6

Approaches to High-Dimensional Covariance and Precision Matrix Estimations

Jianqing Fan[1], Yuan Liao[2] and Han Liu[3]
[1]*Bendheim Center for Finance, Princeton University, USA*
[2]*Department of Mathematics, University of Maryland, USA*
[3]*Department of Operations Research and Financial Engineering, Princeton University, USA*

6.1 Introduction

Large covariance and precision (inverse covariance) matrix estimations have become fundamental problems in multivariate analysis that find applications in many fields, ranging from economics and finance to biology, social networks, and health sciences. When the dimension of the covariance matrix is large, the estimation problem is generally challenging. It is well-known that the sample covariance based on the observed data is singular when the dimension is larger than the sample size. In addition, the aggregation of a huge amount of estimation errors can make considerable adverse impacts on the estimation's accuracy. Therefore, estimating large covariance and precision matrices has attracted rapidly growing research attention in the past decade. Many regularized methods have been developed: see Bickel and Levina (2008), El Karoui (2008), Friedman *et al.* (2008), Fryzlewicz (2013), Han *et al.* (2012), Lam and Fan (2009), Ledoit and Wolf (2003), Pourahmadi (2013), Ravikumar *et al.*, (2011b), Xue and Zou (2012), among others.

One of the commonly used approaches to estimating large matrices is to assume the covariance matrix to be sparse, that is, many off-diagonal components are either zero or nearly so. This effectively reduces the total number of parameters to estimate. However, such a sparsity assumption is restrictive in many applications. For example, financial returns depend on the common risk factors, housing prices depend on the economic health, and gene expressions can

Financial Signal Processing and Machine Learning, First Edition.
Edited by Ali N. Akansu, Sanjeev R. Kulkarni and Dmitry Malioutov.

be stimulated by cytokines. Moreover, in many applications, it is more natural to assume that the precision matrix is sparse instead (e.g., in Gaussian graphical models).

In this chapter, we introduce several recent developments for estimating large covariance and precision matrices without assuming the covariance matrix to be sparse. One of the selected approaches assumes the precision matrix to be sparse and applies column-wise penalization for estimations. This method efficiently estimates the precision matrix in Gaussian graphical models. The other method is based on high-dimensional factor analysis. Both methods will be discussed in Sections 6.2 and 6.3, and are computationally more efficient than the existing ones based on penalized maximum likelihood estimation. We present several applications of these methods, including graph estimation for gene expression data, and several financial applications. In particular, we shall see that estimating covariance matrices of high-dimensional asset excess returns plays a central role in applications of portfolio allocations and in risk management.

In Section 6.4, we provide a detailed description of the so-called factor pricing model, which is one of the most fundamental results in finance. It postulates how financial returns are related to market risks, and has many important practical applications, including portfolio selection, fund performance evaluation, and corporate budgeting. In the model, the excess returns can be represented by a factor model. We shall also study a problem of testing "mean–variance efficiency." In such a testing problem, most of the existing methods are based on the Wald statistic, which has two main difficulties when the number of assets is large. First, the Wald statistic depends on estimating a large inverse covariance matrix, which is a challenging problem in a data-rich environment. Second, it suffers from a lower power in a high-dimensional, low-sample-size situation. To address the problem, we introduce a new test, called the *power enhancement test*, which aims to enhance the power of the usual Wald test.

In Section 6.5, we will present recent developments of efficient estimations in panel data models. As we shall illustrate, the usual principal components method for estimating the factor models is not statistically efficient since it treats the idiosyncratic errors as both cross-sectionally independent and homoscedastic. In contrast, using a consistent high-dimensional precision covariance estimator can potentially improve the estimation efficiency. We shall conclude in Section 6.6.

Throughout the chapter, we shall use $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_F$ as the operator and Frobenius norms of a matrix $\mathbf{A}$. We use $\|\mathbf{v}\|$ to denote the Euclidean norm of a vector $\mathbf{v}$.

6.2 Covariance Estimation via Factor Analysis

Suppose we observe a set of stationary data $\{\boldsymbol{Y}_t\}_{t=1}^T$, where each $\boldsymbol{Y}_t = (Y_{1t}, \cdots, Y_{N,t})'$ is a high-dimensional vector; here, T and N respectively denote the sample size and the dimension. We aim to estimate the covariance matrix of $\boldsymbol{Y}_t$: $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{Y}_t)$, and its inverse $\boldsymbol{\Sigma}^{-1}$, which are assumed to be independent of t. This section introduces a method of estimating $\boldsymbol{\Sigma}$ and its inverse via factor analysis. In many applications, the cross-sectional units often depend on a few common factors. Fan *et al.* (2008) tackled the covariance estimation problem by considering the following factor model:

$$Y_{it} = \mathbf{b}_i' \boldsymbol{f}_t + u_{it}. \tag{6.1}$$

where Y_{it} is the observed response for the ith ($i = 1, \ldots, N$) individual at time $t = 1, \ldots, T$; $\mathbf{b}_i$ is a vector of factor loadings; $\boldsymbol{f}_t$ is a $K \times 1$ vector of common factors; and u_{it} is the error term,

usually called *idiosyncratic component*, uncorrelated with $\boldsymbol{f}_t$. In fact, factor analysis has long been employed in financial studies, where Y_{it} often represents the excess returns of the ith asset (or stock) on time t. The literature includes, for instance, Campbell *et al.* (1997), Chamberlain and Rothschild (1983), Fama and French (1992). It is also commonly used in macroeconomics for forecasting diffusion indices (e.g., Stock and Watson, (2002).

The factor model (6.1) can be put in a matrix form as

$$\boldsymbol{Y}_t = \mathbf{B}\boldsymbol{f}_t + \boldsymbol{u}_t. \tag{6.2}$$

where $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_N)'$ and $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{Nt})'$. We are interested in $\boldsymbol{\Sigma}$, the $N \times N$ covariance matrix of $\boldsymbol{Y}_t$, and its inverse $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, which are assumed to be time-invariant. Under model (6.1), $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = \mathbf{B}\mathrm{Cov}(\boldsymbol{f}_t)\mathbf{B}' + \boldsymbol{\Sigma}_u, \tag{6.3}$$

where $\boldsymbol{\Sigma}_u = (\sigma_{u,ij})_{N\times N}$ is the covariance matrix of $\boldsymbol{u}_t$. Estimating the covariance matrix $\boldsymbol{\Sigma}_u$ of the idiosyncratic components $\{\boldsymbol{u}_t\}$ is also important for statistical inferences. For example, it is needed for large sample inference of the unknown factors and their loadings and for testing the capital asset pricing model (Sentana, 2009).

In the decomposition (6.3), it is natural to consider the *conditional sparsity*: given the common factors, most of the remaining outcomes are mutually weakly correlated. This gives rise to the approximate factor model (e.g., Chamberlain and Rothschild, 1983), in which $\boldsymbol{\Sigma}_u$ is a sparse covariance but not necessarily diagonal, and for some $q \in [0, 1)$,

$$m_N = \max_{i\le N} \sum_{j\le N} |\sigma_{u,ij}|^q \tag{6.4}$$

does not grow too fast as $N \to \infty$. When $q = 0$, m_N measures the maximum number of non zero components in each row.

We would like to emphasize that model (6.3) is related to but different from the problem recently studied in the literature on "low-rank plus sparse representation". In fact, the "low rank plus sparse" representation of (6.3) holds on the population covariance matrix, whereas the model considered by Candès *et al.* (2011) and Chandrasekaran *et al.* (2010) considered such a representation on the data matrix. As there is no $\boldsymbol{\Sigma}$ to estimate, their goal is limited to producing a low-rank plus sparse matrix decomposition of the data matrix, which corresponds to the identifiability issue of our study, and does not involve estimation or inference. In contrast, our ultimate goal is to estimate the population covariance matrices as well as the precision matrices. Our consistency result on $\boldsymbol{\Sigma}_u$ demonstrates that the decomposition (6.3) is identifiable, and hence our results also shed the light of the "surprising phenomenon" of Candès *et al.* (2011) that one can separate fully a sparse matrix from a low-rank matrix when only the sum of these two components is available.

Moreover, note that in financial applications, the common factors $\boldsymbol{f}_t$ are sometimes known, as in Fama and French (1992). In other applications, however, the common factors may be unknown and need to be inferred. Interestingly, asymptotic analysis shows that as the dimensionality grows fast enough (relative to the sample size), the effect of estimating the unknown factors is negligible, and the covariance matrices of $\boldsymbol{Y}_t$ and $\boldsymbol{u}_t$ and their inverses can be estimated as if the factors were known (Fan *et al.*, 2013).

We now divide our discussions into two cases: models with known factors and models with unknown factors.

6.2.1 Known Factors

When the factors are observable, one can estimate $\mathbf{B}$ by the ordinary least squares (OLS): $\hat{\mathbf{B}} = (\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_N)'$, where,

$$\hat{\mathbf{b}}_i = \arg\min_{\mathbf{b}_i} \frac{1}{T} \sum_{t=1}^{T} (Y_{it} - \mathbf{b}_i' \boldsymbol{f}_t)^2, \quad i = 1, \ldots, N.$$

The residuals are obtained using the plug-in method: $\hat{u}_{it} = Y_{it} - \hat{\mathbf{b}}_i' \boldsymbol{f}_t$.

Denote by $\hat{\boldsymbol{u}}_t = (\hat{u}_{1t}, \ldots, \hat{u}_{pt})'$. We then construct the residual covariance matrix as:

$$\mathbf{S}_u = \frac{1}{T} \sum_{t=1}^{T} \hat{\boldsymbol{u}}_t \hat{\boldsymbol{u}}_t' = (s_{u,ij}).$$

Now we apply thresholding on $\mathbf{S}_u$. Define

$$\hat{\boldsymbol{\Sigma}}_u = (\hat{\sigma}_{ij})_{p\times p}, \hat{\sigma}_{ij}^{\mathcal{T}} = \begin{cases} s_{u,ii}, & i = j; \\ th(s_{u,ij}) I(|s_{u,ij}| \geq \tau_{ij}), & i \neq j. \end{cases} \tag{6.5}$$

where $th(\cdot)$ is a generalized shrinkage function of Antoniadis and Fan (2001), employed by Rothman *et al.* (2009) and Cai and Liu (2011), and $\tau_{ij} > 0$ is an entry-dependent threshold. In particular, the hard-thresholding rule $th(x) = xI(|x| \geq \tau_{ij})$ (Bickel and Levina, 2008) and the constant thresholding parameter $\tau_{ij} = \delta$ are allowed. In practice, it is more desirable to have τ_{ij} be entry-adaptive. An example of the threshold is

$$\tau_{ij} = \omega_T (s_{u,ii} s_{u,jj})^{1/2}, \text{ for a given} \omega_T > 0 \tag{6.6}$$

This corresponds to applying the thresholding with parameter ω_T to the correlation matrix of $\mathbf{S}_u$. Cai and Liu (2011) discussed an alternative type of "adaptive threshold." Moreover, we take ω_T to be: some $C > 0$,

$$\omega_T = C \sqrt{\frac{\log\ N}{T}},$$

which is a proper threshold level to overrides the estimation errors.

The covariance matrix $\mathrm{Cov}(\boldsymbol{f}_t)$ can be estimated by the sample covariance matrix

$$\widehat{\mathrm{Cov}}(\boldsymbol{f}_t) = T^{-1} \mathbf{F}'\mathbf{F} - T^{-2} \mathbf{F}'\mathbf{1}\mathbf{1}'\mathbf{F},$$

where $\mathbf{F}' = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)$, and $\mathbf{1}$ is a T-dimensional column vector of ones. Therefore, we obtain a substitution estimator (Fan *et al.*, 2011):

$$\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{B}} \widehat{\mathrm{Cov}}(\boldsymbol{f}_t) \hat{\mathbf{B}}' + \hat{\boldsymbol{\Sigma}}_u. \tag{6.7}$$

By the Sherman–Morrison–Woodbury formula,

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_u^{-1} - \boldsymbol{\Sigma}_u^{-1} \mathbf{B} [\mathrm{Cov}(\boldsymbol{f}_t)^{-1} + \mathbf{B}' \boldsymbol{\Sigma}_u^{-1} \mathbf{B}]^{-1} \mathbf{B}' \boldsymbol{\Sigma}_u^{-1},$$

which is estimated by

$$\hat{\boldsymbol{\Sigma}}^{-1} = \hat{\boldsymbol{\Sigma}}_u^{-1} - \hat{\boldsymbol{\Sigma}}_u^{-1} \hat{\mathbf{B}} [\widehat{\mathrm{Cov}}(\boldsymbol{f}_t)^{-1} + \hat{\mathbf{B}}' \hat{\boldsymbol{\Sigma}}_u^{-1} \hat{\mathbf{B}}]^{-1} \hat{\mathbf{B}}' \hat{\boldsymbol{\Sigma}}_u^{-1}. \tag{6.8}$$

6.2.2 Unknown Factors

When factors are unknown, Fan *et al.* (2013) proposed a nonparametric estimator of $\mathbf{\Sigma}$ based on the principal component analysis. Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_N$ be the ordered eigenvalues of the sample covariance matrix $\mathbf{S}$ of $\boldsymbol{Y}_t$ and $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^N$ be their corresponding eigenvectors. Then the sample covariance has the following spectral decomposition:

$$\mathbf{S} = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \boldsymbol{Q},$$

where $\boldsymbol{Q} = \sum_{i=K+1}^{N} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i'$ is called the *principal orthogonal complement*, and K is the number of common factors. We can apply thresholding on $\boldsymbol{Q}$ as in (6.5) and (6.6). Denote the thresholded $\boldsymbol{Q}$ by $\hat{\mathbf{\Sigma}}_u$. Note that the threshold value in (6.6) now becomes, for some $C > 0$

$$\omega_T = C\left(\sqrt{\frac{\log\ N}{T}} + \frac{1}{\sqrt{N}}\right).$$

The estimator of $\mathbf{\Sigma}$ is then defined as:

$$\hat{\mathbf{\Sigma}}_K = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\mathbf{\Sigma}}_u. \tag{6.9}$$

This estimator is called the principal orthogonal complement thresholding (POET) estimator. It is obtained by thresholding the remaining components of the sample covariance matrix, after taking out the first K principal components. One of the attractiveness of POET is that it is optimization-free, and hence is computationally appealing.

The POET (6.9) has an equivalent representation using a constrained least-squares method. The least-squares method seeks for $\hat{\mathbf{B}} = (\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_N)'$ and $\hat{\mathbf{F}}' = (\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_T)$ such that

$$(\hat{\mathbf{B}}, \hat{\mathbf{F}}) = \arg \min_{\mathbf{b}_i \in \mathbb{R}^K, \boldsymbol{f}_t \in \mathbb{R}^K} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mathbf{b}_i' \boldsymbol{f}_t)^2, \tag{6.10}$$

subject to the normalization

$$\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{f}_t \boldsymbol{f}_t' = \mathbf{I}_K, \text{ and } \frac{1}{N} \sum_{i=1}^{N} \mathbf{b}_i \mathbf{b}_i' \text{ is diagonal.} \tag{6.11}$$

Putting it in a matrix form, the optimization problem can be written as

$$\begin{aligned} &\arg \min_{\mathbf{B},\mathbf{F}} \|\mathbf{Y}' - \mathbf{B}\mathbf{F}'\|_F^2 \\ &T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K, \mathbf{B}'\mathbf{B} \text{ is diagonal.} \end{aligned} \tag{6.12}$$

where $\mathbf{Y}' = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T)$ and $\mathbf{F}' = (\boldsymbol{f}_1, \cdots, \boldsymbol{f}_T)$. For each given $\mathbf{F}$, the least-squares estimator of $\mathbf{B}$ is $\hat{\mathbf{B}} = T^{-1}\mathbf{Y}'\mathbf{F}$, using the constraint (6.11) on the factors. Substituting this into (6.12), the objective function now becomes $\|\mathbf{Y}' - T^{-1}\mathbf{Y}'\mathbf{F}\mathbf{F}'\|_F^2 = \text{tr}[(\mathbf{I}_T - T^{-1}\mathbf{F}\mathbf{F}')\mathbf{Y}\mathbf{Y}']$. The minimizer is now clear: the columns of $\hat{\mathbf{F}}/\sqrt{T}$ are the eigenvectors corresponding to the K largest eigenvalues of the $T \times T$ matrix $\mathbf{Y}\mathbf{Y}'$ and $\hat{\mathbf{B}} = T^{-1}\mathbf{Y}'\hat{\mathbf{F}}$ (see e.g., Stock and Watson, 2002). The

residual is given by $\hat{u}_{it} = Y_{it} - \hat{\mathbf{b}}_i' \hat{\boldsymbol{f}}_t$, based on which we can construct the sample covariance matrix of $\boldsymbol{\Sigma}_u$. Then apply the thresholding to obtain $\hat{\boldsymbol{\Sigma}}_u$. The covariance of $\boldsymbol{Y}_t$ is then estimated by $\hat{\mathbf{B}}\hat{\mathbf{B}}' + \hat{\boldsymbol{\Sigma}}_u$. It can be proved that the estimator in (6.9) satisfies:

$$\hat{\boldsymbol{\Sigma}}_K = \hat{\mathbf{B}}\hat{\mathbf{B}}' + \hat{\boldsymbol{\Sigma}}_u.$$

Several methods have been proposed to consistently estimate the number of factors. For instance, Bai and Ng (2002) proposed to use:

$$\hat{K} = \arg \min_{0 \le k \le M} \frac{1}{N} \mathrm{tr} \left(\sum_{j=k+1}^{N} \hat{\lambda}_j \hat{\boldsymbol{\xi}}_j \hat{\boldsymbol{\xi}}_j' \right) + \frac{k(N+T)}{NT} \log \left(\frac{NT}{N+T} \right), \tag{6.13}$$

where M is a prescribed upper bound. The literature also includes, Ahn and Horenstein (2013), Alessi *et al.* (2010), Hallin and Liška (2007), Kapetanios (2010), among others. Numerical studies in Fan *et al.* (2013) showed that the covariance estimator is robust to overestimating K. Therefore, in practice, we can also choose a relatively large number for K. Consistency can still be guaranteed.

6.2.3 Choosing the Threshold

Recall that the threshold value ω_T depends on a user-specific constant C. In practice, we need to choose C to maintain the positive definiteness of the estimated covariances for any given finite sample. To do so, write the error covariance estimator as $\hat{\boldsymbol{\Sigma}}_u(C)$, which depends on C via the threshold. We choose C in the range where $\lambda_{\min}(\hat{\boldsymbol{\Sigma}}_u) > 0$. Define

$$C_{\min} = \inf \{ C > 0 : \lambda_{\min}(\hat{\boldsymbol{\Sigma}}_u(M)) > 0, \forall M > C \}. \tag{6.14}$$

When C is sufficiently large, the estimator becomes diagonal, while its minimum eigenvalue must retain strictly positive. Thus, $C_{\min}$ is well defined and for all $C > C_{\min}$, $\hat{\boldsymbol{\Sigma}}_u(C)$ is positive definite under finite sample. We can obtain $C_{\min}$ by solving $\lambda_{\min}(\hat{\boldsymbol{\Sigma}}_u(C)) = 0, C \neq 0$. We can also approximate $C_{\min}$ by plotting $\lambda_{\min}(\hat{\boldsymbol{\Sigma}}_u(C))$ as a function of C, as illustrated in Figure 6.1. In practice, we can choose C in the range $(C_{\min} + \epsilon, M)$ for a small ϵ and large enough M. Choosing the threshold in a range to guarantee the finite-sample positive definiteness has also been previously suggested by Fryzlewicz (2013).

6.2.4 Asymptotic Results

Under regularity conditions (e.g., strong mixing, exponential-tail distributions), Fan *et al.* (2011, 2013) showed that for the error covariance estimator, assuming $\omega_T^{1-q} m_N = o(1)$,

$$\| \hat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u \|_2 = O_P(\omega_T^{1-q} m_N),$$

and

$$\| \hat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1} \|_2 = O_P(\omega_T^{1-q} m_N).$$

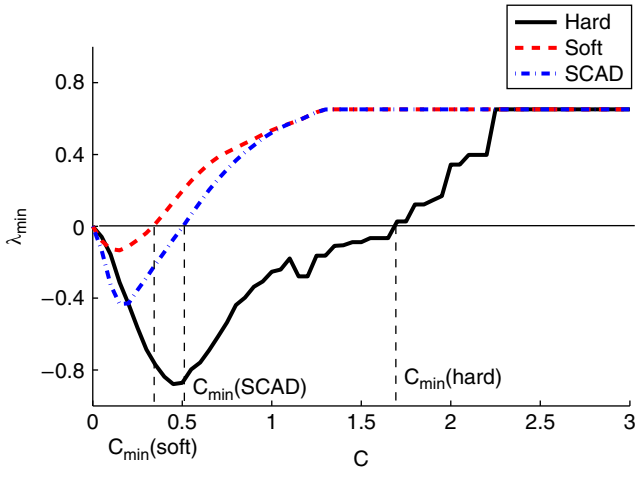

Figure 6.1 Minimum eigenvalue of $\hat{\mathbf{\Sigma}}_u(C)$ as a function of C for three choices of thresholding rules. Adapted from Fan *et al.* (2013).

Here $q \in [0, 1)$ quantifies the level of sparsity as defined in (6.4), and ω_T is given by: for some $C > 0$, when factors are known,

$$\omega_T = \sqrt{\frac{\log N}{T}}$$

when factors are unknown,

$$\omega_T = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}.$$

The dimension N is allowed to grow exponentially fast in T.

As for the convergence of $\hat{\mathbf{\Sigma}}$, because the first K eigenvalues of $\mathbf{\Sigma}$ grow with N, one can hardly estimate $\mathbf{\Sigma}$ with satisfactory accuracy in either the operator norm or the Frobenius norm. This problem arises not from the limitation of any estimation method, but due to the nature of the high-dimensional factor model. We illustrate this in the following example.

Example 6.1 *Consider a simplified case where we know* $\mathbf{b}_i = (1, 0, \ldots, 0)'$ *for each* $i = 1, \ldots, N$, $\mathbf{\Sigma}_u = \mathbf{I}$, *and* $\{\boldsymbol{f}_t\}_{t=1}^T$ *are observable. Then when estimating* $\mathbf{\Sigma}$, *we only need to estimate* $Cov(\boldsymbol{f})$ *using the sample covariance matrix* $\widehat{Cov}(\boldsymbol{f}_t)$, *and obtain an estimator for* $\mathbf{\Sigma}$*:*

$$\hat{\mathbf{\Sigma}} = \mathbf{B}\widehat{Cov}(\boldsymbol{f}_t)\mathbf{B}' + \mathbf{I}.$$

Simple calculations yield to

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = |\frac{1}{T}\sum_{t=1}^{T}(f_{1t} - \bar{f}_1)^2 - Var\ (f_{1t})| \cdot \|\mathbf{1}_N\mathbf{1}_N'\|_2,$$

where $\mathbf{1}_N$ *denotes the* N*-dimensional column vector of ones with* $\|\mathbf{1}_N\mathbf{1}_N'\|_2 = N$. *Therefore, due to the central limit theorem employed on* $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(f_{1t} - \bar{f}_1)^2 - Var\ (f_{1t})$, $\frac{\sqrt{T}}{N}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2$ *is*

asymptotically normal. Hence $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2$ *diverges if* $N \gg \sqrt{T}$*, even for such a simplified toy model.*

As we have seen from the above example, the small error of estimating Var (f_{1t}) is substantially amplified due to the presence of $\|\mathbf{1}_N \mathbf{1}_N'\|_2$; the latter in fact determines the size of the largest eigenvalue of $\mathbf{\Sigma}$. We further illustrate this phenomenon in the following example.

Example 6.2 *Consider an ideal case where we know the spectrum except for the first eigenvector of* $\mathbf{\Sigma}$*. Let* $\{\lambda_j, \boldsymbol{\xi}_j\}_{j=1}^N$ *be the eigenvalues and vectors, and assume that the largest eigenvalue* $\lambda_1 \geq cN$ *for some* $c > 0$*. Let* $\hat{\boldsymbol{\xi}}_1$ *be the estimated first eigenvector, and define the covariance estimator* $\hat{\mathbf{\Sigma}} = \lambda_1 \hat{\boldsymbol{\xi}}_1 \hat{\boldsymbol{\xi}}_1' + \sum_{j=2}^N \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j'$*. Assume that* $\hat{\boldsymbol{\xi}}_1$ *is a good estimator in the sense that* $\|\hat{\boldsymbol{\xi}}_1 - \boldsymbol{\xi}_1\|^2 = O_P(T^{-1})$*. However,*

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = \|\lambda_1(\hat{\boldsymbol{\xi}}_1 \hat{\boldsymbol{\xi}}_1' - \boldsymbol{\xi}_1 \boldsymbol{\xi}_1')\|_2 = \lambda_1 O_P(\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\|) = O_P(\lambda_1 T^{-1/2}),$$

which can diverge when $T = O(N^2)$*.*

On the other hand, we can estimate the precision matrix with a satisfactory rate under the operator norm. The intuition follows from the fact that $\mathbf{\Sigma}^{-1}$ has bounded eigenvalues. Let $\hat{\mathbf{\Sigma}}^{-1}$ denote the inverse of the POET estimator. Fan *et al.* (2013) showed that $\hat{\mathbf{\Sigma}}^{-1}$ has the same rate of convergence as that of $\mathbf{\Sigma}_u^{-1}$. Specifically,

$$\|\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_2 = O_P(\omega_T^{1-q} m_N).$$

Comparing the rates of convergence of known and unknown factors, we see that when the common factors are unobservable, the rate of convergence has an additional term $m_N / N^{(1-q)/2}$, coming from the impact of estimating the unknown factors. This impact vanishes when $N \log N \gg T$, in which case the minimax rate as in Cai and Zhou (2010) is achieved. As N increases, more information about the common factors is collected, which results in more accurate estimation of the common factors $\{\boldsymbol{f}_t\}_{t=1}^T$. Then the rates of convergence in both observable factor and unobservable factor cases are the same.

6.2.5 A Numerical Illustration

We now illustrate the above theoretical results by using a simple three-factor model with a sparse error covariance matrix. The distribution of the data-generating process is taken from Fan *et al.* (2013) (Section 6.7). Specifically, we simulated from a standard Fama–French three-factor model. The factor loadings are drawn from a trivariate normal distribution $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i})' \sim N_3(\boldsymbol{\mu}_B, \mathbf{\Sigma}_B)$, and $\boldsymbol{f}_t$ follows a vector autoregression of the first order (VAR(1)) model $\boldsymbol{f}_t = \boldsymbol{\mu} + \mathbf{\Phi} \boldsymbol{f}_{t-1} + \boldsymbol{\epsilon}_t$. To make the simulation more realistic, model parameters are calibrated from the real data on annualized returns of 100 industrial portfolios, obtained from the website of Kenneth French. As there are three common factors, the largest three eigenvalues of $\mathbf{\Sigma}$ are of the same order as $\sum_{i=1}^N b_{ji}^2, j = 1, 2, 3$, which are approximately $O(N)$, and grow linearly with N.

We generate a sparse covariance matrix $\mathbf{\Sigma}_u$ of the form: $\mathbf{\Sigma}_u = \boldsymbol{D} \mathbf{\Sigma}_0 \boldsymbol{D}$. Here, $\mathbf{\Sigma}_0$ is the error correlation matrix, and $\boldsymbol{D}$ is the diagonal matrix of the standard deviations of the errors. We

Table 6.1 Mean and covariance matrix used to generate $\mathbf{b}_i$

$\boldsymbol{\mu}_B$	$\boldsymbol{\Sigma}_B$		
0.0047	0.0767	−0.00004	0.0087
0.0007	−0.00004	0.0841	0.0013
−1.8078	0.0087	0.0013	0.1649

Table 6.2 Parameters of $\boldsymbol{f}_t$ generating process

$\boldsymbol{\mu}$	$\text{Cov}(\boldsymbol{f}_t)$			$\boldsymbol{\Phi}$		
−0.0050	1.0037	0.0011	−0.0009	−0.0712	0.0468	0.1413
0.0335	0.0011	0.9999	0.0042	−0.0764	−0.0008	0.0646
−0.0756	−0.0009	0.0042	0.9973	0.0195	−0.0071	−0.0544

set $\boldsymbol{D} = \text{diag}(\sigma_1, \ldots, \sigma_p)$, where each σ_i is generated independently from a gamma distribution $G(\alpha, \beta)$, and α and β are chosen to match the sample mean and sample standard deviation of the standard deviations of the errors. The off-diagonal entries of $\boldsymbol{\Sigma}_0$ are generated independently from a normal distribution, with mean and standard deviation equal to the sample mean and sample standard deviation of the sample correlations among the estimated residuals. We then employ hard thresholding to make $\boldsymbol{\Sigma}_0$ sparse, where the threshold is found as the smallest constant that provides the positive definiteness of $\boldsymbol{\Sigma}_0$.

For the simulation, we fix $T = 300$, and let N increase from 20 to 600 in increments of 20. We plot the averages and standard deviations of the distance from $\hat{\boldsymbol{\Sigma}}$ and $\mathbf{S}$ to the true covariance matrix $\boldsymbol{\Sigma}$, under the norm $\|\mathbf{A}\|_\Sigma = \frac{1}{N}\|\boldsymbol{\Sigma}^{-1/2}\mathbf{A}\boldsymbol{\Sigma}^{-1/2}\|_F$ (recall that $\mathbf{S}$ denotes the sample covariance). It is easy to see that

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\Sigma = \frac{1}{N}\|\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}\|_F,$$

which resembles the relative errors. We also plot the means and standard deviations of the distances from $\hat{\boldsymbol{\Sigma}}^{-1}$ and $\mathbf{S}^{-1}$ to $\boldsymbol{\Sigma}^{-1}$ under the spectral norm. Due to invertibility, the operator norm for $\mathbf{S}^{-1}$ is plotted only up to $N = 280$.

We observe that the unobservable factor model performs just as well as the estimator if the factors are known. The cost of not knowing the factors is negligible when N is large enough. As we can see from Figure 6.2, the impact decreases quickly. In addition, when estimating $\boldsymbol{\Sigma}^{-1}$, it is hard to distinguish the estimators with known and unknown factors, whose performances are quite stable compared to that of the sample covariance matrix. Intuitively, as the dimension increases, more information about the common factors becomes available, which helps infer the unknown factors. Indeed, as is shown in Bai (2003) and Fan *et al.* (2014a), the principal components method can estimate the unknown factors at a rate of:

$$\frac{1}{T}\sum_{t=1}^{T}\|\hat{\boldsymbol{f}}_t - \boldsymbol{f}_t\|^2 = O_P\left(\frac{1}{T^2} + \frac{1}{N}\right).$$

Hence, as long as N is relatively large, $\boldsymbol{f}_t$ can be estimated pretty accurately.

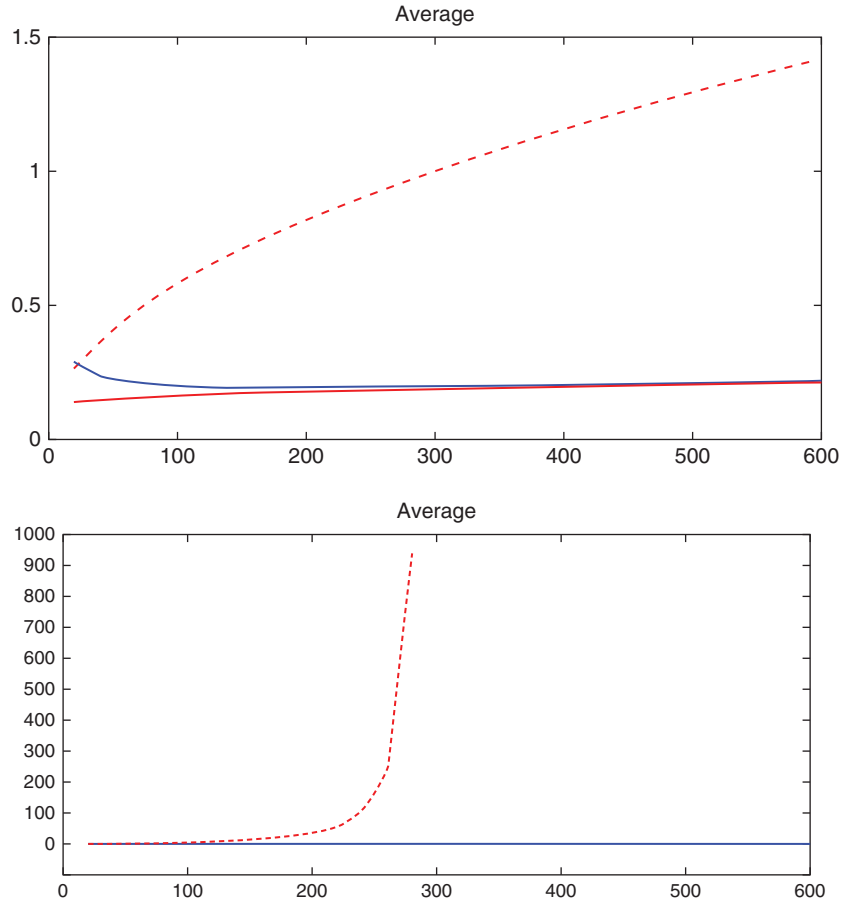

Figure 6.2 Averages of $N^{-1}\|\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}\|_F$ (left panel) and $\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_2$ (right panel) with known factors (solid red curve), unknown factors (solid blue curve), and sample covariance (dashed curve) over 200 simulations, as a function of the dimensionality N. Taken from Fan *et al.* (2013).

6.3 Precision Matrix Estimation and Graphical Models

Let $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_T$ be T data points from an N-dimensional random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_N)'$ with $\boldsymbol{Y} \sim \mathcal{N}_N(\mathbf{0}, \boldsymbol{\Sigma})$. We denote the precision matrix $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$ and define an undirected graph $G = (V, E)$ based on the sparsity pattern of $\boldsymbol{\Theta}$: let $V = \{1, \cdots, N\}$ be the node set corresponding to the N variables in $\boldsymbol{Y}$, an edge $(j, k) \in E$ if and only $\boldsymbol{\Theta}_{jk} \neq 0$.

As we will explain in Section 6.3.1, the graph G describes the conditional independence relationships between $Y_1, \ldots, Y_N$: that is, letting $\boldsymbol{Y}_{\setminus\{j,k\}} := \{Y_\ell : \quad \ell \neq j, k\}$, then Y_j is independent of Y_k given $\boldsymbol{Y}_{\setminus\{j,k\}}$ if and only if $(j, k) \notin E$.

In high-dimensional settings where $N \gg T$, we assume that many entries of $\boldsymbol{\Theta}$ are zero (or, in other words, the graph G is sparse). The problem of estimating a large sparse precision matrix $\boldsymbol{\Theta}$ is called *covariance selection* (Dempster, 1972).

6.3.1 Column-wise Precision Matrix Estimation

A natural approach for estimating $\mathbf{\Theta}$ is by penalizing the likelihood using the L_1-penalty (Banerjee *et al.*, 2008; Friedman *et al.*, 2008; Yuan and Lin, 2007). To further reduce the estimation bias, Jalali *et al.* (2012), Lam and Fan (2009), Shen *et al.*, (2012) propose either greedy algorithms or nonconvex penalties for sparse precision matrix estimation. Under certain conditions, Ravikumar *et al.* (2011a), Rothman *et al.* (2008), Wainwright (2009), Zhao and Yu (2006), Zou (2006), study the theoretical properties of the penalized likelihood methods.

Another approach is to estimate $\mathbf{\Theta}$ in a column-by-column fashion. For this, Yuan (2010) and Cai *et al.* (2011) propose the graphical Dantzig selector and CLIME, respectively, which can be solved by linear programming. More recently, Liu and Luo (2012) and Sun and Zhang (2012) have proposed the SCIO and scaled-lasso methods. Compared to the penalized likelihood methods, the column-by-column estimation methods are computationally simpler and are more amenable to theoretical analysis.

In the rest of this chapter, we explain the main ideas of the column-by-column precision matrix estimation methods. We start with an introduction of notations. Letting $\mathbf{v} := (v_1, \cdots, v_N)' \in \mathbb{R}^N$ and $I(\cdot)$ be the indicator function, for $0 < q < \infty$, we define

$$\|\mathbf{v}\|_q := \left(\sum_{j=1}^{N} |v_j|^q \right)^{1/q}, \|\mathbf{v}\|_0 := \sum_{j=1}^{N} I(v_j \neq 0), \quad \text{and} \quad \|\mathbf{v}\|_\infty := \max_j |v_j|.$$

Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ be a symmetric matrix and $I, J \subset \{1, \cdots, N\}$ be two sets. Denote by $\mathbf{A}_{I,J}$ the submatrix of $\mathbf{A}$ with rows and columns indexed by I and J. Letting $\mathbf{A}_{*j}$ be the j^{th} column of $\mathbf{A}$ and $\mathbf{A}_{*\setminus j}$ be the submatrix of $\mathbf{A}$ with the j^{th} column $\mathbf{A}_{*j}$ removed. We define the following matrix norms:

$$\|\mathbf{A}\|_q := \max_{\|\mathbf{v}\|_q = 1} \|\mathbf{A}\mathbf{v}\|_q \quad , \|\mathbf{A}\|_{\max} := \max_{jk} |\mathbf{A}_{jk}|, \quad \text{and} \quad \|\mathbf{A}\|_F = \left(\sum_{j,k} |\mathbf{A}_{jk}|^2 \right)^{1/2}.$$

We also denote $\Lambda_{\max}(\mathbf{A})$ and $\Lambda_{\min}(\mathbf{A})$ to be the largest and smallest eigenvalues of $\mathbf{A}$.

The column-by-column precision matrix estimation method exploits the relationship between conditional distribution of multivariate Gaussian and linear regression. More specifically, letting $\boldsymbol{Y} \sim \mathcal{N}_N(\mathbf{0}, \mathbf{\Sigma})$, the conditional distribution of Y_j given $\boldsymbol{Y}_{\setminus j}$ satisfies

$$Y_j \mid \boldsymbol{Y}_{\setminus j} \sim \mathcal{N}_{N-1}(\mathbf{\Sigma}_{\setminus j,j}(\mathbf{\Sigma}_{\setminus j,\setminus j})^{-1}\boldsymbol{Y}_{\setminus j}, \mathbf{\Sigma}_{jj} - \mathbf{\Sigma}_{\setminus j,j}(\mathbf{\Sigma}_{\setminus j,\setminus j})^{-1}\mathbf{\Sigma}_{\setminus j,j}).$$

Let $\boldsymbol{\alpha}_j := (\mathbf{\Sigma}_{\setminus j,\setminus j})^{-1}\mathbf{\Sigma}_{\setminus j,j} \in \mathbb{R}^{N-1}$ and $\sigma_j^2 := \mathbf{\Sigma}_{jj} - \mathbf{\Sigma}_{\setminus j,j}(\mathbf{\Sigma}_{\setminus j,\setminus j})^{-1}\mathbf{\Sigma}_{\setminus j,j}$. We have

$$Y_j = \boldsymbol{\alpha}_j' \boldsymbol{Y}_{\setminus j} + \epsilon_j, \tag{6.15}$$

where $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ is independent of $\boldsymbol{Y}_{\setminus j}$. By the block matrix inversion formula, we have

$$\mathbf{\Theta}_{jj} = (\text{Var}\ (\epsilon_j))^{-1} = \sigma_j^{-2}, \tag{6.16}$$

$$\mathbf{\Theta}_{\setminus j,j} = -(\text{Var}\ (\epsilon_j))^{-1}\boldsymbol{\alpha}_j = -\sigma_j^{-2}\boldsymbol{\alpha}_j. \tag{6.17}$$

Therefore, we can recover $\mathbf{\Theta}$ in a column-by-column manner by regressing Y_j on $\boldsymbol{Y}_{\setminus j}$ for $j = 1, 2, \cdots, N$. For example, let $\mathbf{Y} \in \mathbb{R}^{T \times N}$ be the data matrix. We denote by

$\boldsymbol{\alpha}_j := (\alpha_{j1}, \cdots, \alpha_{j(N-1)})' \in \mathbb{R}^{N-1}$. Meinshausen and Bühlmann (2006) propose to estimate each $\boldsymbol{\alpha}_j$ by solving the lasso regression:

$$\hat{\boldsymbol{\alpha}}_j = \arg \min_{\boldsymbol{\alpha}_j \in \mathbb{R}^{N-1}} \frac{1}{2T} \|\mathbf{Y}_{*j} - \mathbf{Y}_{*\backslash j}\boldsymbol{\alpha}_j\|_2^2 + \lambda_j \|\boldsymbol{\alpha}_j\|_1,$$

where λ_j is a tuning parameter. Once $\hat{\boldsymbol{\alpha}}_j$ is given, we get the neighborhood edges by reading out the nonzero coefficients of $\boldsymbol{\alpha}_j$. The final graph estimate $\hat{\mathrm{G}}$ is obtained by either the "AND" or "OR" rule on combining the neighborhoods for all the N nodes. However, the neighborhood pursuit method of Meinshausen and Bühlmann (2006) only estimates the graph G but cannot estimate the inverse covariance matrix $\boldsymbol{\Theta}$.

To estimate $\boldsymbol{\Theta}$, Yuan (2010) proposes to estimate $\boldsymbol{\alpha}_j$ by solving the Dantzig selector:

$$\hat{\boldsymbol{\alpha}}_j = \arg \min_{\boldsymbol{\alpha}_j \in \mathbb{R}^{N-1}} \|\boldsymbol{\alpha}_j\|_1 \quad \text{subject to} \quad \|\mathbf{S}_{\backslash j,j} - \mathbf{S}_{\backslash j,\backslash j}\boldsymbol{\alpha}_j\|_\infty \le \gamma_j,$$

where $\mathbf{S} := T^{-1}\mathbf{Y}'\mathbf{Y}$ is the sample covariance matrix and γ_j is a tuning parameter. Once $\hat{\boldsymbol{\alpha}}_j$ is given, we can estimate σ_j^2 by $\hat{\sigma}_j^2 = [1 - 2\hat{\boldsymbol{\alpha}}_j'\mathbf{S}_{\backslash j,j} + \hat{\boldsymbol{\alpha}}_j'\mathbf{S}_{\backslash j,\backslash j}\hat{\boldsymbol{\alpha}}_j]^{-1}$. We then get the estimator $\hat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$ by plugging $\hat{\boldsymbol{\alpha}}_j$ and $\hat{\sigma}_j^2$ into (6.16) and (6.17). Yuan (2010) analyzes the L_1-norm error $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1$ and shows its minimax optimality over certain model space.

In another work, Sun and Zhang (2012) propose to estimate $\boldsymbol{\alpha}_j$ and σ_j by solving a scaled-lasso problem:

$$\hat{\mathbf{b}}_j, \hat{\sigma}_j = \arg \min_{\mathbf{b}=(b_1,\cdots,b_N)',\sigma} \left\{ \frac{\mathbf{b}_j'\mathbf{S}\mathbf{b}_j}{2\sigma} + \frac{\sigma}{2} + \lambda \sum_{k=1}^{N} \mathbf{S}_{kk}|b_k| \quad \text{subject to} \quad b_j = -1 \right\}.$$

Once $\hat{\mathbf{b}}_j$ is obtained, $\boldsymbol{\alpha}_j = \hat{\mathbf{b}}_{\backslash j}$. Sun and Zhang (2012) provide the spectral-norm rate of convergence of the obtained precision matrix estimator.

Cai *et al.* (2011) proposes the CLIME estimator, which directly estimates the jth column of $\boldsymbol{\Theta}$ by solving

$$\hat{\boldsymbol{\Theta}}_{*j} = \arg \min_{\boldsymbol{\Theta}_{*j}} \|\boldsymbol{\Theta}_{*j}\|_1 \quad \text{subject to} \quad \|\mathbf{S}\boldsymbol{\Theta}_{*j} - \boldsymbol{e}_j\|_\infty \le \delta_j, \quad \text{for} \quad j = 1, \cdots, N,$$

where $\boldsymbol{e}_j$ is the jth canonical vector and δ_j is a tuning parameter. This optimization problem can be formulated into a linear program and has the potential to scale to large problems. In a closely related work of CLIME, Liu and Luo (2012) propose the SCIO estimator, which solves the jth column of $\boldsymbol{\Theta}$ by

$$\hat{\boldsymbol{\Theta}}_{*j} = \arg \min_{\boldsymbol{\Theta}_{*j}} \left\{ \frac{1}{2}\boldsymbol{\Theta}_{*j}'\mathbf{S}\boldsymbol{\Theta}_{*j} - \boldsymbol{e}_j'\boldsymbol{\Theta}_{*j} + \lambda_j \|\boldsymbol{\Theta}_{*j}\|_1 \right\}.$$

The SCIO estimator can be solved efficiently by the pathwise coordinate descent algorithm.

6.3.2 *The Need for Tuning-insensitive Procedures*

Most of the methods described in Section 6.3.1 require choosing some tuning parameters that control the bias–variance tradeoff. Their theoretical justifications are usually built on some

theoretical choices of tuning parameters that cannot be implemented in practice. For example, in the neighborhood pursuit method and the graphical Dantzig selector, the tuning parameter λ_j and γ_j depend on σ_j^2, which is unknown. The tuning parameters of the CLIME and SCIO depend on $\|\mathbf{\Theta}\|_1$, which is unknown.

Choosing the regularization parameter in a data-dependent way remains an open problem. Popular techniques include the C_p-statistic, AIC (Akaike information criterion), BIC (Bayesian information criterion), extended BIC (Chen and Chen, 2008, 2012; Foygel and Drton, 2010), RIC (risk inflation criterion; Foster and George, 1994), cross validation, and covariance penalization (Efron, 2004). Most of these methods require data splitting and have been only justified for low-dimensional settings. Some progress has been made recently on developing likelihood-free regularization selection techniques, including permutation methods (Boos *et al.*, 2009; Lysen, 2009; Wu *et al.*, 2007) and subsampling methods (Bach, 2008; Ben-david *et al.*, 2006; Lange *et al.*, 2004; Meinshausen and Bühlmann, 2010). Meinshausen and Bühlmann (2010), Bach (2008), and Liu *et al.* (2012) also propose to select the tuning parameters using subsampling. However, these subsampling-based methods are computationally expensive and still lack theoretical guarantees.

To handle the challenge of tuning parameter selection, we introduce a "tuning-insensitive" procedure for estimating the precision matrix of high-dimensional Gaussian graphical models. Our method, named TIGER (tuning-insensitive graph estimation and regression), is asymptotically tuning-free and only requires very few efforts to choose the regularization parameter in finite sample settings.

6.3.3 TIGER: A Tuning-insensitive Approach for Optimal Precision Matrix Estimation

The main idea of the TIGER method is to estimate the precision matrix $\mathbf{\Theta}$ in a column-by-column fashion. For each column, the computation is reduced to a sparse regression problem. This idea has been adopted by many methods described in Section 6.3.1. These methods differ from each other mainly by how they solve the sparse regression subproblem. Unlike these existing methods, TIGER solves this sparse regression problem using the SQRT-lasso (Belloni *et al.*, 2012).

The SQRT-lasso is a penalized optimization algorithm for solving high-dimensional linear regression problems. For a linear regression problem $\tilde{\boldsymbol{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\tilde{\boldsymbol{Y}} \in \mathbb{R}^T$ is the response, $\tilde{\mathbf{X}} \in \mathbb{R}^{T\times N}$ is the design matrix, $\boldsymbol{\beta} \in \mathbb{R}^N$ is the vector of unknown coefficients, and $\boldsymbol{\epsilon} \in \mathbb{R}^T$ is the noise vector. The SQRT-lasso estimates $\boldsymbol{\beta}$ by solving

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}\in\mathbb{R}^N} \left\{ \frac{1}{\sqrt{T}} \|\tilde{\boldsymbol{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2 + \lambda \|\boldsymbol{\beta}\|_1 \right\},$$

where λ is the tuning parameter. It is shown in Belloni *et al.* (2012) that the choice of λ for the SQRT-lasso method is asymptotically universal and does not depend on any unknown parameter. In contrast, most other methods, including the lasso and Dantzig selector, rely heavily on a known standard deviation of the noise. Moreover, the SQRT-lasso method achieves near oracle performance for the estimation of $\boldsymbol{\beta}$.

In Liu and Wang (2012), they show that the objective function of the scaled-lasso can be viewed as a variational upper bound of the SQRT-lasso. Thus, the TIGER method is essentially equivalent to the method in Sun and Zhang (2012). However, the SQRT-lasso is more amenable to theoretical analysis and allows us to simultaneously establish optimal rates of convergence for the precision matrix estimation under many different norms.

Let $\hat{\mathbf{\Gamma}} := \text{diag}(\mathbf{S})$ be an N-dimensional diagonal matrix with the diagonal elements the same as those in $\mathbf{S}$. Conditioned on the observed data $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_T$, we define

$$\boldsymbol{Z} := (Z_1, \cdots, Z_N)' = \boldsymbol{Y}\hat{\mathbf{\Gamma}}^{-1/2}.$$

By (6.15), we have

$$Z_j\hat{\mathbf{\Gamma}}_{jj}^{1/2} = \boldsymbol{\alpha}_j'\boldsymbol{Z}_{\setminus j}\hat{\mathbf{\Gamma}}_{\setminus j,\setminus j}^{1/2} + \epsilon_j, \tag{6.18}$$

We define

$$\boldsymbol{\beta}_j := \hat{\mathbf{\Gamma}}_{\setminus j,\setminus j}^{1/2}\hat{\mathbf{\Gamma}}_{jj}^{-1/2}\boldsymbol{\alpha}_j \qquad \text{and} \qquad \tau_j^2 = \sigma_j^2\hat{\mathbf{\Gamma}}_{jj}^{-1}.$$

Therefore, we have

$$Z_j = \boldsymbol{\beta}_j'\boldsymbol{Z}_{\setminus j} + \hat{\mathbf{\Gamma}}_{jj}^{-1/2}\epsilon_j. \tag{6.19}$$

We define $\hat{\mathbf{R}}$ to be the sample correlation matrix: $\hat{\mathbf{R}} := (\text{diag}(\mathbf{S}))^{-1/2}\mathbf{S}(\text{diag}(\mathbf{S}))^{-1/2}$. Motivated by the model in (6.19), we propose the following precision matrix estimator.

TIGER Algorithm

For $j = 1, \ldots, N$, we estimate the jth column of $\mathbf{\Theta}$ by solving:

$$\hat{\boldsymbol{\beta}}_j := \arg\min_{\boldsymbol{\beta}_j \in \mathbb{R}^{N-1}} \left\{ \sqrt{1 - 2\boldsymbol{\beta}'_j\hat{\mathbf{R}}_{\setminus j,j} + \boldsymbol{\beta}'_j\hat{\mathbf{R}}_{\setminus j,\setminus j}\boldsymbol{\beta}_j} + \lambda\|\boldsymbol{\beta}_j\|_1 \right\}, \tag{6.20}$$

$$\hat{\tau}_j := \sqrt{1 - 2\hat{\boldsymbol{\beta}}'_j\hat{\mathbf{R}}_{\setminus j,j} + \hat{\boldsymbol{\beta}}'_j\hat{\mathbf{R}}_{\setminus j,\setminus j}\hat{\boldsymbol{\beta}}_j}, \tag{6.21}$$

$$\hat{\mathbf{\Theta}}_{jj} = \hat{\tau}_j^{-2}\hat{\mathbf{\Gamma}}_{jj}^{-1} \text{ and } \hat{\mathbf{\Theta}}_{\setminus j,j} = -\hat{\tau}_j^{-2}\hat{\mathbf{\Gamma}}_{jj}^{-1/2}\hat{\mathbf{\Gamma}}_{\setminus j,\setminus j}^{-1/2}\hat{\boldsymbol{\beta}}_j.$$

For the estimator in (6.20), λ is a tuning parameter. In Section 6.3.4, we show that by choosing $\lambda = \pi\sqrt{\frac{\log N}{2T}}$, the obtained estimator achieves the optimal rates of convergence in the asymptotic setting. Therefore, the TIGER procedure is asymptotically tuning free. For finite samples, we set

$$\lambda := \zeta\pi\sqrt{\frac{\log N}{2T}} \tag{6.22}$$

with ζ chosen from a range $[\sqrt{2}/\pi, 2]$. Since the choice of ζ does not depend on any unknown parameters, we call the procedure *tuning-insensitive*. Practically, we found that simply setting $\zeta = 1$ gives satisfactory finite sample performance in most applications.

If a symmetric precision matrix estimate is preferred, we conduct the following correction: $\tilde{\mathbf{\Theta}}_{jk} = \min\{\hat{\mathbf{\Theta}}_{jk}, \hat{\mathbf{\Theta}}_{kj}\}$ for all $k \neq j$. Another symmetrization method is

$$\tilde{\mathbf{\Theta}} = \frac{\hat{\mathbf{\Theta}} + \hat{\mathbf{\Theta}}'}{2}.$$

As has been shown by Cai *et al.* (2011), if $\hat{\mathbf{\Theta}}$ is a good estimator, then $\tilde{\mathbf{\Theta}}$ will also be a good estimator: they achieve the same rates of convergence in the asymptotic settings.

Let $\mathbf{Z} \in \mathbb{R}^{T \times N}$ be the normalized data matrix, that is, $\mathbf{Z}_{*j} = \mathbf{Y}_{*j}\mathbf{\Sigma}_{jj}^{-1/2}$ for $j = 1, \cdots, N$. An equivalent form of (6.20) and (6.21) is

$$\hat{\boldsymbol{\beta}}_j = \arg \min_{\boldsymbol{\beta}_j \in \mathbb{R}^{N-1}} \left\{ \frac{1}{\sqrt{T}} \|\mathbf{Z}_{*j} - \mathbf{Z}_{*\setminus j}\boldsymbol{\beta}_j\|_2 + \lambda \|\boldsymbol{\beta}_j\|_1 \right\}, \tag{6.23}$$

$$\hat{\tau}_j = \frac{1}{\sqrt{T}} \|\mathbf{Z}_{*j} - \mathbf{Z}_{*\setminus j}\hat{\boldsymbol{\beta}}_j\|_2. \tag{6.24}$$

Once $\hat{\mathbf{\Theta}}$ is estimated, we can also estimate the graph $\hat{\mathrm{G}} := (V, \hat{\mathrm{E}})$ based on the sparsity pattern of $\hat{\mathbf{\Theta}}_{jk} \neq 0$.

6.3.4 Computation

Instead of directly solving (6.20) and (6.21), we consider the following optimization:

$$\hat{\boldsymbol{\beta}}_j, \hat{\tau}_j := \arg \min_{\boldsymbol{\beta}_j \in \mathbb{R}^{N-1}, \tau_j \geq 0} \left\{ \frac{1 - 2\boldsymbol{\beta}_j'\hat{\mathbf{R}}_{\setminus j,j} + \boldsymbol{\beta}_j'\hat{\mathbf{R}}_{\setminus j,\setminus j}\boldsymbol{\beta}_j}{2\tau_j} + \frac{\tau_j}{2} + \lambda \|\boldsymbol{\beta}_j\|_1 \right\}, \tag{6.25}$$

Liu and Wang (2012) show that the solution to (6.20) and (6.21) is the same as that to (6.25). Equation (6.25) is jointly convex with respect to $\boldsymbol{\beta}_j$ and τ_j and can be solved by a coordinate-descent procedure. In the tth iteration, for a given $\tau_j^{(t)}$, we first solve a subproblem

$$\boldsymbol{\beta}_j^{(t+1)} := \arg \min_{\boldsymbol{\beta}_j \in \mathbb{R}^{N-1}} \left\{ \frac{1 - 2\boldsymbol{\beta}_j'\hat{\mathbf{R}}_{\setminus j,j} + \boldsymbol{\beta}_j'\hat{\mathbf{R}}_{\setminus j,\setminus j}\boldsymbol{\beta}_j}{2\tau_j^{(t)}} + \lambda \|\boldsymbol{\beta}_j\|_1 \right\},$$

This is a lasso problem and can be efficiently solved by the coordinate-descent algorithm developed by Friedman *et al.* (2007). Once $\boldsymbol{\beta}_j^{(t+1)}$ is obtained, we can calculate $\tau_j^{(t+1)}$ as

$$\tau_j^{(t+1)} = \sqrt{1 - 2(\boldsymbol{\beta}_j^{(t+1)})'\hat{\mathbf{R}}_{\setminus j,j} + (\boldsymbol{\beta}_j^{(t+1)})'\hat{\mathbf{R}}_{\setminus j,\setminus j}(\boldsymbol{\beta}_j^{(t+1)})}.$$

We iterate these two steps until the algorithm converges.

6.3.5 Theoretical Properties of TIGER

Liu and Wang (2012) establish the rates of convergence of the TIGER estimator $\hat{\mathbf{\Theta}}$ to the true precision matrix $\mathbf{\Theta}$ under different norms. In particular, let $\|\mathbf{\Theta}\|_{\max} := \max_{jk} |\mathbf{\Theta}_{jk}|$ and

$\|\boldsymbol{\Theta}\|_1 := \max_j \sum_k |\boldsymbol{\Theta}_{jk}|$. Under the assumption that the condition number of $\boldsymbol{\Theta}$ is bounded by a constant, they establish the element-wise sup-norm rate of convergence:

$$\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{\max} = O_P\left(\|\boldsymbol{\Theta}\|_1 \sqrt{\frac{\log N}{T}}\right). \tag{6.26}$$

Under mild conditions, the obtained rate in (6.26) is minimax optimal over the model class consisting of precision matrices with bounded condition numbers.

Let $I(\cdot)$ be the indicator function and $s := \sum_{j \neq k} I(\boldsymbol{\Theta}_{jk} \neq 0)$ be the number of nonzero off-diagonal elements of $\boldsymbol{\Theta}$. The result in (6.26) implies that the Frobenious norm error between $\hat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}$ satisfies:

$$\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F := \sqrt{\sum_{i,j} |\hat{\boldsymbol{\Theta}}_{jk} - \boldsymbol{\Theta}_{jk}|^2} = O_P\left(\|\boldsymbol{\Theta}\|_1 \sqrt{\frac{(N+s)\log N}{T}}\right). \tag{6.27}$$

The rate in (6.27) is the minimax optimal rate for the Frobenious norm error in the same model class consisting of precision matrices with bounded condition numbers.

Let $\|\boldsymbol{\Theta}\|_2$ be the largest eigenvalue of $\boldsymbol{\Theta}$ (i.e., $\|\boldsymbol{\Theta}\|_2$ is the spectral norm of $\boldsymbol{\Theta}$) and $k := \max_{i=1,\cdots,N} \sum_j I(\boldsymbol{\Theta}_{ij} \neq 0)$. Liu and Wang (2010) also show that

$$\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_2 \leq \|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1 = O_P\left(k\|\boldsymbol{\Theta}\|_2 \sqrt{\frac{\log N}{T}}\right). \tag{6.28}$$

This spectral norm rate in (6.28) is also minimax optimal over the same model class as before.

6.3.6 *Applications to Modeling Stock Returns*

We apply the TIGER method to explore a stock price dataset collected from Yahoo! Finance (finance.yahoo.com). More specifically, the daily closing prices were obtained for 452 stocks that were consistently in the S&P 500 index between January 1, 2003, through January 1, 2011. This gives us altogether 2015 data points, and each data point corresponds to the vector of closing prices on a trading day. With $S_{t,j}$ denoting the closing price of stock j on day t, we consider the log-return variable $Y_{jt} = \log\left(S_{t,j}/S_{t-1,j}\right)$ and build graphs over the indices j.

We Winsorize (or truncate) every stock so that its data points are within six times the mean absolute deviation from the sample average. In Figure 6.3, we show boxplots for 10 randomly chosen stocks. We see that the data contain outlier even after Winsorization; the reasons for these outliers include splits in a stock, which increase the number of shares. It is known that the log-return data are heavy-tailed. To suitably apply the TIGER method, we Gaussianize the marginal distribution of the data by the normal-score transformation. In Figure 6.3b, we compare the boxplots of the data before and after Gaussianization. We see that Gaussianization alleviates the effect of outliers.

In this analysis, we use the subset of the data between January 1, 2003, and January 1, 2008, before the onset of the financial crisis. The 452 stocks are categorized into 10 Global Industry Classification Standard (GICS) sectors, including `Consumer Discretionary` (70 stocks), `Consumer Staples` (35 stocks), `Energy` (37 stocks), `Financials` (74 stocks),

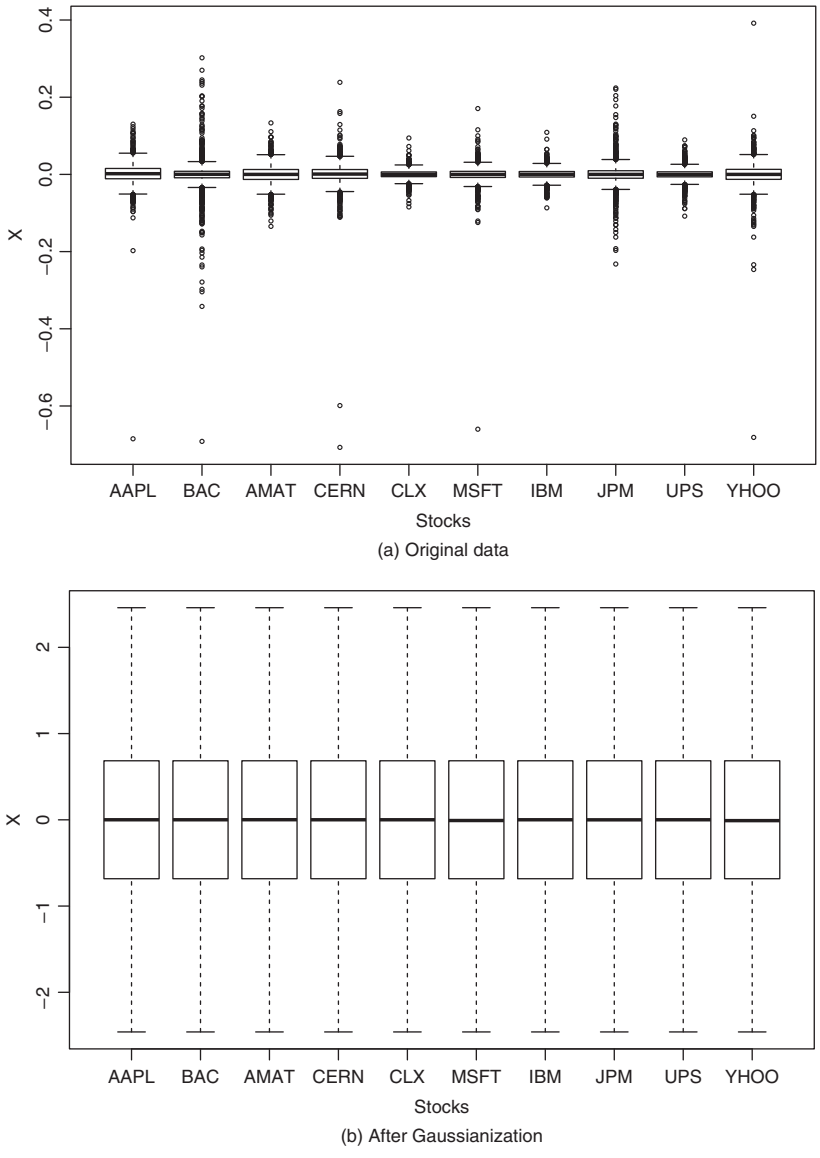

Figure 6.3 Boxplots of $Y_{jt} = \log\,(S_{t,j}/S_{t-1,j})$ for 10 stocks. As can be seen, the original data has many outliers, which is addressed by the normal-score transformation on the rescaled data (right).

`Health Care` (46 stocks), `Industrials` (59 stocks), `Information Technology` (64 stocks), `Materials` (29 stocks), `Telecommunications Services` (6 stocks), and `Utilities` (32 stocks). It is expected that stocks from the same GICS sectors should tend to be clustered together in the estimated graph, since stocks from the same GICS sector tend to interact more with each other. In Figure 6.4, the nodes are colored according to the GICS sector of the corresponding stock.

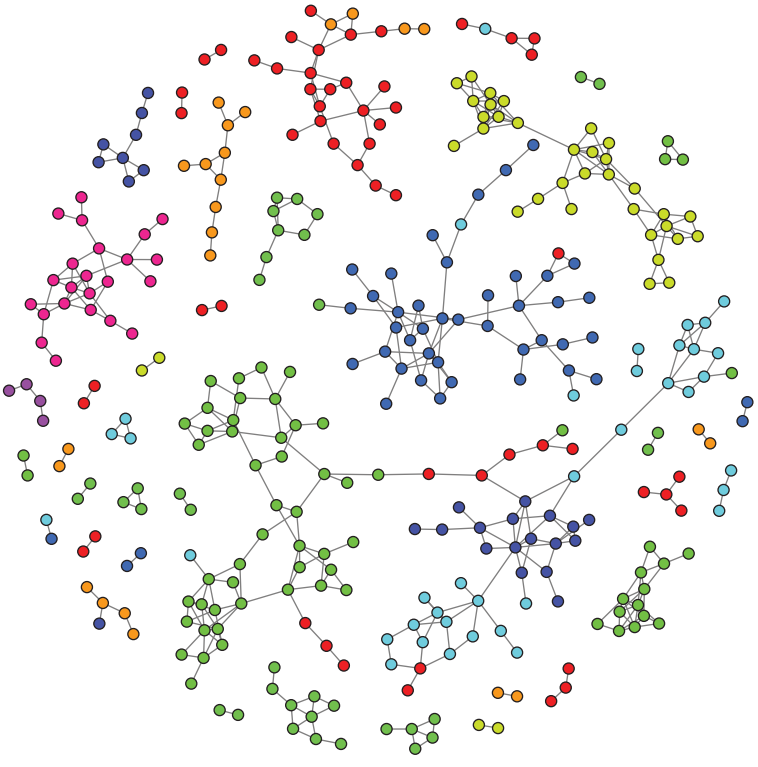

Figure 6.4 The estimated TIGER graph using the S&P 500 stock data from January 1, 2003, to January 1, 2008.

In Figure 6.4 we visualize the estimated graph using the TIGER method on the data from January 1, 2003, to January 1, 2008. There are altogether $T = 1257$ data points and $N = 452$ dimensions. Even though the TIGER procedure is asymptotically tuning-free, Liu and Wang (2010) show that a fine-tune step can further improve its finite sample performance. To fine-tune the tuning parameter, we adopt a variant of the stability selection method proposed by Meinshausen and Bühlmann (2010). As suggested in (6.22), we consider 10 equal-distance values of ζ chosen from a range $[\sqrt{2}/\pi, 2]$. We randomly sample 100 sub-datasets, each containing $B = \lfloor 10\sqrt{T} \rfloor = 320$ data points. On each of these 100 subsampled datasets, we estimate a TIGER graph for each tuning parameter. In the final graph shown in Figure 6.4, we use $\zeta = 1$, and an edge is present only if it appears more than 80% of the time among the 100 subsampled datasets (with all the singleton nodes removed).

From Figure 6.4, we see that stocks from the same GICS sectors are indeed close to each other in the graph. We refrain from drawing any hard conclusions about the effectiveness of the estimated TIGER graph—how it is used will depend on the application. One potential application of such a graph could be for portfolio optimization. When designing a portfolio, we may want to choose stocks with large graph distances to minimize the investment risk.

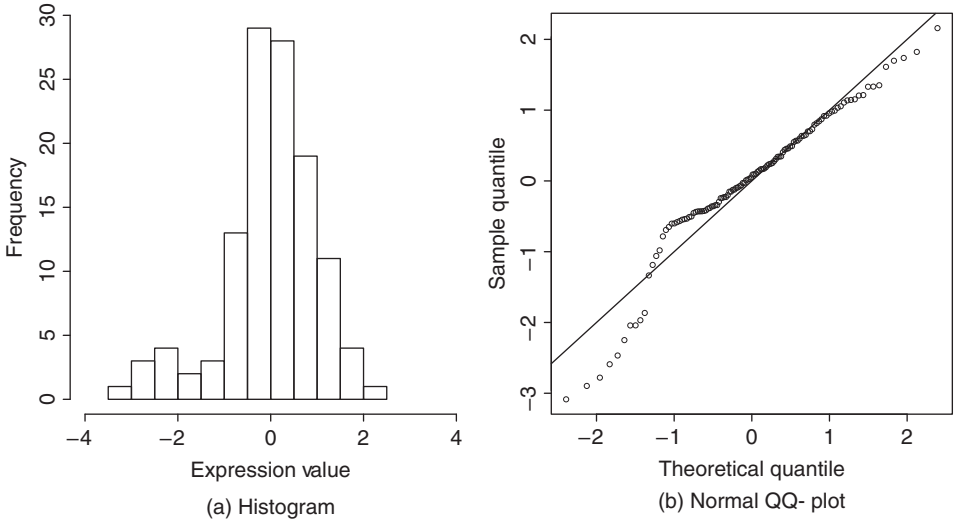

Figure 6.5 The histogram and normal QQ plots of the marginal expression levels of the gene MECPS. We see the data are not exactly Gaussian distributed. Adapted from Liu and Wang (2012).

6.3.7 Applications to Genomic Network

As discussed in this chapter, an important application of precision matrix estimation is to estimate high-dimensional graphical models. In this section, we apply the TIGER method on a gene expression dataset to reconstruct the conditional independence graph of the expression levels of 39 genes.

This dataset, which includes 118 gene expression arrays from *Arabidopsis thaliana*, originally appeared in Wille *et al.* (2004). Our analysis focuses on gene expression from 39 genes involved in two isoprenoid metabolic pathways: 16 from the mevalonate (MVA) pathway are located in the cytoplasm, 18 from the plastidial (MEP) pathway are located in the chloroplast, and 5 are located in the mitochondria. While the two pathways generally operate independently, crosstalk is known to happen (Wille *et al.* 2004). Our scientific goal is to recover the gene regulatory network, with special interest in crosstalk.

We first examine whether the data actually satisfy the Gaussian distribution assumption. In Figure 6.5, we plot the histogram and normal QQ plot of the expression levels of a gene named MECPS. From the histogram, we see the distribution is left-skewed compared to the Gaussian distribution. From the normal QQ plot, we see the empirical distribution has a heavier tail compared to Gaussian. To suitably apply the TIGER method on this dataset, we need to first transform the data so that its distribution is closer to Gaussian. Therefore, we Gaussianize the marginal expression values of each gene by converting them to the corresponding normal-scores. This is automatically done by the `huge.npn` function in the `R` package `huge` (Zhao *et al.*, 2012).

We apply the TIGER on the transformed data using the default tuning parameter $\zeta = \sqrt{2}/\pi$. The estimated network is shown in Figure 6.6. We note that the estimated network is very sparse with only 44 edges. Prior investigations suggest that the connections from genes AACT1 and HMGR2 to gene MECPS indicate a primary source of the crosstalk between the MEP and

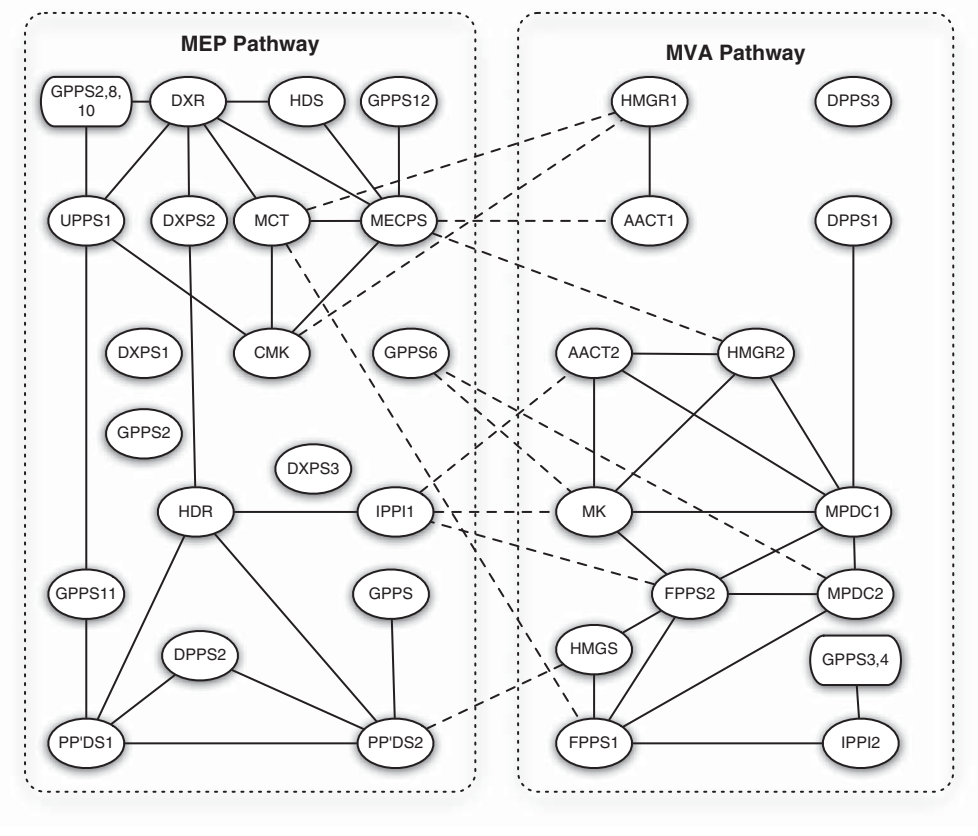

Figure 6.6 The estimated gene networks of the *Arabadopsis* dataset. The within-pathway edges are denoted by solid lines, and between-pathway edges are denoted by dashed lines. From Liu and Wang (2012).

MVA pathways, and these edges are presented in the estimated network. MECPS is clearly a hub gene for this pathway.

For the MEP pathway, the genes DXPS2, DXR, MCT, CMK, HDR, and MECPS are connected as in the true metabolic pathway. Similarly, for the MVA pathway, the genes AACT2, HMGR2, MK, MPDC1, MPDC2, FPPS1, and FPP2 are closely connected. Our analysis suggests 11 cross-pathway links. This is consistent to previous investigation in Wille *et al.* (2004). This result suggests that there might exist rich interpathway crosstalks.

6.4 Financial Applications

6.4.1 Estimating Risks of Large Portfolios

Estimating and assessing the risk of a large portfolio are important topics in financial econometrics and risk management. The risk of a given portfolio allocation vector $\mathbf{w}_N$ is conveniently

measured by $(\mathbf{w}_N'\boldsymbol{\Sigma}\mathbf{w}_N)^{1/2}$, in which $\boldsymbol{\Sigma}$ is a volatility (covariance) matrix of the assets' returns. Often multiple portfolio risks are of interest, and hence it is essential to estimate the volatility matrix $\boldsymbol{\Sigma}$. On the other hand, assets' excess returns are often driven by a few common factors. Hence $\boldsymbol{\Sigma}$ can be estimated via factor analysis as described in Section 6.1.

Let$\{\boldsymbol{Y}_t\}_{t=1}^T$ be a strictly stationary time-series of an $N \times 1$ vector of observed asset returns and $\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{Y}_t)$. We assume that $\boldsymbol{Y}_t$ satisfies an approximate factor model:

$$\boldsymbol{Y}_t = \mathbf{B}\boldsymbol{f}_t + \boldsymbol{u}_t, t \leq T, \tag{6.29}$$

where $\mathbf{B}$ is an $N \times K$ matrix of factor loadings; $\boldsymbol{f}_t$ is a $K \times 1$ vector of common factors; and $\boldsymbol{u}_t$ is an $N \times 1$ vector of idiosyncratic error components. In contrast to N and T, here K is assumed to be fixed. The common factors may or may not be observable. For example, Fama and French (1993) identified three known factors that have successfully described the US stock market. In addition, macroeconomic and financial market variables have been thought to capture systematic risks as observable factors. On the other hand, in an empirical study, Bai and Ng (2002) determined two unobservable factors for stocks traded on the New York Stock Exchange during 1994–1998.

As described in Section 6.1, the factor model implies the following decomposition of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \mathbf{B}\text{Cov}(\boldsymbol{f}_t)\mathbf{B}' + \boldsymbol{\Sigma}_u. \tag{6.30}$$

In the case of observable factors, an estimator of $\boldsymbol{\Sigma}$ is constructed based on thresholding the covariance matrix of idiosyncratic errors, as in (6.7), denoted by $\hat{\boldsymbol{\Sigma}}_f$. In the case of unobservable factors, $\boldsymbol{\Sigma}$ can be estimated by POET as in (6.9), denoted by $\hat{\boldsymbol{\Sigma}}_P$. Because K, the number of factors, might also be unknown, this estimator uses a data-driven number of factors $\hat{K}$. Based on the factor analysis, the risk for a given portfolio $\mathbf{w}_N$ can be estimated by either $\sqrt{\mathbf{w}_N'\hat{\boldsymbol{\Sigma}}_f\mathbf{w}_N}$ or $\sqrt{\mathbf{w}_N'\hat{\boldsymbol{\Sigma}}_P\mathbf{w}_N}$, depending on whether $\boldsymbol{f}_t$ is observable.

6.4.1.1 Estimating a Minimum Variance Portfolio

There are also many methods proposed to choose data-dependent portfolios. For instance, estimated portfolio vectors can arise when the ideal portfolio $\mathbf{w}_N$ depends on the inverse of the large covariance $\boldsymbol{\Sigma}$ (Markowitz, 1952), by consistently estimating $\boldsymbol{\Sigma}^{-1}$. Studying the effects of estimating $\boldsymbol{\Sigma}$ is also important for portfolio allocations. In these problems, estimation errors in estimating $\boldsymbol{\Sigma}$ can have substantial implications (see discussions in El Karoui, 2010). For illustration, consider the following example of estimating the global minimum variance portfolio.

The *global minimum variance* portfolio is the solution to the problem:

$$\mathbf{w}_N^{gmv} = \arg \min_{\mathbf{w}}(\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}), \text{ such that } \mathbf{w}'\mathbf{e} = 1$$

where $\mathbf{e} = (1, \ldots, 1)$, yielding $\mathbf{w}_N^{gmv} = \boldsymbol{\Sigma}^{-1}\mathbf{e}/(\mathbf{e}'\boldsymbol{\Sigma}^{-1}\mathbf{e})$. Although this portfolio does not belong to the efficient frontier, Jagannathan and Ma (2003) showed that its performance is comparable with those of other tangency portfolios.

The factor model yields a positive definite covariance estimator for $\mathbf{\Sigma}$, which then leads to a data-dependent portfolio:

$$\hat{\mathbf{w}}_N^{gmv} = \frac{\hat{\mathbf{\Sigma}}^{-1}\mathbf{e}}{\mathbf{e}'\hat{\mathbf{\Sigma}}^{-1}\boldsymbol{e}}, \quad \hat{\mathbf{\Sigma}}^{-1} = \begin{cases} \hat{\mathbf{\Sigma}}_f^{-1} & \text{known factors;} \\ \hat{\mathbf{\Sigma}}_P^{-1}, & \text{unknown factors} \end{cases}.$$

It can be shown that $\hat{\mathbf{w}}_N^{gmv}$ is L_1-consistent, in the sense that

$$\|\hat{\mathbf{w}}_N^{gmv} - \mathbf{w}_N^{gmv}\|_1 = o_P(1).$$

We refer to El Karoui (2010) and Ledoit and Wolf (2003) for further discussions on the effects of estimating large covariance matrices for portfolio selections.

6.4.1.2 Statistical Inference of the Risks

Confidence intervals of the true risk $\mathbf{w}_N'\mathbf{\Sigma}\mathbf{w}_N$ can be constructed based on the estimated risk $\mathbf{w}_N'\hat{\mathbf{\Sigma}}\mathbf{w}_N$, where $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}_f$ or $\hat{\mathbf{\Sigma}}_P$, depending on whether the factors are known or not. Fan *et al.* (2014a) showed that, under some regularity conditions, respectively,

$$\left[\text{Var}\left(\sum_{t=1}^{T}(\mathbf{w}_N'\mathbf{B}\boldsymbol{f}_t)^2\right)\right]^{-1/2} T\hat{\mathbf{w}}_N'(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma})\hat{\mathbf{w}}_N \to^d \mathcal{N}(0,1), \ \ \hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}_f \text{or } \hat{\mathbf{\Sigma}}_P,$$

where $\hat{\mathbf{w}}_N$ is an L_1-consistent estimator of $\mathbf{w}_N$.

An important implication is that the asymptotic variance is the same regardless of whether the factors are observable or not. Therefore, the impact of estimating the unknown factors is asymptotically negligible. In addition, it can also be shown that the asymptotic variance is slightly smaller than that of $\mathbf{w}_N'\mathbf{S}\mathbf{w}_N$, the sample covariance-based risk estimator. The asymptotic variance $\text{Var}\left(\sum_{t=1}^{T}(\mathbf{w}_N'\mathbf{B}\boldsymbol{f}_t)^2\right)$ can be consistently estimated, using the heteroscedasticity and autocorrelation consistent covariance estimator of Newey and West (1987) based on the truncated sum of estimated autocovariance functions. Therefore, the above limiting distributions can be employed to assess the uncertainty of the estimated risks by, for example, constructing asymptotic confidence intervals for $(\mathbf{w}_N'\mathbf{\Sigma}\mathbf{w}_N)^{1/2}$. Fan *et al.* (2014a) showed that the confidence interval is practically accurate even at the finite sample.

6.4.2 Large Panel Test of Factor Pricing Models

The content of this section is adapted from the recent work by Fan *et al.* (2014b), including graphs and tables. We consider a *factor-pricing model*, in which the excess return has the following decomposition:

$$Y_{it} = \alpha_i + \mathbf{b}_i'\boldsymbol{f}_t + u_{it}, \quad i = 1, \ldots, N, t = 1, \ldots, T. \tag{6.31}$$

In this subsection, we shall focus on the case in which $\boldsymbol{f}_t$'s are observable.

Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$ be the vector of intercepts for all N financial assets. The key implication from the multifactor pricing theory is that $\boldsymbol{\alpha}$ should be zero, known as *mean-variance efficiency*, for any asset i. An important question is then if such a pricing theory can be validated by empirical data, namely whether the null hypothesis

$$H_0 : \boldsymbol{\alpha} = 0, \tag{6.32}$$

is consistent with empirical data.

Most of the existing tests to the problem (6.32) are based on the quadratic statistic $W = \hat{\boldsymbol{\alpha}}' \hat{\boldsymbol{\Sigma}}_u^{-1} \hat{\boldsymbol{\alpha}}$, where $\hat{\boldsymbol{\alpha}}$ is the OLS estimator for $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\Sigma}}_u^{-1}$ is the estimated inverse of the error covariance, and a_T is a positive number that depends on the factors $\boldsymbol{f}_t$ only. Prominent examples are the test given by Gibbons *et al.* (1989), the GMM test in MacKinlay and Richardson (1991), and the likelihood ratio test in Beaulieu *et al.* (2007), all in quadratic forms. Recently, Pesaran and Yamagata (2012) studied the limiting theory of the normalized W assuming $\boldsymbol{\Sigma}_u^{-1}$ were known. They also considered a quadratic test where $\hat{\boldsymbol{\Sigma}}_u^{-1}$ is replaced with its diagonalized matrix.

There are, however, two main challenges in the quadratic statistic W. The first is that estimating $\boldsymbol{\Sigma}_u^{-1}$ is a challenging problem when $N > T$, as described previously. Secondly, even though $\boldsymbol{\Sigma}_u^{-1}$ were known, this test suffers from a lower power in a high-dimensional-low-sample-size situation, as we now explain.

For simplicity, let us temporarily assume that $\{\boldsymbol{u}_t\}_{t=1}^T$ are independent and identically distributed (i.i.d.) Gaussian and $\boldsymbol{\Sigma}_u = \text{Cov}(\boldsymbol{u}_t)$ is known, where $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{Nt})$. Under H_0, W is χ_N^2 distributed, with the critical value $\chi_{N,q}^2$, which is of order N, at significant level q. The test has no power at all when $T\boldsymbol{\alpha}'\boldsymbol{\Sigma}_u\boldsymbol{\alpha} = o(N)$ or $\|\boldsymbol{\alpha}\|^2 = o(N/T)$, assuming that $\boldsymbol{\Sigma}_u$ has bounded eigenvalues. This is not unusual for the high-dimension-low-sample-size situation we encounter, where there are thousands of assets to be tested over a relatively short time period (e.g. 60 monthly data). And it is especially the case when there are only a few significant alphas that arouse market inefficiency. By a similar argument, this problem can not be rescued by using any genuine quadratic statistic, which are powerful only when a non-negligible fraction of assets are mispriced. Indeed, the factor N above reflects the noise accumulation in estimating N parameters of $\boldsymbol{\alpha}$.

6.4.2.1 High-dimensional Wald Test

Suppose $\{\boldsymbol{u}_t\}$ is i.i.d. $\mathcal{N}(0, \boldsymbol{\Sigma}_u)$. Then as $N, T \to \infty$, Pesaran and Yamagata (2012) showed that

$$\frac{Ta\hat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\hat{\boldsymbol{\alpha}} - N}{\sqrt{2N}} \to^d \mathcal{N}(0, 1)$$

where $a = 1 - \frac{1}{T}\sum_t \boldsymbol{f}_t'(\frac{1}{T}\sum_t \boldsymbol{f}_t\boldsymbol{f}_t')^{-1}\frac{1}{T}\sum_t \boldsymbol{f}_t$. This normalized Wald test is infeasible unless $\boldsymbol{\Sigma}_u^{-1}$ is consistently estimable. Under the sparse assumption of $\boldsymbol{\Sigma}_u$, this can be achieved by thresholding estimation as previously described. Letting $\hat{\boldsymbol{\Sigma}}_u^{-1}$ be the thresholding estimator, then a feasible high-dimensional Wald test is

$$J_{sw} \equiv \frac{Ta\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\Sigma}}_u^{-1}\hat{\boldsymbol{\alpha}} - N}{\sqrt{2N}}.$$

With further technical arguments (see Fan *et al.*, 2014b), it can be shown that $J_{sw} \to^d \mathcal{N}(0,1)$. Note that it is very technically involved to show that substituting $\hat{\mathbf{\Sigma}}_u^{-1}$ for $\mathbf{\Sigma}_u^{-1}$ is asymptotically negligible when $N/T \to \infty$.

6.4.2.2 Power Enhancement Test

Traditional tests of factor pricing models are not powerful unless there are enough stocks that have nonvanishing alphas. Even if some individual assets are significantly mispriced, their nontrivial contributions to the test statistic are insufficient to reject the null hypothesis. This problem can be resolved by introducing a power enhancement component (PEM) J_0 to the normalized Wald statistic J_{sw}. The PEM J_0 is a screening statistic, designed to detect sparse alternatives with significant individual alphas.

Specifically, for some predetermined threshold value $\delta_T > 0$, define a set

$$\hat{S} = \left\{ j : \frac{|\hat{\alpha}_j|}{\hat{\sigma}_j} > \delta_T, j = 1, \ldots, N \right\}, \tag{6.33}$$

where $\hat{\alpha}_j$ is the OLS estimator and $\hat{\sigma}_j^2 = \frac{1}{T}\sum_{t=1}^T \hat{u}_{jt}^2/a$ is T times the estimated variance of $\hat{\alpha}_j$, with $\hat{u}_{jt}$ being the regression residuals. Denote a subvector of $\hat{\boldsymbol{\alpha}}$ by

$$\hat{\boldsymbol{\alpha}}_{\hat{S}} = (\hat{\alpha}_j : j \in \hat{S}),$$

the screened-out alpha estimators, which can be interpreted as estimated alphas of mispriced stocks. Let $\hat{\mathbf{\Sigma}}_{\hat{S}}$ be the submatrix of $\hat{\mathbf{\Sigma}}_u$ formed by the rows and columns whose indices are in $\hat{S}$. So $\hat{\mathbf{\Sigma}}_{\hat{S}}/(Ta)$ is an estimated conditional covariance matrix of $\hat{\boldsymbol{\alpha}}_{\hat{S}}$, given the common factors and $\hat{S}$.

With the above notation, we define the screening statistic as

$$J_0 = \sqrt{N}Ta\hat{\boldsymbol{\alpha}}_{\hat{S}}'\hat{\mathbf{\Sigma}}_{\hat{S}}^{-1}\hat{\boldsymbol{\alpha}}_{\hat{S}}. \tag{6.34}$$

The choice of δ_T must suppress most of the noises, resulting in an empty set of $\hat{S}$ under the null hypothesis. On the other hand, δ_T cannot be too large to filter out important signals of alphas under the alternative. For this purpose, noting that the maximum noise level is $O_P(\sqrt{\log\ N/T})$, we let

$$\delta_T = \log\ (\log\ T)\sqrt{\frac{\log\ N}{T}}.$$

This is a high criticism test. When $N = 500$ and $T = 60$, $\delta_T = 3.514$. With this choice of δ_T, if we define, for $\sigma_j^2 = (\mathbf{\Sigma}_u)_{jj}/(1 - E\boldsymbol{f}_t'(E\boldsymbol{f}_t\boldsymbol{f}_t')^{-1}E\boldsymbol{f}_t)$,

$$S = \left\{ j : \frac{|\alpha_j|}{\sigma_j} > 2\delta_T, j = 1, \ldots, N \right\}, \tag{6.35}$$

then under mild conditions, $P(S \subset \hat{S}) \to 1$, with some additional conditions, $P(S = \hat{S}) \to 1$, and $\hat{\boldsymbol{\alpha}}_{\hat{S}}$ behaves like $\boldsymbol{\alpha}_S = (\boldsymbol{\alpha}_j : j \in S)$.

The power enhancement test is then defined to be

$$J_0 + J_{sw},$$

Table 6.3 Variable descriptive statistics for the Fama–French three-factor model (Adapted from Fan *et al.*, 2014b)

Variables	Mean	Std dev.	Median	Min	Max
N_τ	617.70	26.31	621	574	665
$\lvert\hat{S}\rvert_0$	5.49	5.48	4	0	37
$\lvert\tilde{\alpha}\rvert_i^\tau(\%)$	0.9973	0.1630	0.9322	0.7899	1.3897
$\lvert\tilde{\alpha}\rvert_{i\in\hat{S}}^\tau(\%)$	4.3003	0.9274	4.1056	1.7303	8.1299
p-value of J_{wi}	0.2844	0.2998	0.1811	0	0.9946
p-value of J_{sw}	0.1861	0.2947	0.0150	0	0.9926
p-value of PEM	0.1256	0.2602	0.0003	0	0.9836

whose detectable region is the union of those of J_0 and J_{sw}. Note that under the null hypothesis, $S = \emptyset$, so by the selection consistency, $J_0 = 0$ with probability approaching one. Thus, the null distribution of the power enhancement test is that of J_{sw}, which is standard normal. This means adding J_0 does not introduce asymptotic size distortion. On the other hand, since $J_0 \geq 0$, the power of $J_0 + J_{sw}$ is always enhanced. Fan *et al.* (2014b) showed that the test is consistent against the alternative as any subset of:

$$\{\boldsymbol{\alpha} \in \mathbb{R}^N : \max_{j\leq N} |\alpha_j| > 2\delta_T \max_{j\leq N} \sigma_j\} \cup \{\boldsymbol{\alpha} \in \mathbb{R}^N : \|\boldsymbol{\alpha}\|^2 \gg (N \log\ N)/T\}.$$

6.4.2.3 Empirical Study

We study monthly returns on all the S&P 500 constituents from the CRSP database for the period January 1980 to December 2012, during which a total of 1170 stocks have entered the index for our study. Testing of market efficiency is performed on a rolling window basis: for each month from December 1984 to December 2012. The test statistics are evaluated using the preceding 60 months' returns ($T = 60$). The panel at each testing month consists of stocks without missing observations in the past 5 years, which yields a cross-sectional dimension much larger than the time-series dimension ($N > T$). For testing months $\tau = 12/1984, \ldots, 12/2012$, we fit the Fama–French three-factor (FF-3) model:

$$r_{it}^\tau - r_{ft}^\tau = \alpha_i^\tau + \beta_{i,\ \mathrm{MKT}}^\tau(\mathrm{MKT}_t^\tau - r_{ft}^\tau) + \beta_{i,\ \mathrm{SMB}}^\tau \mathrm{SMB}_t^\tau + \beta_{i,\ \mathrm{HML}}^\tau \mathrm{HML}_t^\tau + u_{it}^\tau, \tag{6.36}$$

for $i = 1, \ldots, N_\tau$ and $t = \tau - 59, \ldots, \tau$, where r_{it} represents the return for stock i at month t; r_{ft} is the risk-free rate; and MKT, SMB, and HML constitute the FF-3 model's market, size, and value factors.

Table 6.3 summarizes descriptive statistics for different components and estimates in the model. On average, 618 stocks (which is more than 500 because we are recording stocks that have *ever* become the constituents of the index) enter the panel of the regression during each 5-year estimation window, of which 5.5 stocks are selected by $\hat{S}$. The threshold $\delta_T = \sqrt{\log\ N/T}\log\ (\log\ T)$ is about 0.45 on average, which changes as the panel size N changes for every window of estimation. The selected stocks have much larger alphas than other stocks do, as expected. As far as the signs of those alpha estimates are concerned, 61.84% of all the estimated alphas are positive, and 80.66% of all the selected alphas are positive. This

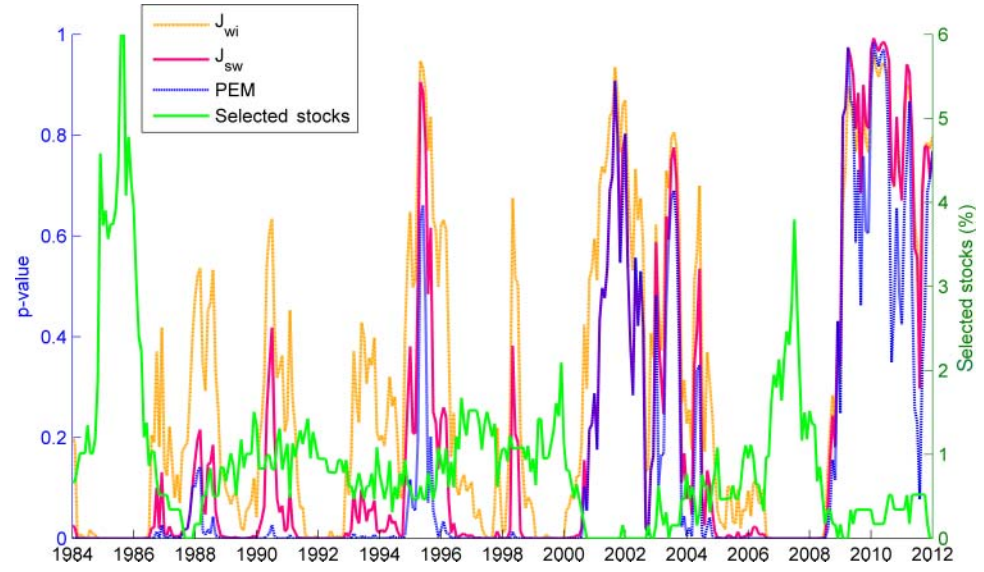

Figure 6.7 Dynamics of p-values and selected stocks (%, from Fan *et al.*, 2014b).

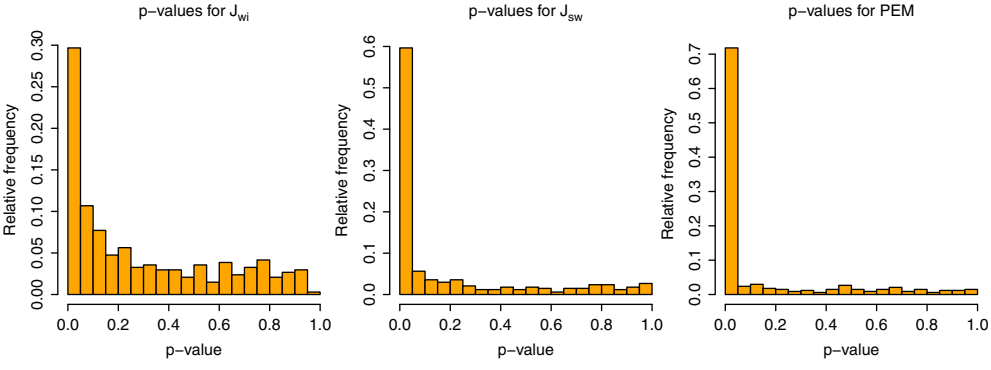

Figure 6.8 Histograms of p-values for J_{wi}, J_{sw}, and PEM (from Fan *et al.*, 2014b).

indicates that market inefficiency is primarily contributed by stocks with extra returns, instead of a large portion of stocks with small alphas, demonstrating the sparse alternatives. In addition, we notice that the p-values of the thresholded Wald test J_{sw} is generally smaller than that of the test J_{wi} given by Pesaran and Yamagata (2012).

We plot the running p-values of J_{wi}, J_{sw}, and the PEM test (augmented from J_{sw}) from December 1984 to December 2012. We also add the dynamics of the percentage of selected stocks ($|\hat{S}|_0/N$) to the plot, as shown in Figure 6.7. There is a strong negative correlation between the stock selection percentage and the p-values of these tests. This shows that the degree of market efficiency is influenced not only by the aggregation of alphas, but also by those extreme ones. We also observe that the p-value line of the PEM test lies beneath those of

J_{sw} and J_{wi} tests as a result of enhanced power, and hence it captures several important market disruptions ignored by the latter two (e.g. Black Monday in 1987, collapse of the Japanese bubble in late 1990, and the European sovereign debt crisis after 2010). Indeed, the null hypothesis of market efficiency is rejected by the PEM test at the 5% level during almost all financial crises, including major financial crises such as Black Wednesday in 1992, the Asian financial crisis in 1997, and the financial crisis in 2008, which are also detected by J_{sw} and J_{wi} tests. For 30%, 60%, and 72% of the study period, J_{wi}, J_{sw}, and the PEM test conclude that the market is inefficient, respectively. The histograms of the p-values of the three test statistics are displayed in Figure 6.8.

6.5 Statistical Inference in Panel Data Models

6.5.1 Efficient Estimation in Pure Factor Models

The sparse covariance estimation can also be employed to improve the estimation efficiency in factor models. Consider:

$$Y_{it} = \mathbf{b}_i' \boldsymbol{f}_t + u_{it}, i \leq N, t \leq T,$$

In the model, only Y_{it} is observable. In most literature, the factors and loadings are estimated via the principal components (PC) method, which solves a constraint minimization problem:

$$\min_{\mathbf{B}, \boldsymbol{f}_t} \sum_{t=1}^{T} (\boldsymbol{Y}_t - \mathbf{B}\boldsymbol{f}_t)'(\boldsymbol{Y}_t - \mathbf{B}\boldsymbol{f}_t)$$

subject to some identifiability constraints so that the solution is unique. The PC method does not incorporate the error covariance $\boldsymbol{\Sigma}_u$, hence it essentially treats the error terms $\boldsymbol{u}_t$ as cross-sectionally homoscedastic and uncorrelated. It is well known that under either cross-sectional heteroscedasticity or correlations, the PC method is not efficient. On the other hand, when $\boldsymbol{\Sigma}_u$ is assumed to be sparse and estimated via thresholding, we can incorporate this covariance estimator into the estimation, and improve the estimation efficiency.

6.5.1.1 Weighted Principal Components

We can estimate the factors and loadings via the weighted least squares. For some $N \times N$ positive definite weight matrix $\mathbf{W}$, solve the following optimization problem:

$$\min_{\mathbf{B}, \boldsymbol{f}_t} \sum_{t=1}^{T} (\boldsymbol{Y}_t - \mathbf{B}\boldsymbol{f}_t)'\mathbf{W}(\boldsymbol{Y}_t - \mathbf{B}\boldsymbol{f}_t),$$

subject to:

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}_t \boldsymbol{f}_t' = \mathbf{I}, \quad \mathbf{B}'\mathbf{W}\mathbf{B} \text{ is diagonal.}$$

Here, $\mathbf{W}$ can be either stochastic or deterministic. When $\mathbf{W}$ is stochastic, it can be understood as a consistent estimator of some deterministic matrix.

Solving the constrained optimization problem gives the WPC estimators: $\hat{\mathbf{b}}_j$ and $\hat{\boldsymbol{f}}_t$ are both $K \times 1$ vectors such that the columns of the $T \times K$ matrix $\hat{\mathbf{F}}/\sqrt{T} = (\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_T)'/\sqrt{T}$ are the eigenvectors corresponding to the largest K eigenvalues of $\mathbf{YWY}'$, and $\hat{\mathbf{B}} = T^{-1}\mathbf{Y}'\hat{\mathbf{F}} = (\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_N)'$. This method is called *weighted principal components* (WPC; see Bai and Liao, 2013), to distinguish from the traditional PC method that uses $\mathbf{W} = \mathbf{I}$. Note that PC does not encounter the problem of estimating large covariance matrices, and is not efficient when $\{u_{it}\}$'s are cross-sectionally correlated across i.

Bai and Liao (2013) studied the inferential theory of the WPC estimators. In particular, they showed that for the estimated common component, as $T, N \to \infty$,

$$\frac{\hat{\mathbf{b}}_i'\hat{\boldsymbol{f}}_t - \mathbf{b}_i'\boldsymbol{f}_t}{(\mathbf{b}_i'\Xi_{\mathbf{W}}\mathbf{b}_i/N + \boldsymbol{f}_t'\boldsymbol{\Omega}_i\boldsymbol{f}_t/T)^{1/2}} \to^d \mathcal{N}(0,1). \tag{6.37}$$

with $\Xi_{\mathbf{W}} = \boldsymbol{\Sigma}_\Lambda^{-1}\mathbf{B}'\mathbf{W}\boldsymbol{\Sigma}_u\mathbf{W}\mathbf{B}\boldsymbol{\Sigma}_\Lambda^{-1}/N$ and $\boldsymbol{\Omega}_i = \text{Cov}(\boldsymbol{f}_t)^{-1}\Phi_i\text{Cov}(\boldsymbol{f}_t)^{-1}$, where

$$\Phi_i = E(\boldsymbol{f}_t\boldsymbol{f}_t'u_{it}^2) + \sum_{t=1}^{\infty} E[(\boldsymbol{f}_1\boldsymbol{f}_{1+t}' + \boldsymbol{f}_{1+t}\boldsymbol{f}_1')u_{i1}u_{i,1+t}],$$

and $\boldsymbol{\Sigma}_\Lambda = \lim_{N\to\infty}\mathbf{B}'\mathbf{W}\mathbf{B}/N$, assumed to exist. Note that although the factors and loadings are not individually identifiable, $\hat{\mathbf{b}}_i'\hat{\boldsymbol{f}}_t$ can consistently estimate the common component $\mathbf{b}_i'\boldsymbol{f}_t$, without introducing a rotational transformation.

6.5.1.2 Optimal Weight Matrix

There are three interesting choices for the weight matrix $\mathbf{W}$. The most commonly seen weight is the identity matrix, which leads to the regular PC estimator. The second choice of the weight matrix takes $\mathbf{W} = \text{diag}^{-1}\{\text{Var}(u_{1t}), \ldots, \text{Var}(u_{Nt})\}$. The third choice is the optimal weight. Note that the asymptotic variance of the estimated common component in (6.37) depends on $\mathbf{W}$ only through

$$\Xi_{\mathbf{W}} = \boldsymbol{\Sigma}_\Lambda^{-1}\mathbf{B}'\mathbf{W}\boldsymbol{\Sigma}_u\mathbf{W}\mathbf{B}\boldsymbol{\Sigma}_\Lambda^{-1}/N.$$

It is straightforward to show that when $\mathbf{W}^* = \boldsymbol{\Sigma}_u^{-1}$, the asymptotic variance is minimized, that is, for any positive definite matrix $\mathbf{W}$, $\Xi_{\mathbf{W}} - \Xi_{\mathbf{W}^*}$ is semipositive definite. In other words, the choice $\mathbf{W} = \boldsymbol{\Sigma}_u^{-1}$ as the weight matrix of the WPC estimator yields the minimum asymptotic variance of the estimated common component.

Table 6.4 gives the estimators and the corresponding weight matrix. The heteroscedastic WPC uses $\mathbf{W} = \mathbf{I}$, which takes into account the cross-sectional heteroscedasticity of $(u_{1t}, \ldots, u_{Nt})$, while the efficient WPC uses the optimal weight matrix $\boldsymbol{\Sigma}_u^{-1}$. Under the sparsity assumption, the optimal weight matrix can be estimated using the POET estimator as described in Section 6.3.

6.5.2 Panel Data Model with Interactive Effects

A closely related model is the panel data with a factor structure in the error term:

$$Y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \quad \varepsilon_{it} = \mathbf{b}_i'\boldsymbol{f}_t + u_{it}, \quad i \le N, t \le T, \tag{6.38}$$

Table 6.4 Three interesting choices of the weight matrix.

	Eigenvectors of	**W**
Regular PC	$\mathbf{YY}'$	$\mathbf{I}$
Heteroscedastic WPC	$\mathbf{Y}\text{diag}(\mathbf{\Sigma}_u)^{-1}\mathbf{Y}'$	$\text{diag}(\mathbf{\Sigma}_u)^{-1}$
Efficient WPC	$\mathbf{Y}\mathbf{\Sigma}_u^{-1}\mathbf{Y}'$	$\mathbf{\Sigma}_u^{-1}$

The estimated $\hat{\mathbf{F}}/\sqrt{T}$ is the eigenvectors of the largest r eigenvalues of $\mathbf{YWY}'$, and $\hat{\mathbf{B}} = T^{-1}\mathbf{Y}'\hat{\mathbf{F}}$.

Table 6.5 Canonical correlations for simulation study (from Bai and Liao, 2013)

		Loadings			Factors			$(\frac{1}{NT}\sum_{i,t}(\hat{\mathbf{b}}_i'\hat{\boldsymbol{f}}_t - \mathbf{b}_i'\boldsymbol{f}_t)^2)^{1/2}$		
T	N	PC	HWPC	EWPC	PC	HWPC	EWPC	PC	HWPC	EWPC
		(The larger the better)			(The larger the better)			(The smaller the better)		
100	80	0.433	0.545	0.631	0.427	0.551	0.652	0.570	0.540	0.496
100	150	0.613	0.761	0.807	0.661	0.835	0.902	0.385	0.346	0.307
100	200	0.751	0.797	0.822	0.827	0.882	0.924	0.333	0.312	0.284
150	100	0.380	0.558	0.738	0.371	0.557	0.749	0.443	0.394	0.334
150	200	0.836	0.865	0.885	0.853	0.897	0.942	0.313	0.276	0.240
150	300	0.882	0.892	0.901	0.927	0.946	0.973	0.257	0.243	0.222

The columns of loadings and factors report the canonical correlations.

where $\boldsymbol{x}_{it}$ is a $d \times 1$ vector of regressors; $\boldsymbol{\beta}$ is a $d \times 1$ vector of unknown coefficients. The regression noise ε_{it} has a factor structure with unknown loadings and factors, regarded as an *interactive effect* of the individual and time effects. In the model, the only observables are $(Y_{it}, \boldsymbol{x}_{it})$. This model has been considered by many researchers, such as Ahn *et al.* (2001), Pesaran (2006), Bai (2009), and Moon and Weidner (2010), and has broad applications in social sciences. For example, in the income studies, Y_{it} represents the income of individual i at age t, and $\boldsymbol{x}_{it}$ is a vector of observable characteristics that are associated with income. Here $\mathbf{b}_i$ represents a vector of unmeasured skills, such as innate ability, motivation, and hardworking; $\boldsymbol{f}_t$ is a vector of unobservable prices for the unmeasured skills, which can be time-varying.

The goal is to estimate the structural parameter $\boldsymbol{\beta}$, whose dimension is fixed. Because the regressor and factor can be correlated, simply regressing Y_{it} on $\boldsymbol{x}_{it}$ is not consistent. Let $\boldsymbol{X}_t = (\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{Nt})'$. The least-squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \arg \min_{\mathbf{B}\boldsymbol{f}_t} \min \sum_{t=1}^{T} (\boldsymbol{Y}_t - \boldsymbol{X}_t\boldsymbol{\beta} - \mathbf{B}\boldsymbol{f}_t)'\mathbf{W}(\boldsymbol{Y}_t - \boldsymbol{X}_t'\boldsymbol{\beta} - \mathbf{B}\boldsymbol{f}_t), \tag{6.39}$$

with a high-dimensional weight matrix $\mathbf{W}$. In particular, it allows a consistent estimator for $\mathbf{\Sigma}_u^{-1}$ as the optimal weight matrix, which takes into account both cross-sectional correlation and heteroscedasticity of u_{it} over i. The minimization is subjected to the constraint $\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}_t\boldsymbol{f}_t'/T = \mathbf{I}$ and $\mathbf{B}'\mathbf{WB}$ being diagonal.

The estimated $\boldsymbol{\beta}$ for each given $(\mathbf{B}, \{\boldsymbol{f}_t\})$ is simply

$$\boldsymbol{\beta}(\mathbf{B}, \{\boldsymbol{f}_t\}) = \left(\sum_{t=1}^{T} \boldsymbol{X}_t'\mathbf{W}\boldsymbol{X}_t\right)^{-1} \sum_{t=1}^{T} \boldsymbol{X}_t'\mathbf{W}(\boldsymbol{Y}_t - \mathbf{B}\boldsymbol{f}_t).$$

On the other hand, given $\boldsymbol{\beta}$, the variable $\boldsymbol{Y}_t - \boldsymbol{X}_t\boldsymbol{\beta}$ has a factor structure. Hence the estimated $(\mathbf{B},\boldsymbol{f}_t)$ are the weighted principal components estimators: let $\boldsymbol{X}(\hat{\boldsymbol{\beta}})$ be an $N \times T$ matrix $\boldsymbol{X}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}_1\hat{\boldsymbol{\beta}}, \ldots, \boldsymbol{X}_T\hat{\boldsymbol{\beta}})$. The columns of the $T \times r$ matrix $\hat{\mathbf{F}}/\sqrt{T} = (\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_T)'/\sqrt{T}$ are the eigenvectors corresponding to the largest r eigenvalues of $(\mathbf{Y}' - \boldsymbol{X}(\hat{\boldsymbol{\beta}}))'\mathbf{W}(\mathbf{Y}' - \boldsymbol{X}(\hat{\boldsymbol{\beta}}))$, and $\hat{\mathbf{B}} = T^{-1}(\mathbf{Y}' - \boldsymbol{X}(\hat{\boldsymbol{\beta}}))\hat{\mathbf{F}}$. Therefore, given $(\mathbf{B},\boldsymbol{f}_t)$, we can estimate $\boldsymbol{\beta}$, and given $\boldsymbol{\beta}$, we can estimate $(\mathbf{B},\boldsymbol{f}_t)$. So $\hat{\boldsymbol{\beta}}$ can be simply obtained by iterations, with an initial value. The inversion $(\sum_{t=1}^{T} \boldsymbol{X}_t'\mathbf{W}\boldsymbol{X}_t)^{-1}$ does not update during iterations.

6.5.2.1 Optimal Weight Matrix

To present the inferential theory of $\hat{\boldsymbol{\beta}}$, additional notation are needed. Rearrange the design matrix

$$\boldsymbol{Z} = (\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1T}, \boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2T}, \ldots, \boldsymbol{X}_{N1}, \ldots, \boldsymbol{X}_{NT})', NT \times \dim\ (\boldsymbol{\beta}).$$

Let

$$\mathbf{A}_{\mathbf{W}} = [\mathbf{W} - \mathbf{WB}(\mathbf{B}'\mathbf{WB})^{-1}\mathbf{B}'\mathbf{W}] \otimes (\mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'/T).$$

Under regularity conditions, Bai and Liao (2013) showed that

$$\sqrt{NT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to^{d} \mathcal{N}(0, \mathbf{V}_{\mathbf{W}}),$$

where, for $\boldsymbol{\Sigma}_u = \mathrm{Cov}(\boldsymbol{u}_t)$,

$$\mathbf{V}_{\mathbf{W}} = \mathrm{plim}_{N,T\to\infty}(\frac{1}{NT}\boldsymbol{Z}'\mathbf{A}_{\mathbf{W}}\boldsymbol{Z})^{-1} \frac{1}{NT}\boldsymbol{Z}'\mathbf{A}_{\mathbf{W}}(\boldsymbol{\Sigma}_u \otimes \mathbf{I})\mathbf{A}_{\mathbf{W}}\boldsymbol{Z}(\frac{1}{NT}\boldsymbol{Z}'\mathbf{A}_{\mathbf{W}}\boldsymbol{Z})^{-1}$$

assuming the right-hand side converges in probability.

It is not difficult to show that $\mathbf{W}^* = \boldsymbol{\Sigma}_u^{-1}$ is the optimal weight matrix, in the sense that $\mathbf{V}_{\mathbf{W}} - \mathbf{V}_{\mathbf{W}^*}$ is semipositive definite for all positive definite weight matrix $\mathbf{W}$. With $\mathbf{W} = \mathbf{W}^*$, the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is

$$\mathbf{V}_{\mathbf{W}^*} = \mathrm{plim}_{N,T\to\infty} = (\frac{1}{NT}\boldsymbol{Z}'\mathbf{A}_{\mathbf{W}^*}\boldsymbol{Z})^{-1}.$$

Assuming $\boldsymbol{\Sigma}_u$ to be sparse, one can estimate $\mathbf{W}^*$ based on an initial estimator of $\boldsymbol{\beta}$. Specifically, define $\hat{\boldsymbol{\beta}}_0$ as in (6.39) with $\mathbf{W} = \mathbf{I}$, which is the estimator used in Bai (2009) and Moon and Weidner (2010). Apply the singular value decomposition to

$$\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{Y}_t - \boldsymbol{X}_t\hat{\boldsymbol{\beta}}_0)(\boldsymbol{Y}_t - \boldsymbol{X}_t\hat{\boldsymbol{\beta}}_0)' = \sum_{i=1}^{N} \nu_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i',$$

where $(\nu_j, \boldsymbol{\xi}_j)_{j=1}^N$ are the eigenvalues–eigenvectors of $\frac{1}{T}\sum_{t=1}^T (\boldsymbol{Y}_t - \boldsymbol{X}_t\hat{\boldsymbol{\beta}}_0)(\boldsymbol{Y}_t - \boldsymbol{X}_t\hat{\boldsymbol{\beta}}_0)'$ in a decreasing order such that $\nu_1 \geq \nu_2 \geq \ldots \geq \nu_N$. Then $\hat{\boldsymbol{\Sigma}}_u = (\hat{\Sigma}_{u,ij})_{N\times N}$,

$$\hat{\Sigma}_{u,ij} = \begin{cases} \tilde{R}_{ii}, & i = j \\ th_{ij}(\tilde{R}_{ij}), & i \neq j \end{cases}, \quad \tilde{R}_{ij} = \sum_{k=r+1}^{N} \nu_k \xi_{ki}\xi_{kj},$$

where $th_{ij}(\cdot)$ is the same thresholding function. The optimal weight matrix $\mathbf{W}^*$ can then be estimated by $\hat{\boldsymbol{\Sigma}}_u^{-1}$, and the resulting estimator $\hat{\boldsymbol{\beta}}$ achieves the asymptotic variance $\mathbf{V}_{\mathbf{W}^*}$.

6.5.3 Numerical Illustrations

We present a simple numerical example to compare the weighted principal components with the popular methods in the literature. The idiosyncratic error terms are generated as follows: let $\{\epsilon_{it}\}_{i\leq N, t\leq T}$ be i.i.d. $\mathcal{N}(0,1)$ in both t, i. Let

$$u_{1t} = \epsilon_{1t},\ u_{2t} = \epsilon_{2t} + a_1\epsilon_{1t},\ u_{3t} = \epsilon_{3t} + a_2\epsilon_{2t} + b_1\epsilon_{1t},$$
$$u_{i+1,t} = \epsilon_{i+1,t} + a_i\epsilon_{it} + b_{i-1}\epsilon_{i-1,t} + c_{i-2}\epsilon_{i-2,t},$$

where $\{a_i, b_i, c_i\}_{i=1}^N$ are i.i.d. $\mathcal{N}(0,1)$. Then $\boldsymbol{\Sigma}_u$ is a banded matrix, with both cross-sectional correlation and heteroscedasticity. Let the two factors $\{f_{1t}, f_{2t}\}$ be i.i.d. $\mathcal{N}(0,1)$, and $\{b_{i,1}, b_{i,2}\}_{i\leq N}$ be uniform on $[0,1]$.

6.5.3.1 Pure Factor Model

Consider the pure factor model $Y_{it} = b_{i1}f_{1,t} + b_{i,2}f_{2t} + u_{it}$. Estimators based on three weight matrices are compared: PC using $\mathbf{W} = \mathbf{I}$; HWPC using $\mathbf{W} = \text{diag}(\boldsymbol{\Sigma}_u)^{-1}$ and EWPC using $\mathbf{W} = \boldsymbol{\Sigma}_u^{-1}$. Here $\boldsymbol{\Sigma}_u$ is estimated using the POET estimator. The smallest canonical correlation (the larger the better) between the estimators and parameters are calculated, as an assessment of the estimation accuracy. The simulation is replicated 100 times, and the average canonical correlations are reported in Table 6.5. The mean squared error of the estimated common components are also compared.

We see that the estimation becomes more accurate when we increase the dimensionality. HWPC improves the regular PC, while the EWPC gives the best estimation results.

6.5.3.2 Interactive Effects

Adding a regression term, we consider the panel data model with interactive effect: $Y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + b_{i1}f_{1,t} + b_{i,2}f_{2t} + u_{it}$, where the true $\boldsymbol{\beta} = (1,3)'$. The regressors are generated to be dependent on $(\boldsymbol{f}_t, \mathbf{b}_i)$:

$$x_{it,1} = 2.5b_{i1}f_{1,t} - 0.2b_{i2}f_{2,t} - 1 + \eta_{it,1}, x_{it,2} = b_{i1}f_{1,t} - 2b_{i2}f_{2,t} + 1 + \eta_{it,2}$$

where $\eta_{it,1}$ and $\eta_{it,2}$ are independent i.i.d. standard normal.

Both methods, PC (Bai, 2009; Moon and Weidner, 2010) and WPC with $\mathbf{W} = \hat{\boldsymbol{\Sigma}}_u^{-1}$, are carried out to estimate $\boldsymbol{\beta}$ for the comparison. The simulation is replicated 100 times; results

Table 6.6 Method comparison for the panel data with interactive effects (from Bai and Liao, 2013)

		$\beta_1 = 1$				$\beta_2 = 3$			
		Mean		Normalized SE		Mean		Normalized SE	
T	N	WPC	PC	WPC	PC	WPC	PC	WPC	PC
100	100	1.002	1.010	0.550	1.418	3.000	3.003	0.416	1.353
100	150	1.003	1.007	0.681	1.626	2.999	3.000	0.611	1.683
100	200	1.002	1.005	0.631	1.800	3.000	3.000	0.774	1.752
150	100	1.003	1.006	0.772	1.399	3.000	2.999	0.714	1.458
150	150	1.001	1.005	0.359	1.318	3.000	3.001	0.408	1.379
150	200	1.001	1.003	0.547	1.566	3.000	3.000	0.602	1.762

"Mean" is the average of the estimators; "Normalized SE" is the standard error of the estimators multiplied by $\sqrt{NT}$.

are summarized in Table 6.6. We see that both methods are almost unbiased, while the efficient WPC indeed has significantly smaller standard errors than the regular PC method in the panel model with interactive effects.

6.6 Conclusions

Large covariance and precision (inverse covariance) matrix estimations have become fundamental problems in multivariate analysis that find applications in many fields, ranging from economics and finance to biology, social networks, and health sciences.

We introduce two efficient methods for estimating large covariance matrices and precision matrices. The introduced precision matrix estimator assumes the precision matrix to be sparse, which is immediately applicable for Gaussian graphical models. It is tuning-parameter insensitive, and simultaneously achieves the minimax optimal rates of convergence in precision matrix estimation under different matrix norms. On the other hand, the estimator based on factor analysis imposes a conditional sparsity assumption. Computationally, our procedures are significantly faster than existing methods. Both theoretical properties and numerical performances of these methods are presented and illustrated. In addition, we also discussed several financial applications of the proposed methods, including risk management, testing high-dimensional factor pricing models. We also illustrate how the proposed covariance estimators can be used to improve statistical efficiency in estimating factor models and panel data models.

References

Ahn, S., Lee, Y. and Schmidt, P. (2001). Gmm estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics*, 101, 219–255.

Ahn S.C. and Horenstein, A.R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81 (3), 1203–1227.

Alessi, L., Barigozzi, M. and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, 80 (23), 1806–1813.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96 (455), 939–955.

Bach, F.R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML). http://www.di.ens.fr/~fbach/fbach_bolasso_icml2008.pdf

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71 (1), 135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77, 1229–1279.

Bai, J. and Liao, Y. (2013). Statistical inferences using large estimated covariances for panel data and factor models. Technical report, University of Maryland.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70, 191–221.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9, 485–516.

Beaulieu, M., Dufour, J. and Khalaf, L. (2007). Multivariate tests of mean-variance efficiency with possibly non-Gaussian errors: an exact simulation based approach. *Journal of Business and Economic Statistics*, 25, 398–410.

Belloni, A., Chernozhukov, V. and Wang, L. (2012). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98, 791–806.

Ben-david, S., Luxburg, U.V. and Pal, D. (2006). A sober look at clustering stability. In *Proceedings of the Annual Conference of Learning Theory*. New York: Springer, pp. 5–19.

Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, 36 (6), 2577–2604.

Boos, D.D., Stefanski, L.A. and Wu, Y. (2009). Fast FSR variable selection with applications to clinical trials. *Biometrics*, 65 (3), 692–700.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106 (494), 672–684.

Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106 (494), 594–607.

Campbell, J.Y., Lo, A.W.C., MacKinlay, A.C., *et al.* (1997). *The econometrics of financial markets*, vol. 2. Princeton, NJ: Princeton University Press.

Candès, E.J., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58 (3), 11.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica*, 51, 1305–1324.

Chandrasekaran, V., Parrilo, P.A. and Willsky, A.S. (2010). Latent variable graphical model selection via convex optimization. In *48th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2010*. Piscataway, NJ: IEEE, pp. 1610–1613.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95 (3), 759–771.

Chen, J. and Chen, Z. (2012). Extended BIC for small-*n*-large-*p* sparse GLM. *Statistica Sinica*, 22, 555–574.

Dempster, A. (1972). Covariance selection. *Biometrics*, 28, 157–175.

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

El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Annals of Statistics*, 36 (6), 2717–2756.

El Karoui, N. (2010). High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: risk underestimation. *Annals of Statistics*, 38 (6), 3487–3566.

Fama, E. and French, K. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47, 427–465.

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147 (1), 186–197.

Fan, J., Liao, Y. and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics*, 39, 3320–3356.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B*, 75, 603–680.

Fan, J., Liao, Y. and Shi, X. (2014a). *Risks of large portfolios*. Technical report. Princeton, NJ: Princeton University.

Fan, J., Liao, Y. and Yao, J. (2014b). Large panel test of factor pricing models. Technical report. Princeton, NJ: Princeton University.

Foster, D.P. and George, E.I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, 22 (4), 1947–1975.

Foygel, R. and Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems*, 23, 604–612.

Friedman, J., Hastie, T. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, 1 (2), 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9 (3), 432–441.

Fryzlewicz, P. (2013). High-dimensional volatility matrix estimation via wavelets and thresholding. *Biometrika*, 100, 921–938.

Gibbons, M., Ross, S. and Shanken, J. (1989). A test of the efficiency of a given portfolio. *Econometrica*, 57, 1121–1152.

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, 102 (478), 603–617.

Han, F., Zhao, T. and Liu, H. (2012). *High dimensional nonparanormal discriminant analysis*. Technical report, Department of Biostatistics. Baltimore: Johns Hopkins Bloomberg School of Public Health.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: why imposing the wrong constraints helps. *Journal of Finance*, 58, 1651–1684.

Jalali, A., Johnson, C. and Ravikumar, P. (2012). High-dimensional sparse inverse covariance estimation using greedy methods. International Conference on Artificial Intelligence and Statistics.

Kapetanios, G. (2010). A testing procedure for determining the number of factors in approximate factor models with large datasets. *Journal of Business & Economic Statistics*, 28 (3), 397–409.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37 (6B), 4254.

Lange, T., Roth, V., Braun, M.L. and Buhmann, J.M. (2004). Stability-based validation of clustering solutions. *Neural Computation*, 16 (6), 1299–1323.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10 (5), 603–621.

Liu, H., Roeder, K. and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS).

Liu, H. and Wang, L. (2012). TIGER: a tuning-insensitive approach for optimally estimating gaussian graphical models. Technical report, Department of Operations Research and Financial Engineering, Princeton University.

Liu, W. and Luo, X. (2012). High-dimensional sparse precision matrix estimation via sparse column inverse operator. arXiv/1203.3896.

Lysen, S. (2009). Permuted inclusion criterion: A variable selection technique. Publicly accessible Penn Dissertations. Paper 28.

MacKinlay, A. and Richardson, M. (1991). Using generalized method of moments to test mean-variance efficiency. *Journal of Finance*, 46, 511–527.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7, 77–91.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34 (3), 1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B, Methodological*, 72, 417–473.

Moon, R. and Weidner, M. (2010). *Linear regression for panel with unknown number of factors as interactive fixed effects*. Technical report. Los Angeles: University of South California.

Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55, 703–708.

Pesaran, H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74, 967–1012.

Pesaran, H. and Yamagata, T. (2012). *Testing capm with a large number of assets*. Technical report. Los Angeles: University of South California.

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*. Hoboken, NJ: John Wiley & Sons.

Ravikumar, P., Wainwright, M.J., Raskutti, G. and Yu, B. (2011a). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5, 935–980.

Ravikumar, P., Wainwright, M.J., Raskutti, G., Yu, B., *et al.* (2011b). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5, 935–980.

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 Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485), 177–186.

Sentana, E. (2009). The econometrics of mean-variance efficiency tests: a survey. *Econometrics Journal*, 12, 65–101.

Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497), 223–232.

Stock, J.H. and Watson, M.W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460), 1167–1179.

Sun, T. and Zhang, C.H. (2012). *Sparse matrix inversion with scaled lasso*. Technical report, Department of Statistics. New Brunswick, NJ: Rutgers University.

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming. *IEEE Transactions on Information Theory*, 55(5), 2183–2201.

Wille, A., Zimmermann, P., Vranova, E., Frholz, A., Laule, O., *et al.* (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biology*, 5 (11), R92.

Wu, Y., Boos, D.D. and Stefanski, L.A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association*, 102, 235–243.

Xue, L. and Zou, H. (2012). Positive-definite ℓ_1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107, 1480–1491.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(8), 2261–2286.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94 (1), 19–35.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7 (11), 2541–2563.

Zhao, T., Liu, H., Roeder, K., Lafferty, J. and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in *r*. *Journal of Machine Learning Research*, forthcoming.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101 (476), 1418–1429.