# Selected Topics of High-dimensional Sparse Modeling

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Feng Yi

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Hui Zou and
Yuhong Yang, Advisers

November 2013

# Acknowledgements

First of all, I would like to thank my advisor Professor Hui Zou for his perceptive guidance and continuous support throughout my graduate study and research. I benefit so much from his deep insight into statistics, his constant encouragement and his help over my career development. I feel very lucky to have Zou as my thesis advisor. At the same time, I am deeply grateful to Professor Yuhong Yang for his invaluable advice and helpful comments as my co-advisor.

I wish to thank other faculty members in the statistics department for their excellent lectures, especially Professor Lan Wang, from who I took three theoretical courses. I am also very grateful to Professor Charles Geyer for teaching me advanced statistical computing and helping me solve many programming problems. My sincere thanks also go to Professor Snigdhansu Chatterjee and Professor Wei Pan (Division of Biostatistics, School of Public Health) for joining my oral committee and providing many suggestions on my thesis writings.

I would like to extend my thanks to Qing Mai, Yi Yang, Lingzhou Xue, Ying Lu, Benjamin Sherwood, Gang Cheng, Changqing Ye, Mark Holland and many other friends from School of Statistics for their sincere help, advice and stimulating discussions in the last five years.

Finally, I want to thank my girlfriend Pingyan Lei for her consistent support and patience, walking with me through joy and difficult times for the most beautiful four years so far in my life.

# Dedication

This dissertation is dedicated to my parents

*Xiaobin Yi*
*Yingjun Chi*

## Abstract

In this thesis we study three problems over high-dimensional sparse modeling.

We first discuss the problem of high-dimensional covariance matrix estimation. Nowadays, massive high-dimensional data are more and more common in scientific investigations. Here we focus on one type of covariance matrices - bandable covariance matrices in which the dependence structure of variables follows a nature order. Many off-diagonal elements are very small, especially when they are far away from diagonal, which technically makes the covariance matrix very sparse. It has been shown in Cai et al. (2010) that the tapering covariance estimator attains the optimal minimax rates of convergence for estimating large bandable covariance matrices. The estimation risk critically depends on the choice of tapering parameter. We develop a Steins Unbiased Risk Estimation (SURE) theory for estimating the Frobenius risk of the tapering estimator. SURE tuning selects the minimizer of SURE curve as the chosen tapering parameter. Covariance matrix is finally estimated according to the selected tapering parameter in the tapering covariance estimator.

The second part of the thesis is about high-dimensional varying-coefficient model. Varying-coefficient model is used when the effects of some variables depend on the values of other variables. One interesting and useful varying-coefficient model is that the coefficients of all variables are changing over time. Non-parametric method based on B-splines is used to estimate marginal coefficient of each variable, and varing-coefficient Independence Screening (VIS) is proposed to screen important variables. To improve the performance of the algorithm, Iterative VIS (IVIS) procedure is proposed.

In the third part of the thesis, we study a high-dimensional extension of traditional

factor analysis by relaxing the independence assumption of the error term. In the new model, we assume that the inverse covariance is sparse but not necessarily diagonal. We propose a generalized E-M algorithm to fit the extended factor analysis model. Our new model not only makes factor analysis more flexible, but also could be used to discover the hidden conditional structure of variables after common factors are discovered and removed.

A summary of the thesis is given in Chapter 5.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

High-dimensional data analysis has becomed a very hot topic in statistics in the last 10 years or so. It's easier and cheaper to collect massive amount of high-dimensional data due to advancement of modern technology. Statisticians keep working on new methodologies for high-dimensional data analysis. In many applications, there are hundreds or even thousands of features available from which information could be discovered, but in many situations most of the features are not very useful, or even completely irrelevant. In sparse modeling, we try to figure out a small number of features without losing important information for analyzing high-dimensional data. In different problems, distinct methods have been used to implement the idea of sparse modeling, including sparse regression, Tibshirani (1996), Fan and Li (2001) and Zou and Hastie (2005); sparse classification, Zhu et al. (2004) and Chan et al. (2007); sparse graphical model selection, Ravikumar et al. (2008), Meinshausen and Buehlmann (2006) and Bani Asadi et al. (2009); and sparse dimensionality reduction, Zou et al. (2006), d'Aspremont et al. (2007), Hoyer (2004) and Kim and Park (2007).

Here in this thesis, we study three different sparse modeling problems:

1. Large covariance matrix estimation,

2. Varying-coefficient regression model,

3. Extended factor analysis.

In Chapter 2, we discuss covariance matrix estimation when dimension is very large. When there are $p$ variables, there are about $p^2/2$ elements to be estimated. In this thesis we consider a situation that variables are ordered, and the covariance matrix has a bandable structure, which means that the magnitude of matrix elements decays when the elements are further and further away from the diagonal. The tapering estimator is shown to be minimax rate optimal for estimating the bandable covariance matrix in Cai et al. (2010). But their theoretical results assumed the elements have a polynomial rate of decay as they are moving away from the diagonal, and the optimal solution depends on $\alpha$, which specifies the polynomial rate. To make the tapering estimator useful, we develop a new method, named "SURE-tuned Tapering Estimation" to select $\alpha$ according to unbiased evaluation of matrix estimation error, based on the idea in Steins unbiased risk estimation (SURE) theory Stein (1981), Efron (1986) and Efron (2004). And we demonstrate through simulations that SURE-tuned tapering estimate performs competitively with oracle estimate.

In Chapter 3, we study high-dimensional varying-coefficient regression model. We consider the coefficients of variables are time varying. Non-parametric methods based on B-spline are used to estimate the marginal effects. Then we use independence screening to filter out majority of variables with low marginal effects with response. Afterwards, we use group variable selection methods (e.g. group SCAD Wang et al. (2007)) to select and estimate time-varying effects. To overcome the limitation of marginal screening in the first step, iterative varying-coefficient independence screening (IVIS) is proposed, in which screening and group variable selection is implemented iteratively until the variable selection is stabilized. Three simulations models are used to show the variable selection and estimation power of IVIS.

In unsupervised learning, sparse modeling is also very powerful. In Chapter 4, we extend the traditional factor analysis. In our new model, the factors are still

unobserved, but the errors are allowed to be correlated. We argue that traditional factor model is powerful sometimes but very restrictive in other cases. There could be interesting conditional structure of variables after common factors are discovered. We proposed a generalized version of EM algorithm with graphical-lasso algorithm incorporated to estimate extended factor analysis model. Our numerical results show that our method performs much better when error structure is complicated, and is comparable to traditional factor analysis when diagonal error structure is valid.

A summary of the thesis is given in Chapter 5.

# Chapter 2

# SURE-tuned Tapering Estimation for Large Covariance Matrices

## 2.1 Introduction

Suppose we observe independent and identically distributed $p$-dimensional random variables $X_1, \ldots, X_n$ with covariance matrix $\Sigma_{p\times p}$. The usual sample covariance matrix is an excellent estimator for $\Sigma_{p\times p}$ in the conventional setting where $p$ is small and fixed and the sample size $n$ diverges to infinity. Nowadays, massive high-dimensional data are more and more common in scientific investigations, such as imaging, web mining, microarrays, risk management, spatial and temporal data, and so on. In high-dimensional settings, the sample covariance matrix performs very poorly; see Johnstone (2001) and references therein. To overcome the difficulty imposed by high dimensions, many regularized estimates of large covariance matrices have been proposed in the recent literature. These regularization methods include Cholesky-based penalization Huang et al. (2006); Lam and Fan (2007); Rothman et al. (2010a), thresholding Bickel and Levina (2008a); El Karoui (2008); Rothman et al. (2009), banding Bickel and Levina (2008b); Wu and Pourahmadi (2009), tapering Furrer and Bengtsson (2007); Cai et al. (2010). In particular, the tapering estimator is shown to be minimax rate optimal for estimating the bandable covariance matrices that are

often used to model the dependence structure of variables that follow a nature order Cai et al. (2010); Cai and Zhou (2012). Much of the published theoretical work assumes the data follow a normal distribution, although some have relaxed the normality assumption to a tail probability condition such as sub-Gaussian distribution assumption. Nevertheless, the lower bound results in the minimax estimation theory were actually established for a family of multivariate normal distributions Cai et al. (2010); Cai and Zhou (2012). In this chapter we consider the tapering estimator under the normal distribution assumption.

We begin with some notation and definitions. Let $\|A\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$ denote the Frobenius norm of $A$. Let $\|A\|_q$ denote the $\ell_q$ operator norm of $A$. When $q = 1$, the $\ell_1$ norm is $\max_i \sum_j |a_{ij}|$; when $q = 2$, the $\ell_2$ norm is equal to the largest singular value of $A$. Consider the following parameter spaces:

$$\mathcal{F}_\alpha = \{\Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \leq Mk^{-\alpha} \text{ for all } k, \text{ and } \lambda_{max}(\Sigma) \leq M_0\},$$

$$\mathcal{F}'_\alpha = \{\Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \leq Mk^{-\alpha} \text{ for all } k, \text{ and } \max_i \sigma_{ii} \leq M_0\},$$

where $\alpha, M, M_0$ are positive constants. The parameter $\alpha$ specifies the rate of decay of the off-diagonal elements of $\Sigma$ as they move away from the diagonal. A larger $\alpha$ parameter indicates a higher degree of "sparsity". Thus we can also regard $\alpha$ as a *sparsity index* of the parameter space. Let $\tilde{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T - \bar{X}\bar{X}^T$ be the MLE of $\Sigma$. The tapering estimator Cai et al. (2010) is defined as

$$\breve{\Sigma}^{(k)} = (\breve{\sigma}_{ij}^{(k)})_{1\leq i,j\leq p} = (w_{ij}^{(k)}\tilde{\sigma}_{ij})_{1\leq i,j\leq p},$$

where, for a tapering parameter $k$,

$$w_{ij}^{(k)} = \begin{cases} 1, & \text{when } |i-j| \leq k/2 \\ 2 - \frac{|i-j|}{k/2}, & \text{when } k/2 < |i-j| < k \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

Tapering is a generalization of banding where $\hat{\sigma}_{ij}^{B(k)} = I(|i-j| \leq k)\tilde{\sigma}_{ij}$. We assume $p \geq n$ and $\log(p) = o(n)$ in the sequel. We cite the following results Cai et al. (2010); Cai and Zhou (2012):

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_\alpha} p^{-1}\mathbb{E}\|\hat{\Sigma} - \Sigma\|_F^2 \asymp n^{-(2\alpha+1)/(2\alpha+2)}, \tag{2.2}$$

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_\alpha} \mathbb{E}\|\hat{\Sigma} - \Sigma\|_2^2 \asymp n^{-2\alpha/(2\alpha+1)} + \frac{\log(p)}{n}, \tag{2.3}$$

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}'_\alpha} \mathbb{E}\|\hat{\Sigma} - \Sigma\|_1^2 \asymp n^{-\alpha/(\alpha+1)} + \frac{\log(p)}{n}, \tag{2.4}$$

where $a_n \asymp b_n$ if there are positive constants $c_1$ and $c_2$ independent of $n$ such that $c_1 \leq a_n/b_n \leq c_2$. Furthermore, define three tapering parameters as following

$$k_F = n^{1/(2\alpha+2)}, \quad k_2 = n^{1/(2\alpha+1)} \tag{2.5}$$
$$k_1 = \min\{n^{1/(2\alpha+2)}, (n/\log(p))^{1/(2\alpha+1)}\}.$$

Then the tapering estimator with $k = k_F$, $k = k_2$ and $k = k_1$ attains the minimax bound in (2.2), (2.3) and (2.4), respectively.

The minimax rate optimal choices of $k$ shed light on the importance of choosing the right tapering parameter. However, there are at least two difficulties in using the minimax theory to construct the tapering parameter. First, the minimax tapering estimators depend on $\alpha$. If $\alpha$ is unknown, which is often the case in reality, then

the minimax optimal tapering "estimators" are not real estimators. Second, the minimax rate optimal tapering estimators can be conservative for estimating some covariance matrices. For instance, assume that the data are generated from a normal distribution with a $MA(1)$ covariance where $\sigma_{ij} = I(i = j) + 0.5I(|i - j| = 1)$. Although this covariance matrix is in $\mathcal{F}_\alpha$ for $\alpha > 0$, the optimal $k$ should be 2 no matter which matrix norm is used. Therefore, it is desirable to have a reliable data-driven method to choose the tapering parameter. Tuning is usually done by first constructing an estimate of the risk for each $k$ and then picking the minimizer of the estimated risk curve. Cross-validation and Bootstrap are the popular nonparametric techniques for that purpose. Bickel and Levina (2008a,b) discussed the use of two-fold cross-validation for selecting the banding parameter of the banding estimator. They claimed that although cross-validation estimates the risk very poorly, it can still select the banding parameter quite well.

Here we suggest a different tuning method by borrowing the idea in Stein's unbiased risk estimation (SURE) theory Stein (1981), Efron (1986, 2004). Compared with cross-validation, the SURE approach is computationally less expensive and provides a much better estimate of the Frobenius risk. The explicit form of SURE formula is derived in Section 2.2. Here we demonstrate the effectiveness of SURE tuning in Figure 2.1 where we compare the true Frobenius risk curve (as a function of $k$) and the SURE curves. We generated the data from the simulation model used in Cai et al. (2010). Two $\alpha$ values were used: $\alpha = 0.1$ corresponds to a dense covariance model and $\alpha = 0.5$ corresponds to a sparse covariance model. Figure 2.1 clearly shows three important points. First, the average of 100 SURE curves is virtually identical to the Frobenius risk curve, which agrees with the SURE theory as shown in Section 2.2. Second, the minimizer of each SURE curve is very close to the minimizer of the true risk curve. Third, the minimizer of each cross-validation curve is also close to the minimizer of the true risk curve, but the cross-validation estimator of the Frobenius

risk is way too large. The true risk is within $[100, 500]$ while the cross-validation risk is within $[5000, 5500]$. In practice we not only want to select a good model but also want to understand how well the model performs. Efron (2004) did a careful comparison between SURE and cross-validation and concluded that with minimal modeling SURE can significantly outperform cross-validation. Figure 2.1 suggests that Efron's conclusion continues to hold in the covariance matrix estimation problem.

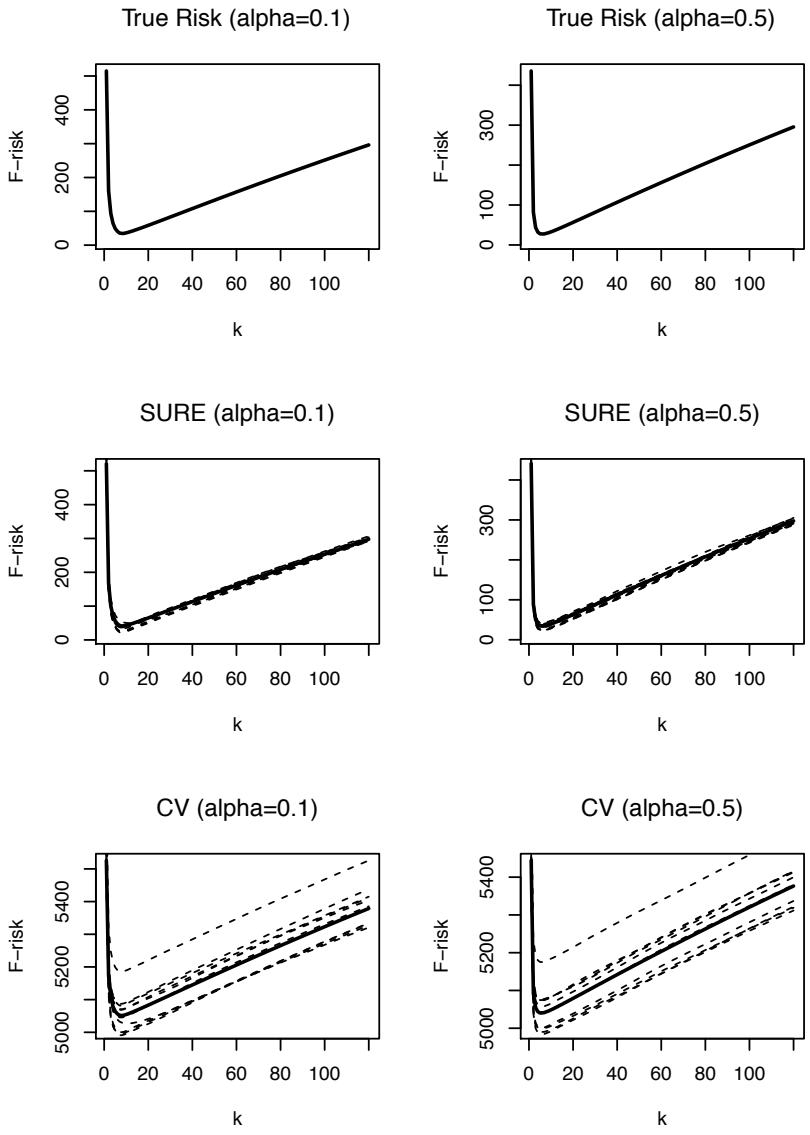



Figure 2.1: Comparing the true risk curve, the SURE curve and the CV curve under the Frobenius norm. The data are generated from the simulation model 1 in Section 2.3 with $n = 250, p = 500, \alpha = 0.1$ and 0.5. In the second row we plot 10 SURE curves (dashed lines) and the average of 100 SURE curves (the solid line). Similar plots are shown in the third row for cross-validation.

## 2.2 Stein's Unbiased Risk Estimation in Covariance Matrix Estimation

In this section we develop a SURE theory for estimating the Frobenius risk of a weighted MLE, denoted by $\widehat{\Sigma}^{(k)}$, which has the expression $\widehat{\Sigma}^{(k)}_{ij} = w^{(k)}_{i,j}\tilde{\sigma}_{ij}$ where $w^{(k)}_{i,j}$ only depends on $i, j, k$. The tapering and banding estimators are special examples of the weighted MLE. Tapering weights are defined in (2.1). The banding estimator Bickel and Levina (2008b) uses simpler weights $w^{(k)}_{i,j} = I(|i-j| \leq k)$.

The basic idea in SURE can be traced back to the James-Stein estimator of multivariate normal mean. Efron (1986, 2004) studied the use of SURE in estimating prediction error and he named it covariance penalty method. Shen and Ye (2002) applied the covariance penalty idea to perform adaptive model selection. Donoho and Johnstone (1995) developed SureShrink for adaptive wavelet thresholding. Efron et al. (2004) and Zou and Hastie (2007) applied SURE to Lasso model selection.

### 2.2.1 SURE identity

For an arbitrary estimator $\widehat{\Sigma}$ of the covariance matrix, the Frobenius risk ($\mathbb{E}\|\widehat{\Sigma}-\Sigma\|_F^2$) is equivalent to the squared $\ell_2$ risk for estimating the vector $(\sigma_{11}, \ldots, \sigma_{1p}, \ldots, \sigma_{p1}, \ldots, \sigma_{pp})^T$. As the first step of SURE, we derive a covariance penalty identity for the matrix Frobenius risk of an arbitrary estimator of $\Sigma$.

**Lemma 2.2.1**

Let $\tilde{\Sigma}^s = \frac{n}{n-1}\tilde{\Sigma}$ be the usual sample covariance matrix. For an arbitrary estimator of $\Sigma$, denoted by $\widehat{\Sigma} = (\hat{\sigma}_{ij})$, its Frobenius risk can be written as

$$\mathbb{E}\|\widehat{\Sigma} - \Sigma\|_F^2 = \mathbb{E}\|\widehat{\Sigma} - \tilde{\Sigma}^s\|_F^2 - \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Var}(\tilde{\sigma}^s_{ij}) + 2\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}(\hat{\sigma}_{ij}, \tilde{\sigma}^s_{ij}). \quad (2.6)$$

□

The second term in the right hand of (2.6) is the same for all estimators of $\Sigma$. Thus, if we only care of comparing the Frobenius risk of different estimators, the second term can be dropped and we can write

$$\begin{aligned} PR(\widehat{\Sigma}) &= \mathbb{E}\|\widehat{\Sigma} - \tilde{\Sigma}^s\|_F^2 + 2\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}(\hat{\sigma}_{ij}, \tilde{\sigma}_{ij}^s) \\ &= \text{Apparent error} + \text{Optimism}, \end{aligned} \tag{2.7}$$

where $PR$ stands for prediction risk and we have borrowed Efron's terminology 'apparent error' and 'optimism' Efron (2004). The optimism is expressed by a covariance penalty term. Since $\|\widehat{\Sigma} - \tilde{\Sigma}^s\|_F^2$ is an automatic unbiased estimate of the apparent error, it suffices to construct a good estimate of the optimism in order to estimate $PR$.

For the weighted MLE, we observe that $\mathrm{Cov}(\hat{\sigma}_{ij}^{(k)}, \tilde{\sigma}_{ij}^s) = w_{ij}^{(k)}\frac{n-1}{n}\mathrm{Var}(\tilde{\sigma}_{ij}^s)$. The next lemma provides a nice unbiased estimator of $\mathrm{Var}(\tilde{\sigma}_{ij}^s)$.

**Lemma 2.2.2**

If $\{X_i\}_{i=1}^n$ is a random sample from $N(\mu, \Sigma)$, then

$$\mathrm{Var}(\tilde{\sigma}_{ij}^s) = \frac{\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}}{n-1}, \tag{2.8}$$

and an unbiased estimate of $\mathrm{Var}(\tilde{\sigma}_{ij}^s)$ is given by $\widehat{\mathrm{Var}}(\tilde{\sigma}_{ij}^s)$ which equals

$$\frac{n^2(n^2-n-4)}{(n-1)^2(n^3+n^2-2n-4)}\tilde{\sigma}_{ij}^2 + \frac{n^3}{(n-1)(n^3+n^2-2n-4)}\tilde{\sigma}_{ii}\tilde{\sigma}_{jj}. \tag{2.9}$$

□

From (2.8) we see the MLE for $\mathrm{Var}(\tilde{\sigma}_{ij}^s)$ is $\frac{\tilde{\sigma}_{ij}^2 + \tilde{\sigma}_{ii}\tilde{\sigma}_{jj}}{n-1}$, which is almost identical to the unbiased estimator in (2.9). We prefer to use an exact unbiased estimate of the optimism. In addition,the unbiased estimator in (2.9) is the UMVUE of $\mathrm{Var}(\tilde{\sigma}_{ij}^s)$.

Lemma 2.2.2 shows that an unbiased estimator for $PR(\widehat{\Sigma}^{(k)})$ is given by

$$\widehat{PR}(k) = \|\widehat{\Sigma}^{(k)} - \tilde{\Sigma}^s\|_F^2 + \sum_{1\le i,j\le p} (2w_{ij}^{(k)}\frac{n-1}{n})\widehat{\mathrm{Var}}(\tilde{\sigma}_{ij}^s). \tag{2.10}$$

Similarly, an unbiased estimator for $\mathbb{E}\|\widehat{\Sigma}^{(k)} - \Sigma\|_F^2$ is given by

$$\begin{aligned} SURE(k) &= \|\widehat{\Sigma}^{(k)} - \tilde{\Sigma}^s\|_F^2 + \sum_{1\le i,j\le p} (2w_{ij}^{(k)}\frac{n-1}{n} - 1)\widehat{\mathrm{Var}}(\tilde{\sigma}_{ij}^s) \\ &= \sum_{1\le i,j\le p} (\frac{n}{n-1} - w_{ij}^{(k)})^2\tilde{\sigma}_{ij}^2 + \sum_{1\le i,j\le p} (2w_{ij}^{(k)} - \frac{n}{n-1})(a_n\tilde{\sigma}_{ij}^2 + b_n\tilde{\sigma}_{ii}\tilde{\sigma}_{jj}) \end{aligned} \tag{2.11}$$

with $a_n = \frac{n(n^2-n-4)}{(n-1)(n^3+n^2-2n-4)}$ and $b_n = \frac{n^2}{n^3+n^2-2n-4}$.

### 2.2.2 SURE tuning

Once the tapering estimator is constructed, the SURE formula automatically provides a good estimate of its Frobenius risk. Naturally we use $\hat{k}^{sure}$ as the tapering parameter under the Frobenius norm where

$$\hat{k}^{sure} = \arg\min_k SURE(k). \tag{2.12}$$

Unfortunately we do not have a direct SURE formula for the matrix $\ell_q$ norm, $q = 1, 2$. We suggest using $\hat{k}^{sure}$ as the tapering parameter for both $\ell_1$ and $\ell_2$ norm as well. We list several good reasons for using this selection strategy.

1. One can expect the optimal tapering parameter should be the same under different matrix norm if the underlying covariance matrix is an exactly banded matrix, i.e., there is a constant $k_0$ such that $\sigma_{ij} = 0$ whenever $|i-j| > k_0$. Hence, it is reasonable to expect that the optimal choices of tapering param-

eter under the Frobenius norm and the matrix $\ell_1, \ell_2$ norms stay close if the underlying covariance model is very sparse.

2. Cai and Zhou (2012) showed that as long as $\log(p) \leq n^{1/(2\alpha+2)}$, the minimax optimal tapering parameters under the $\ell_1$ norm and the Frobenius norm are the same. This can be easily seen from (2.5).

3. The $\ell_2$ norm is the most popular matrix operator norm. We argue that minimizing the Frobenius norm leads to a good estimator, although may not be the best, under the $\ell_2$ norm. From Cai et al. (2010) we know that

$$\sup_{\mathcal{F}_\alpha} \mathbb{E}\|\breve{\Sigma}^{(k)} - \Sigma\|_2^2 \;\; \leq \;\; C\left[k^{-2\alpha} + \frac{k+\log(p)}{n}\right] \equiv C \cdot R_2(k),$$

Letting $k = k_F = n^{1/(2\alpha+2)}$ yields that

$$R_2(k_F) = O(n^{-\alpha/(\alpha+1)} + \log(p)/n).$$

Compare the rate to the minimax optimal rate $n^{-2\alpha/(2\alpha+1)} + \log(p)/n$.

4. As shown in simulation, SURE selection is very stable, although it is biased under the $\ell_1, \ell_2$ norms. Selection stability is a very important concern in model selection Breiman (1996). In contrast, even the oracle tuning under the $\ell_1, \ell_2$ norms can show very high variability when the underlying covariance matrix is not very sparse.

## 2.3 Monte Carlo Study

In this section we conduct extensive simulation to compare SURE tuning with cross-validation and oracle tuning.

### 2.3.1 Models and tuning methods

The data are generated from $N(0, \Sigma)$. Three covariance models are considered.

**Model 2.1.** This model is adopted from Cai et al. (2010). The covariance matrix has the form

$$\sigma_{ij} = \begin{cases} 1, & 1 \leq i = j \leq p \\ \rho|i-j|^{-(\alpha+1)}. & 1 \leq i \neq j \leq p \end{cases}$$

We let $\rho = 0.6$, $\alpha = 0.1, 0.5$, $n = 250$ and $p = 250, 500, 1000$.

**Model 2.2.** The covariance matrix has the form $\sigma_{ij} = \rho^{|i-j|}$, $1 \leq i, j \leq p$. We let $\rho = 0.95, 0.5$, $n = 250$ and $p = 250, 500, 1000$. This is a commonly used autoregressive covariance matrix for modeling spatial-temporal dependence.

**Model 2.3.** This simulation model is a truncated version of model 2.1. The covariance matrix has the form

$$\sigma_{ij} = \begin{cases} 1, & 1 \leq i = j \leq p \\ \rho|i-j|^{-(\alpha+1)} I(|i-j| \leq 6). & 1 \leq i \neq j \leq p \end{cases}$$

We let $\rho = 0.6$, $\alpha = 0.1, 0.5$, $n = 250$ and $p = 250, 500, 1000$. Model 2.3 represents an exactly banded covariance matrix. It is the sparest among all three simulation models.

**Model 2.4.** This model is a modification of Model 2.1 - allows negative correlation. Specifically

$$\sigma_{ij} = \begin{cases} 1, & 1 \leq i = j \leq p \\ \rho|i-j|^{-(\alpha+1)}(-1)^{|i-j|}. & 1 \leq i \neq j \leq p \end{cases}$$

The parameters are the same as Model 2.1.

**Model 2.5.** With negative correlation, $\sigma_{ij}$ has the form of $\sigma_{ij} = \rho^{|i-j|}(-1)^{|i-j|}$, $1 \le i, j \le p$. Everything else follows from Model 2.2.

**Model 2.6.** Similarly, negative correlation is introduced into Model 2.3, then we have

$$\sigma_{ij} = \begin{cases} 1, & 1 \le i = j \le p \\ \rho|i-j|^{-(\alpha+1)} I(|i-j| \le 6)(-1)^{|i-j|}. & 1 \le i \ne j \le p \end{cases}$$

It has identical parameterization as Model 2.3.

For each covariance model, the theoretical optimal tapering parameters are defined as $k_a^{opt} = \arg\min_k \mathbb{E}\|\breve{\Sigma}^{(k)} - \Sigma\|_a^2$, where $a = F, 1, 2$. In our simulation study the risk curves can be computed numerically, and thus we can find the numerical values of $k_a^{opt}$ for $a = F, 1, 2$.

We considered three tuning techniques in the simulation study: SURE, cross-validation and oracle tuning. The oracle tuning is defined as

$$\hat{k}_a^{oracle} = \arg\min_k \|\breve{\Sigma}^{(k)} - \Sigma\|_a^2$$

where $a = F, 1, 2$. The idea of oracle tuning is intuitive. Suppose that we could use an independent validation data set of size $m$ ($m \ge n$) for tuning. The chosen $k$ is then found by comparing $\widehat{\Sigma}^{(k)}$ and $\tilde{\Sigma}_m$ under a given matrix norm, where $\tilde{\Sigma}_m$ is the MLE of $\Sigma$ using the independent validation set. Now imagine $m$ could be as large as we wish. The oracle tuning is basically the independent-validation-set tuning with infinitely many data. The oracle tuning is not realistic but serves as a golden benchmark to check the performance of practical tuning methods.

Cross-validation is a commonly-used practical tuning method. Randomly split the training data in to $V$ parts. For $v = 1, \ldots, V$, we leave observations in the $v$th part as validation data and compute a MLE of $\Sigma$, denoted by $\tilde{\Sigma}_v$. Let $\breve{\Sigma}_{-v}^{(k)}$ denote the tapering estimator computed on the rest $V-1$ parts. Then the cross-validation choices of $k$ under the Frobenius norm and the matrix $\ell_1, \ell_2$ norm are defined as $\hat{k}_a^{cv} = \arg\min_k \frac{1}{V}\sum_{v=1}^{V} \|\breve{\Sigma}_{-v}^{(k)} - \tilde{\Sigma}_v\|_a^2$ where $a = F, 1, 2$, denoting the Frobenius, $\ell_1$, $\ell_2$ norms. Five-fold cross-validation was used in our simulation.

We also considered an unconventional cross-validation called cv-F that always uses Frobenius-norm for tuning even when the $\ell_1$ or $\ell_2$ norm is used to evaluate the risk of the tapering estimator. Note that cv-F is a direct analogue of SURE tuning.

### 2.3.2 Results and conclusions

For each model we compared the chosen tapering parameters by oracle, SURE and cross-validation to the optimal tapering parameter and compared the estimation risk of the three tuned tapering covariance estimators. Table 2.1 – Table 2.12 summarize the simulation results. We have the following remarks.

1. Under the Frobenius norm, SURE works as well as the oracle tuning. Cross-validation is slightly worse than SURE.

2. Under the $\ell_1$ norm, SURE is very close to the oracle tuning. Cross-validation is the worst in all cases.

3. The story under the $\ell_2$ norm case is more intriguing. When the covariance matrix is sparse, which corresponds to model 1 with $\alpha = 0.5$, model 2.2 with $\rho = 0.5$ and model 2.3, SURE is very close to the oracle tuning. But when the covariance matrix is dense, which corresponds to model 1 with $\alpha = 0.1$ and model 2.2 with $\rho = 0.95$, SURE is significantly worse than the oracle tuning.

| Model 2.1: Tapering parameter selection | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $k^{opt}$ | | | $\hat{k}^{oracle}$ | | | $\hat{k}^{sure}$ | $\hat{k}^{cv}$ | | |
| $p$ | $\alpha$ | F | $\ell_1$ | $\ell_2$ | F | $\ell_1$ | $\ell_2$ | F,$\ell_1$,$\ell_2$ | F | $\ell_1$ | $\ell_2$ |
| 250 | 0.1 | 11 | 9 | 30 | 10.70 | 10.46 | 36.29 | 10.63 | 9.66 | 18.34 | 48.97 |
| | | | | | (0.56) | (3.03) | (8.52) | (1.18) | (1.02) | (9.50) | (27.15) |
| 250 | 0.5 | 6 | 5 | 9 | 5.99 | 5.88 | 10.56 | 6.15 | 5.46 | 10.28 | 20.41 |
| | | | | | (0.41) | (1.60) | (2.21) | (0.73) | (0.67) | (6.24) | (11.8) |
| 500 | 0.1 | 11 | 9 | 39 | 10.83 | 9.96 | 44.57 | 10.52 | 9.35 | 19.75 | 50.56 |
| | | | | | (0.43) | (2.60) | (8.37) | (0.88) | (0.73) | (10.40) | (23.76) |
| 500 | 0.5 | 6 | 5 | 10 | 6.04 | 5.52 | 10.64 | 6.11 | 5.29 | 12.08 | 21.08 |
| | | | | | (0.28) | (1.72) | (2.02) | (0.60) | (0.46) | (5.48) | (11.30) |
| 1000 | 0.1 | 11 | 9 | 51 | 10.92 | 9.60 | 55.91 | 10.65 | 9.22 | 18.67 | 70.68 |
| | | | | | (0.31) | (2.37) | (8.02) | (0.64) | (0.54) | (10.09) | (29.88) |
| 1000 | 0.5 | 6 | 5 | 10 | 6.00 | 5.24 | 11.03 | 6.14 | 5.17 | 10.74 | 28.25 |
| | | | | | (0.14) | (1.45) | (1.83) | (0.47) | (0.38) | (5.67) | (14.88) |

Table 2.1: Simulation model 2.1: tapering parameter selection. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

On the other hand, the oracle tuning is not a true tuning method. Comparing SURE and cross-validation, we see that SURE performs much better than cross-validation, except for model 2.1 with $\alpha = 0.1$ which corresponds to the densest covariance matrix in the simulation study.

4. We can understand the failure of cross-validation under the $\ell_1, \ell_2$ norms by looking at its selection variability, as evident in Table 2.1, Table 2.3, Table 2.5, Table 2.7, Table 2.9 and Table 2.11. Even the oracle tuning exhibits high variability when the covariance matrix is dense. Cross-validation has even higher variability.

| Model 2.1: Estimation risk | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $p$ | $\alpha$ | Oracle | | SURE | | CV | | CV-F | |
| Frobenius | 250 | 0.1 | 26.04 | (0.11) | 26.23 | (0.11) | 26.30 | (0.10) | 26.30 | (0.10) |
| Norm | 250 | 0.5 | 13.63 | (0.07) | 13.77 | (0.07) | 13.83 | (0.07) | 13.83 | (0.07) |
| | 500 | 0.1 | 53.33 | (0.14) | 53.54 | (0.14) | 53.82 | (0.14) | 53.82 | (0.14) |
| | 500 | 0.5 | 27.48 | (0.11) | 27.65 | (0.11) | 27.87 | (0.11) | 27.87 | (0.11) |
| | 1000 | 0.1 | 108.11 | (0.21) | 108.29 | (0.22) | 109.15 | (0.21) | 109.15 | (0.21) |
| | 1000 | 0.5 | 55.03 | (0.14) | 55.25 | (0.14) | 55.04 | (0.15) | 55.04 | (0.15) |
| $\ell_1$ Norm | 250 | 0.1 | 14.17 | (0.12) | 14.78 | (0.15) | 17.84 | (0.50) | 14.78 | (0.15) |
| | 250 | 0.5 | 3.67 | (0.05) | 3.87 | (0.06) | 5.22 | (0.34) | 3.86 | (0.05) |
| | 500 | 0.1 | 18.94 | (0.14) | 19.58 | (0.17) | 24.20 | (0.71) | 19.51 | (0.15) |
| | 500 | 0.5 | 4.22 | (0.04) | 4.43 | (0.06) | 5.62 | (0.22) | 4.40 | (0.05) |
| | 1000 | 0.1 | 24.08 | (0.13) | 24.88 | (0.17) | 29.85 | (0.88) | 24.73 | (0.16) |
| | 1000 | 0.5 | 4.64 | (0.04) | 4.87 | (0.05) | 6.49 | (0.24) | 4.78 | (0.04) |
| $\ell_2$ Norm | 250 | 0.1 | 2.96 | (0.05) | 5.35 | (0.07) | 4.29 | (0.16) | 5.71 | (0.07) |
| | 250 | 0.5 | 0.88 | (0.01) | 1.09 | (0.02) | 1.48 | (0.08) | 1.19 | (0.02) |
| | 500 | 0.1 | 4.26 | (0.05) | 7.87 | (0.07) | 5.27 | (0.16) | 8.45 | (0.06) |
| | 500 | 0.5 | 0.99 | (0.01) | 1.23 | (0.01) | 1.59 | (0.07) | 1.37 | (0.01) |
| | 1000 | 0.1 | 5.82 | (0.05) | 10.56 | (0.06) | 7.36 | (0.19) | 11.40 | (0.05) |
| | 1000 | 0.5 | 1.08 | (0.01) | 1.33 | (0.01) | 2.09 | (0.10) | 1.52 | (0.01) |

Table 2.2: Simulation model 2.1: Frobenius, $\ell_1$ $\ell_2$ risk. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

| Model 2.2: Tapering parameter selection | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $k^{opt}$ | | | $\hat{k}^{oracle}$ | | | $\hat{k}^{sure}$ | $\hat{k}^{cv}$ | | |
| $p$ | $\rho$ | F | $\ell_1$ | $\ell_2$ | F | $\ell_1$ | $\ell_2$ | F,$\ell_1$,$\ell_2$ | F | $\ell_1$ | $\ell_2$ |
| 250 | 0.95 | 71 | 71 | 76 | 70.79 | 72.84 | 77.36 | 71.23 | 68.64 | 80.07 | 88.24 |
| | | | | | (4.53) | (11.93) | (17.32) | (12.45) | (12.92) | (28.30) | (33.14) |
| 250 | 0.50 | 5 | 5 | 5 | 5.00 | 4.84 | 5.13 | 5.03 | 5.00 | 7.87 | 13.18 |
| | | | | | (0.00) | (0.93) | (1.02) | (0.17) | (0.00) | (6.09) | (11.93) |
| 500 | 0.95 | 70 | 68 | 69 | 70.10 | 69.50 | 72.51 | 70.76 | 68.04 | 88.77 | 107.52 |
| | | | | | (3.08) | (12.17) | (17.00) | (6.14) | (6.41) | (30.46) | (33.82) |
| 500 | 0.50 | 5 | 5 | 5 | 5.00 | 4.89 | 5.17 | 5.00 | 5.00 | 8.60 | 16.68 |
| | | | | | (0.00) | (0.90) | (1.00) | (0.00) | (0.00) | (4.55) | (15.84) |
| 1000 | 0.95 | 69 | 67 | 71 | 69.71 | 69.83 | 73.83 | 70.66 | 67.48 | 92.29 | 117.41 |
| | | | | | (2.16) | (11.95) | (11.68) | (3.86) | (3.83) | (30.56) | (33.84) |
| 1000 | 0.50 | 5 | 5 | 5 | 5.00 | 4.73 | 5.00 | 5.00 | 5.00 | 8.85 | 21.08 |
| | | | | | (0.00) | (0.93) | (0.94) | (0.00) | (0.00) | (6.04) | (20.90) |

Table 2.3: Simulation model 2.2: tapering parameter selection. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

| Model 2.2: Estimation risk | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $p$ | $\rho$ | Oracle | | SURE | | CV | | CV-F | |
| Frobenius | 250 | 0.95 | 118.09 | (2.66) | 125.00 | (2.88) | 126.19 | (2.86) | 126.19 | (2.86) |
| Norm | 250 | 0.50 | 9.88 | (0.06) | 9.91 | (0.07) | 9.88 | (0.06) | 9.88 | (0.06) |
| | 500 | 0.95 | 250.53 | (3.54) | 256.94 | (3.62) | 258.10 | (3.59) | 258.10 | (3.59) |
| | 500 | 0.50 | 19.10 | (0.08) | 19.81 | (0.08) | 19.81 | (0.08) | 19.81 | (0.08) |
| | 1000 | 0.95 | 512.13 | (4.90) | 517.94 | (4.92) | 519.26 | (4.90) | 519.26 | (4.90) |
| | 1000 | 0.50 | 39.72 | (0.11) | 39.72 | (0.11) | 39.72 | (0.11) | 39.72 | (0.11) |
| $\ell_1$ Norm | 250 | 0.95 | 142.91 | (5.17) | 158.36 | (5.80) | 176.09 | (8.29) | 159.29 | (5.79) |
| | 250 | 0.50 | 1.33 | (0.03) | 1.39 | (0.03) | 2.29 | (0.27) | 1.37 | (0.03) |
| | 500 | 0.95 | 183.55 | (5.21) | 198.28 | (5.97) | 233.56 | (9.67) | 197.97 | (5.79) |
| | 500 | 0.50 | 1.43 | (0.02) | 1.46 | (0.03) | 2.54 | (0.17) | 1.46 | (0.03) |
| | 1000 | 0.95 | 210.56 | (3.98) | 223.65 | (4.76) | 279.71 | (12.01) | 222.86 | (4.58) |
| | 1000 | 0.50 | 1.58 | (0.03) | 1.64 | (0.03) | 3.04 | (0.33) | 1.64 | (0.03) |
| $\ell_2$ Norm | 250 | 0.95 | 36.90 | (1.61) | 42.98 | (1.95) | 44.87 | (2.02) | 43.77 | (1.98) |
| | 250 | 0.50 | 0.47 | (0.01) | 0.49 | (0.01) | 0.89 | (0.07) | 0.49 | (0.01) |
| | 500 | 0.95 | 47.09 | (1.41) | 54.45 | (2.06) | 66.64 | (2.96) | 54.82 | (2.04) |
| | 500 | 0.50 | 0.51 | (0.01) | 0.53 | (0.01) | 1.18 | (0.10) | 0.53 | (0.01) |
| | 1000 | 0.95 | 56.70 | (1.40) | 62.31 | (1.79) | 78.59 | (2.85) | 62.76 | (1.80) |
| | 1000 | 0.50 | 0.59 | (0.01) | 0.61 | (0.01) | 1.58 | (0.14) | 0.61 | (0.01) |

Table 2.4: Simulation model 2.2: Frobenius, $\ell_1$, $\ell_2$ risk. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

| Model 2.3: Tapering parameter selection | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $k^{opt}$ | | | $\hat{k}^{oracle}$ | | | $\hat{k}^{sure}$ | $\hat{k}^{cv}$ | | |
| $p$ | $\alpha$ | F | $\ell_1$ | $\ell_2$ | F | $\ell_1$ | $\ell_2$ | F,$\ell_1$,$\ell_2$ | F | $\ell_1$ | $\ell_2$ |
| 250 | 0.1 | 8 | 7 | 7 | 7.91 | 7.21 | 7.56 | 7.93 | 7.35 | 11.15 | 17.19 |
| | | | | | (0.29) | (0.77) | (1.12) | (0.26) | (0.48) | (5.81) | (12.54) |
| 250 | 0.5 | 6 | 5 | 5 | 5.97 | 5.57 | 5.91 | 6.13 | 5.47 | 8.76 | 13.79 |
| | | | | | (0.41) | (1.30) | (1.14) | (0.68) | (0.64) | (4.64) | (9.34) |
| 500 | 0.1 | 8 | 7 | 7 | 8.00 | 7.06 | 7.29 | 7.93 | 7.22 | 11.21 | 19.49 |
| | | | | | (0.00) | (0.81) | (1.09) | (0.26) | (0.42) | (5.87) | (18.70) |
| 500 | 0.5 | 6 | 5 | 5 | 5.97 | 5.49 | 5.59 | 6.18 | 5.41 | 9.95 | 15.39 |
| | | | | | (0.17) | (1.10) | (1.01) | (0.59) | (0.59) | (8.39) | (10.43) |
| 1000 | 0.1 | 8 | 7 | 7 | 8.00 | 6.77 | 6.99 | 8.00 | 7.12 | 11.26 | 21.79 |
| | | | | | (0.00) | (0.90) | (1.12) | (0.61) | (0.33) | (6.10) | (17.94) |
| 1000 | 0.5 | 6 | 5 | 5 | 6.00 | 5.13 | 5.31 | 6.13 | 5.20 | 8.96 | 18.24 |
| | | | | | (0.00) | (1.28) | (1.20) | (0.37) | (0.40) | (5.72) | (13.66) |

Table 2.5: Simulation model 2.3: tapering parameter selection. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

| Model 2.3: Estimation risk | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $p$ | $\alpha$ | Oracle | | SURE | | CV | | CV-F | |
| Frobenius Norm | 250 | 0.1 | 13.89 | (0.09) | 13.93 | (0.09) | 14.09 | (0.09) | 14.09 | (0.09) |
| | 250 | 0.5 | 11.63 | (0.07) | 11.75 | (0.07) | 11.82 | (0.07) | 11.82 | (0.07) |
| | 500 | 0.1 | 27.68 | (0.13) | 27.73 | (0.13) | 28.08 | (0.13) | 28.08 | (0.13) |
| | 500 | 0.5 | 23.42 | (0.10) | 23.59 | (0.11) | 23.78 | (0.10) | 23.78 | (0.10) |
| | 1000 | 0.1 | 55.79 | (0.22) | 55.79 | (0.22) | 56.68 | (0.22) | 56.68 | (0.22) |
| | 1000 | 0.5 | 46.95 | (0.16) | 47.06 | (0.16) | 47.70 | (0.14) | 47.70 | (0.14) |
| $\ell_1$ Norm | 250 | 0.1 | 1.98 | (0.04) | 2.10 | (0.04) | 3.42 | (0.30) | 2.05 | (0.04) |
| | 250 | 0.5 | 1.47 | (0.03) | 1.60 | (0.03) | 2.38 | (0.18) | 1.59 | (0.03) |
| | 500 | 0.1 | 2.18 | (0.04) | 2.36 | (0.05) | 3.79 | (0.34) | 2.26 | (0.04) |
| | 500 | 0.5 | 1.65 | (0.02) | 1.78 | (0.03) | 3.62 | (0.55) | 1.75 | (0.03) |
| | 1000 | 0.1 | 2.49 | (0.04) | 2.72 | (0.05) | 4.34 | (0.48) | 2.55 | (0.05) |
| | 1000 | 0.5 | 1.88 | (0.03) | 2.07 | (0.05) | 3.34 | (0.30) | 1.98 | (0.04) |
| $\ell_2$ Norm | 250 | 0.1 | 0.67 | (0.01) | 0.72 | (0.02) | 1.33 | (0.09) | 0.71 | (0.02) |
| | 250 | 0.5 | 0.53 | (0.01) | 0.58 | (0.01) | 0.94 | (0.06) | 0.57 | (0.01) |
| | 500 | 0.1 | 0.78 | (0.02) | 0.85 | (0.02) | 1.66 | (0.16) | 0.82 | (0.02) |
| | 500 | 0.5 | 0.59 | (0.01) | 0.63 | (0.01) | 1.18 | (0.08) | 0.62 | (0.01) |
| | 1000 | 0.1 | 0.88 | (0.01) | 0.98 | (0.02) | 2.02 | (0.14) | 0.93 | (0.02) |
| | 1000 | 0.5 | 0.69 | (0.01) | 0.76 | (0.02) | 1.54 | (0.10) | 0.73 | (0.01) |

Table 2.6: Simulation model 2.3: Frobenius, $\ell_1$ $\ell_2$ risk. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

| Model 2.4: Tapering parameter selection | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $k^{opt}$ | | | $\hat{k}^{oracle}$ | | | $\hat{k}^{sure}$ | $\hat{k}^{cv}$ | | |
| $p$ | $\alpha$ | F | $\ell_1$ | $\ell_2$ | F | $\ell_1$ | $\ell_2$ | F,$\ell_1$,$\ell_2$ | F | $\ell_1$ | $\ell_2$ |
| 250 | 0.1 | 11 | 9 | 31 | 10.76 | 10.49 | 36.88 | 10.44 | 9.50 | 18.03 | 46.96 |
| | | | | | (0.55) | (2.94) | (8.62) | (1.21) | (0.97) | (9.28) | (24.06) |
| 250 | 0.5 | 6 | 5 | 9 | 5.99 | 5.63 | 10.64 | 6.04 | 5.44 | 10.11 | 20.84 |
| | | | | | (0.44) | (1.40) | (2.29) | (0.76) | (0.64) | (5.86) | (14.70) |
| 500 | 0.1 | 11 | 9 | 38 | 10.78 | 9.66 | 44.15 | 10.47 | 9.36 | 18.88 | 56.91 |
| | | | | | (0.46) | (2.29) | (8.37) | (0.85) | (0.70) | (10.07) | (24.31) |
| 500 | 0.5 | 6 | 5 | 10 | 6.01 | 5.51 | 10.76 | 6.11 | 5.29 | 11.35 | 20.58 |
| | | | | | (0.22) | (1.58) | (2.22) | (0.63) | (0.50) | (6.81) | (13.10) |
| 1000 | 0.1 | 11 | 9 | 51 | 10.92 | 9.10 | 56.00 | 10.79 | 9.26 | 19.12 | 63.46 |
| | | | | | (0.27) | (2.73) | (7.28) | (0.46) | (0.57) | (12.11) | (31.95) |
| 1000 | 0.5 | 6 | 5 | 10 | 6.00 | 5.20 | 10.41 | 6.05 | 5.19 | 10.31 | 27.61 |
| | | | | | (0.14) | (1.44) | (2.03) | (0.46) | (0.39) | (6.04) | (19.52) |

Table 2.7: Simulation model 2.4: tapering parameter selection. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

| Model 2.4: Estimation risk | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $p$ | $\alpha$ | Oracle | | SURE | | CV | | CV-F | |
| Frobenius Norm | 250 | 0.1 | 26.07 | (0.09) | 26.28 | (0.09) | 26.38 | (0.10) | 26.38 | (0.10) |
| | 250 | 0.5 | 13.59 | (0.07) | 13.75 | (0.07) | 13.80 | (0.07) | 13.80 | (0.07) |
| | 500 | 0.1 | 53.36 | (0.14) | 53.54 | (0.15) | 53.81 | (0.14) | 53.81 | (0.14) |
| | 500 | 0.5 | 27.57 | (0.11) | 27.76 | (0.11) | 27.99 | (0.11) | 27.99 | (0.11) |
| | 1000 | 0.1 | 108.44 | (0.21) | 108.51 | (0.21) | 109.35 | (0.20) | 109.35 | (0.20) |
| | 1000 | 0.5 | 55.42 | (0.18) | 55.63 | (0.18) | 56.22 | (0.17) | 56.22 | (0.17) |
| $\ell_1$ Norm | 250 | 0.1 | 14.14 | (0.10) | 14.64 | (0.12) | 17.62 | (0.47) | 14.58 | (0.11) |
| | 250 | 0.5 | 3.59 | (0.04) | 3.80 | (0.05) | 4.95 | (0.24) | 3.76 | (0.05) |
| | 500 | 0.1 | 18.74 | (0.11) | 19.35 | (0.14) | 23.31 | (0.63) | 19.34 | (0.12) |
| | 500 | 0.5 | 4.24 | (0.05) | 4.47 | (0.06) | 6.38 | (0.51) | 4.41 | (0.06) |
| | 1000 | 0.1 | 24.15 | (0.13) | 24.97 | (0.17) | 30.44 | (1.15) | 24.80 | (0.16) |
| | 1000 | 0.5 | 4.60 | (0.04) | 4.87 | (0.06) | 6.31 | (0.24) | 4.74 | (0.04) |
| $\ell_2$ Norm | 250 | 0.1 | 2.98 | (0.05) | 5.49 | (0.07) | 4.21 | (0.15) | 5.84 | (0.07) |
| | 250 | 0.5 | 0.88 | (0.01) | 1.11 | (0.02) | 1.44 | (0.09) | 1.20 | (0.02) |
| | 500 | 0.1 | 4.23 | (0.05) | 7.90 | (0.06) | 5.55 | (0.18) | 8.45 | (0.06) |
| | 500 | 0.5 | 1.01 | (0.01) | 1.26 | (0.01) | 1.57 | (0.09) | 1.39 | (0.01) |
| | 1000 | 0.1 | 5.66 | (0.04) | 10.44 | (0.05) | 7.07 | (0.20) | 11.34 | (0.05) |
| | 1000 | 0.5 | 1.10 | (0.01) | 1.36 | (0.01) | 2.18 | (0.13) | 1.52 | (0.01) |

Table 2.8: Simulation model 2.4: Frobenius, $\ell_1$ $\ell_2$ risk. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

| Model 2.5: Tapering parameter selection | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $k^{opt}$ | | | $\hat{k}^{oracle}$ | | | $\hat{k}^{sure}$ | $\hat{k}^{cv}$ | | |
| $p$ | $\rho$ | F | $\ell_1$ | $\ell_2$ | F | $\ell_1$ | $\ell_2$ | F,$\ell_1$,$\ell_2$ | F | $\ell_1$ | $\ell_2$ |
| 250 | 0.95 | 71 | 71 | 76 | 70.79 | 72.84 | 77.36 | 71.01 | 68.59 | 80.93 | 89.33 |
| | | | | | (4.53) | (11.93) | (17.32) | (12.38) | (12.80) | (28.25) | (33.80) |
| 250 | 0.50 | 5 | 5 | 5 | 5.00 | 4.99 | 5.18 | 5.02 | 5.00 | 8.93 | 12.34 |
| | | | | | (0.00) | (0.92) | (0.97) | (0.14) | (0.00) | (6.76) | (10.86) |
| 500 | 0.95 | 70 | 70 | 71 | 70.39 | 71.40 | 74.86 | 70.32 | 67.13 | 87.43 | 110.37 |
| | | | | | (3.17) | (12.76) | (18.99) | (7.15) | (7.23) | (31.87) | (39.78) |
| 500 | 0.50 | 5 | 5 | 5 | 5.00 | 4.80 | 5.11 | 5.00 | 5.00 | 8.97 | 15.95 |
| | | | | | (0.00) | (0.90) | (1.05) | (0.00) | (0.00) | (4.88) | (13.79) |
| 1000 | 0.95 | 69 | 68 | 72 | 69.87 | 68.65 | 75.06 | 70.31 | 67.37 | 90.49 | 119.22 |
| | | | | | (2.48) | (11.11) | (12.49) | (4.23) | (4.42) | (28.50) | (38.16) |
| 1000 | 0.50 | 5 | 5 | 5 | 5.00 | 4.65 | 4.86 | 5.00 | 5.00 | 8.03 | 19.02 |
| | | | | | (0.00) | (0.97) | (0.92) | (0.00) | (0.00) | (5.65) | (17.53) |

Table 2.9: Simulation model 2.5: tapering parameter selection. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

| Model 2.5: Estimation risk | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $p$ | $\rho$ | Oracle | | SURE | | CV | | CV-F | |
| Frobenius Norm | 250 | 0.95 | 118.09 | (2.66) | 124.96 | (2.88) | 126.19 | (2.87) | 126.19 | (2.87) |
| | 250 | 0.50 | 9.92 | (0.06) | 9.93 | (0.06) | 9.92 | (0.06) | 9.92 | (0.06) |
| | 500 | 0.95 | 247.49 | (3.90) | 254.18 | (4.22) | 256.02 | (4.17) | 256.02 | (4.17) |
| | 500 | 0.50 | 19.81 | (0.08) | 19.81 | (0.08) | 19.81 | (0.08) | 19.81 | (0.08) |
| | 1000 | 0.95 | 511.21 | (6.22) | 519.52 | (6.53) | 520.79 | (6.34) | 520.79 | (6.34) |
| | 1000 | 0.50 | 39.80 | (0.12) | 39.80 | (0.12) | 39.80 | (0.12) | 39.80 | (0.12) |
| $\ell_1$ Norm | 250 | 0.95 | 142.91 | (5.17) | 158.30 | (5.80) | 174.46 | (7.75) | 159.24 | (5.82) |
| | 250 | 0.50 | 1.31 | (0.02) | 1.36 | (0.03) | 2.66 | (0.33) | 1.36 | (0.03) |
| | 500 | 0.95 | 184.75 | (5.36) | 201.05 | (6.86) | 236.85 | (10.41) | 201.38 | (6.72) |
| | 500 | 0.50 | 1.62 | (0.03) | 1.68 | (0.03) | 2.74 | (0.18) | 1.50 | (0.03) |
| | 1000 | 0.95 | 209.75 | (4.26) | 225.51 | (5.81) | 275.02 | (11.77) | 223.53 | (5.29) |
| | 1000 | 0.50 | 1.62 | (0.03) | 1.68 | (0.03) | 2.80 | (0.34) | 1.68 | (0.03) |
| $\ell_2$ Norm | 250 | 0.95 | 36.90 | (1.61) | 43.01 | (1.95) | 45.23 | (2.05) | 43.74 | (1.99) |
| | 250 | 0.50 | 0.45 | (0.01) | 0.48 | (0.01) | 0.83 | (0.06) | 0.47 | (0.01) |
| | 500 | 0.95 | 48.20 | (1.72) | 55.50 | (2.33) | 68.21 | (3.84) | 56.20 | (2.31) |
| | 500 | 0.50 | 0.51 | (0.01) | 0.54 | (0.01) | 1.15 | (0.08) | 0.54 | (0.01) |
| | 1000 | 0.95 | 57.00 | (1.56) | 63.66 | (2.00) | 82.40 | (3.70) | 63.86 | (1.90) |
| | 1000 | 0.50 | 0.59 | (0.01) | 0.62 | (0.01) | 1.48 | (0.11) | 0.62 | (0.01) |

Table 2.10: Simulation model 2.5: Frobenius, $\ell_1$, $\ell_2$ risk. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

| Model 2.6: Tapering parameter selection | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $k^{opt}$ | | | $\hat{k}^{oracle}$ | | | $\hat{k}^{sure}$ | $\hat{k}^{cv}$ | | |
| $p$ | $\alpha$ | F | $\ell_1$ | $\ell_2$ | F | $\ell_1$ | $\ell_2$ | F,$\ell_1$,$\ell_2$ | F | $\ell_1$ | $\ell_2$ |
| 250 | 0.1 | 8 | 7 | 7 | 7.91 | 7.01 | 7.57 | 7.89 | 7.28 | 10.78 | 16.28 |
| | | | | | (0.29) | (0.77) | (1.08) | (0.31) | (0.45) | (7.22) | (11.39) |
| 250 | 0.5 | 6 | 5 | 5 | 5.99 | 5.59 | 5.96 | 5.99 | 5.34 | 8.93 | 14.78 |
| | | | | | (0.41) | (1.22) | (1.37) | (0.70) | (0.57) | (4.90) | (10.48) |
| 500 | 0.1 | 8 | 7 | 7 | 7.97 | 7.15 | 7.18 | 7.92 | 7.19 | 10.59 | 19.79 |
| | | | | | (0.17) | (0.86) | (0.98) | (0.27) | (0.39) | (3.94) | (16.91) |
| 500 | 0.5 | 6 | 5 | 5 | 6.00 | 5.53 | 5.64 | 6.07 | 5.36 | 9.50 | 16.49 |
| | | | | | (0.25) | (1.34) | (1.38) | (0.62) | (0.56) | (7.25) | (14.40) |
| 1000 | 0.1 | 8 | 7 | 7 | 7.99 | 6.93 | 6.98 | 7.99 | 7.11 | 11.43 | 24.50 |
| | | | | | (0.10) | (0.88) | (1.06) | (0.10) | (0.31) | (6.87) | (20.40) |
| 1000 | 0.5 | 6 | 5 | 5 | 5.99 | 5.13 | 5.52 | 6.07 | 5.22 | 9.86 | 20.23 |
| | | | | | (0.10) | (1.21) | (1.19) | (0.46) | (0.42) | (6.15) | (15.90) |

Table 2.11: Simulation model 2.6: tapering parameter selection. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

| Model 2.6: Estimation risk | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $p$ | $\alpha$ | Oracle | | SURE | | CV | | CV-F | |
| Frobenius Norm | 250 | 0.1 | 13.89 | (0.09) | 13.95 | (0.09) | 14.09 | (0.09) | 14.09 | (0.09) |
| | 250 | 0.5 | 11.61 | (0.07) | 11.76 | (0.07) | 11.82 | (0.07) | 11.82 | (0.07) |
| | 500 | 0.1 | 27.82 | (0.14) | 27.90 | (0.14) | 28.25 | (0.14) | 28.25 | (0.14) |
| | 500 | 0.5 | 23.35 | (0.10) | 23.54 | (0.10) | 23.77 | (0.10) | 23.77 | (0.10) |
| | 1000 | 0.1 | 56.08 | (0.21) | 56.10 | (0.21) | 56.95 | (0.21) | 56.95 | (0.21) |
| | 1000 | 0.5 | 46.96 | (0.16) | 47.13 | (0.17) | 47.74 | (0.15) | 47.74 | (0.15) |
| $\ell_1$ Norm | 250 | 0.1 | 1.99 | (0.04) | 2.13 | (0.05) | 3.51 | (0.43) | 2.05 | (0.05) |
| | 250 | 0.5 | 1.46 | (0.03) | 1.58 | (0.03) | 2.46 | (0.20) | 1.56 | (0.03) |
| | 500 | 0.1 | 2.18 | (0.04) | 2.35 | (0.05) | 3.42 | (0.20) | 2.26 | (0.04) |
| | 500 | 0.5 | 1.66 | (0.03) | 1.79 | (0.04) | 3.23 | (0.45) | 1.77 | (0.04) |
| | 1000 | 0.1 | 2.41 | (0.04) | 2.64 | (0.05) | 4.53 | (0.48) | 2.49 | (0.04) |
| | 1000 | 0.5 | 1.85 | (0.03) | 2.03 | (0.04) | 3.64 | (0.35) | 1.96 | (0.03) |
| $\ell_2$ Norm | 250 | 0.1 | 0.70 | (0.02) | 0.74 | (0.02) | 1.25 | (0.08) | 0.73 | (0.02) |
| | 250 | 0.5 | 0.53 | (0.01) | 0.57 | (0.01) | 0.98 | (0.06) | 0.56 | (0.01) |
| | 500 | 0.1 | 0.78 | (0.02) | 0.84 | (0.02) | 1.66 | (0.14) | 0.82 | (0.02) |
| | 500 | 0.5 | 0.62 | (0.01) | 0.67 | (0.02) | 1.24 | (0.10) | 0.67 | (0.01) |
| | 1000 | 0.1 | 0.86 | (0.01) | 0.97 | (0.02) | 2.17 | (0.16) | 0.91 | (0.02) |
| | 1000 | 0.5 | 0.68 | (0.01) | 0.73 | (0.02) | 1.61 | (0.10) | 0.71 | (0.01) |

Table 2.12: Simulation model 2.6: Frobenius, $\ell_1$ $\ell_2$ risk. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

## 2.4 Rock Sonar Spectrum Data

In this section we use the sonar data to illustrate the efficacy of SURE tuning and to further demonstrate the conclusions made in the simulation study. The sonar data is publicly available from the UCI repository of machine learning databases Frank and Asuncion (2010). We consider its subset consisting of 97 sonar spectra bounced off from rocks. Each spectrum has 60 frequency band energy measurements. Although the dimension is 60, this is still a relative large dimension scenario, because the sample size is 97. We examined the entries of sample covariance matrix and found there is a quite obvious decay pattern as the entries move away from the diagonal. Hence we used tapering to regularize the sample covariance matrix. SURE and cross-validation were used to select the tapering parameter. Bootstrap was used to assess the variability of each tuning procedure.

In Figure 2.2 we plot SURE and cross-validated estimates of the Frobenius risk and also show the bootstrap histogram of the selected tapering parameter by SURE and cross-validation. Some interesting phenomena are evident in the figure. First, the two bootstrap histograms clearly show that SURE tuning is less variable than cross-validation. Second, SURE tuning selected the high peak of the SURE bootstrap histogram but cross-validation selected a left tail value of its bootstrap histogram. Third, the cross-validation estimate of the Frobenius risk is much larger that the SURE estimate.

Figure 2.3 shows the cross-validation tuning results under the $\ell_1, \ell_2$ norms. The selected tapering parameters under the $\ell_1, \ell_2$ norms are not very different from those under the Frobenius norm. The significant difference is that cross-validation tuning under the $\ell_1, \ell_2$ norms has much flatter bootstrap histograms, indicating much larger variability in selection.

We also repeated the above analysis on the other subset consisting of 111 sonar

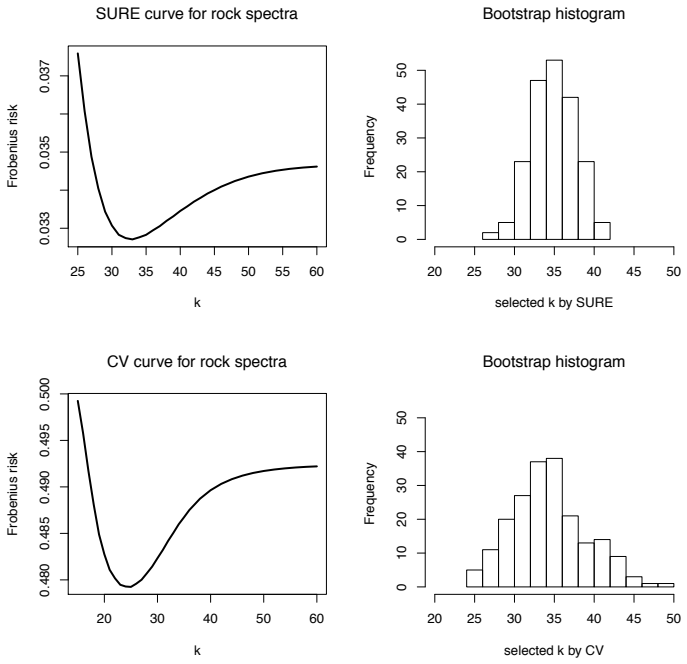



Figure 2.2: Rock sonar spectrum data: SURE and cross-validation tuning under the Frobenius norm. The right panels display the bootstrap histograms of the selected tapering parameter by SURE and cross-validation.

spectra bounced off from metal cylinders and the conclusions are basically the same. For the sake of space consideration, we opt to present the analysis results and figures in a technical report version of the original manuscript.

In conclusion, what we have observed in this real data example is consistent with the simulation conclusions.

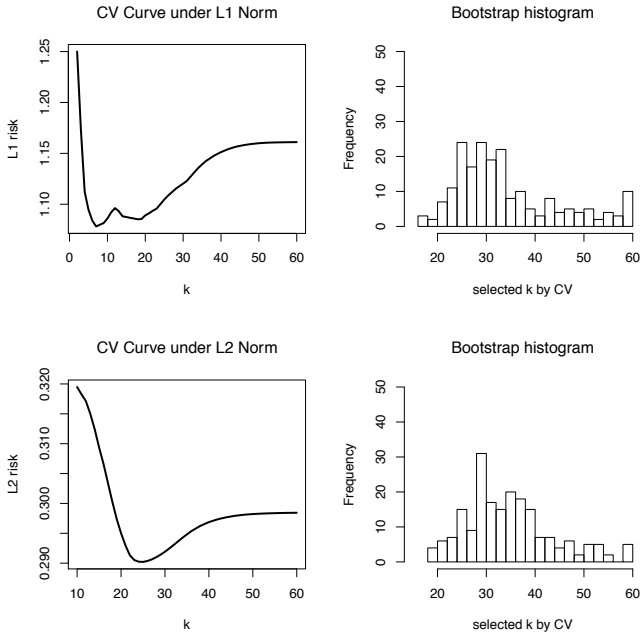



Figure 2.3: Rock sonar spectrum data: cross-validation tuning under the $\ell_1, \ell_2$ norms. The right panels display the bootstrap histograms of the selected tapering parameter by cross-validation.

## 2.5 Discussion

There are two important issues in any regularized estimation procedure:

1. how to select the regularization parameter?

2. how to estimate the accuracy of a regularized estimator?

In traditional vector-estimation problems such as nonparametric regression or classification, cross-validation is a routinely used method for answering both questions and perform well in general. Efron (2004) has shown that SURE can be more accurate than cross-validation for estimating the risk of a vector estimator. We have found that cross-validation does not perform satisfactorily for tuning the tapering covariance estimator when the objective loss function is the matrix $\ell_1$ or $\ell_2$ norm. Cross-validation can capture the shape of the Frobenius risk, but the cross-validated estimate of the Frobenius risk tends to be too large to be a good estimate. Our empirical study suggests that the Frobenius norm is better for tuning a covariance matrix estimator even when the objective loss is the $\ell_1$ or $\ell_2$ norm. To that end, the proposed SURE formula is very useful: it is computationally economic, stable and provides a reliable estimate of the Frobenius risk.

## 2.6 Appendix

**Proof 2.1 (Proof of Lemma 2.2.1)**

We start with the Stein's identity Efron (2004)

$$(\hat{\sigma}_{ij} - \sigma_{ij})^2 = (\hat{\sigma}_{ij} - \tilde{\sigma}^s_{ij})^2 - (\tilde{\sigma}^s_{ij} - \sigma_{ij})^2 + 2(\hat{\sigma}_{ij} - \sigma_{ij})(\tilde{\sigma}^s_{ij} - \sigma_{ij}). \tag{2.13}$$

Taking expectation at both side of (2.13) and summing over $i, j = 1$ yields

$$\mathbb{E}\|\widehat{\Sigma} - \Sigma\|_F^2 = \mathbb{E}\|\widehat{\Sigma} - \tilde{\Sigma}^s\|_F^2 - \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Var}(\tilde{\sigma}_{ij}^s) + 2\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}(\hat{\sigma}_{ij}, \tilde{\sigma}_{ij}^s).$$

Note that $\mathbb{E}[(\hat{\sigma}_{ij} - \sigma_{ij})(\tilde{\sigma}_{ij}^s - \sigma_{ij})] = \mathrm{Cov}(\hat{\sigma}_{ij}, \tilde{\sigma}_{ij}^s)$ because $\mathbb{E}\tilde{\sigma}_{ij}^s = \sigma_{ij}$. □

**Proof 2.2 (Proof of Lemma 2.2.2)**

The estimators under consideration are translational invariant. Without loss of generality, we can let $\mu = \mathbb{E}(x) = 0$. By straightforward calculation based on bivariate normal distribution, we have

$$\mathbb{E}(x_i^2 x_j^2) = \sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2, \tag{2.14}$$

which holds for both $i = j$ and $i \neq j$.

(2.15)

$$\begin{aligned}
\mathbb{E}((\tilde{\sigma}_{ij}^s)^2) &= \mathbb{E}((n-1)^{-2}(\sum_{k=1}^{n} x_{k,i}x_{k,j} - n\bar{x}_i\bar{x}_j)^2) \\
&= (n-1)^{-2}\left\{\mathbb{E}((\sum_{k=1}^{n} x_{k,i}x_{k,j})^2) - 2n^{-1}\sum_{k=1}^{n}\mathbb{E}(n\bar{x}_i n\bar{x}_j x_{k,i}x_{k,j}) + n^2\mathbb{E}(\bar{x}_i^2\bar{x}_j^2)\right\}.
\end{aligned}$$

We also have

$$\begin{aligned}
\mathbb{E}((n^{-1}\sum_{k=1}^{n} x_{k,i}x_{k,j})^2) &= \frac{1}{n}\mathrm{Var}(x_i x_j) + (\mathbb{E}(x_i x_j))^2 \\
&= \frac{1}{n}(\sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2 - \sigma_{ij}^2) + \sigma_{ij}^2 \\
&= \frac{1}{n}\sigma_{ii}\sigma_{jj} + \frac{1+n}{n}\sigma_{ij}^2.
\end{aligned} \tag{2.16}$$

Note that $\bar{X} \sim N(0, \Sigma/n)$. Using (2.14) we have

$$n^2\mathbb{E}(\bar{x}_i^2\bar{x}_j^2) = 2\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}. \tag{2.17}$$

$$\begin{aligned}\mathbb{E}(n\bar{x}_i n\bar{x}_j x_{k,i}x_{k,j}) &= \sum_{1\le l,l'\le n}\left\{I(l=l'\neq k)\mathbb{E}(x_{l,i}x_{l,j}x_{k,i}x_{k,j}) + I(l=l'=k)\mathbb{E}(x_{k,i}^2x_{k,j}^2)\right\} \\ &= (n-1)\sigma_{12}^2 + (\sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2). \end{aligned}\tag{2.18}$$

Substituting (2.16)–(2.18) into (2.15) gives

$$\mathbb{E}((\tilde{\sigma}_{ij}^s)^2) = \frac{n\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}}{n-1}. \tag{2.19}$$

Thus, $\mathrm{Var}(\tilde{\sigma}_{ij}^s) = \mathbb{E}((\tilde{\sigma}_{ij}^s)^2) - \sigma_{ij}^2 = \frac{\sigma_{ij}^2+\sigma_{ii}\sigma_{jj}}{n-1}$.

We now show (2.9) by deriving an expression for $\mathbb{E}(\tilde{\sigma}_{ii}^s\tilde{\sigma}_{jj}^s)$.

$$(n-1)^2\mathbb{E}(\tilde{\sigma}_{ii}^s\tilde{\sigma}_{jj}^s) = \sum_{1\le k,k'\le n}\mathbb{E}(x_{k,i}^2x_{k',j}^2) - \sum_{1\le k'\le n}\mathbb{E}(\bar{x}_i^2x_{k',j}^2) - \sum_{1\le k\le n}\mathbb{E}(\bar{x}_j^2x_{k,i}^2) + n^2\mathbb{E}(\bar{x}_i^2\bar{x}_j^2). \tag{2.20}$$

Repeatedly using (2.14) we have

$$\sum_{1\le k,k'\le n}\mathbb{E}(x_{k,i}^2x_{k',j}^2) = n^2\sigma_{ii}\sigma_{jj} + 2n\sigma_{ij}^2, \tag{2.21}$$

$$\begin{aligned}n^2\mathbb{E}(\bar{x}_i^2x_{k',j}^2) &= \sum_{1\le l,l'\le n}\left\{I(l=l'\neq k')\mathbb{E}(x_{l,i}^2x_{k',j}^2) + I(l=l'=k')\mathbb{E}(x_{k',i}^2x_{k',j}^2)\right\} \\ &= n\sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2, \end{aligned}\tag{2.22}$$

$$n^2\mathbb{E}(\bar{x}_j^2 x_{k,i}^2) \;=\; n\sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2. \tag{2.23}$$

Substituting (2.17) and (2.21)–(2.23) into (2.20) gives

$$\mathbb{E}(\tilde{\sigma}_{ii}^s \tilde{\sigma}_{jj}^s) = \frac{n+1}{n-1}\sigma_{ii}\sigma_{jj} + \frac{2(n+2)}{n(n-1)}\sigma_{ij}^2. \tag{2.24}$$

□

Combining (2.19) and (2.24) gives (2.9).

# Chapter 3

# Varying-coefficient Independence Screening for High-dimensional Varying-coefficient Model

## 3.1 Introduction

In modern scientific research, it is more and more common to confront the situation when the number of predictor variables $p$ is of tens of thousands, potentially much larger than the number of observations $n$. Examples include data from microarrays, proteomics, brain images and etc. Variable selection hence becomes an increasingly important task in statistical research. There are vast literature on variable selection for regression problems under linear regression settings. Recent developments mostly focus on penalized methods, including the LASSO Tibshirani (1996), SCAD Fan and Li (2001), the Dantzig selector Candes and Tao (2007) and their variations. These methods have been thoroughly studied for variable selection with high-dimensional data van de Geer (2008); Bickel et al. (2009); Meinshausen and Yu (2009). A much computationally simpler method that can work well in practice for very high dimensional data is the sure independence screening (SIS), demonstrated in Fan and Lv (2008) in the classical regression context. Specifically, the sure independence screen-

ing recruits the features with best marginal utility, which corresponds to the largest marginal absolute Pearson correlation between the response and predictor in the context of least-squares regression for linear model. Fan and Lv (2008) showed that SIS has a sure screening property, that is, with probability very close to 1, it can retain all of the important features in the model. After sure screening, the remaining covariates are use to fit a penalized linear regression model. Recent works on sure screening include Fan et al. (2009), Fan and Song (2010), Fan et al. (2011), Zhu et al. (2011), Li et al. (2012b), Li et al. (2012a), among others.

This chapter concerns variable selection in the varying coefficient model which is an very important and useful generalization of the linear regression model. In practice, it is common to present the data as longitudinal observations $\{Y_{ij}, X_i(t_{ij}), t_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, n_i\}$, where $t_{ij}$ and $n_i$ are the time of the $j$th measurement and the number of repeated measurement for the $i$th subject, respectively; and $Y_{ij}$ and $\mathbf{X}_i(t_{ij}) = (X_{1i}(t_{ij}), \ldots, X_{pi}(t_{ij}))'$ are the $i$th subject's observed outcome and covariates at time $t_{ij}$. Examples include longitudinal data analysis Hoover et al. (1998) and functional response models Rice (2004) among others. The research interest mostly focuses on investigating the time-dependent effects of the covariates on responses measured repeatedly and/or longitudinally. Different regression models are proposed for this type of data, among them the varying-coefficient model, as an important generalization of the linear regression model, has gained a lot of popularity. For variable selection with varying-coefficients models, Wang et al. (2008b) and Wang et al. (2008a) both proposed a group penalization method in the fixed $p$ case, and Wei et al. (2011) recently extended this work to the case of diverging $p$. However, for very large $p$, these penalized methods remain computationally demanding.

Here we consider screening of the important covariates in varying-coefficient models by ranking the magnitude of nonparametric marginal correlations. The magnitude of the proposed screener can preserve the non-sparsity of the varying-coefficient mod-

els under some reasonable conditions, even with converging minimum strength of signals. Our work can be regarded as an important and nontrivial extension of SIS procedures proposed in Fan and Lv (2008) and Fan et al. (2011), with differences and our contributions highlighted as follows. First, compared with these screening literature, the minimum distinguishable signal is related with the stochastic error in estimating the nonparametric components, approximation errors in modeling nonparametric components and the number of observations within each subject. More efforts were taken to study the influence of the longitudinal observations on the sure screening property. This brings significant challenges to the theoretical development and leads to an interesting result on the extent to which the dimensionality can be reduced by varying-coefficient independence screening. The dimensionality of the model is allowed to grow near exponentially with the sample size. Second, we also propose an iterative nonparametric independence screening procedure, IVIS-gSCAD, to reduce the false positive rate and stabilize the computation. Additionally, unlike Li and Liang (2008) and Lam and Fan (2008), which are based on local polynomial regression, we use B-spline to approximate the nonparametric coefficients, which is computationally easier.

The outline of the chapter is as follows. In Section 3.2, we propose the varying-coefficient independence screening method based on B-spline to approximation. Theoretical results are shown in Section 3.3. In Section 3.4, an iterative varying-coefficient screening (IVIS) method is proposed. In Section 3.5, simulation studies are carried out to demonstrate the performance of the proposed method. In addition, a real data set is used as an illustration of varying-coefficient regression models.

## 3.2 Varying-coefficient Independence Screening

Consider the population $\{\mathbf{X}(t), Y(t)\}$ from the following time-varying coefficient model

$$Y(t) = \mathbf{X}(t)'\boldsymbol{\alpha}(t) + \epsilon(t),\ t \in \mathcal{T}, \tag{3.1}$$

where $\mathbf{X}(t) = (X_1(t), \ldots, X_p(t))'$ are the covariates, $\boldsymbol{\alpha}(t) = (\alpha_1(t), \ldots, \alpha_p(t))'$ are the time-varying coefficients, $\epsilon(t)$ is a mean zero stochastic process, $Y(t)$ is a mean zero outcome function and $\mathcal{T}$ is the time interval when the measurements are taken.

Our purpose is to select nonzero time-varying coefficients among $\boldsymbol{\alpha}(t)$, i.e., to identify the set $\mathcal{M}_\star = \{l : \alpha_l(t) \neq 0\}$. We consider $p$ marginal nonparametric regression problems:

$$\min_{\beta(t) \in L_2(P)} \mathbb{E}(Y(t) - X_l(t)\beta(t))^2, \tag{3.2}$$

where $P$ denotes the joint distribution of $\mathbf{X}(t)$ and $Y(t)$ and $L_2(P)$ is the class of square integrable functions under the measure $P$. The minimizer of (3.2) $\beta_{l0}(t) = \mathbb{E}X_l(t)Y(t)$. The population version of VIS is to screen the time-varying coefficients $\alpha_l(t)$ in model (3.1) according to $|\mathbb{E}X_l(t)Y(t)|$ to select a small group of covariates via thresholding.

Suppose that there is a random sample of $n$ independent subjects $\{\mathbf{X}_i(t), Y_i(t)\}_{i=1}^n$ from model (3.1). Let $t_{ij}$ and $n_i$ be the time of the $j$th measurement and the number of repeated measurement for the $i$th subject. $Y_{ij} = Y_i(t_{ij})$ and $\mathbf{X}_i(t_{ij}) = (X_{1i}(t_{ij}), \ldots, X_{pi}(t_{ij}))'$ are the $i$th subject's observed outcome and covariates at time $t_{ij}$. Based on longitudinal observations $\{Y_{ij}, \mathbf{X}_i(t_{ij}), t_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, n_i\}$, the model can be written as:

$$Y_i(t_{ij}) = \mathbf{X}_i(t_{ij})'\boldsymbol{\alpha}(t_{ij}) + \epsilon_i(t_{ij}). \tag{3.3}$$

For $l = 1, \ldots, p$, let $\{B_{lk}(\cdot), k = 1, \ldots, K_l\}$ denote a basis of B-spline functions. Each $\beta_l(t)$ can be approximated by a linear combination of B-spline basis functions. We consider marginal weighted least square estimation based on B-spline expansion, for $l = 1, \ldots, p$, by minimizing $e_l = \sum_{i=1}^{n} \omega_i \sum_{j=1}^{n_i} \Big( Y_{ij} - \sum_{k=1}^{K_l} X_{lij} B_{lk}(t_{ij}) \gamma_{lk} \Big)^2$, with respect to $\gamma_{lk}$. Choices of $\omega_i$ can be 1 or $1/n_i$, corresponding to equal weight to each single observation and equal weight to each subject respectively.

Let $\boldsymbol{\gamma}_l = (\gamma_{l1}, \ldots, \gamma_{lK_l})'$. Define $\mathbf{B}_l(t) = (B_{l1}(t), \ldots, B_{lK_l}(t))'$, $\mathbf{U}_{lij} = X_{lij}\mathbf{B}_l(t_{ij})$, $\mathbf{U}_{li} = (\mathbf{U}_{li1}, \ldots, \mathbf{U}_{lin_i})'$, $\mathbf{W}_i = \text{diag}(w_i, \ldots, w_i)$ with size $n_i$. Denote $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ and $X_{lij} = X_{li}(t_{ij})$ for $j = 1, \ldots, n_i$ and $\mathbf{X}_{li} = (X_{li1}, \ldots, X_{lin_i})'$. We can further express $e_l$ as $e_l = e_l(\boldsymbol{\gamma}_l) = \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{U}_{li}\boldsymbol{\gamma}_l)' \mathbf{W}_i (\mathbf{Y}_i - \mathbf{U}_{li}\boldsymbol{\gamma}_l)$. Let $\mathbf{U}_l'\mathbf{W}\mathbf{Y} = \sum_i \mathbf{U}_{li}'\mathbf{W}_i\mathbf{Y}_i$ and $\mathbf{U}_l'\mathbf{W}\mathbf{U}_l = \sum_i \mathbf{U}_{li}'\mathbf{W}_i\mathbf{U}_{li}$. Since $\mathbf{U}_l'\mathbf{W}\mathbf{U}_l$ is invertible with probability approaching one (which will be established in Lemma 3 from Song et al. (2013)), the unique minimizer of $e_l(\boldsymbol{\gamma}_l)$ is

$$\hat{\boldsymbol{\gamma}}_l = \Big( \mathbf{U}_l'\mathbf{W}\mathbf{U}_l \Big)^{-1} \mathbf{U}_l'\mathbf{W}\mathbf{Y}. \tag{3.4}$$

Let $\hat{\beta}_l(t) = \mathbf{B}_l'(t)\hat{\boldsymbol{\gamma}}_l = \sum_k \hat{\gamma}_{lk} B_{lk}(t)$. Define the set

$$\hat{\mathcal{M}}_{\nu_n} = \{ l : \ \omega_l = \frac{1}{|\mathcal{T}|} \int_{\mathcal{T}} \hat{\beta}_l(t)^2 dt \geq \nu_n \}$$

as the selected set, where $|\mathcal{T}|$ is the length of $\mathcal{T}$. $\nu_n$ is a pre-specified threshold. To compute $\int_{\mathcal{T}} \hat{\beta}_l(t)^2 dt / |\mathcal{T}|$, we take $N$ equally spaced time points $t_1 \leq \ldots \leq t_N$ in $\mathcal{T}$, and then we compute $\omega_{Nl} = \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}_l(t_i)^2$. As long as $N$ is large enough, $\omega_{Nl}$ can be used as $\omega_l$. In our numerical study we let $N = 10,000$.

We correspondingly define the population version of the marginal least square regression,

$$u_l = u_l(\boldsymbol{\gamma}_l) = E(\mathbf{Y} - \mathbf{U}_l\boldsymbol{\gamma}_l)'\mathbf{W}(\mathbf{Y} - \mathbf{U}_l\boldsymbol{\gamma}_l). \tag{3.5}$$

It can be shown that the unique minimizer of $u_l(\boldsymbol{\gamma}_l)$ is

$$\tilde{\boldsymbol{\gamma}}_l = \left(\mathbb{E}\mathbf{U}_l'\mathbf{W}\mathbf{U}_l\right)^{-1}\mathbb{E}\mathbf{U}_l'\mathbf{W}\mathbf{Y}.$$

Let $\tilde{\beta}_l(t) = \mathbf{B}_l'(t)\tilde{\boldsymbol{\gamma}}_l = \sum_k \tilde{\gamma}_{lk}B_{lk}(t)$. It can be shown that $\tilde{\beta}_l(t)$ is the projection of $\beta_{l0}(t)$ onto the space $\mathcal{G}_l$, a linear space of spline functions on $\mathcal{T}$ with a fixed degree and knot sequence.

Let $\mathbf{X}_{li} = \text{diag}(X_{li1}, \ldots, X_{lin_i})$. Define

$$\mathbf{B}_{li} = \begin{pmatrix} B_{l1}(t_{i1}) & \ldots & B_{l1}(t_{in_i}) \\ \vdots & & \vdots \\ B_{lK_l}(t_{i1}) & \ldots & B_{lK_l}(t_{in_i}) \end{pmatrix}.$$

It can be seen that $\mathbf{U}_{li} = \mathbf{X}_{li}\mathbf{B}_{li}'$. With some algebra, we can rewrite (3.4) into the following form.

$$\hat{\boldsymbol{\gamma}}_l = \left(\sum_i \mathbf{B}_{li}\mathbf{X}_{li}\mathbf{W}_i\mathbf{X}_{li}\mathbf{B}_{li}'\right)^{-1}\sum_i \mathbf{B}_{li}\mathbf{X}_{li}\mathbf{W}_i\mathbf{Y}_i.$$

When $n_i = 1$ for $i = 1, \ldots, n$, i.e., there is no longitudinal observations or repeated measures for each subject, the model (3.3) boils down to the linear model. In this case, $\hat{\beta}_l(t)$ boils down to the marginal correlation proposed in Fan and Lv (2008).

## 3.3 Theoretical results

To establish the sure screening property, we decompose

$$\hat{\beta}_l(t) - \beta_{l0}(t) = \hat{\beta}_l(t) - \tilde{\beta}_l(t) + \tilde{\beta}_l(t) - \beta_{l0}(t)$$

corresponding to the estimation error and the approximation error respectively. Define

$$\begin{aligned} \omega &= \max_i \omega_i \\ N &= \max_i n_i \\ K_s &= \min_l K_l \\ K_m &= \max_l K_l \end{aligned}$$

and

$$\text{dist}(\beta_l, \mathbb{G}_l) = \inf_{g_l \in \mathbb{G}_l} \sup_{t \in \mathcal{T}} |\beta_l(t) - g_l(t)|$$

as the $L_\infty$ distance between $\beta_l(\cdot)$ and $\mathbb{G}_l$, where $\mathbb{G}_l$ is a linear space of spline functions on $\mathcal{T}$. Let $\rho_n = \max_l \text{dist}(\beta_{l0}, \mathbb{G}_l)$. The following conditions will be needed.

A. The observation times $\{t_{ij}\}$ , $j = 1, \ldots, n_i$, $i = 1, \ldots, n$, are chosen independently according to a distribution $F_T$ on a finite interval $\mathcal{T}$. $L_1 \leq |\mathcal{T}| \leq L_2$. In addition, they are independent of the response and covariate processes $(Y_i(t), \mathbf{X}_i(t)), i = 1, \ldots, n$. Its Lebesgue density $f_T(t)$ satisfies $M_1 \leq f_T(t) \leq M_2$ uniformly over $t \in \mathcal{T}$ for some positive constants $M_1$ and $M_2$.

B. There is a positive constant $M_3$ such that $|X_l(t)| \leq M_3$ for $t \in \mathcal{T}$ and $l = 1, \ldots, p$.

C. $\min_{l \in \mathcal{M}_\star} \frac{1}{|\mathcal{T}|} \int_{\mathcal{T}} (EX_l(t)Y(t))^2 dt \geq c_1 n^{-2\kappa}$, for some $\kappa \in (0, 1/2)$.

The following lemma shows that the minimum signal $\{\int_{\mathcal{T}} \tilde{\beta}_l(t)^2 dt / |\mathcal{T}|\}_{j \in \mathcal{M}_\star}$ is at the same level of the integrated marginal correlation, provided that the approximation error is negligible.

**Lemma 3.3.1**

Under conditions A–C, we have

$$\min_{l\in\mathcal{M}_\star}\frac{1}{|\mathcal{T}|}\int_{\mathcal{T}}\tilde{\beta}_l(t)^2dt \geq c_1\xi n^{-2\kappa},$$

if $\rho_n^2 \leq c_1M_1(1-\xi)n^{-2\kappa}K_mM_2^{-1}L_2^{-1}$ for some $\xi\in(0,1)$.

(Please refer to Song et al. (2013) for proof.) □

Now we establish the sure screening properties of the varying-coefficient independence screening (VIS). Let $\tilde{Y}_{ij} = \mathbf{X}_i(t_{ij})'\boldsymbol{\alpha}(t_{ij})$, $\tilde{\mathbf{Y}}_i = (\tilde{Y}_{i1},\ldots,\tilde{Y}_{in_i})'$, $\tilde{\mathbf{Y}} = (\tilde{\mathbf{Y}}_1,\ldots,\tilde{\mathbf{Y}}_n)'$. We need the following additional conditions:

D. $\|\tilde{\mathbf{Y}}\|_\infty < B_1$ for some positive constant $B_1$, where $\|\cdot\|_\infty$ is the sup norm.

E. The random error $\{\varepsilon_i(t)\}_{i=1}^n$ are i.i.d. with conditional mean zero and for any $B_2>0$, there exists a positive constant $B_3$ such that $E[\exp(B_2|\varepsilon_i(t)|)|\mathbf{X}_i(t)] < B_3$, for $t\in\mathcal{T}$.

F. There exist a positive constant $c_1$ and $\xi\in(0,1)$ such that $\rho_n^2 \leq c_1M_1(1-\xi)n^{-2\kappa}K_mM_2^{-1}L_2^{-1}$.

**Theorem 3.3.2**

Suppose that Conditions A-F hold, by taking $\nu_n = c_6n^{-2\kappa}$ with $c_6 \leq c_1\xi/2$, we have

$$\begin{aligned}P(\mathcal{M}_\star\subset\widehat{\mathcal{M}}_{\nu_n}) \;\geq\; & 1 - s_nK_m\Big\{(8+2K_m)\exp\Big(-c_3N^{-2}\omega^{-2}n^{1-4\kappa}K_m^{-3}\Big)\\ & +6c_5K_m\exp\Big(-c_4nK_m^{-1}\Big)\Big\}.\end{aligned}$$

(Please refer to Song et al. (2013) for proof.) □

This theorem implies that we can handle the NP-dimensionality:

$$\log p_n = o(N^{-2}\omega^{-2}n^{1-4\kappa}K_m^{-3} + nK_m^{-1}). \tag{3.6}$$

For $p_n$ and $n$ satisfying this condition, sure screening property holds, i.e., $P(\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\nu_n}) \to 1$. We note that the maximal number of spline basis $K_m$, the maximal number of observational time points and the maximal weights affect the order of dimensionality. In equation (3.6), it shows that the larger the minimum signal level, the smaller the number of basis functions, the smaller the weights, or the smaller the number of observational time points, the higher dimensionality the varying-coefficient independence screening (VIS) can handle. Meanwhile, the approximation rate $\rho_n$ also affects the dimensionality that the VIS can handle, through its relation with the choice of $K_m$ as required in Condition F. Since the approximation error can not be too large, the number of basis functions can not be too small. When the $\beta_l$ have bounded second derivatives and the number of observations for each subject is bounded, we have $\rho_n = O(K_m^{-2})$ Schumaker (1981), by taking $K_l = n^{1/5}$, the optimal rate for nonparametric regression Stone (1985), we have $\log p_n = o(n^{2/5})$. The second term in the right-hand side of (3.6) is improved compared with Fan et al. (2011) due to a technical improvement.

In addition to the sure screening property, controlling false selection rates is also an important criteria. To achieve the vanishing false selection rate, we bound the size of the selected set in the following theorem.

**Theorem 3.3.3**

Suppose Conditions A–F hold and $\mathrm{Var}(\mathbf{Y}) = O(1)$. Then, for any $\nu_n = c_6 K_m n^{-2\kappa}$, there exist positive constants $c_3$, $c_4$ and $c_5$ such that

$$\begin{aligned} & P[|\widehat{\mathcal{M}}_{\nu_n}| \le O\{n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma})\}] \\ \ge\ & 1 - p_n K_m \Big\{(8+2K_m)\exp\Big(-c_3 N^{-2}\omega^{-2}n^{1-4\kappa}K_m^{-3}\Big) + 6c_5 K_m \exp\Big(-c_4 n K_m^{-1}\Big)\Big\}, \end{aligned}$$

where $\mathbf{\Sigma} = E(\mathbf{U}\mathbf{W}\mathbf{U}')$.

(Please refer to Song et al. (2013) for proof.) □

This theorem implies that the correlation within the basis functions, i.e., the design matrix of the basis functions, instead of the covariance matrix as in the aforementioned two papers, will lead to the dimension reduction with varying-coefficient models. When the number of observations for each subject and the weights are bounded, $K_m = 1$ and $\lambda_{\max}(\Sigma) = O(n^{\tau})$, the size of the selected variable is of order $O(n^{2\kappa+\tau})$. This is of the same order as in Fan and Lv (2008) for the i.i.d. case.

## 3.4 Iterative VIS Procedure

In the next section, we will discuss the performance of IVIS using numerical examples.

As the independence screening procedure with marginal utilities uses only the marginal information of the covariates instead of the true model, its sure screening property may fail when its required technical conditions are not satisfied. Fan and Lv (2008) summarizes potential problems for SIS with linear models. Similar problems will be possible issues for the proposed screening methods as well:

1. A covariate that is jointly important but marginally unimportant to the response cannot be picked by independent screening methods. This issue will make the sure screening property fail.

2. Unimportant covariates that are highly correlated with the important covariates can have higher priority to be selected by independent screening methods than important covariates that are relatively weakly related to the response. This issue will not affect sure screening property, but will increase the false positive selection rates.

To address these issues while maintaining the computational expediency, Fan and Lv (2008) proposed iterative screening procedure to jointly employ the large-scale

screening and moderate-scale selection strategy for linear models. We adapt the idea and propose iterative screening procedures for VIS as follows:

1. Initial selection with marginal VIS and moderate-size variable selection: for every $l \in \{1, \cdots, p\}$, we apply the independence VIS procedure to pick a set $\mathcal{A}_1$ of indices of size $k_1$, which can be taken as $\lfloor 2n/(3\log(n)) \rfloor$ to guarantee it will take at least two iterations. We apply further some existing penalized algorithm for grouped-variables selection, such as group lasso in Yuan and Lin (2006), or group SCAD in Wang et al. (2007), on the set $\mathcal{A}_1$ to select a subset $\mathcal{M}_1$. Inside the penalized method, the penalty parameter can be selected by Bayes information type of criterion or (generalized) cross validation.

2. Forward large-scale conditional marginal screening: for every $l \in \mathcal{M}_1^c = \{1, \cdots, p\} \backslash \mathcal{M}_1$, we can compute the conditional marginal least squares with the set of features $\mathcal{M}_1$ pertained in the model:

$$\min \sum_{i=1}^{n} (\mathbf{Y}_i - \sum_{m \in \mathcal{M}_1} \mathbf{U}_{mi}\boldsymbol{\gamma}_m - \mathbf{U}_{li}\boldsymbol{\gamma}_l)' \mathbf{W}_i (\mathbf{Y}_i - \sum_{m \in \mathcal{M}_1} \mathbf{U}_{mi}\boldsymbol{\gamma}_m - \mathbf{U}_{li}\boldsymbol{\gamma}_l).$$

   This regression reflects the additional contribution of the $l$-th covariate conditioning on the existence of the variable set $\mathcal{M}_1$. After marginally screening as in the first step, we can pick a set $\mathcal{A}_2$ of indices of size $k_2 = 1$.

3. Backward moderate-size variable selection: We can apply further the penalized method used in the first step on the set $\mathcal{M}_1 \bigcup \mathcal{A}_2$ to select a subset $\mathcal{M}_2$.

4. Iteration until stabilization: iterate steps 2 and 3 until $|\mathcal{M}_l|$ beyond a pre-specified number or $\mathcal{M}_l = \mathcal{M}_{l-1}$.

## 3.5 Numerical Examples

### 3.5.1 Simulations

We used three simulation models to examine the finite-sample performance of IVIS and VIS+gSCAD. We fixed the sample to be 200 and the dimension to be 500 in all simulation examples. For each model we run 100 independent replicates.

**Model 3.1**: This model is taken from Wei et al. (2011). The response variable is generated by

$$y_i(t_{ij}) = \sum_{l=1}^{p} x_{li}(t_{ij})\beta_l(t_{ij}) + \epsilon_i(t_{ij})$$

The time points $t_{ij}$ are taken from $\{1, 2, 3, \ldots, 30\}$ with probability 0.4. Note that the number of actually observed time points $n_i$ for different subjects are different. Only the first six variables have nonzero coefficient functions. The coefficient functions are given by:

$$\begin{aligned} \beta_1(t) = 15 + 20\sin(\frac{\pi t}{15}) \quad &, \quad \beta_2(t) = 15 + 20\cos(\frac{\pi t}{15}), \\ \beta_3(t) = 2 - 3\sin(\frac{\pi(t-25)}{15}) \quad &, \quad \beta_4(t) = 2 - 3\cos(\frac{\pi(t-25)}{15}), \\ \beta_5(t) = 6 - 0.2t^2 \quad &, \quad \beta_6(t) = -4 + \frac{(20-t)^3}{2000}. \end{aligned}$$

The variables are generated as follows:

$$\begin{aligned} x_1(t) \sim \text{Unif}[t/10, 2 + t/10] \quad &, \quad \{x_l(t)\}_{l=2}^{5} \sim N(0, \frac{1 + x_1(t)}{2 + x_1(t)}), \\ x_6(t) \sim N(3\exp(t/30), 1) \quad &, \quad \{x_l(t)\}_{l=7}^{500} \sim \text{MVN}(\mathbf{0}, \Sigma), \end{aligned}$$

where

$$\Sigma_{t,s} = \mathrm{Cov}(x_l(t), x_l(s)) = 4\exp(-|t-s|)$$

The random error $\epsilon(t) = Z(t) + E(t)$, where $Z(t)$ has the same distribution as $\{x_l(t)\}_{l=7}^{500}$ and $E(t)$ is $N(0,4)$.

**Model 3.2**: Similar to Model 3.1, the response variable is generated by

$$y_i(t_{ij}) = \sum_{l=1}^{p} x_{li}(t_{ij})\beta_l(t_{ij}) + \epsilon_i(t_{ij})$$

where time points are taken from $\{1, 2, \ldots, 30\}$ with probability 0.5. The variables $(x_1, x_2, \ldots, x_p)$ are simulated as follows:

$$x_l = \frac{W_l + U}{2},\ l = 1, \ldots, p,$$

where $W_1, W_2, \ldots, W_p$ and $U$ are i.i.d. $Unif(0,1)$. The random error $\epsilon \sim N(0,1)$. The coefficient functions are designed to be

$$\begin{aligned} \beta_1(t) = 7\cos^2(\frac{t-10}{7}) + 0.1t \quad &, \quad \beta_2(t) = -0.5t, \\ \beta_3(t) = \frac{(t-15)^2}{20} \quad &, \quad \beta_4(t) = 15\sin(\frac{t+5}{3.5})\exp(-\frac{t}{30}), \\ \beta_l(t) = 0,\ \text{for } l \geq 5. & \end{aligned}$$

**Model 3.3**: The response variable is generated by the same way as in Models 3.1 and 3.2, except that the time points are taken from $\{1, 2, \ldots, 30\}$ with probability 0.3. Only the first six coefficient functions are nonzero:

$$\beta_1 = \beta_3 = \beta_5 = 1 \text{ and } \beta_2 = \beta_4 = \beta_6 = -1.$$

| | IVIS | | | | | VIS+gSCAD | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | SA | ES | MS | OS1 | OS2+ | SA | ES | MS | OS1 | OS2+ |
| Model 3.1 | 100% | 99% | 0% | 0% | 1% | 88% | 88% | 12% | 0% | 0% |
| Model 3.2 | 100% | 78% | 0% | 17% | 5% | 0% | 0% | 100% | 0% | 0% |
| Model 3.3 | 100% | 88% | 0% | 11% | 1% | 100% | 100% | 0% | 0% | 0% |

Table 3.1: Variable selection performance of IVIS and VIS+gSCAD.

Let $\{x_k(t)\}_{l=1}^{450}$ be i.i.d Gaussian process with mean zero and variance one and

$$x_l(t) = \sum_{j=1}^{6} x_j(t)(-1)^{(j+1)}/5 + \sqrt{1 - \frac{6}{25}}\epsilon_l(t), \ k = 451, \ldots, 500,$$

where $\{\epsilon_l(t)\}_{k=451}^{500}$ are Gaussian with mean zero and variance three. The random error $\epsilon(t) \sim N(0, 1)$. Note that this model is in fact a parametric linear model. We want to use this model to examine whether doing nonparametric screening and estimation does much worse than using the parametric screening and estimation.

As shown in Table 3.1 we used several quantities to measure the variable selection performance. "SA" is the percentage of occasions on which all the correct variables are included in the selected model; "ES" is the frequency of exactly selecting all true variables and nothing else; "MS" is the percentage of occasions on which some correct variables are missed; "OS1" is the frequency of exactly 1 false variable is selected and "OS2" is the frequency of selecting 2 or more false variables. We see that VIS+gSCAD tends to be too greedy in Models 3.1 and 3.2, missing some true variables. But IVIS fixes this problem nicely: it always selects all true variables. 0% for "MS" indicates extremely low false negative rates for all 3 model using IVIS. On the other hand, small values for "OS1" and "OS2+" shows low false positive rates of variable selection. Overall, IVIS has very good variable selection performance.

We note that the computation time heavily depends on the implementation of group SCAD algorithm, because the screening process is pretty fast. So if IVIS can

| Iterations | 2 | 3 | 4 | 5 | >5 |
|---|---|---|---|---|---|
| Model 3.1 | 43 | 40 | 16 | 0 | 1 |
| Model 3.2 | 1 | 23 | 69 | 7 | 0 |
| Model 3.3 | 88 | 11 | 0 | 1 | 0 |

Table 3.2: The number of iterations needed to achieve stabilization in IVIS.

achieve stability within a few iterations, the computation will be reasonably fast. Here we report the number of iterations in Table 3.2 needed to achieve stability. Obviously, 2 is the minimum number of iterations needed to confirm that variables selected in current iteration match the previous selection, and we observe that almost all the time stability is reached within 5 iterations. Interestingly, in Model 3.3 which has 6 constant coefficients, IVIS converges within 3 steps 99% of times.

In our simulations of 3 different models, after the first step (VIS+groupSCAD) usually majority of true variables will be selected, and occasionally a few false variables may get in, which could bring false positive errors eventually. The missing true variables will be selected in later forward screening steps, and stability will be achieved very quickly. Even when a couple of false variables are selected, the accuracy of coefficient estimation for true variables is not compromised.

Two quantities are used to measure the estimation accuracy of IVIS in Table 3.3. For each coefficient function estimator $\hat{\beta}_j(t)$, we define its integrated mean squared error (IMSE) as $\int(\hat{\beta}_j(t) - \beta_j(t))^2 dt$, which can be computed by numeric integration. We also report the relative IMSE (RIMSE) which is defined as the ratio of the IMSE of an estimator relative to the IMSE of the oracle estimator. Note that the oracle estimator knows the true variables and only needs to estimate the true coefficient functions. In Models 3.1 and 3.2, the oracle estimator uses 5 and 10 B-spline basis functions to estimate each true coefficient function, just like IVIS. The RIMSE of IVIS is very close to 1 in these two models, which is expected given the variable selection

| | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ |
|---|---|---|---|---|---|---|
| Model 3.1 (IMSE) | 3.29 | 22.41 | 1.83 | 0.69 | 0.76 | 0.47 |
| | (1.18) | (0.71) | (0.47) | (0.44) | (0.44) | (0.26) |
| Model 3.1 (RIMSE) | 1.14 | 1.00 | 1.00 | 1.02 | 1.01 | 1.27 |
| | (0.27) | (0.00) | (0.01) | (0.18) | (0.05) | (0.54) |
| Model 3.2 (IMSE) | 4.27 | 3.75 | 3.84 | 4.24 | NA | NA |
| | (2.29) | (1.70) | (2.28) | (1.69) | NA | NA |
| Model 3.2 (RIMSE) | 1.19 | 1.10 | 1.10 | 1.05 | NA | NA |
| | (0.64) | (0.35) | (0.39) | (0.18) | NA | NA |
| Model 3.3 (IMSE) | 0.16 | 0.21 | 0.25 | 0.23 | 0.17 | 0.16 |
| | (0.22) | (0.27) | (0.43) | (0.34) | (0.22) | (0.23) |
| Model 3.3 (RIMSE) | 2.17 | 2.61 | 3.14 | 2.94 | 2.22 | 1.96 |
| | (3.15) | (4.84) | (5.54) | (5.24) | (3.70) | (3.69) |

Table 3.3: IMSE and relative IMSE for estimating true $\beta$'s.

results in Table 3.1. In Model 3.3 the RIMSE of IVIS is larger than 2, although IVIS still does very good variable selection. This disparity can be explained by the fact that in Model 3.3 we actually allow the oracle to use the knowledge that the true coefficient functions are constant so that the oracle estimator directly estimates these constants and did not use 5 B-splines basis functions.

In Table 3.4 we further compare the prediction accuracy of the oracle estimator, IVIS and VIS+gSCAD. The prediction errors were computed on an independent test dataset. We see that IVIS and the oracle have nearly identical prediction performance in all three models. It is also interesting to see that in Models 3.1 and 3.3 VIS+gSCAD performs very similarly to IVIS and the oracle estimator but it has significantly worse prediction in Model 3.2 which is consistent with its unsatisfactory variable selection performance as shown in Table 3.1.

In Figure 3.1–Figure 3.3 we depict the estimated coefficient functions by IVIS compared to the ground truth.

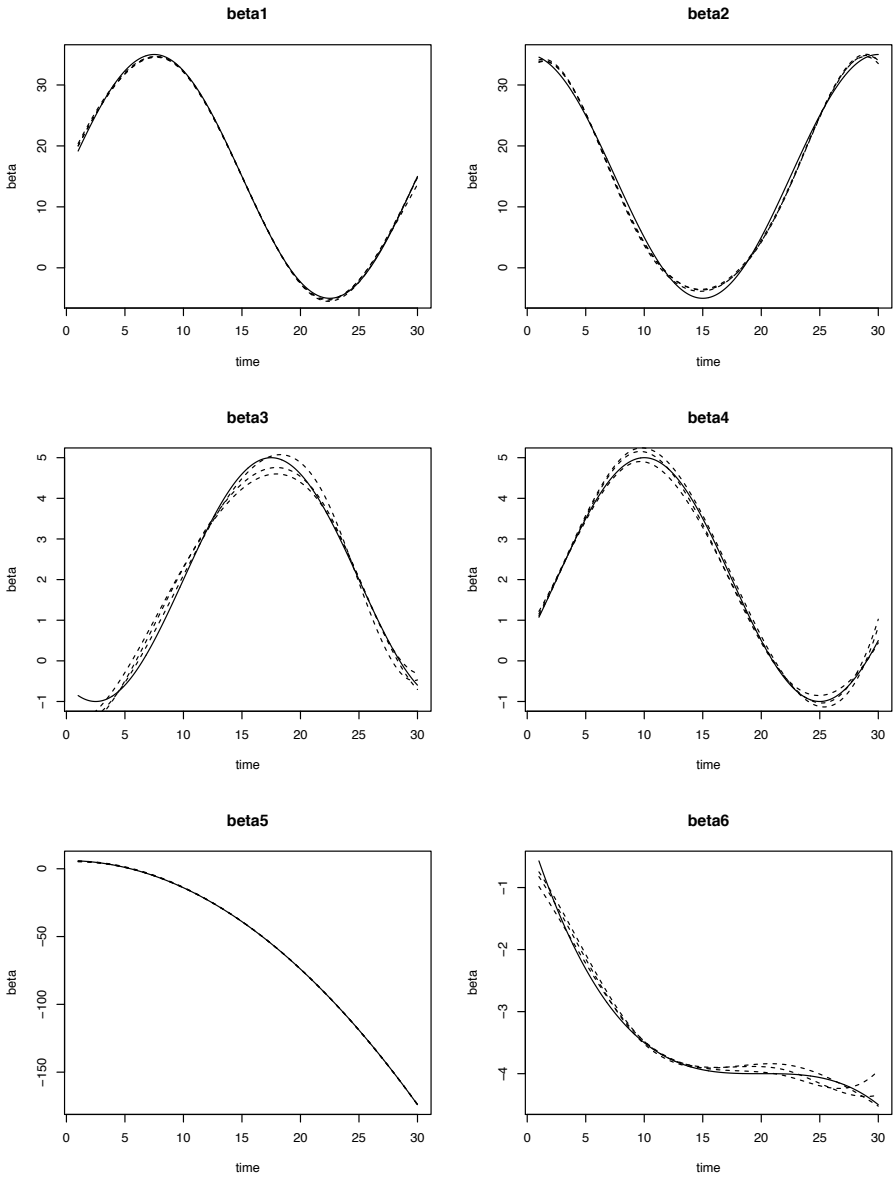



Figure 3.1: Model 3.1: real line is true $\beta$ curve, three dash lines are estimated curves by IVIS in three runs .

| | Oracle | IVIS | VIS+gSCAD |
|---|---|---|---|
| Mode 3.1 | 8.93 | 8.95 | 9.75 |
| | (0.28) | (0.29) | (2.20) |
| Model 3.2 | 1.04 | 1.05 | 3.58 |
| | (0.03) | (0.03) | (0.71) |
| Model 3.3 | 1.01 | 1.04 | 1.04 |
| | (0.03) | (0.04) | (0.04) |

Table 3.4: Prediction error comparison.

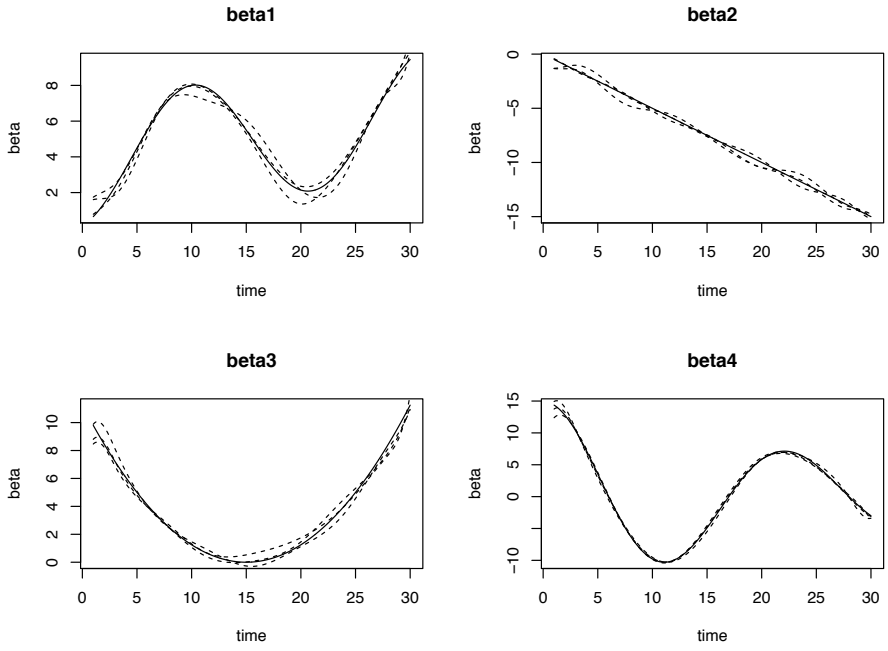



Figure 3.2: Model 3.2: real line is true $\beta$ curve, three dash lines are estimated curves by IVIS in three runs.

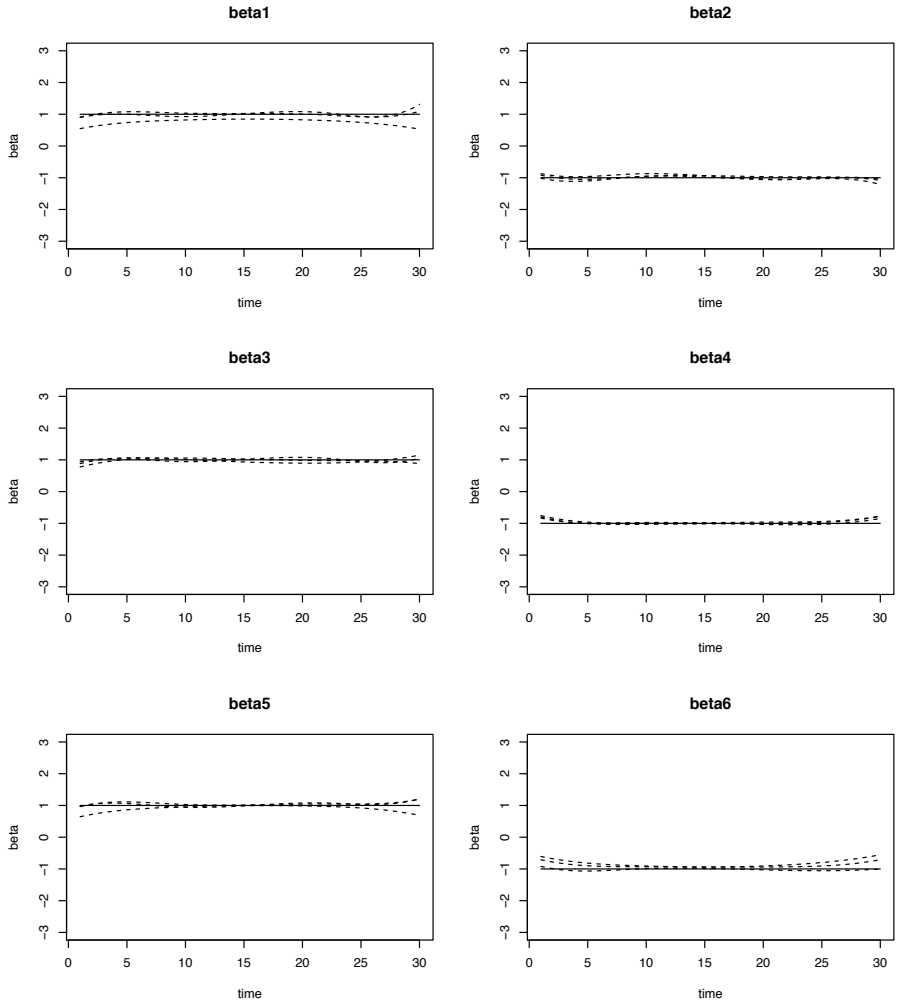



Figure 3.3: Model 3.3: real line is true $\beta$ curve, three dash lines are estimated curves by IVIS in three runs.

### 3.5.2 Real data

The experiment by Spellman et al. (1998) recorded genome-wide mRNA levels for 6178 yeast ORFs (open reading frames) simultaneously over approximately two cell cycle periods at 7-minutes intervals for 119 minutes with a total of 18 time points. The cell cycle is an ordered set of events and the cell cycle process is commonly divided into G1-S-G2-M stages, where the G1 stage stands for "GAP 1", the S stage stands for "Synthesis" during which DNA replication occurs, the G2 stage stands for "GAP 2" and the M stage stands for "mitosis" during which nuclear and cytoplasmic division occur. The experiment identified approximately 800 genes which vary in a periodic fashion during the yeast cell cycle, however little was known about the regulation of most of these genes. Transcription factors (TFs) play critical roles in gene expression regulation. A transcription factor is a protein that binds to specific DNA sequences, thereby controlling the flow of genetic information from DNA to mRNA.

We apply our IVIS method to investigate the transcription factors (TFs) involved in the yeast cell cycle. We consider 240 genes without missing values, and there are 96 transcriptional factors with at least one nonzero binding probability. Let $y_i(t_j)$ denote the log-expression level for gene $i$ at time point $t_j$ during the cell cycle process, and then the chromatin immunoprecipitation (ChIP-chip) data of Lee et al. (2002) is used to derive the binding probabilities. This dataset has been analyzed by Wang et al. (2008b) and Wei et al. (2011) who used a varying coefficient model defined as follows to link the binding probabilities to the log-gene expression levels

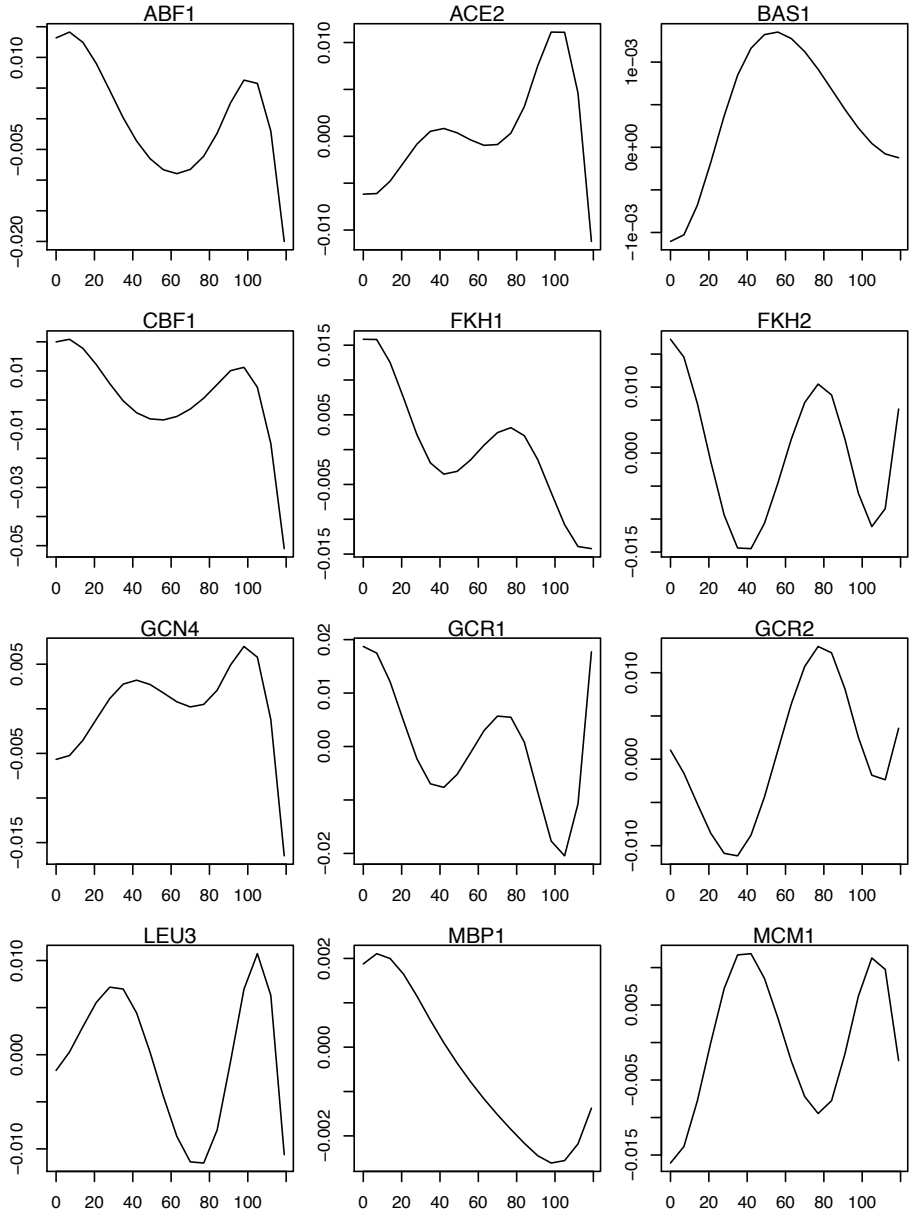
$$y_i(t_j) = \mu(t_j) + \sum_{l=1}^{96} x_{i,l}\beta_l(t_j) + \epsilon_i(t_j).$$

This dataset has a moderately high dimension, $p = 96$ with $n = 240$. We first used gSCAD to obtain a sparse estimator of the varying coefficient model. Figure 3.4

shows the estimated $\beta$ curves over time for 21 known yeast TFs.

In order to demonstrate the performance of IVIS in a really high dimensional case, we added extra 384 pure noise variables to the original data to have a total 480 variables. Now we can test IVIS in the high-dimensional setting, as the total number of variables double the number of subjects. These 384 noise variables for each subject are independently sampled from the standard normal distribution. We applied IVIS to the augmented dataset and repeated the process 100 times. Among the 21 known important TFs, IVIS on average identified 14 TFs with stand deviation 0.84. Figure 3.5 shows the estimated $\beta$ curves of 14 TFs identified by IVIS in one trial. Although the curves are not exactly the same as those in Figure 3.4, but very similar patterns are shown for most of the 14 TFs. We compare the estimated transcriptional effects side-by-side for 5 TFs in Figure 3.6.

We also compare the prediction error of IVIS with estimation for the full model without variable selection. Five-fold cross validation is used to calculate prediction error. We run 100 replicates for each method. In Table 3.5, we record the prediction error of SCAD, IVIS and that of no variable selection with and without adding 384 noise variables. We can see that, IVIS significantly outperforms the estimation without variable selection in terms of prediction when noise variables are added. Interestingly the performance of IVIS in high dimensional setting is very close to the SCAD in much lower dimensional setting. The prediction error of IVIS is also much lower than that of the full model without variable selection and without noise variables. This demonstrates the prediction power of IVIS in real data with very high dimensional covariates.




(Plot continued on the next page)

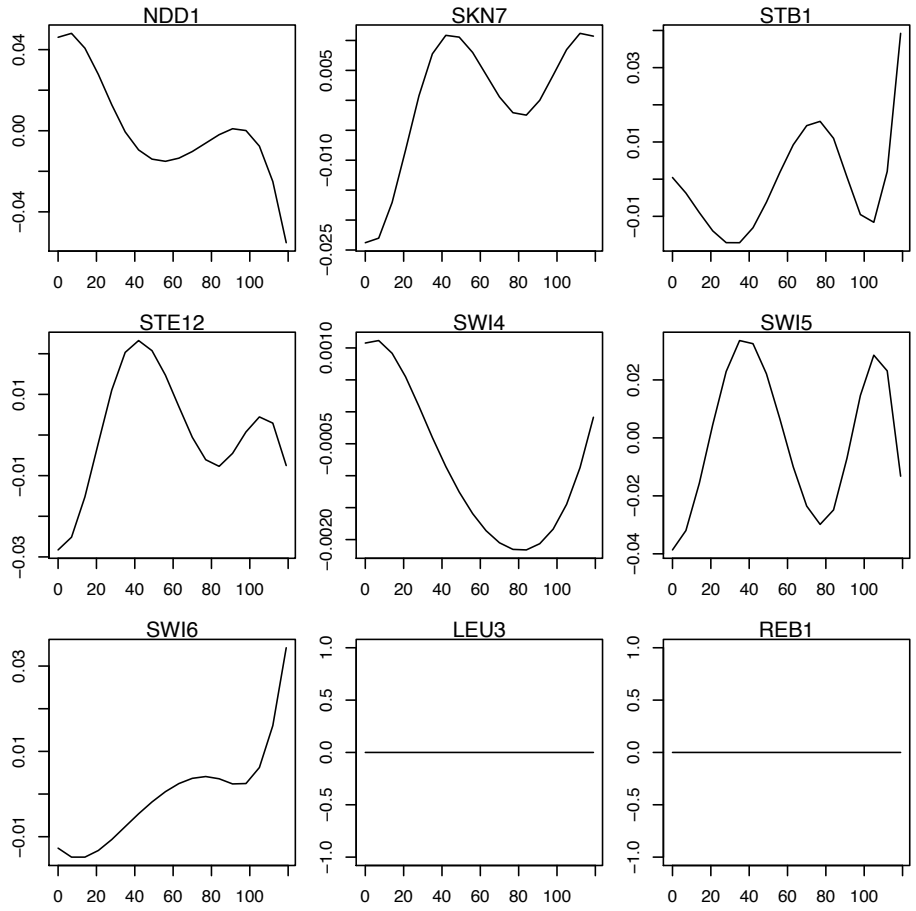
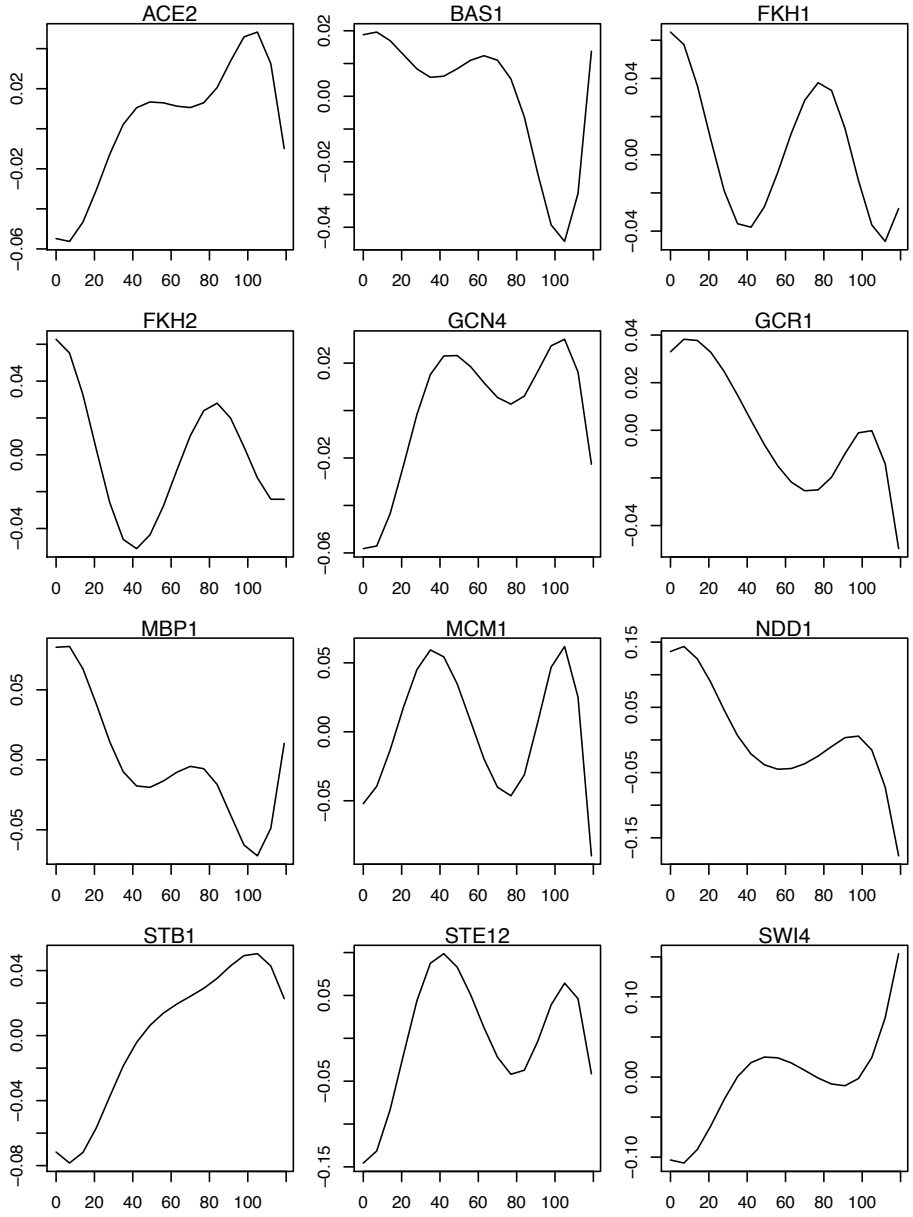



Figure 3.4: Estimated time-varying transcriptional effects for 21 known yeast TFs related to cell cycle process. LEU3 and REB1 are not selected, so there are no estimates for these two.

| | w/o noise variables | | w/ 384 noise variables | |
|---|---|---|---|---|
| | SCAD | no variable selection | IVIS | no variable selection |
| prediction error | 0.225 | 0.507 | 0.294 | 0.782 |
| (standard deviation) | (0.004) | (0.019) | (0.017) | (0.037) |

Table 3.5: Prediction error comparison.




(Plot continued on the next page)

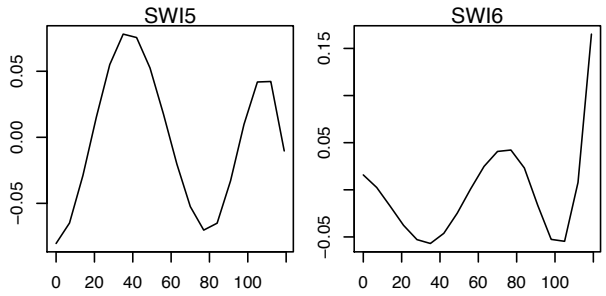



Figure 3.5: Estimated time-varying transcriptional effects for 14 TFs identified by IVIS on an augmented higher dimension dataset.

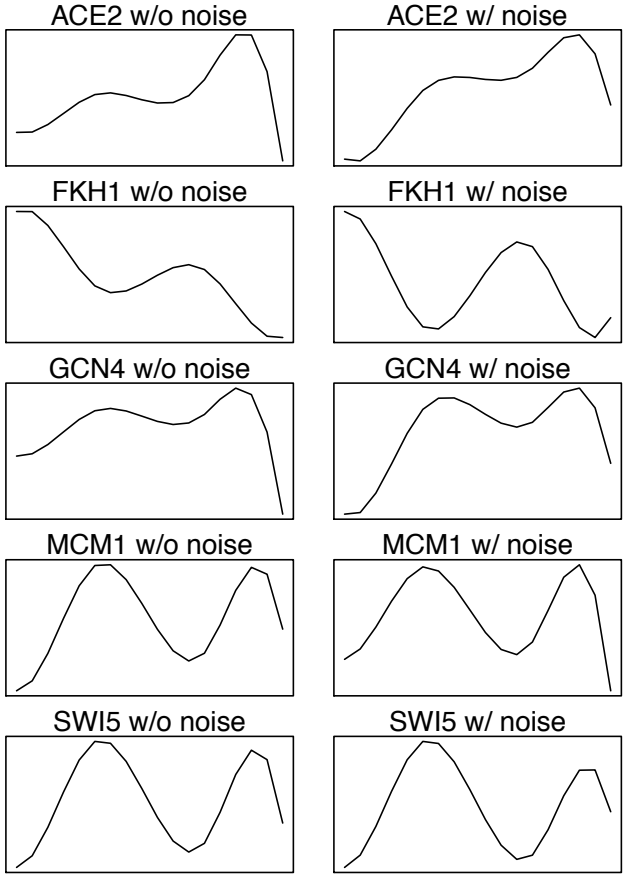



Figure 3.6: Comparison of estimated time-varying transcriptional effects for 5 TFs identified by SCAD w/o noise (left column) and IVIS w/ 384 noise (right column).

## 3.6 Conclusion

we have studied VIS for variable screening in varying-coefficient models and established its sure screening property. We have further proposed IVIS for fitting varying-coefficient models in ultra-high dimensions by iterating between a greedy conditional VIS step and a gSCAD penalized fitting step. The proposed methodology is well supported by numeric examples.

We now make a remark on the similarity and difference between varying-coefficient independence screening (VIS) in this chapter and nonparametric independence screening (NIS) in Fan et al. (2011). Both methods are flexible extensions of the marginal correlation ranking idea in Fan and Lv (2008) and both methods use B-splines to compute their marginal ranking statistics. The marginal ranking statistics is the fundamental quantity in a marginal screening method. There two methods use different marginal statistics, as they are designed for different data structure. Specifically, NIS is applied to the additive models, while VIS is applied to varying-coefficient models. Because their targeted models are different, VIS and NIS use very different marginal ranking statistics. NIS uses marginal correlation of the response variable and the estimated marginal nonparametric regression function. VIS uses $\frac{1}{|\mathcal{T}|}\int_{\mathcal{T}} \hat{\beta}_j(t)^2 dt$ to rank the $j$-th covariate, where $\hat{\beta}_j(t)$ is the estimated marginal coefficient function of the $j$-th covariate. This can be viewed as the integrated marginal correlation of the time-varying response variable and the $j$-th time-varying covariate projected onto the B-Spline space. More efforts were taken in VIS to analyze the influence of longitudinal observations on the dimensionality that VIS can handle, such as the number of observations points within subjects and the weight functions.

# Chapter 4

# Extended Factor Analysis

## 4.1 Introduction

A factor analysis model Gorsuch (1983), Johnson and Wichern (2007) is a traditional and very popular statistical model. One weakness of the factor model, however, is the restriction of orthogonality of errors. In this chapter, the penalized extended factor model is introduced. The new model is more applicable due to its flexibility. It is the same as the traditional factor analysis (FA) model in that the factors are unobserved, but different in that correlated errors are assumed. The estimation is improved by imposing the lasso penalty on the errors.

Suppose we have $n$ independent and identically distributed $p$-dimensional observed vectors $Y_i's, i = 1, \ldots, n$. Without loss of generality, we can assume the mean of $Y_i$ is zero and its covariance is $\mathbf{\Sigma}(Y_i)$ . The extended factor model is given by

$$Y_i = \boldsymbol{\beta}^T S_i + \epsilon_i, \tag{4.1}$$

where $S_i$ is an *unobserved* random vector of length $q$, $\boldsymbol{\beta}$ is a $q \times p$ matrix and $\epsilon_i$ represents a $p$-dimensional random error vector whose mean is zero and covariance is $\mathbf{\Sigma}^e$. It is assumed that $S_i$'s are normally distributed, and have zero mean and identity

covariance matrix. Therefore, the covariance of $Y_i$ can be expressed as

$$\boldsymbol{\Sigma}(Y_i) = \boldsymbol{\beta}^T\boldsymbol{\beta} + \boldsymbol{\Sigma}^e$$

Observe that the model (4.1) is the same as the traditional model if $\epsilon_i$ is not correlated. The model can be represented in a matrix form. If we let the $i$-th rows of $\mathbf{Y}$, $\mathbf{S}$ and $\boldsymbol{\epsilon}$ are $Y_i$, $S_i$ and $e_i$, respectively, the model is written as

$$\mathbf{Y}_{n\times p} = \mathbf{S}_{n\times q}\boldsymbol{\beta}_{q\times p} + \boldsymbol{\epsilon}_{n\times p}$$

As in the traditional model, $\mathbf{S}$ is called the factor matrix and $\boldsymbol{\beta}$ is the factor loading matrix. As in many statistical literatures, we assume

$$S_i \sim N(0, \boldsymbol{I}_q)$$

$$Y_i \sim N(0, \boldsymbol{\beta}^T\boldsymbol{\beta} + \boldsymbol{\Sigma}^e)$$

The maximum likelihood estimation can be implemented by using the Expectation-Maximization algorithm.

$\ell_1$ penalization method was introduced to matrix analysis by many researchers. For example, the $\ell_1$ penalization to the covariance estimation of multivariate regression were proposed in Friedman et al. (2008) and Rothman et al. (2008). Also in Rothman et al. (2010b) new methodology is invented for the simultaneous penalized estimation of the regression coefficients and the covariances of error. In this chapter we apply the penalization idea to the new model and propose to fit an $\ell_1$ penalized Extended Factor Analysis (EFA) model with $\ell_1$ penalty imposed on the Inverse-Covariance-Matrix-of-Error (ICE). Graphical lasso proposed by Friedman et al. (2008) can be carried out. Due to the sparse shrinkage property of $\ell_1$ penalty, some elements

of ICE are estimated by exact zero, which indicates there are no dependence between these variables given other variables. Other penalties such as Adaptive lasso Zou (2006) and SCAD Fan and Li (2001) may be applied to the coefficient estimation.

For the following two cases, we believe that EFA will outperform traditional FA in estimation of factors and covariance of error:

- There are only a few common factors, and the $p$ variables can be divided into small groups after accounting for the common factors, where there are almost no connection between groups, but there are some strong dependence within each group. In this case, if you want to use traditional Factor Analysis, many factors are needed, and still may not be able to make good estimation.

- The dependence structure of $p$ variables are complex, - it is very hard to find a small set of factors to capture all the correlations between these $p$ variables. In real applications of Factor Analysis, people prefer fewer factors which could make interpretation easier.

In EFA, factors are still independent, since we constrain $S_i \sim N(0, \boldsymbol{I}_q)$, while $\boldsymbol{\Sigma}^e$ may not be strictly diagonal. People may be interested in factors only, or people may also be interested in dependence structure of errors. But if factors are estimated without considering dependence structure of errors, or if dependence structure of errors is estimated without taking out common factors ahead of your calculation, it is possible that the estimation is unreliable or inaccurate. On the other hand, estimated likelihood of traditional FA and EFA can be compared to see if the orthogonality assumption of error term is reasonable. We are going to show the power of EFA through simulations and real data.

The rest of this chapter is organized as follows. Section 4.2 presents the model of the $\ell_1$ penalized Extended Factor Analysis. Section 4.3 develops a generalized Expectation-Maximization algorithm to compute the factor loadings and $\ell_1$ penalized

inverse covariance matrix of errors in EFA. Six simulations models and one read data example are shown in Section 4.4 to demonstrate the power of EFA. Section 4.5 concludes this chapter with discussion.

## 4.2 Extended Factor Analysis

In this section we define the Extended Factor Analysis model. Assume the mean of $Y_i$ is zero which is done in practice by centering the data matrix. Then, under the normality assumption, the log-likelihood can be written as

$$LL(\mathbf{\Sigma}^e, \boldsymbol{\beta}) = -\frac{n}{2}\left(\log|\mathbf{\Sigma}^e + \boldsymbol{\beta}^T\boldsymbol{\beta}| + tr\left((\mathbf{\Sigma}^e + \boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\mathbf{\Sigma}^s\right)\right),$$

where $\mathbf{\Sigma}^s = \frac{1}{n}Y_iY_i^T$ is the sample covariance matrix of $Y_i$.

If the $\ell_1$ penalty is added to the log-likelihood, the penalized log-likelihood is defined by

$$\begin{aligned} LL_p(\mathbf{\Theta}, \boldsymbol{\beta}) &= -\frac{n}{2}\log|\mathbf{\Theta}^{-1} + \boldsymbol{\beta}^T\boldsymbol{\beta}| - \frac{n}{2}tr[\mathbf{\Sigma}^s(\mathbf{\Theta}^{-1} + \boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}] \\ &\quad -\frac{1}{2}\sum_{j=1}^{p}\sum_{j'=1}^{p}\lambda|\theta_{jj'}| \end{aligned} \tag{4.2}$$

where ICE $= \mathbf{\Theta} = (\mathbf{\Sigma}^e)^{-1}$ and $\theta_{ij}$ is the $(i,j)$th entry of $\mathbf{\Theta}$. In (4.2) $\ell_1$ penalty is added to ICE. The sparse estimator, denoted by $(\widehat{\mathbf{\Theta}}, \widehat{\boldsymbol{\beta}})$, is then defined as

$$\begin{aligned} (\widehat{\mathbf{\Theta}}, \widehat{\boldsymbol{\beta}}) &= \arg\min \log|\mathbf{\Theta}^{-1} + \boldsymbol{\beta}^T\boldsymbol{\beta}| + tr[\mathbf{\Sigma}^s(\mathbf{\Theta}^{-1} + \boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}] \\ &\quad +\frac{\lambda}{n}\sum_{j=1}^{p}\sum_{j'=1}^{p}|\theta_{jj'}|. \end{aligned}$$

## 4.3 Algorithm

Rubin and Thayer (1982) derived an E-M algorithm for computing the MLE for the factor model. We apply a lasso penalty to Inverse-Covariance-Matrix-of-Error(ICE) and derive a new E-M algorithm for computing the factor loadings and $\ell_1$ penalized errors.

The common factor $\mathbf{S}$ is unobserved data so we take $\mathbf{S}$ as the missing data in the E-M algorithm. By $S_i \sim N(0, \boldsymbol{I}_q)$ we write down the joint likelihood of $(\mathbf{Y}, \mathbf{S})$ as

$$\begin{aligned} L_{\mathbf{Y},\mathbf{S}}(\boldsymbol{\Theta}, \boldsymbol{\beta}) &= P(\mathbf{Y}|\mathbf{S}, \boldsymbol{\Theta}, \boldsymbol{\beta}) \times P(\mathbf{S}|\boldsymbol{\Theta}, \boldsymbol{\beta}) \\ &= [1/(2\pi)|\boldsymbol{\Theta}|]^{n/2} \exp\{-\frac{1}{2} tr[(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{S}\boldsymbol{\beta})\boldsymbol{\Theta}]\} \\ &\quad \times [2\pi|\mathbf{I}|]^{-n/2} \exp[-\frac{1}{2} tr(\mathbf{S}\mathbf{S}^T)] \end{aligned}$$

EM algorithms iterate between the E-step and the M-step. Let $(\boldsymbol{\beta}_{(k)}, \boldsymbol{\Theta}_{(k)})$ be the estimates of step $k$.

### 4.3.1 E-step

At the E-step, we need compute the conditional expectation of the log-likelihood given $\mathbf{Y}$ and $(\boldsymbol{\beta}_{(k)}, \boldsymbol{\Theta}_{(k)})$. Let $CELL_{(k)}$ be the conditional expectation of the log-likelihood.

We have

$$\begin{aligned} CELL_{(k)}(\boldsymbol{\beta},\boldsymbol{\Theta}) &= \mathbb{E}(\log P(\mathbf{S},\mathbf{Y}|\boldsymbol{\beta},\boldsymbol{\Theta})|\mathbf{Y},\boldsymbol{\beta}_{(k)},\boldsymbol{\Theta}_{(k)}) \\ &= \frac{n}{2}\log|\boldsymbol{\Theta}| - \frac{1}{2}\mathbb{E}\{tr[(\mathbf{Y}-\mathbf{S}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{S}\boldsymbol{\beta})\boldsymbol{\Theta}]\} + \text{const} \\ &= \frac{n}{2}\log|\boldsymbol{\Theta}| - \frac{1}{2}\{tr[\mathbf{Y}^T\mathbf{Y}\boldsymbol{\Theta} - 2\mathbf{Y}^T\mathbb{E}(\mathbf{S}|\mathbf{Y},\boldsymbol{\beta}_{(k)},\boldsymbol{\Theta}_{(k)})\boldsymbol{\beta}\boldsymbol{\Theta} \\ &\quad +\boldsymbol{\beta}^T\mathbb{E}(\mathbf{S}^T\mathbf{S}|\mathbf{Y},\boldsymbol{\beta}_{(k)},\boldsymbol{\Theta}_{(k)})\boldsymbol{\beta}\boldsymbol{\Theta}]\} \\ &\quad -\frac{1}{2}tr\{\mathbb{E}(\mathbf{S}^T\mathbf{S}|\mathbf{Y},\boldsymbol{\beta}_{(k)},\boldsymbol{\Theta}_{(k)})\} + \text{const} \end{aligned}$$

Since

$$S_i^T|Y_i^T,\boldsymbol{\beta},\boldsymbol{\Theta} \sim N(\boldsymbol{\beta}(\boldsymbol{\Theta}^{-1}+\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}Y_i^T, \mathbf{I}-\boldsymbol{\beta}(\boldsymbol{\Theta}^{-1}+\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T)$$

we can write

$$\begin{aligned} \mathbb{E}(S_i^T|Y_i^T,\boldsymbol{\beta}_{(k)},\boldsymbol{\Theta}_{(k)}) &= \boldsymbol{\delta}_{(k)}^T Y_i^T \\ \text{Var}(S_i^T|Y_i^T,\boldsymbol{\beta}_{(k)},\boldsymbol{\Theta}_{(k)}) &= \boldsymbol{\Delta}_{(k)} \\ \mathbb{E}(S_i^T S_i|Y_i^T,\boldsymbol{\beta}_{(k)},\boldsymbol{\Theta}_{(k)}) &= \boldsymbol{\Delta}_{(k)} + \boldsymbol{\delta}_{(k)}^T Y_i^T Y_i \boldsymbol{\delta}_{(k)} \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\delta}_{(k)} &= (\boldsymbol{\Theta}_{(k)}^{-1}+\boldsymbol{\beta}_{(k)}^T\boldsymbol{\beta}_{(k)})^{-1}\boldsymbol{\beta}_{(k)}^T, \\ \boldsymbol{\Delta}_{(k)} &= \mathbf{I}-\boldsymbol{\beta}_{(k)}(\boldsymbol{\Theta}_{(k)}^{-1}+\boldsymbol{\beta}_{(k)}^T\boldsymbol{\beta}_{(k)})^{-1}\boldsymbol{\beta}_{(k)}^T. \end{aligned}$$

then we have

$$\begin{aligned}
\mathbb{E}(\mathbf{S}|\mathbf{Y}, \boldsymbol{\beta}_{(k)}, \boldsymbol{\Theta}_{(k)}) &= \mathbf{Y}\boldsymbol{\delta} \\
\mathbb{E}(\mathbf{S}^T\mathbf{S}|\mathbf{Y}, \boldsymbol{\beta}_{(k)}, \boldsymbol{\Theta}_{(k)}) &= \sum_{i=1}^{n} \mathbb{E}(S_i^T S_i | Y_i^T, \boldsymbol{\beta}_{(k)}, \boldsymbol{\Theta}_{(k)}) \\
&= n(\boldsymbol{\Delta} + \boldsymbol{\delta}^T \boldsymbol{\Sigma}^s \boldsymbol{\delta})
\end{aligned}$$

So we have expression for $CELL_{(k)}(\boldsymbol{\beta}, \boldsymbol{\Theta})$:

$$\begin{aligned}
CELL_{(k)}(\boldsymbol{\beta}, \boldsymbol{\Theta}) &= \frac{n}{2}\log|\boldsymbol{\Theta}| - \frac{n}{2}tr(\boldsymbol{\Sigma}^s\boldsymbol{\Theta}) + \frac{n}{2}tr(2\boldsymbol{\Sigma}^s\boldsymbol{\delta}_{(k)}\boldsymbol{\beta}\boldsymbol{\Theta}) \\
&\quad -\frac{n}{2}tr(\boldsymbol{\beta}^T(\boldsymbol{\Delta}_{(k)} + \boldsymbol{\delta}_{(k)}^T\boldsymbol{\Sigma}^s\boldsymbol{\delta}_{(k)})\boldsymbol{\beta}\boldsymbol{\Theta}) \\
&\quad -\frac{n}{2}tr(\boldsymbol{\Delta}_{(k)} + \boldsymbol{\delta}_{(k)}^T\boldsymbol{\Sigma}^s\boldsymbol{\delta}_{(k)}) + \text{const}
\end{aligned}$$

### 4.3.2 M-step

At the M step, we maximize the so-called $R$ function defined as

$$R(\boldsymbol{\beta}, \boldsymbol{\theta}) = CELL_{(k)}(\boldsymbol{\beta}, \boldsymbol{\theta}) - \frac{1}{2}P_\lambda(\boldsymbol{\Theta}).$$

where

$$P_\lambda(\boldsymbol{\Theta}) = \lambda \sum_{j=1}^{p}\sum_{j'=1}^{p} |\theta_{jj'}|$$

However, it would take another iterative process to find the maximizer of the $R$ function. To mitigate the computation difficulty, we just find an update to increase the $R$ function rather than maximize it. This idea was introduced in the original EM paper Dempster et al. (1977). First, we compute $\boldsymbol{\beta}_{(k+1)}$ by

$$\boldsymbol{\beta}_{(k+1)} = \arg\max_{\beta}[R(\boldsymbol{\beta}, \boldsymbol{\Theta})|\boldsymbol{\Theta} = \boldsymbol{\Theta}_{(k)}] \tag{4.3}$$

which is easy to compute because $\boldsymbol{\beta}_{(k+1)}$ has a closed form. Then we find $\boldsymbol{\Theta}_{(k+1)}$ by letting

$$\boldsymbol{\Theta}_{(k+1)} = \arg\max_{\Theta}[R(\boldsymbol{\beta}, \boldsymbol{\Theta})|\boldsymbol{\beta} = \boldsymbol{\beta}_{(k+1)}]. \tag{4.4}$$

Considering the terms involving $\boldsymbol{\beta}$ in $CELL_{(k)}(\boldsymbol{\beta}, \boldsymbol{\Theta})$, and define

$$\begin{aligned} f(\boldsymbol{\beta}) &= \frac{2}{n} tr[\mathbf{Y}^T \mathbb{E}(\mathbf{S}|\mathbf{Y}, \boldsymbol{\beta}_{(k)}, \boldsymbol{\Theta}_{(k)})\boldsymbol{\beta}\boldsymbol{\Theta}] - \frac{1}{n} tr[\boldsymbol{\beta}^T \mathbb{E}(\mathbf{S}^T\mathbf{S}|\mathbf{Y}, \boldsymbol{\beta}_{(k)}, \boldsymbol{\Theta}_{(k)})\boldsymbol{\beta}\boldsymbol{\Theta}] \\ &= 2tr[\boldsymbol{\beta}\boldsymbol{\Theta}\boldsymbol{\Sigma}^s\boldsymbol{\delta}_{(k)}] - tr[\boldsymbol{\beta}\boldsymbol{\Theta}\boldsymbol{\beta}^T(\boldsymbol{\Delta}_{(k)} + \boldsymbol{\delta}_{(k)}^T\boldsymbol{\Sigma}^s\boldsymbol{\delta}_{(k)})] \end{aligned}$$

Take derivative over $\boldsymbol{\beta}$, we have

$$\mathbf{0} = 2\boldsymbol{\delta}_{(k)}^T\boldsymbol{\Sigma}^s\boldsymbol{\Theta} - 2(\boldsymbol{\Delta}_{(k)} + \boldsymbol{\delta}_{(k)}^T\boldsymbol{\Sigma}^s\boldsymbol{\delta}_{(k)})\boldsymbol{\beta}\boldsymbol{\Theta}$$

So we have closed form for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_{(k+1)} = (\boldsymbol{\Delta}_{(k)} + \boldsymbol{\delta}_{(k)}^T\boldsymbol{\Sigma}^s\boldsymbol{\delta}_{(k)})^{-1}\boldsymbol{\delta}_{(k)}^T\boldsymbol{\Sigma}^s \tag{4.5}$$

We cannot calculate $\boldsymbol{\Theta}$ directly because it doesn't have a closed form but we can see from graphical lasso or *glasso* Friedman et al. (2008) that

$$\boldsymbol{\Theta}_{(k+1)} = \arg\max_{\Theta} \log\|\boldsymbol{\Theta}\| - tr(\mathbb{E}\mathbf{S}^e\boldsymbol{\Theta}) - \frac{\lambda}{n}\sum_{j=1}^{p}\sum_{j'=1}^{p}|\theta_{jj'}|. \tag{4.6}$$

where $\mathbb{E}\mathbf{S}^e = \{\boldsymbol{\Sigma}^s - 2\boldsymbol{\Sigma}^s\boldsymbol{\delta}_{(k)}\boldsymbol{\beta}_{(k+1)} + \boldsymbol{\beta}_{(k+1)}^T(\boldsymbol{\Delta}_{(k)} + \boldsymbol{\delta}_{(k)}^T\boldsymbol{\Sigma}^s\boldsymbol{\delta}_{(k)})\boldsymbol{\beta}_{(k+1)}\}$.

The above procedure is summarized in Algorithm 1 (glasso-GEM). It is call "generalized" EM (GEM) algorithm, because in the M-step the $R$ function is a penalized conditional log-likelihood and we increase the $R$ function rather than maximize it.

Algorithm 1 can also be used to compute the penalized estimator using a general penalty function $P_\lambda(|\mathbf{\Theta}|)$. The $\ell_1$ penalty enjoys great computational advantages, for we can use *glasso* - Lasso algorithm to solve the $\ell_1$-penalized least squares problem in very efficient manners Efron et al. (2004); Friedman et al. (2008).

As a generalized E-M algorithm, Algorithm 1 enjoys a nice ascent property which is proven in the Appendix. In calculation, the glasso-GEM algorithm can start from random matrix or Factor Analysis estimation. We should also point out that the ascent property has nothing to do with the normality assumption of the data, although we interpret the objective function as penalized log-likelihood of normal data.

**Algorithm 1: glasso-GEM for Extended Factor Analysis**

Step 0. Compute $\mathbf{\Sigma}^s = \mathbf{Y}^T\mathbf{Y}/n$.

Step 1 : Set initial values for $\boldsymbol{\beta}$ and $\mathbf{\Sigma}^e$.

Step 2 : Calculate $\boldsymbol{\delta}$, $\mathbf{\Delta}$ :

$$\begin{aligned} \boldsymbol{\delta} &= (\mathbf{\Sigma}^e + \boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T \\ \mathbf{\Delta} &= \mathbf{I} - \boldsymbol{\beta}(\mathbf{\Sigma}^e + \boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^T \end{aligned}$$

Step 3 : Update the estimation of $\boldsymbol{\beta}$ and $\mathbf{\Theta}$:

(3.a) Compute $\boldsymbol{\beta}$ by (4.5).

(3.b) Compute $\mathbf{\Theta}$ by *glasso* (4.6).

Step 4 : Repeat Steps 2-3 till convergence.

### 4.3.3 Tune parameter and select number of factors $q$

When we know the tuning parameter $\lambda$ in *glasso* and number of factors $q$, glasso-GEM can be applied directly on data to estimate factor loadings and ICE. But we need to find out $\lambda$ and $q$ before glasso-GEM is used on data. Suppose we have data $\mathbf{Y}$ with size $n$, and we could split the data into two parts: training data ($Y_t$) and validation data ($Y_v$). Suppose a method $\mu$ produces an estimator $\widehat{\boldsymbol{\beta}}(\mu)$ and $\widehat{\boldsymbol{\Sigma}^e}(\mu)$ by using training data. Write

$$\boldsymbol{\Sigma}(\mu) \equiv \widehat{\boldsymbol{\beta}}(\mu)^T \widehat{\boldsymbol{\beta}}(\mu) + \widehat{\boldsymbol{\Sigma}^e}(\mu).$$

We can evaluate performance using negative log-likelihood:

$$nLL(\mu)_{\mathrm{v}} = \log(\det(\boldsymbol{\Sigma}(\mu))) + tr(\boldsymbol{\Sigma}^{-1}(\mu)\boldsymbol{\Sigma}_{\mathrm{v}}) \tag{4.7}$$

where $\boldsymbol{\Sigma}_{\mathrm{v}}$ is the sample covariance matrix computed using the validation data. For each candidate $q$, choose $\lambda$ with minimum $nLL(\mu)_{\mathrm{v}}$. After the $\lambda$'s are selected for all candidate $q$'s, choose $q$ with minimum $nLL(\mu)_{\mathrm{v}}$. With selected $\lambda$ and $q$, we could implement glasso-GEM algorithm over the full data to make the estimation of $\boldsymbol{\beta}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$.

## 4.4 Numerical Examples

In this section we use both simulated and real data to demonstrate the proposed estimators.

### 4.4.1 Simulation data

We examine the performance of the factor loadings and the lasso estimator of the errors in the situation when covariance of $\mathbf{Y}$ is not fully explained by common factors. The simulation data are generated by taking i.i.d. random vectors $Y_i$ of dimension $p$ from normal distribution with zero mean and covariance $\mathbf{\Sigma} = \boldsymbol{\beta}^T\boldsymbol{\beta} + \mathbf{\Sigma}^e = \{\sigma_{ij}\}$ where $\boldsymbol{\beta}$ is generated for each entry with independent draw from $N(0,1)$ in each replication, and $\{\theta_{ij}\} = \mathbf{\Theta} = (\mathbf{\Sigma}^e)^{-1}$. We are considering the following models:

- Model 4.1: AR(1)

$$\theta_{ij} = I(|i-j| = 0) + 0.5 \cdot I(|i-j| = 1)$$

- Model 4.2: AR(2)

$$\theta_{ij} = I(|i-j| = 0) + 0.5 \cdot I(|i-j| = 1) + 0.25 \cdot I(|i-j| = 2)$$

- Model 4.3: AR(3)

$$\theta_{ij} = I(|i-j| = 0) + 0.4 \cdot I(|i-j| = 1) + 0.2 \cdot I(|i-j| = 2) + 0.2 \cdot I(|i-j| = 3)$$

- Model 4.4:

$$\Theta = 0.1 \cdot I + 0.9 \cdot diag(A_1, A_2, \cdots, A_{p/2})$$

  where $I$ is diagonal matrix, and $A_1, A_2, \cdots, A_{p/2}$ are $2 \times 2$ matrices with all one elements.

- Model 4.5:

$$\Theta = 0.1 \cdot I + 0.9 \cdot diag(A_1, A_2, \cdots, A_{p/5})$$

where $I$ is diagonal matrix, and $A_1, A_2, \cdots, A_{p/5}$ are $5 \times 5$ matrices with all one elements.

- Model 4.6: Diagonal

$$\mathbf{\Sigma} = diag(1, 2, ..., p)$$

Within each of 50 replications in simulation, we generated a dataset of size $n = 200$, of which half is used as training data and the other half is used as validation data to tune $\lambda$ in *glasso* and number of factors $q$. After $\lambda$ and $q$ are selected, use the full dataset to estimate $\mathbf{\Sigma}(Y_i)$, $\boldsymbol{\beta}$ and $\mathbf{\Theta}$.

For each model, we make simulations with $p = 20$ and $p = 50$, and true $q$ changes from 1 to 3. The results are summarized in Table 4.2 - Table 4.13 at the end of the chapter.

First, we compare the $q$ selection: For all of the 6 models, we find that the traditional FA tends to select large $q$, and EFA has much better performance for selecting $q$. This is no surprise, since covariance matrix of error in FA is restricted to be diagonal, so more factors are needed to explain the data when the diagonalization assumption does not hold. On the other hand, when $p$ increases from 20 to 50, the $q$ selection of EFA gets better (closer to the true $q$) except for Model 4.6 which has diagonal covariance matrix of error.

Next, we compare the estimation of $\mathbf{\Sigma}(Y_i)$, $\boldsymbol{\beta}$ and $\mathbf{\Theta}$ from FA and EFA through the following 3 expressions:

$$\mathbf{\Sigma}(Y_i) \text{ error} = |\mathbf{\Sigma}(Y_i).estimation - \mathbf{\Sigma}(Y_i).true|$$

$$(\boldsymbol{\beta}^T\boldsymbol{\beta}) \text{ error} = |(\boldsymbol{\beta}^T\boldsymbol{\beta}).estimation - (\boldsymbol{\beta}^T\boldsymbol{\beta}).true|$$

$$\mathbf{\Theta} \text{ error} = |\mathbf{\Theta}.estimation - \mathbf{\Theta}.true|$$

and we consider three different matrix norms: Frobenius norm, $\ell_1$ norm and $\ell_2$ norm. The assumption of EFA is that ICE is sparse but not necessarily diagonal, so we expect significant improvement of estimation of $\mathbf{\Theta}$ will be achieved under EFA, and that is the truth for Model 4.1-4.5 under all three norms. And in Model 4.6, the advantage of EFA gets weaker, and is gone when $p$ increases from 20 to 50, but the results from EFA are not much worse than those from FA.

We hope to make the better estimation of $\mathbf{\Theta}$, but we do not want to sacrifice estimation precision on $\boldsymbol{\beta}$ to get that. In our simulation, we observe that even if we have fewer factors in EFA than in FA, the estimation of $\boldsymbol{\beta}^T\boldsymbol{\beta}$ from EFA is better in most cases, and is very close to that from FA in other cases. That means no matter your main focus is finding out common factors or predicting correlation among some variables given others, you probably will get more accurate estimation by using EFA, even if your true model has diagonal covariance matrix of error.

Last, although estimation of $\mathbf{\Sigma}(Y_i)$ is not the major concern in Factor Analysis - since other methods, regularized and non-regularized, could do a good job - we hope our method EFA do better than FA on that estimation. From the 6 tables, we show that EFA performs much better for all models under the three norms for estimating $\mathbf{\Sigma}(Y_i)$.

### 4.4.2 Cancer data

We apply the proposed penalized extended factor analysis method to analyze cancer data. The data is available at UCI Data Repository Frank and Asuncion (2010). The data contain two groups of observations: recur of cancer and non-recur of cancer after 2 years. Number of observations are 47 and 151 respectively. Ten real-valued features are recorded for each cell nucleus, and the mean, standard error, and the largest of these features are computed, resulting in 30 features. Before fitting the model, we standardized the data such that each feature will have mean zero and

| Nonrecur data | | | | | | |
|---|---|---|---|---|---|---|
| q | 0 | 1 | 2 | 3 | 4 | 5 |
| $nLL$ | -19.63 | -19.84 | -19.83 | -19.66 | -19.57 | -19.46 |
| | (0.39) | (0.37) | (0.39) | (0.40) | (0.40) | (0.42) |
| Recur data | | | | | | |
| q | 0 | 1 | 2 | 3 | 4 | 5 |
| $nLL$ | -6.80 | -8.68 | -8.53 | -8.14 | -7.66 | -7.24 |
| | (0.61) | (0.62) | (0.65) | (0.68) | (0.69) | (0.70) |

Table 4.1: Cancer data - averaged negative log-Likelihood estimation for Nonrecur and Recur.

standard deviation one.

For 151 observations of non-recur, divide them into two parts - part one has 76 and part two has the rest 75. For some $q$, estimate $\mathbf{\Sigma}(\mu)$ by using part one for a set of $\lambda$ values, and then use part two to select $\lambda$ with minimum $nLL(\mu)_v$ in Equation (4.7). $\mu$ here stands for the method EFA. For $q = 0, 1, \cdots, 6$, this process is done for 100 times, and average value of $\lambda$ is saved for each $q$. With ($q$, $\lambda$) combination, use part one data to estimate $\mathbf{\Sigma}(\mu)$, and use part two to calculate $nLL(\mu)_v$, and this procedure is done 100 times. The ($q$, $\lambda$) combination with minimum averaged $nLL(\mu)_v$ is selected. Finally the selected ($q$, $\lambda$) is used in estimation of $\widehat{\boldsymbol{\beta}}(\mu)$ and $\widehat{\mathbf{\Sigma}}(\mu)$ and $\widehat{\mathbf{\Theta}}(\mu)$ by using the full dataset.

Same procedure is applied for recur data. For both recur and non-recur data, we find that 1 factor is selected (as shown in Table 4.1 and Figure 4.1), but the estimated $\mathbf{\Theta}$ are quite different which indicates the difference between non-recur and recur groups of people. (as shown in Figure 4.2)

In the Appendix, we provide a proof of the ascent property of the glasso-GEM algorithm. Here for cancer data, we use Figure 4.3 to show the values of Penalized loglikehood function (PLL) which is defined in Appendix. It is obvious that the PLL curve is monotonically increasing over the algorithm iterations.

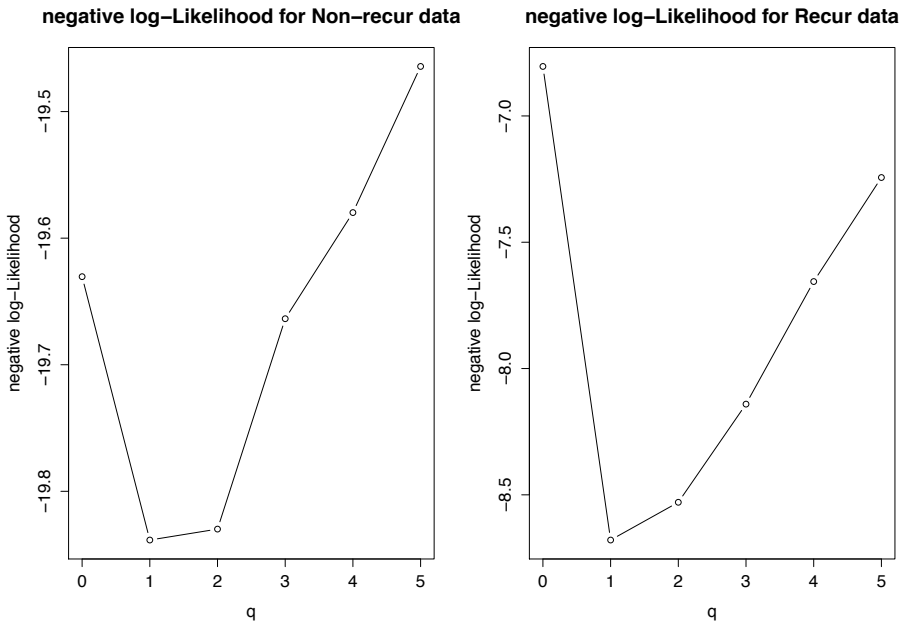



Figure 4.1: Cancer data - averaged negative log-Likelihood estimation for Nonrecur and Recur

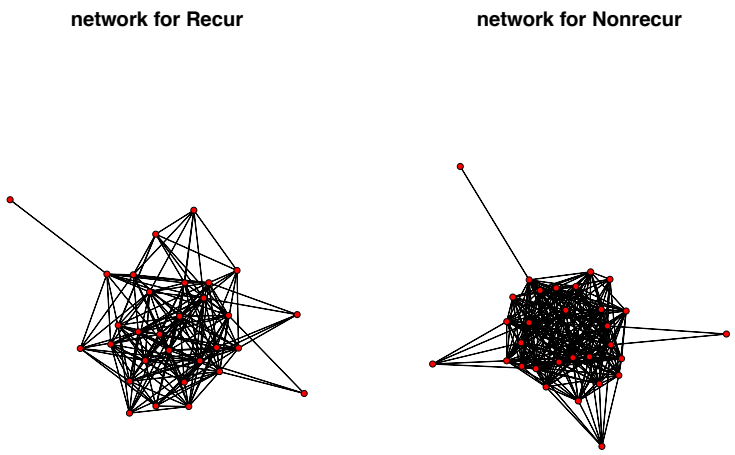



Figure 4.2: Cancer data - connections among 30 features estimated by EFA

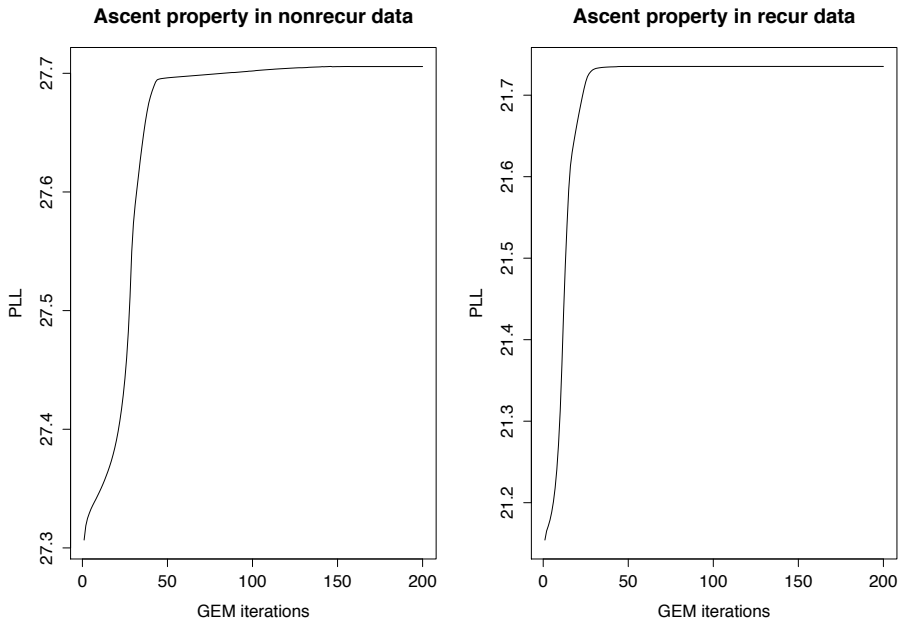



Figure 4.3: Cancer data - A numerical demonstration of glasso-GEM algorithm's ascent property

## 4.5 Discussion

We proposed Extended Factor Analysis for data with sparse but not necessarily diagonal inverse covariance matrix of error. We have shown that the $\ell_1$ penalized maximum likelihood estimation can be done via a generalized E-M algorithm by incorporating *glasso* algorithm to achieve sparse estimation of inverse covariance matrix of error. From simulations, we observe the advantages of EFA on accurately selecting number of factors and on better estimating factor loadings $\boldsymbol{\beta}$ and connections of variables through $\boldsymbol{\Theta}$. In real data analysis, we demonstrated the usefulness of EFA in estimating factors and predicting different structures of different dataset for comparison.

## 4.6 Appendix

**Ascent property of glasso-GEM Algorithm.** The E-M algorithm is usually used for maximum likelihood estimation. Green (1990) showed that the E-M algorithm can also be used for penalized maximum likelihood estimation. Here we provide a self-contained proof of the ascent property of the generalized E-M algorithm considered in Section 3.

Let the conditional density of $\mathbf{S}$ given $\mathbf{Y}$ and $\boldsymbol{\Theta}, \boldsymbol{\beta}$ be

$$f(\mathbf{S}|\boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{Y}) = \frac{f(\mathbf{Y},\mathbf{S}|\boldsymbol{\Theta}, \boldsymbol{\beta})}{f(\mathbf{Y}|\boldsymbol{\Theta}, \boldsymbol{\beta})}$$

and Penalized log-likelihood (PLL) of $\mathbf{Y}$ be

$$PLL(\boldsymbol{\Theta}, \boldsymbol{\beta}) = LL(\boldsymbol{\Theta}, \boldsymbol{\beta}) - \frac{1}{2} P_\lambda(\boldsymbol{\Theta}) \tag{4.8}$$

where $P_\lambda(\cdot)$ is the penalty function. Then we have

$$\begin{aligned}
R_{(k)}(\boldsymbol{\Theta}, \boldsymbol{\beta}) &= CELL_{(k)}(\boldsymbol{\Theta}, \boldsymbol{\beta}) - \frac{1}{2} P_\lambda(\boldsymbol{\Theta}) \\
&= \int \log f(\mathbf{S}, \mathbf{Y} | \boldsymbol{\Theta}, \boldsymbol{\beta}) f(\mathbf{S} | \mathbf{Y}, \boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k)}) d\mathbf{S} - \frac{1}{2} P_\lambda(\boldsymbol{\Theta}) \\
&= \int \log f(\mathbf{S} | \boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{Y}) f(\mathbf{S} | \mathbf{Y}, \boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k)}) d\mathbf{S} - \frac{1}{2} P_\lambda(\boldsymbol{\Theta}) + \log f(\mathbf{Y} | \boldsymbol{\Theta}, \boldsymbol{\beta}) \\
&= \int \log f(\mathbf{S} | \boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{Y}) f(\mathbf{S} | \mathbf{Y}, \boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k)}) d\mathbf{S} + PLL(\boldsymbol{\Theta}, \boldsymbol{\beta}) \\
&= PLL(\boldsymbol{\Theta}, \boldsymbol{\beta}) \\
&\quad + \int \log f(\mathbf{S} | \boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k)}, \mathbf{Y}) f(\mathbf{S} | \mathbf{Y}, \boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k)}) d\mathbf{S} \\
&\quad + \int \log \frac{f(\mathbf{S} | \boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{Y})}{f(\mathbf{S} | \boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k)}, \mathbf{Y})} f(\mathbf{S} | \mathbf{Y}, \boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k)}) d\mathbf{S}
\end{aligned}$$

Define $C_{(k)} = \int \log f(\mathbf{S} | \boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k)}, \mathbf{Y}) f(\mathbf{S} | \mathbf{Y}, \boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k)}) d\mathbf{S}$ which is constant, then it is easy to see that

$$R_{(k)}(\boldsymbol{\Theta}, \boldsymbol{\beta}) \leq PLL(\boldsymbol{\Theta}, \boldsymbol{\beta}) + C_{(k)}$$

By (4.3) and (4.4) we have

$$\begin{aligned}
R_{(k)}(\boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k+1)}) &\geq R_{(k)}(\boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k)}) \\
R_{(k)}(\boldsymbol{\Theta}_{(k+1)}, \boldsymbol{\beta}_{(k+1)}) &\geq R_{(k)}(\boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k+1)})
\end{aligned}$$

Thus, we can conclude

$$\begin{aligned}
PLL(\boldsymbol{\Theta}_{(k+1)}, \boldsymbol{\beta}_{(k+1)}) &\geq R_{(k)}(\boldsymbol{\Theta}_{(k+1)}, \boldsymbol{\beta}_{(k+1)}) - C_{(k)} \\
&\geq \boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k)} - C_{(k)} \\
&= PLL(\boldsymbol{\Theta}_{(k)}, \boldsymbol{\beta}_{(k)})
\end{aligned}$$

| Model 4.1: Estimation summary, $p = 20$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q.true | q selection | | norm | $\boldsymbol{\Sigma}(\mathbf{Y})$ error | | $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ error | | $\boldsymbol{\Theta}$ error | |
| | EFA | FA | | EFA | FA | EFA | FA | EFA | FA |
| 1 | 1.12 | 5.20 | | | | | | | |
| | (0.05) | (0.52) | | | | | | | |
| | | | F | 11.92 | 307.80 | 14.14 | 24.61 | 1.04 | 5.31 |
| | | | | (0.50) | (6.06) | (2.80) | (0.50) | (0.04) | (<0.01) |
| | | | $\ell_1$ | 17.08 | 185.13 | 21.75 | 31.20 | 0.94 | 1.99 |
| | | | | (0.77) | (2.78) | (3.75) | (0.62) | (0.05) | (<0.01) |
| | | | $\ell_2$ | 10.41 | 119.98 | 13.53 | 23.19 | 0.58 | 1.98 |
| | | | | (0.57) | (2.76) | (2.81) | (0.54) | (0.04) | (<0.01) |
| 2 | 1.86 | 5.64 | | | | | | | |
| | (0.08) | (0.45) | | | | | | | |
| | | | F | 13.38 | 349.47 | 16.01 | 19.51 | 1.15 | 5.33 |
| | | | | (0.59) | (6.96) | (2.52) | (0.40) | (0.04) | (<0.01) |
| | | | $\ell_1$ | 18.43 | 246.43 | 25.75 | 25.11 | 1.22 | 1.99 |
| | | | | (0.87) | (4.35) | (3.49) | (0.33) | (0.08) | (<0.01) |
| | | | $\ell_2$ | 11.57 | 165.67 | 15.03 | 15.47 | 0.68 | 1.98 |
| | | | | (0.67) | (4.32) | (2.56) | (0.51) | (0.04) | (<0.01) |
| 3 | 2.92 | 6.36 | | | | | | | |
| | (0.10) | (0.51) | | | | | | | |
| | | | F | 16.01 | 521.81 | 23.00 | 35.55 | 1.40 | 5.36 |
| | | | | (0.66) | (10.19) | (1.83) | (0.11) | (0.04) | (<0.01) |
| | | | $\ell_1$ | 21.78 | 340.05 | 33.61 | 48.44 | 1.48 | 2.00 |
| | | | | (1.00) | (6.22) | (2.45) | (0.25) | (0.08) | (<0.01) |
| | | | $\ell_2$ | 13.10 | 257.42 | 20.15 | 32.64 | 0.85 | 1.98 |
| | | | | (0.75) | (6.17) | (1.85) | (0.10) | (0.04) | (<0.01) |

Table 4.2: Comparing estimation results from Factor Analysis and Extended Factor Analysis in Model 4.1 for $p$=20.

| Model 4.1: Estimation summary, $p = 50$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q.true | q selection | | norm | $\mathbf{\Sigma}(\mathbf{Y})$ error | | $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ error | | $\mathbf{\Theta}$ error | |
| | EFA | FA | | EFA | FA | EFA | FA | EFA | FA |
| 1 | 1.00 | 6.82 | | | | | | | |
| | (<0.01) | (0.45) | | | | | | | |
| | | | F | 72.20 | 2986.1 | 24.12 | 142.29 | 1.73 | 8.56 |
| | | | | (3.25) | (64.61) | (1.80) | (2.94) | (0.02) | (<0.01) |
| | | | $\ell_1$ | 108.43 | 1299.9 | 46.36 | 182.72 | 1.06 | 2.00 |
| | | | | (5.47) | (22.54) | (3.65) | (3.63) | (0.01) | (<0.01) |
| | | | $\ell_2$ | 64.40 | 858.19 | 21.81 | 140.18 | 0.58 | 2.00 |
| | | | | (3.65) | (25.00) | (1.81) | (2.99) | (0.01) | (<0.01) |
| 2 | 2.00 | 5.46 | | | | | | | |
| | (<0.01) | (0.51) | | | | | | | |
| | | | F | 76.60 | 3389.7 | 37.41 | 122.60 | 1.79 | 8.58 |
| | | | | (3.03) | (67.47) | (2.00) | (2.35) | (0.02) | (<0.01) |
| | | | $\ell_1$ | 114.45 | 1455.8 | 59.00 | 156.13 | 1.07 | 2.00 |
| | | | | (5.21) | (22.58) | (3.77) | (2.15) | (0.01) | (<0.01) |
| | | | $\ell_2$ | 65.81 | 961.76 | 29.80 | 97.35 | 0.60 | 2.00 |
| | | | | (3.51) | (22.76) | (2.07) | (2.79) | (0.01) | (<0.01) |
| 3 | 3.05 | 7.50 | | | | | | | |
| | (0.05) | (0.72) | | | | | | | |
| | | | F | 80.20 | 3803.1 | 46.71 | 105.79 | 1.83 | 8.58 |
| | | | | (4.39) | (112.50) | (4.03) | (2.47) | (0.03) | (<0.01) |
| | | | $\ell_1$ | 127.00 | 1738.5 | 78.81 | 149.06 | 1.07 | 2.00 |
| | | | | (7.57) | (48.53) | (8.10) | (2.00) | (0.02) | (<0.01) |
| | | | $\ell_2$ | 70.26 | 1197.2 | 38.68 | 77.30 | 0.62 | 2.00 |
| | | | | (4.95) | (48.57) | (4.34) | (2.14) | (0.01) | (<0.01) |

Table 4.3: Comparing estimation results from Factor Analysis and Extended Factor Analysis in Model 4.1 for $p$=50.

| Model 4.2: Estimation summary, $p = 20$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q.true | q selection | | norm | $\boldsymbol{\Sigma}(\mathbf{Y})$ error | | $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ error | | $\boldsymbol{\Theta}$ error | |
| | EFA | FA | | EFA | FA | EFA | FA | EFA | FA |
| 1 | 1.28 | 8.46 | | | | | | | |
| | (0.10) | (0.33) | | | | | | | |
| | | | F | 3.20 | 39.33 | 5.82 | 6.73 | 1.67 | 4.76 |
| | | | | (0.06) | (1.18) | (0.45) | (0.09) | (0.04) | (0.01) |
| | | | $\ell_1$ | 4.67 | 48.12 | 12.04 | 14.80 | 1.72 | 2.36 |
| | | | | (0.18) | (1.14) | (1.06) | (0.20) | (0.07) | (<0.01) |
| | | | $\ell_2$ | 2.08 | 37.38 | 5.36 | 6.64 | 0.95 | 2.21 |
| | | | | (0.08) | (1.22) | (0.46) | (0.09) | (0.04) | (<0.01) |
| 2 | 1.48 | 8.12 | | | | | | | |
| | (0.11) | (0.40) | | | | | | | |
| | | | F | 3.95 | 61.43 | 9.22 | 11.44 | 1.73 | 4.98 |
| | | | | (0.08) | (1.62) | (0.48) | (0.10) | (0.03) | (<0.01) |
| | | | $\ell_1$ | 5.57 | 69.20 | 19.59 | 20.37 | 2.07 | 2.46 |
| | | | | (0.20) | (1.68) | (1.24) | (0.23) | (0.06) | (<0.01) |
| | | | $\ell_2$ | 2.59 | 53.78 | 8.73 | 9.79 | 1.01 | 2.28 |
| | | | | (0.09) | (1.73) | (0.53) | (0.11) | (0.02) | (<0.01) |
| 3 | 1.70 | 6.76 | | | | | | | |
| | (0.15) | (0.41) | | | | | | | |
| | | | F | 6.87 | 115.39 | 22.41 | 25.90 | 2.12 | 5.31 |
| | | | | (0.20) | (2.14) | (1.33) | (0.20) | (0.02) | (<0.01) |
| | | | $\ell_1$ | 9.79 | 90.76 | 34.75 | 34.72 | 2.17 | 2.48 |
| | | | | (0.34) | (1.59) | (1.84) | (0.27) | (0.04) | (<0.01) |
| | | | $\ell_2$ | 5.34 | 66.12 | 20.72 | 21.71 | 1.22 | 2.37 |
| | | | | (0.23) | (1.62) | (1.33) | (0.23) | (0.02) | (<0.01) |

Table 4.4: Comparing estimation results from Factor Analysis and Extended Factor Analysis in Model 4.2 for $p$=20.

| Model 4.2: Estimation summary, $p = 50$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q.true | q selection | | norm | $\boldsymbol{\Sigma}(\mathbf{Y})$ error | | $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ error | | $\boldsymbol{\Theta}$ error | |
| | EFA | FA | | EFA | FA | EFA | FA | EFA | FA |
| 1 | 1.02 | 10.98 | | | | | | | |
| | (0.02) | (0.17) | | | | | | | |
| | | | F | 8.33 | 62.17 | 6.73 | 10.60 | 2.78 | 7.74 |
| | | | | (0.17) | (1.72) | (0.21) | (0.34) | (0.03) | (0.01) |
| | | | $\ell_1$ | 13.20 | 64.75 | 12.88 | 21.86 | 1.77 | 2.48 |
| | | | | (0.55) | (1.05) | (0.57) | (0.79) | (0.02) | (<0.01) |
| | | | $\ell_2$ | 5.86 | 43.80 | 5.93 | 10.15 | 0.98 | 2.29 |
| | | | | (0.25) | (1.59) | (0.25) | (0.37) | (0.01) | (<0.01) |
| 2 | 2.00 | 8.54 | | | | | | | |
| | (0.04) | (0.39) | | | | | | | |
| | | | F | 13.85 | 167.87 | 14.58 | 25.82 | 2.93 | 8.26 |
| | | | | (0.32) | (2.96) | (0.99) | (0.52) | (0.03) | (<0.01) |
| | | | $\ell_1$ | 21.49 | 120.63 | 23.70 | 41.25 | 1.85 | 2.49 |
| | | | | (0.77) | (1.26) | (1.76) | (0.97) | (0.03) | (<0.01) |
| | | | $\ell_2$ | 10.47 | 86.03 | 12.02 | 22.49 | 1.05 | 2.37 |
| | | | | (0.42) | (2.10) | (1.05) | (0.56) | (0.01) | (<0.01) |
| 3 | 2.84 | 7.88 | | | | | | | |
| | (0.09) | (0.43) | | | | | | | |
| | | | F | 16.60 | 260.76 | 21.88 | 38.03 | 3.11 | 8.39 |
| | | | | (0.42) | (4.42) | (2.31) | (0.50) | (0.04) | (<0.01) |
| | | | $\ell_1$ | 26.67 | 170.99 | 38.80 | 60.90 | 2.07 | 2.49 |
| | | | | (0.97) | (2.36) | (4.36) | (0.96) | (0.05) | (<0.01) |
| | | | $\ell_2$ | 12.12 | 125.90 | 18.11 | 29.33 | 1.13 | 2.40 |
| | | | | (0.54) | (3.00) | (2.31) | (0.60) | (0.02) | (<0.01) |

Table 4.5: Comparing estimation results from Factor Analysis and Extended Factor Analysis in Model 4.2.

| Model 4.3: Estimation summary, $p = 20$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q.true | q selection | | norm | $\boldsymbol{\Sigma}(\mathbf{Y})$ error | | $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ error | | $\boldsymbol{\Theta}$ error | |
| | EFA | FA | | EFA | FA | EFA | FA | EFA | FA |
| 1 | 1.26 | 9.04 | | | | | | | |
| | (0.08) | (0.33) | | | | | | | |
| | | | F | 2.79 | 37.87 | 4.97 | 5.67 | 1.77 | 4.09 |
| | | | | (0.06) | (0.95) | (0.38) | (0.08) | (0.02) | (0.02) |
| | | | $\ell_1$ | 4.26 | 46.40 | 10.56 | 12.29 | 1.68 | 2.43 |
| | | | | (0.16) | (0.86) | (0.93) | (0.22) | (0.04) | (0.03) |
| | | | $\ell_2$ | 1.90 | 37.01 | 4.62 | 5.16 | 0.96 | 2.16 |
| | | | | (0.07) | (0.97) | (0.40) | (0.09) | (0.01) | (0.01) |
| 2 | 1.72 | 8.20 | | | | | | | |
| | (0.13) | (0.34) | | | | | | | |
| | | | F | 3.53 | 53.68 | 8.18 | 9.03 | 1.94 | 4.42 |
| | | | | (0.06) | (1.34) | (0.52) | (0.09) | (0.02) | (0.01) |
| | | | $\ell_1$ | 5.26 | 60.84 | 16.81 | 17.09 | 1.96 | 2.55 |
| | | | | (0.15) | (1.34) | (1.27) | (0.23) | (0.05) | (<0.01) |
| | | | $\ell_2$ | 2.49 | 47.89 | 7.65 | 7.53 | 1.05 | 2.27 |
| | | | | (0.08) | (1.44) | (0.57) | (0.10) | (0.02) | (<0.01) |
| 3 | 1.88 | 6.96 | | | | | | | |
| | (0.12) | (0.49) | | | | | | | |
| | | | F | 6.38 | 107.04 | 16.24 | 23.04 | 2.13 | 4.95 |
| | | | | (0.21) | (2.18) | (1.10) | (0.14) | (0.02) | (<0.01) |
| | | | $\ell_1$ | 9.19 | 86.87 | 27.05 | 32.71 | 2.14 | 2.58 |
| | | | | (0.37) | (1.77) | (1.72) | (0.30) | (0.03) | (<0.01) |
| | | | $\ell_2$ | 4.96 | 64.60 | 14.60 | 19.70 | 1.15 | 2.41 |
| | | | | (0.25) | (1.85) | (1.10) | (0.16) | (0.01) | (<0.01) |

Table 4.6: Comparing estimation results from Factor Analysis and Extended Factor Analysis in Model 4.3 for $p$=20.

| Model 3: Estimation summary, $p = 50$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q.true | q selection | | norm | $\boldsymbol{\Sigma}(\mathbf{Y})$ error | | $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ error | | $\boldsymbol{\Theta}$ error | |
| | EFA | FA | | EFA | FA | EFA | FA | EFA | FA |
| 1 | 1.00 | 11.28 | | | | | | | |
| | (<0.01) | (0.14) | | | | | | | |
| | | | F | 7.20 | 54.33 | 6.14 | 9.23 | 3.32 | 6.75 |
| | | | | (0.17) | (1.58) | (0.19) | (0.24) | (0.02) | (0.01) |
| | | | $\ell_1$ | 12.22 | 56.89 | 11.93 | 17.38 | 2.03 | 2.58 |
| | | | | (0.51) | (0.99) | (0.48) | (0.63) | (0.01) | (<0.01) |
| | | | $\ell_2$ | 5.37 | 40.18 | 5.43 | 7.94 | 1.25 | 2.28 |
| | | | | (0.23) | (1.45) | (0.23) | (0.29) | (0.01) | (<0.01) |
| 2 | 1.96 | 8.54 | | | | | | | |
| | (0.05) | (0.31) | | | | | | | |
| | | | F | 12.12 | 148.44 | 13.42 | 21.92 | 3.41 | 7.56 |
| | | | | (0.35) | (2.75) | (1.03) | (0.54) | (0.02) | (0.01) |
| | | | $\ell_1$ | 18.94 | 107.62 | 22.42 | 35.78 | 2.07 | 2.59 |
| | | | | (0.71) | (1.09) | (2.04) | (1.04) | (0.03) | (<0.01) |
| | | | $\ell_2$ | 9.29 | 77.82 | 11.13 | 19.13 | 1.30 | 2.42 |
| | | | | (0.44) | (2.07) | (1.10) | (0.59) | (0.01) | (<0.01) |
| 3 | 2.78 | 8.44 | | | | | | | |
| | (0.08) | (0.31) | | | | | | | |
| | | | F | 14.90 | 244.97 | 20.76 | 31.59 | 3.48 | 7.77 |
| | | | | (0.38) | (3.90) | (1.95) | (0.41) | (0.02) | (0.01) |
| | | | $\ell_1$ | 23.76 | 153.21 | 39.17 | 51.64 | 2.32 | 2.59 |
| | | | | (0.91) | (1.85) | (4.30) | (0.80) | (0.07) | (<0.01) |
| | | | $\ell_2$ | 10.94 | 115.69 | 17.32 | 24.12 | 1.32 | 2.46 |
| | | | | (0.47) | (2.47) | (1.95) | (0.47) | (0.01) | (<0.01) |

Table 4.7: Comparing estimation results from Factor Analysis and Extended Factor Analysis in Model 4.3.

| Model 4.4: Estimation summary, $p = 20$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q.true | q selection | | norm | $\mathbf{\Sigma}(\mathbf{Y})$ error | | $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ error | | $\mathbf{\Theta}$ error | |
| | EFA | FA | | EFA | FA | EFA | FA | EFA | FA |
| 1 | 1.20 | 5.56 | | | | | | | |
| | (0.06) | (0.46) | | | | | | | |
| | | | F | 8.26 | 152.63 | 4.62 | 10.04 | 0.92 | 5.91 |
| | | | | (0.12) | (2.44) | (0.24) | (<0.01) | (0.02) | (<0.01) |
| | | | $\ell_1$ | 9.97 | 123.12 | 8.46 | 22.92 | 0.71 | 1.89 |
| | | | | (0.27) | (3.27) | (0.40) | (<0.01) | (0.01) | (<0.01) |
| | | | $\ell_2$ | 4.76 | 101.77 | 4.00 | 10.04 | 0.49 | 1.88 |
| | | | | (0.13) | (3.24) | (0.22) | (<0.01) | (0.01) | (<0.01) |
| 2 | 2.14 | 5.82 | | | | | | | |
| | (0.06) | (0.53) | | | | | | | |
| | | | F | 8.76 | 186.11 | 6.01 | 15.22 | 0.99 | 5.92 |
| | | | | (0.12) | (2.73) | (0.17) | (<0.01) | (0.02) | (<0.01) |
| | | | $\ell_1$ | 10.57 | 145.66 | 9.39 | 28.45 | 0.74 | 1.89 |
| | | | | (0.27) | (3.48) | (0.31) | (<0.01) | (0.01) | (<0.01) |
| | | | $\ell_2$ | 4.85 | 119.80 | 4.50 | 13.77 | 0.53 | 1.89 |
| | | | | (0.11) | (3.43) | (0.17) | (<0.01) | (0.01) | (<0.01) |
| 3 | 1.44 | 5.78 | | | | | | | |
| | (0.09) | (0.49) | | | | | | | |
| | | | F | 15.55 | 521.67 | 28.48 | 37.22 | 4.42 | 9.19 |
| | | | | (0.19) | (5.51) | (0.94) | (<0.01) | (0.07) | (<0.01) |
| | | | $\ell_1$ | 18.73 | 260.57 | 41.95 | 52.18 | 4.35 | 4.60 |
| | | | | (0.61) | (4.55) | (1.53) | (<0.01) | (0.06) | (<0.01) |
| | | | $\ell_2$ | 8.87 | 225.62 | 26.26 | 34.05 | 3.32 | 4.59 |
| | | | | (0.22) | (4.59) | (1.10) | (<0.01) | (0.03) | (<0.01) |

Table 4.8: Comparing estimation results from Factor Analysis and Extended Factor Analysis in Model 4.4 for $p$=20.

| Model 4.4: Estimation summary, $p = 50$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q.true | q selection | | norm | $\mathbf{\Sigma}(\mathbf{Y})$ error | | $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ error | | $\mathbf{\Theta}$ error | |
| | EFA | FA | | EFA | FA | EFA | FA | EFA | FA |
| 1 | 1.02 | 6.82 | | | | | | | |
| | (0.02) | (0.43) | | | | | | | |
| | | | F | 20.29 | 245.05 | 10.52 | 33.18 | 2.08 | 9.35 |
| | | | | (0.14) | (2.67) | (0.21) | (0.02) | (0.02) | (<0.01) |
| | | | $\ell_1$ | 25.48 | 171.51 | 19.57 | 71.03 | 0.98 | 1.89 |
| | | | | (0.55) | (3.48) | (0.54) | (0.04) | (0.01) | (<0.01) |
| | | | $\ell_2$ | 10.03 | 104.80 | 8.89 | 33.18 | 0.63 | 1.88 |
| | | | | (0.24) | (3.24) | (0.26) | (0.02) | (0.01) | (<0.01) |
| 2 | 2.10 | 7.40 | | | | | | | |
| | (0.04) | (0.50) | | | | | | | |
| | | | F | 25.14 | 426.61 | 18.79 | 70.82 | 2.20 | 9.39 |
| | | | | (0.26) | (4.76) | (0.34) | (0.55) | (0.03) | (<0.01) |
| | | | $\ell_1$ | 33.15 | 279.27 | 29.12 | 119.95 | 0.98 | 1.89 |
| | | | | (0.82) | (4.02) | (0.85) | (1.11) | (0.01) | (<0.01) |
| | | | $\ell_2$ | 14.82 | 166.56 | 14.28 | 65.65 | 0.65 | 1.89 |
| | | | | (0.43) | (4.37) | (0.45) | (0.59) | (0.01) | (<0.01) |
| 3 | 3.04 | 6.86 | | | | | | | |
| | (0.03) | (0.46) | | | | | | | |
| | | | F | 27.81 | 556.89 | 22.44 | 84.45 | 2.25 | 9.40 |
| | | | | (0.34) | (6.25) | (0.42) | (0.47) | (0.03) | (<0.01) |
| | | | $\ell_1$ | 37.42 | 339.87 | 35.05 | 134.54 | 0.99 | 1.90 |
| | | | | (0.91) | (4.51) | (1.05) | (1.84) | (0.01) | (<0.01) |
| | | | $\ell_2$ | 15.98 | 225.74 | 15.79 | 73.77 | 0.66 | 1.89 |
| | | | | (0.52) | (4.03) | (0.53) | (0.53) | (0.01) | (<0.01) |

Table 4.9: Comparing estimation results from Factor Analysis and Extended Factor Analysis in Model 4.4.

| Model 4.5: Estimation summary, $p = 20$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q.true | q selection | | norm | $\mathbf{\Sigma}(\mathbf{Y})$ error | | $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ error | | $\mathbf{\Theta}$ error | |
| | EFA | FA | | EFA | FA | EFA | FA | EFA | FA |
| 1 | 1.06 | 5.12 | | | | | | | |
| | (0.03) | (0.42) | | | | | | | |
| | | | F | 12.52 | 313.47 | 6.02 | 10.04 | 2.03 | 9.18 |
| | | | | (0.13) | (3.66) | (0.21) | (<0.01) | (0.08) | (<0.01) |
| | | | $\ell_1$ | 14.23 | 183.69 | 11.21 | 22.92 | 1.96 | 4.59 |
| | | | | (0.29) | (5.42) | (0.39) | (<0.01) | (0.05) | (<0.01) |
| | | | $\ell_2$ | 6.29 | 161.28 | 5.13 | 10.04 | 1.52 | 4.59 |
| | | | | (0.14) | (5.38) | (0.21) | (<0.01) | (0.05) | (<0.01) |
| 2 | 1.24 | 5.02 | | | | | | | |
| | (0.06) | (0.44) | | | | | | | |
| | | | F | 13.02 | 364.74 | 9.32 | 15.22 | 2.24 | 9.18 |
| | | | | (0.14) | (3.66) | (0.19) | (<0.01) | (0.08) | (<0.01) |
| | | | $\ell_1$ | 14.62 | 216.27 | 15.22 | 28.45 | 2.41 | 4.59 |
| | | | | (0.24) | (4.84) | (0.29) | (<0.01) | (0.07) | (<0.01) |
| | | | $\ell_2$ | 6.29 | 190.15 | 6.98 | 13.77 | 1.71 | 4.59 |
| | | | | (0.12) | (4.80) | (0.12) | (<0.01) | (0.05) | (<0.01) |
| 3 | 1.44 | 5.78 | | | | | | | |
| | (0.09) | (0.49) | | | | | | | |
| | | | F | 15.55 | 521.67 | 28.48 | 37.22 | 4.42 | 9.19 |
| | | | | (0.19) | (5.51) | (0.94) | (<0.01) | (0.07) | (<0.01) |
| | | | $\ell_1$ | 18.73 | 260.57 | 41.95 | 52.18 | 4.35 | 4.60 |
| | | | | (0.41) | (4.55) | (1.53) | (<0.01) | (0.06) | (<0.01) |
| | | | $\ell_2$ | 8.87 | 225.62 | 26.26 | 34.05 | 3.32 | 4.59 |
| | | | | (0.22) | (4.59) | (1.10) | (<0.01) | (0.03) | (<0.01) |

Table 4.10: Comparing estimation results from Factor Analysis and Extended Factor Analysis in Model 4.5 for $p$=20.

| Model 4.5: Estimation summary, $p = 50$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q.true | q selection | | norm | $\boldsymbol{\Sigma}(\mathbf{Y})$ error | | $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ error | | $\boldsymbol{\Theta}$ error | |
| | EFA | FA | | EFA | FA | EFA | FA | EFA | FA |
| 1 | 1.00 | 6.18 | | | | | | | |
| | (<0.01) | (0.54) | | | | | | | |
| | | | F | 29.23 | 509.36 | 13.42 | 33.20 | 8.93 | 14.51 |
| | | | | (0.14) | (3.34) | (0.23) | (<0.01) | (0.11) | (<0.01) |
| | | | $\ell_1$ | 32.82 | 233.80 | 24.92 | 71.08 | 3.46 | 4.59 |
| | | | | (0.52) | (4.39) | (0.67) | (<0.01) | (0.02) | (<0.01) |
| | | | $\ell_2$ | 11.89 | 167.33 | 11.56 | 33.20 | 3.13 | 4.59 |
| | | | | (0.21) | (4.13) | (0.30) | (<0.01) | (0.03) | (<0.01) |
| 2 | 1.82 | 6.32 | | | | | | | |
| | (0.05) | (0.47) | | | | | | | |
| | | | F | 34.33 | 740.59 | 25.45 | 81.00 | 9.98 | 14.52 |
| | | | | (0.33) | (6.95) | (0.74) | (<0.01) | (0.11) | (<0.01) |
| | | | $\ell_1$ | 40.98 | 378.89 | 41.37 | 139.60 | 3.76 | 4.60 |
| | | | | (0.77) | (5.55) | (1.92) | (<0.01) | (0.04) | (<0.01) |
| | | | $\ell_2$ | 17.36 | 249.14 | 19.39 | 76.48 | 3.41 | 4.59 |
| | | | | (0.53) | (5.20) | (0.83) | (<0.01) | (0.03) | (<0.01) |
| 3 | 2.58 | 6.30 | | | | | | | |
| | (0.10) | (0.52) | | | | | | | |
| | | | F | 37.14 | 887.40 | 36.10 | 86.87 | 10.74 | 14.52 |
| | | | | (0.34) | (8.20) | (1.68) | (<0.01) | (0.13) | (<0.01) |
| | | | $\ell_1$ | 46.61 | 434.57 | 59.86 | 138.90 | 3.97 | 4.60 |
| | | | | (0.92) | (6.18) | (3.71) | (<0.01) | (0.05) | (<0.01) |
| | | | $\ell_2$ | 19.12 | 314.90 | 27.82 | 76.52 | 3.58 | 4.59 |
| | | | | (0.51) | (5.81) | (1.80) | (<0.01) | (0.04) | (<0.01) |

Table 4.11: Comparing estimation results from Factor Analysis and Extended Factor Analysis in Model 4.5.

| Model 4.6: Estimation summary, $p = 20$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q.true | q selection | | norm | $\boldsymbol{\Sigma}(\mathbf{Y})$ error | | $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ error | | $\boldsymbol{\Theta}$ error | |
| | EFA | FA | | EFA | FA | EFA | FA | EFA | FA |
| 1 | 1.22 | 7.38 | | | | | | | |
| | (0.07) | (0.48) | | | | | | | |
| | | | F | 12.19 | 948.42 | 8.51 | 10.01 | 0.36 | 0.66 |
| | | | | (0.22) | (9.92) | (0.26) | (0.03) | (0.01) | (0.01) |
| | | | $\ell_1$ | 16.19 | 612.44 | 15.14 | 22.84 | 0.35 | 0.39 |
| | | | | (0.41) | (11.00) | (0.54) | (0.06) | (0.01) | (0.02) |
| | | | $\ell_2$ | 7.11 | 595.25 | 7.00 | 10.00 | 0.34 | 0.38 |
| | | | | (0.20) | (10.93) | (0.25) | (0.03) | (0.01) | (0.02) |
| 2 | 1.40 | 6.26 | | | | | | | |
| | (0.10) | (0.51) | | | | | | | |
| | | | F | 13.31 | 1035.7 | 10.45 | 15.22 | 0.39 | 0.78 |
| | | | | (0.22) | (11.93) | (0.23) | (<0.01) | (0.01) | (0.01) |
| | | | $\ell_1$ | 17.94 | 682.26 | 17.12 | 28.45 | 0.36 | 0.50 |
| | | | | (0.39) | (16.32) | (0.45) | (<0.01) | (0.01) | (0.01) |
| | | | $\ell_2$ | 7.73 | 660.59 | 7.59 | 13.77 | 0.36 | 0.50 |
| | | | | (0.20) | (16.27) | (0.17) | (<0.01) | (0.01) | (0.01) |
| 3 | 2.24 | 7.10 | | | | | | | |
| | (0.10) | (0.54) | | | | | | | |
| | | | F | 16.66 | 1258.6 | 15.59 | 37.22 | 0.40 | 0.86 |
| | | | | (0.23) | (16.69) | (0.49) | (<0.01) | (0.01) | (0.01) |
| | | | $\ell_1$ | 22.17 | 725.73 | 22.67 | 52.18 | 0.36 | 0.52 |
| | | | | (0.57) | (19.11) | (0.86) | (<0.01) | (0.01) | (0.01) |
| | | | $\ell_2$ | 10.17 | 702.26 | 10.92 | 34.04 | 0.35 | 0.52 |
| | | | | (0.27) | (19.14) | (0.58) | (<0.01) | (0.01) | (0.01) |

Table 4.12: Comparing estimation results from Factor Analysis and Extended Factor Analysis in Model 4.6 for $p$=20.

| Model 4.6: Estimation summary, $p = 50$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| q.true | q selection | | norm | $\boldsymbol{\Sigma}(\mathbf{Y})$ error | | $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ error | | $\boldsymbol{\Theta}$ error | |
| | EFA | FA | | EFA | FA | EFA | FA | EFA | FA |
| 1 | 1.32 | 5.26 | | | | | | | |
| | (0.08) | (0.48) | | | | | | | |
| | | | F | 72.53 | 8344.1 | 32.29 | 33.16 | 0.48 | 0.68 |
| | | | | (1.51) | (54.13) | (0.83) | (0.03) | (0.02) | (0.01) |
| | | | $\ell_1$ | 90.02 | 3085.3 | 58.11 | 71.01 | 0.45 | 0.40 |
| | | | | (2.03) | (58.15) | (1.64) | (0.06) | (0.02) | (0.01) |
| | | | $\ell_2$ | 29.65 | 3044.5 | 25.47 | 33.16 | 0.43 | 0.40 |
| | | | | (0.62) | (57.36) | (0.62) | (0.03) | (0.02) | (0.01) |
| 2 | 1.16 | 6.54 | | | | | | | |
| | (0.05) | (0.43) | | | | | | | |
| | | | F | 68.62 | 8953.6 | 46.51 | 80.96 | 0.67 | 0.86 |
| | | | | (0.74) | (70.76) | (0.60) | (0.04) | (0.01) | (0.01) |
| | | | $\ell_1$ | 91.25 | 3359.9 | 75.36 | 139.52 | 0.61 | 0.53 |
| | | | | (1.81) | (66.05) | (1.81) | (0.08) | (0.01) | (0.01) |
| | | | $\ell_2$ | 33.63 | 3272.39 | 32.68 | 76.45 | 0.61 | 0.53 |
| | | | | (0.78) | (65.71) | (0.64) | (0.04) | (0.01) | (0.01) |
| 3 | 1.78 | 5.32 | | | | | | | |
| | (0.12) | (0.48) | | | | | | | |
| | | | F | 74.96 | 9232.4 | 55.32 | 86.87 | 0.66 | 0.93 |
| | | | | (1.04) | (69.28) | (0.69) | (<0.01) | (0.01) | (0.01) |
| | | | $\ell_1$ | 99.45 | 3578.9 | 86.54 | 138.90 | 0.58 | 0.58 |
| | | | | (1.87) | (83.63) | (1.86) | (<0.01) | (0.01) | (0.01) |
| | | | $\ell_2$ | 35.41 | 3485.3 | 36.66 | 76.52 | 0.57 | 0.58 |
| | | | | (0.70) | (81.94) | (0.83) | (<0.01) | (0.01) | (0.01) |

Table 4.13: Comparing estimation results from Factor Analysis and Extended Factor Analysis in Model 4.6.

# Chapter 5

# Summary of Thesis

In this thesis, I discuss sparse modeling in three problems: large covariance matrix estimation, varying-coefficient model and extended factor analysis. Different methods are implemented to achieve sparsity.

In Chapter 2, for large bandable covariance matrix, we proposed SURE-tuned tapering estimation to select sparsity parameter according to estimated risk. We argue that even if the risk can not be estimated perfectly, the risk curve always has the same shape as the truth, which makes the selection of sparsity parameter is very reliable in most cases, and consequently ensure the accuracy of final matrix estimation due to the optimality of tapering estimate given that the sparsity of the matrix is already known.

In Chapter 3, an algorithm named "Iteratively Varying-coefficient Independence Screening" (IVIS) is invented to make variable selection and estimation in high-dimensional varying-coefficient models. First of all, non-parametric methods based on B-spline is used to estimate the marginal correlation between each explanatory variable and the response variable. Then screening step filters out most of irrelevant variables, before other variable selection methods, such as regularization methods, are implemented. Direct application of regularization methods is not acceptable, since there are several estimates of coefficients for each variable at different time points.

Group variable selection - group SCAD - is chosen to do this job. Due to the theoretical drawback of one-step algorithm which ignores joint effects of explanatory variables, iterative algorithm is used to achieve good performance in real data analysis.

In Chapter 4, traditional factor analysis is extended to allow correlated errors. On one hand, we think strict assumption of orthogonality of error may not be met sometimes in reality; on the other hand, we argue that estimation from extended factor analysis and traditional factor analysis can be compared to observe if the orthogonality assumption is valid or approximately good. In terms of model estimation, "glasso-GEM" is proposed to estimate the factor loadings and (inverse) covariance matrix of error. Through several simulation models, the estimation accuracy of number of factors, factor loadings and inverse covariance matrix is very convincing.

In summary, this thesis demonstrates the power of sparse modeling in three areas: estimation of high-dimensional statistics, supervised learning and unsupervised learning.

# References


Bani Asadi, N., Rish, I., Scheinberg, K., Kanevsky, D., and Ramabhadran, B. (2009). Map approach to learning sparse gaussian markov networks. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009.*, pages 1721–1724.

Bickel, P. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 36:2577–2604.

Bickel, P. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36:199–227.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37:1705–1732.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24:2350–2383.

Cai, T., Zhang, C.-H., and Zhou, H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38:2118–2144.

Cai, T. and Zhou, H. (2012). Minimax estimation of large covariance matrices under $\ell_1$-norm. *Statistica Sinica*, 22:1319–1378.

Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics*, 35:2313–2351.

Chan, A., Vasconcelos, N., and Lanckriet, G. (2007). Direct convex relaxations of sparse svm. In *Proceedings of the 24th international conference on Machine learning*, pages 145–153.

d'Aspremont, A., El Ghaoui, L., Jordan, M., and Lanckriet, G. (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM review*, pages 434–448.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, series B*, 39:1–38.

Donoho, D. and Johnstone, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90:1200–1224.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule. *Journal of the American Statistical Association*, 81:461–470.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99:619–632.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression (with discussion). *The Annals of Statistics*, 32:407–499.

El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *The Annals of Statistics*, 36:2717–2756.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106:544–557.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

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

Fan, J., Samworth, R., and Wu, Y. (2009). Ultra-dimensional variable selection via independent learning: beyond the linear model. *Journal of Machine Learning Research*, 10:1829–1853.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38:3657–3604.

Frank, A. and Asuncion, A. (2010). Uci machine learning repository [http://archive.ics.uci.edu/ml]. irvine, ca: University of california, school of information and computer science.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441.

Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98:227–255.

Gorsuch, R. (1983). *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85:809–822.

Hoyer, P. (2004). Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, pages 1457–1469.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93:85–98.

Johnson, R. and Wichern, D. (2007). *Applied Multivariate Statistical Analysis*. New Jersey: Pearson Education, Inc.

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29:295–327.

Kim, H. and Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, pages 1495–1502.

Lam, C. and Fan, J. (2007). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37:4254–4278.

Lam, C. and Fan, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. *The Annals of Statistics*, 36:2232–2260.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., and Simon, I., e. a. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298:799–804.

Li, G., Peng, H., Zhang, J., and Zhu, L.-X. (2012a). Robust rank correlation based screening. *The Annals of Statistics*, 40:1846–1877.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *The Annals of Statistics*, 36:261–286.

Li, R., Zhong, W., and Zhu, L.-P. (2012b). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107:1129–1139.

Meinshausen, N. and Buehlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37:246–270.

Ravikumar, P., Wainwright, M., and Lafferty, J. (2008). High-dimensional graphical model selection using $\ell 1$-regularized logistic regression. *Technical Report 750, Department of Statistics, University of California, Berkeley.*

Rice, J. A. (2004). Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, 14:631–637.

Rothman, A., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104:177–186.

Rothman, A., Levina, E., and Zhu, J. (2010a). A new approach to cholesky-based covariance regularization in high dimensions. *Biometrika*, 97:539–550.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Rothman, A. J., Levina, E., and J., Z. (2010b). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19:69–76.

Rubin, D. and Thayer, D. (1982). Em algorithms for ml factor analysis. *Psychometrika*, 47:69–76.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

Shen, X. and Ye, J. (2002). Adaptive model selection. *Journal of the American Statistical Association*, 97:210–221.

Song, R., Yi, F., and Zou, H. (2013). Web appendix of on varying-coefficient independence screening for high-dimensional varying-coefficient models. *submitted to Statistica Sinica*.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycleregulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of Cell*, 9:3273–3297.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151.

Stone, C. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13:689–705.

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

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36:614–645.

Wang, L., Chen, G., and Li, H. (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23:1486–1494.

Wang, L., Li, H., and Huang, J. (2008a). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104:747–757.

Wang, L. F., Li, H. Z., and Huang, J. H. (2008b). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103:1556–1569.

Wei, F., Huang, J., and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, 21:1515–1540.

Wu, W. and Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, 19:1755–1768.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68:49–67.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2004). 1-norm support vector machines. *Advances in neural information processing systems*, pages 49–56.

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultra-high dimensional data. *Journal of the American Statistical Association*, 106:1464–1475.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

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

Zou, H. and Hastie, T. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics*, 35:2173–2192.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, pages 265–286.