \\
u_2 &= a_2 - \mathbf{proj}_{\Omega,q_1} a_2, \quad q_2 = u_2/||u_2||_{\Omega}; \\
u_3 &= a_3 - \mathbf{proj}_{\Omega,q_1} a_3 - \mathbf{proj}_{\Omega,q_2} a_3, \quad q_3 = u_3/||u_3||_{\Omega}; \\
&\vdots \\
u_n &= a_n - \mathbf{proj}_{\Omega,q_1} a_n - \cdots - \mathbf{proj}_{\Omega,q_{n-1}} a_n, \quad q_n = u_n/||u_n||_{\Omega}.
\end{aligned}$$

Hence, the solution of the Ω-QR decomposition is $Q = [q_1, \cdots, q_n]$. (It is trivial to prove that $q_i'\Omega q_i = 1$ and $q_i'\Omega q_j = 0$ if $i \neq j$, i.e. $Q'\Omega Q = I$.)



(a) (b)

(c) (d)

(e) (f)

Figure B.5: Eight slides of the brain images for the first three GMD factors of the Starplus data for FIT-SGMD. $\mathbf{u}_1$: (a)-(b), $\mathbf{u}_2$: (c)-(d); $\mathbf{u}_3$: (e)-(f).