Estimation of Covariance, Correlation and Precision Matrices for High-dimensional Data

Na Huang

Department of Statistics

London School of Economics and Political Science

A thesis submitted for the degree of

Doctor of Philosophy

March 2016

To Qian, Yimo and my loving parents.

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I confirm that Chapter 2-4 are jointly co-authored with Professor Piotr Fryzlewicz.

Acknowledgements

First of all, I would like to express my heartfelt gratitude to my supervisor Professor Piotr Fryzlewicz for his immense knowledge, inspiring guidance and invaluable encouragement throughout my PhD study. I appreciate his endless stream of ideas, all his contribution of time and feedback to make my PhD experience inspiring and stimulating. His consistent support and understanding help me through many difficulties. I could not have asked for a better supervisor. I am also thankful to my second supervisor, Dr. Matto Barigozzi, not only for all his recommendations on my research, but also for holding an engaging Time Series Reading Group which I had opportunity to attend. My research would not have been possible without my sponsors, the Economic and Social Research Council and the London School of Economics, whose generous financial support is gratefully acknowledged.

I would like to thank the staff and the colleagues at my department, who make my time at LSE a enjoyable experience. I especially thank Ian Marshall for his genuine support. I would like to thank Dr. Haeran Cho for discussions on my research and Dr. Oliver Ledoit for sharing their codes of a competing method. Moreover, I would like to extend my gratitude to Prof. Rainer von Sachs and Dr. Clifford Lam for being my thesis examin-

ers, providing valuable comments and making my viva voce examination pleasurable.

I am deeply indebted to my husband Qian and my parents for their boundless love and consistent support even at the hardest of times, and to my son Yimo who made this experience a much bigger challenge but even inspiring. Further, I wish to thank all my loving friends and relatives here in the UK, back in China, and elsewhere in the world to whom I am indebted for their constant encouragement.

Abstract

The thesis concerns estimating large correlation and covariance matrices and their inverses. Two new methods are proposed. First, tilting-based methods are proposed to estimate the precision matrix of a p-dimensional random variable, X, when p is possibly much larger than the sample size n. Each 2 by 2 block indexed by (i, j) of the precision matrix can be estimated by the inversion of the pairwise sample conditional covariance matrix of X_i and X_j controlling for all the other variables. However, in the high dimensional setting, including too many or irrelevant controlling variables may distort the results. To determine the controlling subsets, the tilting technique is applied to measure the contribution of each remaining variable to the covariance matrix of X_i and X_j, and only puts the (hopefully) highly relevant remaining variables into the controlling subsets. Four types of tilting-based methods are introduced and the properties are demonstrated. The simulation results are presented under different scenarios for the underlying precision matrix. The second method NOVEL Integration of the Sample and Thresholded covariance estimators (NOVELIST) performs shrinkage of the sample covariance (correlation) towards its thresholded version. The sample covariance (correlation) component is non-sparse and can be low-rank in high dimensions. The thresholded sample covariance (correlation) component is sparse, and its addition en-

sures the stable invertibility of NOVELIST. The benefits of the NOVELIST estimator include simplicity, ease of implementation, computational efficiency and the fact that its application avoids eigenanalysis. We obtain an explicit convergence rate in the operator norm over a large class of covariance (correlation) matrices when p and n satisfy $\log p/n \to 0$. In empirical comparisons with several popular estimators, the NOVELIST estimator performs well in estimating covariance and precision matrices over a wide range of models. An automatic algorithm for NOVELIST is developed. Comprehensive applications and real data examples of NOVELIST are presented. Moreover, intensive real data applications of NOVELIST are presented.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Literature review

Estimating the covariance matrix and its inverse, also known as the concentration or precision matrix, has always been an important part of multivariate analysis, and arises prominently. In particular, covariance matrix and its inverse play a central role in portfolio selection and financial risk management. The adequacy of diversification of a portfolio, which is highly related to "risk", is quantified by the covariance matrix of the assets [Markowitz, 1952]. For example, the largest and smallest eigenvalues of the covariance matrix provide the boundary for the variance of return of each possible portfolio allocation [Fan et al., 2008; Markowitz, 1952]. See Ledoit and Wolf [2003], Talih [2003], Goldfarb and Iyengar [2003] and Longerstaey et al. [1996] for applications of covariance matrices to portfolio selection and financial risk management. Also, in principal component analysis, where eigenanalysis of covariance matrix is essential for computing principal components [Croux and Haesbroeck, 2000; Jackson, 1991; Johnstone and Lu, 2009; Pearson, 1901], and in linear discriminant analysis,

where common or individual covariance matrix is inverted in discriminant function for classification or dimension reduction purposes [Bickel and Levina, 2004; Fisher, 1936; Guo et al., 2007]. Moreover, graphical modeling [Meinshausen and Bühlmann, 2008; Ravikumar et al., 2011; Yuan, 2010] with its applications in network science [Gardner et al., 2003; Jeong et al., 2001] require a good covariance matrix estimator inverting which does not excessively amplify the estimation error. Naturally, this is also true of the correlation matrix, and the following discussion applies to the correlation matrix, too. The sample covariance matrix is a straightforward and often used estimator of the covariance matrix [Anderson, 1968]. However, estimating large covariance matrices is intrinsically challenging. When the dimension p of the data grows with the sample size n, the sample covariance matrix is no longer a consistent estimate in the sense that its eigenvalues do not converge to those of the true covariance matrix, according to random matrix theory [Chen et al., 2013; Johnstone, 2001; Marčenko and Pastur, 1967]. Moreover, sample precision matrix is not defined because sample covariance matrix is singular in the high-dimensional setting. Even if p is smaller than but of the same order of magnitude as n, the number of parameters to estimate is $p(p+1)/2$, which can significantly exceed n. In this case, the sample covariance matrix is not reliable, and alternative estimation methods are needed.

We would categorise the most commonly used alternative covariance estimators into two broad classes. Estimators in the first class rely on various structural assumptions on the underlying true covariance. One prominent example is ordered covariance matrices, often appearing in time series analysis, spatial statistics and spatio-temporal modelling; these assume that there is a metric on the variable indices. Bickel and Levina [2008a] develop a class of well-conditioned and approximately "bandable" matrices, and use banding to achieve consistent estimation uniformly over the class as

long as $\log p/n \to 0$ under Gaussianity. Furrer and Bengtsson [2007] and Cai et al. [2010] regularise estimated ordered covariance matrices by tapering. Cai et al. [2010] derive the optimal estimation rates for the covariance matrix under the operator and Frobenius norms, a result which implies sub-optimality of the convergence rate of the banding estimator of Bickel and Levina [2008a] in the operator norm. The banding technique is also applied to the estimated Cholesky factorisation of the inverse of the covariance matrices [Bickel and Levina, 2008a; Wu and Pourahmadi, 2003]. Another important example of a structural assumption on the true covariance or precision matrices is sparsity; it is often made e.g. in the statistical analysis of genetic regulatory networks [Gardner et al., 2003; Jeong et al., 2001]. El Karoui [2008] and Bickel and Levina [2008b] simultaneously and independently regularise the estimated sparse covariance matrix by universal thresholding, which is a simple and permutation-invariant method of covariance regularization. El Karoui [2008] develops thresholding under a special notion of sparsity called β-sparsity, and Bickel and Levina [2008b] study thresholding under another class of sparse matrices, which is stronger and parallels to the class of the approximately "bandable" matrices in [Bickel and Levina, 2008a]. Bickel and Levina [2008b] derive the consistency results of the thresholded estimators with Gaussian and sub-Gaussian models, and show that the results are stronger than those in El Karoui [2008] under suitable assumptions. Adaptive thresholding, in which the threshold is a random function of the data [Cai and Liu, 2011; Fryzlewicz, 2013], leads to more natural thresholding rules and hence, potentially, more precise estimation. Cai and Liu [2011] show that adaptive thresholding estimators can achieve the optimal rate of convergence over a class of sparse covariance matrices under operator norm, while the universal thresholding estimators are shown to be sub-optimal under the same conditions. The Lasso penalty is another popular way to regularise the covari-

ance and precision matrices [d'Aspremont et al., 2008; Friedman et al., 2008; Rothman et al., 2008; Yuan and Lin, 2007; Zou, 2006]. Also, Fan and Li [2007]; Lam and Fan [2009]; Zhao and Yu [2001] addresses explicitly the issues of sparsistency and the bias problem due to L_1 penalization. Upon sparsity assumption, a closely related problem is the estimation of the support of the precision matrix which corresponds to the selection of graphical models for Gaussian distributions [Lauritzen, 1996]. Focusing on model selection rather than parameter estimation, Meinshausen and Bühlmann [2008] propose the neighbourhood selection method with the Lasso technique for estimating the pattern of zero entries in the precision matrix of a multivariate normal distribution, based on which Peng et al. [2009] develop a faster algorithm to select the non-zero partial correlations by using a joint sparse regression model. One other commonly occurring structural assumption in covariance estimation is the factor model, often used e.g. in financial applications. Motivated by the Arbitrage Pricing Theory in finance, Fan, Fan, and Lv [2008] impose a multi-factor model on data to reduce dimensionality and to estimate the covariance matrix, where the factors are observable and the number of factors can grow with dimension p. Fan et al. [2013] propose the POET estimator, which assumes that the covariance matrix is the sum of a part derived from a factor model, and a sparse part.

Different from the estimators in the first class which rely on various structural assumptions on the underlying true covariance, estimators in the second broad class do not assume a specific structure of the covariance or precision matrices, but shrink the sample eigenvalues of the sample covariance matrix towards an assumed shrinkage target [Ledoit and Wolf, 2012]. A considerable number of shrinkage estimators have been proposed along these lines. Ledoit and Wolf [2004] derive an optimal linear shrinkage formula, which imposes the same shrinkage intensity on all sample eigenvalues but

leave the sample eigenvectors unchanged. However, Ledoit and Péché [2011] argue that the differences between the eigenvalues of the sample covariance matrix and those of the population covariance matrix are highly nonlinear and derive the asymptotically optimal bias correction for sample eigenvalues. Based on it, Ledoit and Wolf [2012] extend linear shrinkage to nonlinear shrinkage of the eigenvalues of the sample covariance matrix. Ledoit and Wolf [2013] also derive a consistent estimator of the oracle nonlinear shrinkage based on the consistent estimation of the population eigenvalues (also known as the spectrum). Lam [2016] introduces a Nonparametric Eigenvalue-Regularized Covariance Matrix Estimator (NERCOME) through subsampling of the data, which is asymptotically equivalent to the nonlinear shrinkage method of Ledoit and Wolf [2012]. Shrinkage can also be applied on the sample covariance matrix directly. Ledoit and Wolf [2003] propose a weighted average estimator of the covariance matrix with a single-index factor target to account for common market covariance and provide analytic calculation of the optimal shrinkage intensity. Schäfer and Strimmer [2005] review six different shrinkage targets and derive improved covariance estimator based on the optimal shrinkage intensity in Ledoit and Wolf [2003]. Besides, shrinkage techniques are also used for spectral analysis of multivariate time series of high dimensionality. Böhm and von Sachs [2008] shrink the empirical eigenvalues in the frequency domain towards one another to improve upon the smoothed periodogram as an estimator for the multivariate spectrum. Also, Böhm and von Sachs [2009] propose a nonparametric shrinkage estimator of the spectral matrix which has asymptotically minimal risk among all linear combinations of the identity and the averaged periodogram matrix. Naturally related to the shrinkage approach is Bayesian estimation of the covariance and precision matrices. Evans [1965], Chen [1979], and Dickey et al. [1985] use possibly the most natural priors distribution of the covariance

matrix of a multivariate normal distribution, the inverted Wishart distribution. Moreover, Leonard and John [2012] propose a flexible class of covariance matrix prior, which yields more general hierarchical and empirical Bayes smoothing and inference. Alvarez [2014] proposes some alternative distributions, including the scaled inverse Wishart distribution, which gives more flexibility on the variance priors, and separate priors for variances and correlations, which eliminates any prior relationship among covariance matrix elements.

1.2 Organization and Outline of the thesis

The thesis is structured as follows. In Chapter 2 we propose tilting-based precision matrix estimators of a p-dimensional random variable, X, when p is possibly much larger than the sample size n. Four types of tilting-based methods are introduced and the rate of convergence are addressed under certain assumptions. Asymptotic properties of the estimators are studied when p is fixed and p grows with n. For finite p and n, extensive comparisons of thresholding estimators and the proposed methods are demonstrated. Several improvement approaches are made. The simulation results are presented under different models.

Chapter 3 proposes NOVEL Integration of the Sample and Thresholded covariance estimators (NOVELIST), which is shrinkage of the sample covariance (correlation) towards its thresholded version. The linkage between NOVELIST and ridge regression are demonstrated. We obtain an explicit convergence rate in the operator norm over a large class of covariance (correlation) matrices when p and n satisfy $\log p/n \to 0$. Empirical choices of parameters and a data-driven algorithm for NOVELIST estimators which combines Ledoit and Wolf [2003]'s method and cross-validation (LW-CV

algorithm) is presented. Further empirical improvements of NOVELIST are proposed. Comprehensive simulation study is based on a wide range of models and results of comparisons with several popular estimators are presented. Finally, an automatic algorithm is constructed to provide an adaptive choice between the use of LW-CV algorithm and fixed parameters.

Chapter 4 is devoted to explore the applications of NOVELIST estimators and to exhibit the results of applying the estimators on real data, including portfolio optimization using low-frequency and high-frequency FTSE 100 constituents log returns, forecasting the number of calls for a call center and estimating false discovery proportion through a well-known breast cancer study. Chapter 5 concludes the thesis.

1.3 Conclusion

In conclusion, estimating the covariance matrix and its inverse for high-dimensional data has always been an important part of multivariate analysis. This chapter categorises the existing and most commonly used estimators proposed in recent years into two broad classes and provides a brief review of them. Several methods have offered inspirations to the methods introduced in Chapter 3 and 4. This chapter also gives an overview of the organization and outline of the thesis.

Chapter 2

Precision Matrix Estimation via tilting

2.1 Introduction

For multivariate normal distributions, the support of the estimate of the precision matrix is closely related to graphical models. For graphical models, each node corresponds to a random variable, and each non-zero edge between two nodes represents conditional dependence between the corresponding random variables after removing the effects of all the other variables. In this chapter, we consider a p-dimensional multivariate normal distributed random variable $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)$ with n i.i.d. observations, $\mathcal{P} = \{1, 2, \cdots, p\}$, $E\boldsymbol{X} = 0$, covariance matrix is $\Sigma = \{\sigma_{i,j}\} = E(\boldsymbol{X}^T\boldsymbol{X})$, and precision matrix is $\Sigma^{-1} = P = \{p_{i,j}\}$, $i, j \in \mathcal{P}$. For given i and j, the conditional dependence between two variables X_i and X_j given other variables is equivalent to the non-zero corresponding entry of the precision matrix, $p_{i,j}$ [Edward, 2000]. Hence, for Gaussian distributions, recovering the structure of the graphical models is equivalent to the identification and estimation of the non-zero entries in the precision matrix [Lauritzen, 1996]. Moreover, non-zero entries of the precision matrix imply non-zero par-

tial correlations between corresponding variable pairs conditional on the rest of the variables, as partial correlation between X_i and X_j is defined as $\breve{\rho}_{i,j} = \frac{-p_{i,j}}{\sqrt{p_{i,i}}\sqrt{p_{j,j}}}$ [Peng et al., 2009], which is very useful in estimation in Gaussian graphical models.

There exists a well-known link between partial correlations and regression models under Gaussianity, based on which a partial correlation estimation method is introduced by Peng et al. [2009]. Although it is not directly linked to this work, it gives us inspiration for exploring the relationship between precision matrix and regression models, based on which our work is carried on. For given i, by regression X_i on all the other variables in $\mathcal{P}$, we have

$$X_i = \sum_{j \in \mathcal{P} \setminus \{i\}} \beta_{i,j} X_j + \zeta_i, \tag{2.1}$$

where ζ_i are uncorrelated with each X_j, $j \in \mathcal{P} \setminus \{i\}$. From Lemma 1 in Peng et al. [2009], we have $\beta_{i,j} = \breve{\rho}_{i,j}\sqrt{\frac{p_{j,j}}{p_{i,i}}}$. Analogously, by regression X_j on all the other variables, we also have $\beta_{j,i} = \breve{\rho}_{j,i}\sqrt{\frac{p_{i,i}}{p_{j,j}}}$. Since $\breve{\rho}_{i,j} = \breve{\rho}_{j,i}$, we obtain $\breve{\rho}_{i,j} = \text{sign}(\beta_{i,j})\sqrt{\beta_{i,j}\beta_{j,i}}$. Therefore, the search for non-zero partial correlations, i.e. determining the non-zero edges in graphical models, can be viewed as a model selection problem under the Gaussian regression settings.

However, we aim to estimate the precision matrix, which is closely related to the partial correlations but cannot be explicitly expressed by them. Instead, we find another way to link a $p \times p$ precision matrix Σ^{-1} to regression models block by block as follows. For simplicity, we choose $p = 3$ for illustration.

Step one: for any given i and j, for example $i = 1$ and $j = 2$, we obtain the first four elements of Σ^{-1} (indicated as red dots), which is called 2×2 pairwise precision

matrix, and denoted by ${\Sigma^{\circ}_{i,j}}^{-1}$, where

$$\Sigma^{-1} = \begin{pmatrix} \bullet & \bullet & \cdot \\ \bullet & \bullet & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix} \begin{matrix} \leftarrow i = 1 \\ \\ \leftarrow j = 2, \end{matrix} \tag{2.2}$$

$${\Sigma^{\circ}_{1,2}}^{-1} \doteq \begin{pmatrix} \bullet & \bullet \\ & \\ \bullet & \bullet \end{pmatrix} = \text{cov}^{-1}(\boldsymbol{X}_1, \boldsymbol{X}_2 | \boldsymbol{X}_{-(1,2)}). \tag{2.3}$$

Here, $\text{cov}(\boldsymbol{X}_1, \boldsymbol{X}_2 | \boldsymbol{X}_{-(1,2)})$ is a partial covariance matrix, i.e. the covariance matrix of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ given all the other variables. Formula (2.3) indicates that the pairwise precision matrix ${\Sigma^{\circ}_{1,2}}^{-1}$ equals the inverse of the pairwise partial covariance matrix. Since partial covariance matrix can be estimated by using regression models, precision matrix estimation is linked to regression problems. More detailed explanation about this comes later.

Step two: for $i = 1$ and $j = 3$, we obtain another four elements of Σ^{-1} (indicated as green dots).

$$\Sigma^{-1} = \begin{pmatrix} \bullet & \bullet & \bullet \\ \bullet & \bullet & \cdot \\ \bullet & \cdot & \bullet \end{pmatrix} \begin{matrix} \leftarrow i = 1 \\ \\ \leftarrow j = 3, \end{matrix} \tag{2.4}$$

$${\Sigma^{\circ}_{1,3}}^{-1} \doteq \begin{pmatrix} \bullet & \bullet \\ & \\ \bullet & \bullet \end{pmatrix} = \text{cov}^{-1}(\boldsymbol{X}_1, \boldsymbol{X}_3 | \boldsymbol{X}_{-(1,3)}), \tag{2.5}$$

where, $\text{cov}(\boldsymbol{X}_1, \boldsymbol{X}_3 | \boldsymbol{X}_{-(1,3)})$ is the partial covariance matrix of $\boldsymbol{X}_1$ and $\boldsymbol{X}_3$ given all the other variables.

Step three: move i and j around across all the indices in $\mathcal{P}$, we are able to obtain

all the entries of Σ^{-1}. We note that each diagonal involves $(p-1)$ different 2 by 2 blocks as i and j move around across all indices, we use their average values finally.

Now, we focus on formula (2.3) and explain how the last equality is obtained. Actually it comes from the block-wise inversion of matrix [Bernstein, 2009, p.147] as follows,

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}^{-1} = \begin{pmatrix} (A - BC^{-1}B^T)^{-1} & -(A - BC^{-1}B^T)^{-1}BC^{-1} \\ -C^{-1}B^T(A - BC^{-1}B^T)^{-1} & C^{-1} + C^{-1}B^T(A - BC^{-1}B^T)^{-1}BC^{-1} \end{pmatrix}, \tag{2.6}$$

where A, B and C are matrix sub-blocks of arbitrary size, A and C must be square, C and $A - BC^{-1}B^T$ must be nonsingular. The proof of formula (2.6) is given in Section 2.12.1. We note that, $A - BC^{-1}B^T$ is actually in a form closely related to conditional covariance. For illustration, we give a simple example of a 3 by 3 sample precision matrix $\hat{\Sigma}^{-1}$ of a multivariate normal random variable $\boldsymbol{X} = (X_1, X_2, X_3)$ with n i.i.d. observations, $\hat{\Sigma}^{-1} = \boldsymbol{X}^T\boldsymbol{X}$. We partition $\hat{\Sigma}^{-1}$ and apply the top-left part (indicated in red) of the right-hand side of formula (2.6) on it, which leads to

$$\begin{aligned} \hat{\Sigma}^{-1} &= \left[\begin{array}{cc|c} \hat{\sigma}_{1,1} & \hat{\sigma}_{1,2} & \hat{\sigma}_{1,3} \\ \hat{\sigma}_{2,1} & \hat{\sigma}_{2,2} & \hat{\sigma}_{2,3} \\ \hline \hat{\sigma}_{3,1} & \hat{\sigma}_{3,2} & \hat{\sigma}_{3,3} \end{array}\right]^{-1} \\ &= \left[\begin{array}{cc|c} \begin{pmatrix} \hat{\sigma}_{1,1} - \hat{\sigma}_{1,3}\hat{\sigma}_{3,3}^{-1}\hat{\sigma}_{3,1} & \hat{\sigma}_{1,2} - \hat{\sigma}_{1,3}\hat{\sigma}_{3,3}^{-1}\hat{\sigma}_{3,2} \\ \hat{\sigma}_{2,1} - \hat{\sigma}_{2,3}\hat{\sigma}_{3,3}^{-1}\hat{\sigma}_{1,3} & \hat{\sigma}_{2,2} - \hat{\sigma}_{2,3}\hat{\sigma}_{3,3}^{-1}\hat{\sigma}_{3,2} \end{pmatrix}^{-1} & \begin{matrix} \cdot \\ \cdot \end{matrix} \\ \hline \begin{matrix} \cdot & \cdot \end{matrix} & \cdot \end{array}\right], \end{aligned} \tag{2.7}$$

where only the top-left part (indicated in red) are calculated for illustration. We observe

that this part corresponds to the inverse of the pairwise sample conditional covariance matrix of (X_1, X_2) given X_3. For example, we note that

$$\begin{aligned}
&\hat{\sigma}_{1,2} - \hat{\sigma}_{1,3}\hat{\sigma}_{3,3}^{-1}\hat{\sigma}_{3,2} \\
=&X_1^T X_2 - X_1^T X_3 (X_3^T X_3)^{-1} X_3^T X_2 \\
=&X_1^T (\boldsymbol{I}_n - \boldsymbol{H}_3) X_2 \\
=&\widehat{\text{cov}}((\boldsymbol{I}_n - \boldsymbol{H}_3)X_1, (\boldsymbol{I}_n - \boldsymbol{H}_3)X_2) \\
=&\widehat{\text{cov}}(X_1|X_3, X_2|X_3),
\end{aligned} \tag{2.8}$$

where $\boldsymbol{I}_n$ is a n by n diagonal matrix, $\boldsymbol{H}_3$ is the projection matrix onto the space spanned by X_3, $\boldsymbol{H}_3 \doteq X_3(X_3^T X_3)^{-1} X_3^T$ and $\widehat{\text{cov}}(X_1|X_3, X_2|X_3)$ is the sample conditional covariance between X_1 and X_2 given X_3, which can be obtained by computing the sample covariance between the residuals of regressing X_1 and X_2 on X_3. When $\hat{\Sigma}^{-1}$ is partitioned in different combinations of the indices, the results of the remaining part of formula (2.7) (indicated as dots) will be obtained. Actually, this relationship is also true at the population level, which means that any 2 by 2 block indexed by (i, j) of any precision matrix is equivalent to the inversion of the pairwise conditional covariance matrix of (X_i, X_j) given all the other variables, see Lemma 1 in Section 2.3.1. Here we find how precision matrix estimation links to regression models which helps us to estimate precision matrix. Now, we understand that precision matrix estimation can be achieved block by block through 2 simultaneous regression problems for each block under Gaussianity. More details and a generalization are also given in Section 2.3.1. However, in high-dimensional settings, difficulties arise in estimating the regression coefficients and residuals, even individually. Including all or "too many" remaining variables in the regression models would distort the estimation results due

to the large dimensionality and possibly strong collinearity among the remaining variables. Also, in high dimensional geometry, even when variables follow independent Gaussian distributions, spurious sample marginal correlations among variables would be observed [Fan and Lv, 2008], leading to wrong regression models. Over the last two decades, substantial efforts have been made in tackling this high-dimensional variable selection problem. An exhaustive review can be found in Fan and Lv [2010] under the assumption that regression coefficients are assumed to be sparse with many being zero. Among them, one of the intensively studied area is the penalised least squares estimation, such as the Lasso [Tibshirani, 1996], the ridge regression, the SCAD [Fan and Li, 2007] and their extensions [Meinshausen, 2007; Zou, 2006]. Fan and Lv [2008] introduce the Sure Independence Screening (SIS), which ranks the importance of each variable according to the magnitude of the corresponding marginal correlation between the variable and the response, and selects the first d_n variables which have the largest magnitude of correlations. SIS reduces the dimensionality from high or ultra high (for example, $\log p = O(n^a)$ for some $a > 0$) to the scale d_n, which can be less than n, in a computationally efficient way.

Despite of good theoretical properties and empirical performances achieved by these methods, Cho and Fryzlewicz [2012] argue that the results relying on heavy usage of marginal correlation for measuring the contribution of each variable to the response can be misleading with growing dimensionality p. Many iterative algorithms for measures other than marginal correlation are proposed in variable selection problems for high-dimensional regression models. Traditional forward selection [Weisberg, 2005] and forward regression [Wang, 2009] consider the relationship between a new variable and the response after removing the effects of the existing variables in the model at each iteration. Bühlmann et al. [2009] introduced a PC-simple algo-

rithm, where partial correlation instead of marginal correlation is applied in order to iteratively remove irrelevant variables from the model. Cho and Fryzlewicz [2012] introduced tilted correlation to measure the strength association between the variables and the response which takes into account collinearity. The tilted correlation is closely related to partial correlation, but it focuses on regressing the response Y on the variables X_k, and thus Y and X_k are not treated on an equal footing. For any given variable X_k, $k \in \mathcal{P}$, the tilted correlation is designed to capture the linear relationship between X_k and the response Y, after removing the effects of all the highly related remaining variables (not all the other variables), on X_k only instead of on both X_k and Y. A more detailed explanation of the tilted correlation can be found in Section 2.2.

Motivated by the link between precision matrix and regression models, this chapter proposes tilting techniques which are applied to simultaneously select the (hopefully) highly relevant remaining variables for each pair X_i and X_j when p grows with n, which leads to block by block large precision matrix estimation . To tackle the simultaneous variable selection problems for high-dimensional regression models, we introduce four types of tilting methods. The first three methods rely on ranking of the marginal correlations, while the last one apply tilted correlations in order to remove or reduce the effects of collinearity. We investigate the asymptotic properties of the tilting estimators under suitable assumptions as well as small sample inference. Furthermore, empirical choices of parameters and improvements are discussed and algorithms are listed for the estimators. Also, simulation studies are presented afterwards.

The rest of the chapter is organised as follows. Section 2.2 gives preliminary knowledge regarding the tiltied correlation introduced by Cho and Fryzlewicz [2012]. In Section 2.3, we introduce the notations, describe and generalize the building block of the tilting estimators, which comes from the block-wise inversion of covariance

matrix, and illustrate the motivation by a simulation example. In Section 2.4, tilting methodology is formally defined and four types of tilting methods are introduced. Section 2.5 lists the algorithms for the tilting estimators. Section 2.6 establishes the consistency of the tilting estimators under assumptions for fixed p and when p grows with n. Section 2.7 analytically investigates the finite sample performance of tilting estimators and the differences and links between the tilting estimators and soft and hard thresholding estimators. Section 2.8 gives suggestions on choices of parameters. Section 2.9 exploits optional empirical improvements of the tilting estimators. Section 2.10 exhibits practical performances of the tilting estimators in comparison to the thresholding estimators. Section 2.11 concludes the chapter. Section 2.12 is additional lemmas and proofs.

2.2 Preliminary: *tiltied correlation*

Before introducing the proposed methods for precision matrix estimation, we need to briefly describe what is the so called "tilted" correlation introduced by Cho and Fryzlewicz [2012] and how it works. It considers the following linear model:

$$Y = X\beta + \epsilon, \tag{2.9}$$

where $Y = (Y_1, \cdots, Y_n)^T \in R^n$ is an n-vector of the response, $X = (X_1, \cdots, X_p)$ is an $n \times p$ design matrix and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T \in R^n$ is an n-vector of i.i.d. random errors. The aim of the regression problem is to identify $\mathcal{S} = \{1 \leqslant k \leqslant p : \beta_k \neq 0\}$ under the assumption that only a small number of variables actually contribute to the response, i.e., $\mathcal{S}$ is of cardinality $|\mathcal{S}| \ll p$.

The marginal correlation between each variable X_k and Y can be written as the following decomposition,

$$X_k^T Y = X_k^T (\sum_{s=1}^{p} \beta_s X_s + \epsilon) = \beta_k + \underline{\sum_{s \in \mathcal{S} \setminus \{k\}} \beta_s X_k^T X_s} + X_k^T \epsilon, \tag{2.10}$$

which shows that marginal correlation screening is not reliable on selecting $\mathcal{S}$ if the underlined summand in formula (2.10) is non-negligible. For example, irrelevant variables that are highly related with the relevant ones can be selected by using marginal correlation screening. Also, if high collinearity exists among the variables, the results coming from marginal correlation screening could be far away from the true set $\mathcal{S}$. It can even be the case that the relevant variables are ruled out when marginal correlation screening is applied. Consider the following example,

$$Y = \beta X_1 + \beta X_2 - 2\beta\sqrt{\varphi} X_3 + \epsilon, \tag{2.11}$$

where $\epsilon \sim \mathcal{N}(0, \boldsymbol{I}_n)$ and $(X_1, X_2, X_3)^T$ are generated from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$ independently for $i = 1, 2, 3$. The population covariance matrix $\Sigma = \{\sigma_{i,j}\}$ satisfies $\sigma_{i,i} = 1$ and $\sigma_{i,j} = \varphi$, $i \neq j$, except $\sigma_{i,3} = \sqrt{\varphi}$. It is clear that $\text{corr}(X_3, Y) = 0$, which indicates that X_3 is marginally uncorrelated with Y at the population level, and is likely to be ruled out if marginal correlation screening is applied, but X_3 is actually a relevant variable with Y.

In order to find an alternative measurement instead of marginal correlation that can be represented as β_k (plus an negligible term), Cho and Fryzlewicz [2012] introduce

the tilted variable X_k^* for each X_k, which is defined as

$$X_k^* \doteq (\boldsymbol{I}_n - \boldsymbol{H}_k)X_k, \tag{2.12}$$

where $\boldsymbol{H}_k$ is the projection matrix onto the space spanned by $\boldsymbol{X}_{\widetilde{k}}$, i.e. $\boldsymbol{H}_k \doteq \boldsymbol{X}_{\widetilde{k}}(\boldsymbol{X}_{\widetilde{k}}^T \boldsymbol{X}_{\widetilde{k}})^{-1}\boldsymbol{X}_{\widetilde{k}}^T$, and $\boldsymbol{X}_{\widetilde{k}}$ is a submatrix of $\boldsymbol{X}_{\mathcal{P}\setminus\{k\}}$, which contains all the remaining variables that are highly correlated with X_k. It is clear that the tilted variable is a projected version of the original one, which removes the effects of all the highly correlated variables.

Then the tilted correlation is introduced based on the tilted variable. We can decompose $(X_k^*)^T Y$ as

$$\begin{aligned}(X_k^*)^T Y &= X_k^T(\boldsymbol{I}_n - \boldsymbol{H}_k)Y = X_k^T\{\sum_{s=1}^{p}\beta_s(\boldsymbol{I}_n - \boldsymbol{H}_k)X_s + (\boldsymbol{I}_n - \boldsymbol{H}_k)\epsilon\} \\ &= \beta_k X_k^T(\boldsymbol{I}_n - \boldsymbol{H}_k)X_k + \underline{\sum_{s\in\mathcal{S}\setminus\{\widetilde{k}\},s\neq k}\beta_s X_k^T(\boldsymbol{I}_n - \boldsymbol{H}_k)X_s + X_k^T(\boldsymbol{I}_n - \boldsymbol{H}_k)\epsilon}\end{aligned} \tag{2.13}$$

If we rescale $(X_k^*)^T Y$ by dividing $X_k^T(\boldsymbol{I}_n - \boldsymbol{H}_k)X_k$ (rescaling 1 in Cho and Fryzlewicz [2012]), and as long as the second and the third summands in formula (2.13) are negligible in comparison with the first, the rescaled tilted correlation can be represented as β_k plus a small term. We denote $a_k \doteq \|\boldsymbol{H}_k X_k\|_2^2 / \|X_k\|_2^2$, then we have $1 - a_k = X_k^T(\boldsymbol{I}_n - \boldsymbol{H}_k)X_k$ as the rescaling factor of making the norm of the tilted correlation to be 1. From now on, we refer to "tilted correlation" as the rescaled tilted correlation, and denote it by $\widehat{\text{corr}}^*$.

In Cho and Fryzlewicz [2012]'s paper, conditions are used in order to ensure that

the underlined term in formula (2.13) is negligible. For example, condition 1 in Cho and Fryzlewicz [2012] means that if X_s is not highly relevant to X_k itself, it remains not highly relevant to the projected X_k onto the space spanned by $\boldsymbol{X}_{\widetilde{k}}$, i.e. $\boldsymbol{H}_k X_k$, which can be shown to hold asymptotically when each column X_k is generated independently as a random vector on a sphere of radius 1, which is the surface of the Euclidean ball $B_2^n = \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i^2 \leqslant 1\}$ by using Lemma 4 in Section 2.12.2.

To sum up, the tilted correlation measures the rescaled correlation between the response Y and the tilted version of the variable X_i that removes the effects of all the highly relevant remaining variables on X_i. More explanations regarding how it can be applied to precision matrix estimation come later in Section 2.4.2.4 and the algorithm can be found in Section 2.5.2.

2.3 Notations, building block and motivations

2.3.1 Notations and building block $\hat{\Sigma}^{\circ}_{2\times 2}{}^{-1}$

For a given pair of i and j, $i, j \in \mathcal{P}$, we denote $\mathcal{K} = \mathcal{P} \setminus \{i, j\}$. If we partition $\boldsymbol{X}$ as $(\boldsymbol{X}_{ij}, \boldsymbol{X}_{-(ij)})$, where $\boldsymbol{X}_{ij} = (X_i, X_j)$, $\boldsymbol{X}_{-(ij)} = (X_k : k \in \mathcal{K})$, the covariance matrix Σ is decomposed as follows,

$$\Sigma = \begin{pmatrix} \Sigma_{2\times 2} & \Sigma_{2\times(p-2)} \\ \Sigma_{(p-2)\times 2} & \Sigma_{(p-2)\times(p-2)} \end{pmatrix}_{p\times p}, \tag{2.14}$$

where $\Sigma_{2\times 2} = E(\boldsymbol{X}_{ij}^T \boldsymbol{X}_{ij})$, $\Sigma_{2\times(p-2)} = E(\boldsymbol{X}_{ij}^T \boldsymbol{X}_{-(ij)})$, $\Sigma_{(p-2)\times 2} = E(\boldsymbol{X}_{-(ij)}^T \boldsymbol{X}_{ij})$, $\Sigma_{(p-2)\times(p-2)} = E(\boldsymbol{X}_{-(ij)}^T \boldsymbol{X}_{-(ij)})$. Analogously, the precision matrix P can be parti-

tioned as

$$P = \begin{pmatrix} \Sigma_{2\times 2} & \Sigma_{2\times(p-2)} \\ \Sigma_{(p-2)\times 2} & \Sigma_{(p-2)\times(p-2)} \end{pmatrix}^{-1}_{p\times p} = \begin{pmatrix} P_{2\times 2} & P_{2\times(p-2)} \\ P_{(p-2)\times 2} & P_{(p-2)\times(p-2)} \end{pmatrix}_{p\times p} . \tag{2.15}$$

Lemma 1 *If* $\boldsymbol{X} \sim \mathcal{N}(0, \Sigma)$, Σ *and* P *are partitioned as in formula (2.14) and (2.15), and* $\Sigma_{(p-2)\times(p-2)}$ *and* $\Sigma_{2\times 2} - \Sigma_{2\times(p-2)}\Sigma^{-1}_{(p-2)\times(p-2)}\Sigma_{(p-2)\times 2}$ *are nonsingular, then*

$$P_{2\times 2} = (\Sigma_{2\times 2} - \Sigma_{2\times(p-2)}\Sigma^{-1}_{(p-2)\times(p-2)}\Sigma_{(p-2)\times 2})^{-1} = cov^{-1}(\boldsymbol{X}_{ij}|\boldsymbol{X}_{-(ij)}), \tag{2.16}$$

where $\boldsymbol{X}_{ij} = (X_i, X_j)$.

The first equality follows because of the block-wise inversion of matrix (formula (2.6) in Section 2.1). The second equality follows due to properties of marginal and conditional normal distribution [Tong, 2012, p.35].

Lemma 1 shows that $P_{2\times 2}$ is not the inversion of $\Sigma_{2\times 2}$, instead, it is the inversion of the 2 by 2 pairwise conditional covariance matrix which we define as

$$\Sigma^{\circ}_{2\times 2} \doteq \text{cov}(\boldsymbol{X}_{ij}|\boldsymbol{X}_{-(ij)}), \tag{2.17}$$

where $\boldsymbol{X}_{ij} = (X_i, X_j)$. i.e. $\Sigma^{\circ}_{2\times 2}$ is the covariance matrix of X_i and X_j controlling all the other variables, $\boldsymbol{X}_{-(ij)}$, for estimating which, the natural way in practice is to regress X_i and X_j on all the other variables.

By regressing X_i and X_j on all the other variables, $\boldsymbol{X}_{-(ij)}$, respectively, we obtain 2 simultaneous regression models

$$X_i = \sum_{k\in\mathcal{K}} \beta_{i,k} X_k + \epsilon_i, \tag{2.18}$$

$$X_j = \sum_{k \in \mathcal{K}} \beta_{j,k} X_k + \epsilon_j, \tag{2.19}$$

where ϵ_i and ϵ_j are specific terms of X_i and X_j respectively, $E(\epsilon_i) = E(\epsilon_j) = 0$, X_k is uncorrelated with ϵ_i and ϵ_j. We denote $\boldsymbol{\epsilon}_{ij} = (\epsilon_i, \epsilon_j)$, hence we have $\text{cov}(\boldsymbol{\epsilon}_{ij}) = \Sigma^{\circ}_{2\times 2}$. In order to estimate $P_{2\times 2}$, we need to replace $\boldsymbol{\epsilon}_{ij}$ by $\hat{\boldsymbol{\epsilon}}_{ij}$. Typically, $\hat{\boldsymbol{\epsilon}}_{ij}$ can be obtained by Least Squares Estimation,

$$\hat{\epsilon}_i = (\boldsymbol{I}_n - \boldsymbol{H}_{-(ij)}) X_i, \tag{2.20}$$

$$\hat{\epsilon}_j = (\boldsymbol{I}_n - \boldsymbol{H}_{-(ij)}) X_j, \tag{2.21}$$

i.e. $\hat{\boldsymbol{\epsilon}}_{ij} = (\boldsymbol{I}_n - \boldsymbol{H}_{-(ij)})\boldsymbol{X}_{ij}$. Then we obtain $\hat{\Sigma}^{\circ}_{2\times 2} = \widehat{\text{cov}}(\hat{\boldsymbol{\epsilon}}_{ij})$, and $\hat{P}_{2\times 2} = {\hat{\Sigma}^{\circ}_{2\times 2}}^{-1}$. Furthermore, as i and j move around across all indices in $\mathcal{P}$, each pair of i and j yields its ${\hat{\Sigma}^{\circ}_{2\times 2}}^{-1}$, which fills in the corresponding elements of the precision matrix estimator, and eventually the estimation of the entire precision matrix P is obtained. It is clear that ${\hat{\Sigma}^{\circ}_{2\times 2}}^{-1}$ is the building block of the precision matrix estimator.

However, we note that the building block does not have to be a 2 by 2 matrix. If we denote $\mathcal{S}$ as a subset of $\mathcal{P}$, $\mathcal{K} = \mathcal{P} \setminus \mathcal{S}$, $|\mathcal{S}| = m$ and $|\mathcal{K}| = p - m$. In general, for any m satisfying $2 \leqslant m < \min(p, n)$, the link between precision matrix and regression models still exist. Now, we describe the links between precision matrix and regression models in the general notations. We partition $\boldsymbol{X}$ as $(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{-\mathcal{S}})$, where $\boldsymbol{X}_{\mathcal{S}} = (X_s : s \in \mathcal{S})$, $\boldsymbol{X}_{-\mathcal{S}} = \boldsymbol{X}_{\mathcal{K}} = (X_k : k \in \mathcal{K})$. The partitioned covariance matrix is

$$\Sigma = \begin{pmatrix} \Sigma_{m\times m} & \Sigma_{m\times(p-m)} \\ \Sigma_{(p-m)\times m} & \Sigma_{(p-m)\times(p-m)} \end{pmatrix}_{p\times p}, \tag{2.22}$$

and the partitioned precision matrix is

$$P = \begin{pmatrix} P_{m\times m} & P_{m\times(p-m)} \\ P_{(p-m)\times m} & P_{(p-m)\times(p-m)} \end{pmatrix}_{p\times p}, \tag{2.23}$$

By respectively regressing $\boldsymbol{X}_{\mathcal{S}}$ on all the other variables $\boldsymbol{X}_{-\mathcal{S}}$, we obtain m simultaneous regression models. We denote the specific terms $\boldsymbol{\epsilon}_{\mathcal{S}}$ as $\boldsymbol{\epsilon}_{\mathcal{S}} = (\epsilon_s : s \in \mathcal{S})$. Since $m < n$, the projection matrix $\boldsymbol{H}_{-\mathcal{S}} = \boldsymbol{X}_{-\mathcal{S}}(\boldsymbol{X}_{-\mathcal{S}}^T\boldsymbol{X}_{-\mathcal{S}})^{-1}\boldsymbol{X}_{-\mathcal{S}}^T$ is well defined. Hence, we obtain $\hat{\boldsymbol{\epsilon}}_{\mathcal{S}} = (\boldsymbol{I}_n - \boldsymbol{H}_{-\mathcal{S}})\boldsymbol{X}_{\mathcal{S}}$ and

$$\hat{\Sigma}^{\circ}_{m\times m} = \widehat{\mathrm{cov}}(\hat{\boldsymbol{\epsilon}}_{\mathcal{S}}), \tag{2.24}$$

$$\hat{P}_{m\times m} = {\hat{\Sigma}^{\circ}_{m\times m}}^{-1}. \tag{2.25}$$

Theoretically speaking, m can be any integer as long as $2 \leqslant m < p$ with large n. We observe that the size of m has little effect on the precision matrix estimation when n is large enough according to prior experimental numerical results. However, when n is close to p or even smaller than p, there is a trade-off between the size of $\mathcal{S}$ and $\mathcal{K}$ because $|\mathcal{S}|+|\mathcal{K}| = p$. On the one hand, as m increases, the number of regression models needed to be simultaneously solved is getting larger, which leads to computational complexity largely increases and then decreases; on the other hand, the number of all the remaining variables for each regression model, i.e. candidate regressors, is as large as $p - m$, and small m means large $p - m$ which possibly results in high-dimensional regression problems. We choose $m = 2$ for the estimate. See Section 2.8 for other choices of m and numerical results.

2.3.2 Motivation and example illustrations

Choosing $|\mathcal{S}| = 2$ means that $|\mathcal{K}| = p-2$. When $p \gg n$, $\mathcal{K}$ involves too many variables such that the projection matrix $\boldsymbol{H}_{-\mathcal{S}}$ is not well-defined and regression coefficients cannot be solved. Even if $n > p$ but n is close to p, putting all the other variables in the regression models will also distort the estimators. To tackle this problem, it appears natural to replace $\boldsymbol{X}_{\mathcal{K}}$ by a controlling subset $\boldsymbol{X}_{\mathcal{C}}$, where $|\mathcal{C}|$ is not bigger, in most cases much smaller than $|\mathcal{K}|$, and $\boldsymbol{X}_{\mathcal{C}}$ hopefully only contains the highly relevant controlling variables. Figure 2.1 shows that the optimal size of the controlling subset

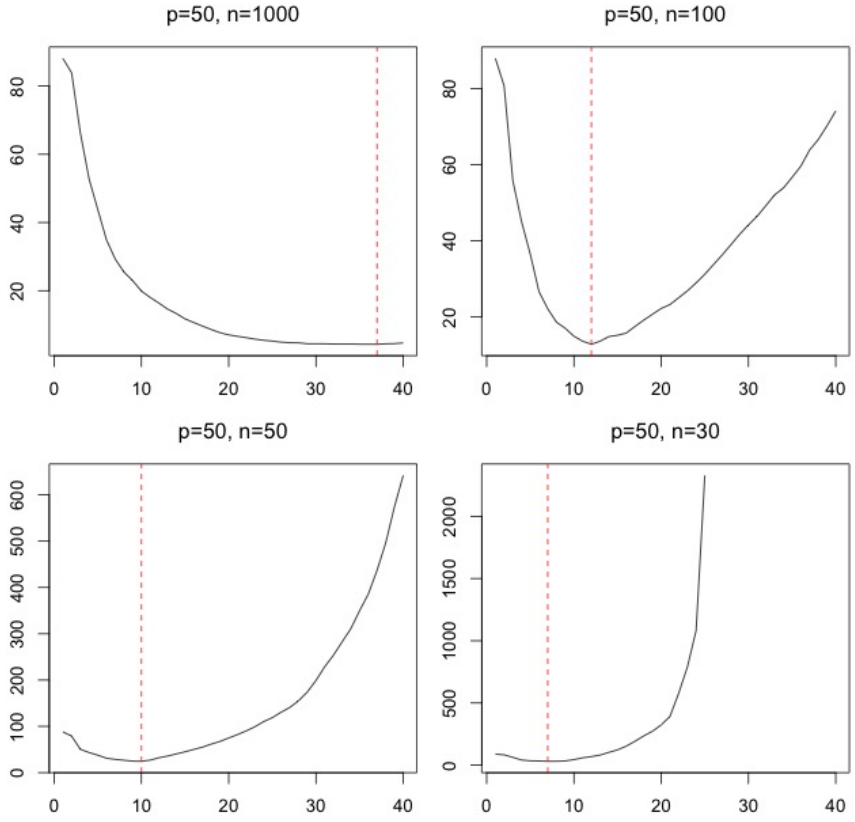


Figure 2.1: Frobenius norm errors of $\hat{\Sigma}^{\circ}_{2\times 2}{}^{-1}$ to $P_{2\times 2}$ with different size of the controlling subsets under model D in Section 2.10.1 , X-axis is the size of controlling subsets $|\mathcal{C}|$, Y-axis is the average Frobenius norm error of $\hat{\Sigma}^{\circ}_{2\times 2}{}^{-1}$ to $P_{2\times 2}$. The red dashed lines are located at the optimal size of $|\mathcal{C}|$. $|\mathcal{S}| = 2$, $|\mathcal{K}| = p-2$. Simulation times=50.

is much smaller than $|\mathcal{K}|$ in most of the cases unless $n \gg p$.

Since, there are 2 regression models to be solved at the same time, the determinations of 2 controlling subsets for X_i and X_j are made simultaneously, which makes the problem more difficult. In the following section, we will introduce the tilting methods of regularizing the controlling subsets for X_i and X_j which take into account the effects of the relationship between X_i and X_j on them.

2.4 Definitions and methods

2.4.1 Definitions

In this section, we formally introduce and define the precision matrix estimation via tilting. It is defined as

$$\hat{T} = \{\hat{t}_{i,j}\}, \quad i, j \in \mathcal{P}, \tag{2.26}$$

where

$$\hat{t}_{i,j} = \begin{cases} [\hat{\Sigma}_{ij}^{\circ\,-1}]_{1,2} & \text{if } i \neq j \\ \frac{1}{p-1} \sum_{l \in \mathcal{P} \setminus \{i\}} [\hat{\Sigma}_{il}^{\circ\,-1}]_{1,1} & \text{if } i = j \end{cases}, \tag{2.27}$$

$\hat{\Sigma}_{ij}^{\circ}$ is an alternative notation to $\hat{\Sigma}_{2\times 2}^{\circ}$ with emphasis on the indices, $\hat{\Sigma}_{ij}^{\circ} = \widehat{\text{cov}}(\boldsymbol{X}_{ij}^{*})$, where $\boldsymbol{X}_{ij}^{*} = (X_i|\boldsymbol{X}_{\mathcal{C}_i}, X_j|\boldsymbol{X}_{\mathcal{C}_j})$, $\boldsymbol{X}_{\mathcal{C}_i}$ and $\boldsymbol{X}_{\mathcal{C}_j}$ are the controlling subsets for X_i and X_j respectively, which can be equal to each other, $[M]_{a,b}$ is a scalar, which is the element indexed by (a, b) in the matrix M. Since each off-diagonal is calculated once, we obtain the estimate of the entry straight away, while each diagonal is involved in

$(p-1)$ different 2 by 2 blocks as i and j move around across all indices in $\mathcal{P}$, we use the average value as the final estimate of each diagonal.

2.4.2 Four types of tilting methods

One key ingredient of this methodology is simultaneous choice of the sets $\mathcal{C}_i$ and $\mathcal{C}_j$ for X_i and X_j, which is the essential for the building block $\hat{\Sigma}^{\circ}{}_{2\times 2}$, especially in high-dimensional cases. Now, four types of tilting methods are introduced for determining the sets $\mathcal{C}_i$ and $\mathcal{C}_j$, which can be identical to each other.

For each pair of specified indices i and j, we can decompose formula (2.18) and (2.19) as follows,

$$X_i = \sum_{b\in\mathcal{B}} \beta_{i,b}X_b + \sum_{e_i\in\mathcal{E}_i} \beta_{i,e_i}X_{e_i} + \underline{\sum_{e_j\in\mathcal{E}_j} \beta_{i,e_j}X_{e_j}} + \underline{\underline{\sum_{u\in\mathcal{U}} \beta_{i,u}X_u}} + \epsilon_i, \tag{2.28}$$

$$X_j = \sum_{b\in\mathcal{B}} \beta_{j,b}X_b + \underline{\sum_{e_i\in\mathcal{E}_i} \beta_{j,e_i}X_{e_i}} + \sum_{e_j\in\mathcal{E}_j} \beta_{j,e_j}X_{e_j} + \underline{\underline{\sum_{u\in\mathcal{U}} \beta_{j,u}X_u}} + \epsilon_j, \tag{2.29}$$

where

$\mathcal{B} = \{b : \beta_{i,b} \neq 0, \text{and } \beta_{j,b} \neq 0\}$, i.e. each X_b is a predictor for both X_i and X_j;

$\mathcal{E}_i = \{e_i : \beta_{i,e_i} \neq 0, \text{and } \beta_{j,e_i} = 0\}$, i.e. each X_{e_i} is a predictor for X_i, but not X_j;

$\mathcal{E}_j = \{e_j : \beta_{i,e_j} = 0, \text{and } \beta_{j,e_j} \neq 0\}$, i.e. each X_{e_j} is a predictor for X_j, but not X_i;

$\mathcal{U} = \{u : \beta_{i,u} = 0, \text{and } \beta_{j,u} = 0\}$, i.e. none of X_u is a predictor for either X_i or X_j;

ϵ_i, and ϵ_j are uncorrelated with each X_b, X_{e_i}, X_{e_j} and X_u.

Zero summands for only X_i or X_j are underlined, and those for both are double-underlined. Based on the decomposition, four different types of tilting methods are defined below. The first three methods rely on ranking of the marginal correlations and

computationally fast. We note that the marginal correlation between variable X_k and X_i, for example if $k \in \mathcal{B}$, has following decomposition,

$$\begin{aligned} X_k^T X_i =& X_k^T (\sum_{b\in\mathcal{B}} \beta_{i,b} X_b + \sum_{e_i\in\mathcal{E}_i} \beta_{i,e_i} X_{e_i} + \underline{\sum_{e_j\in\mathcal{E}_j} \beta_{i,e_j} X_{e_j}} + \underline{\underline{\sum_{u\in\mathcal{U}} \beta_{i,u} X_u}} + \epsilon_i) \\ =& \beta_{k,i} + (\sum_{b\in\mathcal{B}\setminus\{k\}} \beta_{i,b} X_k^T X_b + \sum_{e_i\in\mathcal{E}_i} \beta_{i,e_i} X_k^T X_{e_i}) + X_k^T \epsilon_i, \end{aligned} \tag{2.30}$$

which shows that marginal correlation between two variables is the corresponding regression coefficient plus bias terms (in bracket). But we will show in Section 2.6.2 that under certain assumptions, the bias terms would not contaminate consistency of the tilting estimators at element-wise level. The last tilting method applies tilted correlations [Cho and Fryzlewicz, 2012] instead of marginal correlations in order to make such bias terms zero or negligible.

2.4.2.1 Simple tilting

Simple tilting puts the variables which are highly correlated with either X_i or X_j into the controlling subset $\boldsymbol{X}_{\mathcal{C}_i^s}$ and $\boldsymbol{X}_{\mathcal{C}_j^s}$, where $\mathcal{C}_i^s$ and $\mathcal{C}_j^s$ are defined as

$$\mathcal{C}_i^s = \mathcal{C}_j^s = \mathcal{C}_{ij}^s = \{c : |\widehat{\text{corr}}(X_c, X_i)| > \pi_1 \text{ or } |\widehat{\text{corr}}(X_c, X_j)| > \pi_1, c \in \mathcal{K}\}, \tag{2.31}$$

where π_1 is a threshold, $\pi_1 \in (0, 1)$. Actually, $\mathcal{C}_{ij}^s$ intends to capture $\mathcal{B} \cup \mathcal{E}_i \cup \mathcal{E}_j$. Subject to $|\mathcal{B}| + |\mathcal{E}_i| + |\mathcal{E}_j| < n$, after controlling $\boldsymbol{X}_{\mathcal{B}\cup\mathcal{E}_i\cup\mathcal{E}_j}$ for both X_i and X_j, the remaining parts are

$$R_i = \underline{\underline{\sum_{u\in\mathcal{U}} \beta_{i,u} X_u}} + \epsilon_i, \tag{2.32}$$

$$R_j = \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{j,u} X_u}} + \epsilon_j. \tag{2.33}$$

And their covariance can be written as

$$\text{cov}(R_i, R_j) = \text{cov}(\underline{\underline{\sum_{u \in \mathcal{U}} \beta_{i,n} X_n}}, \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{j,u} X_u}}) + \text{cov}(\epsilon_i, \epsilon_j). \tag{2.34}$$

2.4.2.2 Double tilting

Double tilting only controls the variables which are highly correlated with both X_i and X_j. $\mathcal{C}_i^d$ and $\mathcal{C}_j^d$ are defined as

$$\mathcal{C}_i^d = \mathcal{C}_j^d = \mathcal{C}_{ij}^d = \{c : |\widehat{\text{corr}}(X_c, X_i)| > \pi_1 \text{ and } |\widehat{\text{corr}}(X_c, X_j)| > \pi_1, c \in \mathcal{K}\}. \tag{2.35}$$

It is clear that double tilting intends to control the variables in $\boldsymbol{X}_{\mathcal{B}}$. Subject to $|\mathcal{B}| < n$, only controlling $\boldsymbol{X}_{\mathcal{B}}$ for both X_i and X_j, the remaining terms are

$$R_i = \sum_{e_i \in \mathcal{E}_i} \beta_{i,e_i} X_{e_i} + \underline{\sum_{e_j \in \mathcal{E}_j} \beta_{i,e_j} X_{e_j}} + \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{i,u} X_u}} + \epsilon_i, \tag{2.36}$$

$$R_j = \underline{\sum_{e_i \in \mathcal{E}_i} \beta_{j,e_i} X_{e_i}} + \sum_{e_j \in \mathcal{E}_j} \beta_{j,e_j} X_{e_j} + \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{j,u} X_u}} + \epsilon_j. \tag{2.37}$$

Then the corresponding covariance is

$$\begin{aligned} \text{cov}(R_i, R_j) =& \text{cov}(\sum_{e_i \in \mathcal{E}_i} \beta_{i,e_i} X_{e_i} + \underline{\sum_{e_j \in \mathcal{E}_j} \beta_{i,e_j} X_{e_j}} + \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{i,u} X_u}}, \underline{\sum_{e_i \in \mathcal{E}_i} \beta_{j,e_i} X_{e_i}} \\ &+ \sum_{e_j \in \mathcal{E}_j} \beta_{j,e_j} X_{e_j} + \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{j,u} X_u}}) + \text{cov}(\epsilon_i, \epsilon_j). \end{aligned} \tag{2.38}$$

2.4.2.3 Separate tilting

Separate tilting applies different controlling subsets on X_i and X_j, i.e. $\mathcal{C}_i^{se} \neq \mathcal{C}_j^{se}$. We define $\mathcal{C}_i^{se}$ and $\mathcal{C}_j^{se}$ as follows,

$$\mathcal{C}_i^{se} = \{c_i : |\widehat{\text{corr}}(X_{c_i}, X_i)| > \pi_1, c_i \in \mathcal{K}\}, \tag{2.39}$$

$$\mathcal{C}_j^{se} = \{c_j : |\widehat{\text{corr}}(X_{c_j}, X_j)| > \pi_1, c_j \in \mathcal{K}\}. \tag{2.40}$$

We view $\mathcal{B} \cup \mathcal{E}_i$ and $\mathcal{B} \cup \mathcal{E}_j$ as the population-level counterparts of $\mathcal{C}_i^{se}$ and $\mathcal{C}_j^{se}$ respectively. If we assume that $|\mathcal{B}| + |\mathcal{E}_i| < n$ and $|\mathcal{B}| + |\mathcal{E}_j| < n$, the remaining summands after controlling $\boldsymbol{X}_{\mathcal{B}\cup\mathcal{E}_i}$ for X_i and $\boldsymbol{X}_{\mathcal{B}\cup\mathcal{E}_j}$ for X_j respectively can be written as

$$R_i = \underline{\sum_{e_j \in \mathcal{E}_j} \beta_{i,e_j} X_{e_j}} + \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{i,u} X_u}} + \epsilon_i, \tag{2.41}$$

$$R_j = \underline{\sum_{e_i \in \mathcal{E}_i} \beta_{j,e_i} X_{e_i}} + \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{j,u} X_u}} + \epsilon_j, \tag{2.42}$$

followed by expressing the covariance as

$$\begin{aligned} \text{cov}(R_i, R_j) =& \text{cov}(\underline{\sum_{e_j \in \mathcal{E}_j} \beta_{i,e_j} X_{e_j}} + \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{i,u} X_u}}, \underline{\sum_{e_i \in \mathcal{E}_i} \beta_{j,e_i} X_{e_i}} + \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{j,u} X_u}} \\ &+ \text{cov}(\epsilon_i, \epsilon_j) \end{aligned} \tag{2.43}$$

2.4.2.4 Competing tilting

The last tilting method is an application and extension of the tilted correlation introduced by Cho and Fryzlewicz [2012], as mentioned in Section 2.2. Instead of using sample marginal correlations, competing tilting apply tilted correlations on regulariza-

tion of $\mathcal{C}_i^c$ and $\mathcal{C}_j^c$. We name it as "competing tilting" because at each iteration step, it determines a subset which includes correlated variables and lets them compete to each other according to the conditional correlations between each variable and the response given all the other variables within the subset. We recall the tilted correlation $\widehat{\text{corr}}^*$ in Section 2.2, and define the controlling subsets for competing tilting as follows,

$$\mathcal{C}_i^c = \{c_i : |\widehat{\text{corr}}^*(X_{c_i}, X_i)| > \pi_1, c_i \in \mathcal{P}\}, \tag{2.44}$$

$$\mathcal{C}_j^c = \{c_j : |\widehat{\text{corr}}^*(X_{c_j}, X_j)| > \pi_1, c_j \in \mathcal{P}\}. \tag{2.45}$$

Competing tilting is highly related to separate tilting, as both aim to capture the sets $\mathcal{B} \cup \mathcal{E}_i$ and $\mathcal{B} \cup \mathcal{E}_j$.

For any remaining variable X_k, $k \in \mathcal{K}$, we denotes $\boldsymbol{X}_{\widetilde{k}}$ as a submatrix of $\boldsymbol{X}_{\mathcal{K}\setminus\{k\}}$, which contains $X_{\widetilde{k}}$, $\widetilde{k} \in \mathcal{C}_k$ as its columns, and each of them is highly correlated with X_k, i.e. $\mathcal{C}_k = \{\widetilde{k} : \widehat{\text{corr}}(X_{\widetilde{k}}, X_k) > \pi_n\}$. For considering the linear relationship between X_i and X_k after removing the effects of $\boldsymbol{X}_{\widetilde{k}}$, the tilted correlation between X_i and X_k after appropriate rescaling method (rescaling 1 in Cho and Fryzlewicz [2012]) is defined as

$$\widehat{\text{corr}}^*(X_k, X_i) = (1 - a_k)^{-1} X_k^T (\boldsymbol{I}_n - \boldsymbol{H}_k) X_i, \tag{2.46}$$

where $1 - a_k$ is the rescaling factor of making the norm of the tilted correlation to be 1, $a_k \doteq \|\boldsymbol{H}_k X_k\|_2^2 / \|X_k\|_2^2$, $\boldsymbol{H}_k$ is the projection matrix onto the space spanned by $\boldsymbol{X}_{\widetilde{k}}$, $\boldsymbol{H}_k \doteq \boldsymbol{X}_{\widetilde{k}} (\boldsymbol{X}_{\widetilde{k}}^T \boldsymbol{X}_{\widetilde{k}})^{-1} \boldsymbol{X}_{\widetilde{k}}^T$. The algorithm of model selection via tilted correlation for single regression model can be found in Section 2.5.2. It is straightforward to extend it to this 2-regression-model case as $\mathcal{C}_i^c$ and $\mathcal{C}_j^c$ are chosen separately.

2.5 Algorithm of the tilting estimators for precision matrix

2.5.1 Separate tilting

Here we list the algorithm of the separate tilting estimator. The simple and double tilting estimators can be achieved in the similar manner.

Step 1: Estimate the pairwise precision matrices by applying the separate tilting.

Step 1.1: For a given pair of (i, j), and a chosen threshold π_1, determine the controlling subsets $\mathcal{C}_i^{se} = \{c_i : |\widehat{\text{corr}}(X_{c_i}, X_i)| > \pi_1, c_i \in \mathcal{K}\}$ and $\mathcal{C}_j^{se} = \{c_j : |\widehat{\text{corr}}(X_{c_j}, X_j)| > \pi_1, c_j \in \mathcal{K}\}$.

Step 1.2: Compute the pairwise precision matrix $(\hat{\Sigma}_{ij}^{\circ})^{-1} = \frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_{c_i}(\boldsymbol{X}_{c_i}^T \boldsymbol{X}_{c_i})^{-1} \boldsymbol{X}_{c_i}^T)(\boldsymbol{I}_n - \boldsymbol{X}_{c_j}(\boldsymbol{X}_{c_j}^T \boldsymbol{X}_{c_j})^{-1} \boldsymbol{X}_{c_j}^T) X_j$.

Step 1.3: Repeat 2.1-2.2 for all the combination of i and j.

Step 2: Construct the precision matrix estimation.

Step 2.1: For off-diagonal entries, $\hat{t}_{i,j}^{se} = [(\hat{\Sigma}_{ij}^{\circ})^{-1}]_{1,2}$

Step 2.2: For diagonal entries, $\hat{t}_{i,j}^{se} = \frac{1}{p-1} \sum_{j \neq i} [(\hat{\Sigma}_{ij}^{\circ})^{-1}]_{1,1}$

2.5.2 Competing tilting and the TCS algorithm

The only difference between separate tilting and competing tilting is in Step 1.1, where the marginal correlation $\widehat{\text{corr}}$ is replaced by the tilted correlation $\widehat{\text{corr}}^*$ for competing tilting. The tilted correlation screening algorithm (TCS algorithm) is described in Section 3.1 of Cho and Fryzlewicz [2012]. Below, we list the algorithm which is taken from the paper to make the contents of the thesis coherent and easy to follow.

Consider the following linear model: $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \varrho$, where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T \in$

$\mathbb{R}^n$ is an n-vector of the response, $\boldsymbol{\beta} = \{\beta_1, \beta_2, \cdots, \beta_p\}$ is the coefficient vector, $\boldsymbol{X} = (X_1, X_2 \cdot \cdot, X_p)$ is an $n \times p$ design matrix and $\boldsymbol{\varrho} = (\varrho_1, \cdots, \varrho_n)^T \in \mathbb{R}^n$ is an n-vector of i.i.d. random errors, the aim of the TSC algorithm is to determine an active set denoted as $\mathcal{A}$, that contains the real relevant X variables to $\mathbf{y}$ after effectively removing the non-negligible effects of all the other X variables, i.e. according to the tilted correlation $\widehat{\text{corr}}^*$ between X variables and $\mathbf{y}$.

Step 0: Start with an empty active set $\mathcal{A} = \emptyset$, current residual $\mathbf{z} = \mathbf{y}$, and current design matrix $\boldsymbol{Z} = \boldsymbol{X}$.

Step 1: Find the variable which achieves the maximum marginal correlation with z and let $k = \arg\max_{j \notin \mathcal{A}} |Z_j^T z|$. Identify $\mathcal{C}_k = \{j \notin \mathcal{A}, j \neq k : |Z_j^T z| > \pi_n\}$ and if $\mathcal{C}_k = \emptyset$, let $k^* = k$ and go to Step 3.

Step 2: If $\mathcal{C}_k \neq \emptyset$, screen the tilted correlations between Z_j and z, $\widehat{\text{corr}}^*(Z_j, z)$ in formula (2.46), for $j \in \mathcal{C}_k \cup \{k\}$ and find $k^* = \arg\max_{j \in \mathcal{C}_k \cup \{k\}} |\widehat{\text{corr}}^*(Z_j, z)|$.

Step 3: Add k^* to $\mathcal{A}$ and update the current residual and the current design matrix $\mathbf{z} \leftarrow (\boldsymbol{I}_n - \boldsymbol{H}_{\mathcal{A}})\mathbf{y}$ and $\boldsymbol{Z} \leftarrow (\boldsymbol{I}_n - \boldsymbol{H}_{\mathcal{A}})\boldsymbol{X}$, respectively, where $\boldsymbol{H}_{\mathcal{A}}$ the projection matrix of X_k, $k \in \mathcal{A}$, i.e., $\boldsymbol{H}_{\mathcal{A}} \doteq \boldsymbol{X}_{\mathcal{A}}(\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}})^{-1} \boldsymbol{X}_{\mathcal{A}}^T$. Further, rescale each column $j \notin \mathcal{A}$ of $\boldsymbol{Z}$ to have norm one.

Step 4:: Repeat Step 1 to 3 until the cardinality of active set $|\mathcal{A}|$ reaches a pre-specified $m_1 < n$.

2.6 Asymptotic properties of tilting methods

2.6.1 Fixed p: asymptotic properties of ${\hat{\Sigma}^{\circ}_{m\times m}}^{-1}$

In this section, we briefly show the consistency of the building block of the estimators in the generalized form, i.e. ${\hat{\Sigma}^{\circ}_{m\times m}}^{-1}$ defined in formula (2.24)-(2.25), when p is fixed. Here, the controlling subsets contain all the other variables, that is to say, for fixed p, it is safe to include all the remaining variables in the controlling subsets as long as n is large enough.

Lemma 2 ***Consistency**: If $p < \infty$ and $2 \leqslant m < p$, then ${\hat{\Sigma}^{\circ}_{m\times m}}^{-1} \xrightarrow{p} P_{m\times m}$.*

Proof of Lemma 2: Since each X_i, $i \in \mathcal{P}$, follows i.i.d. Gaussian distribution with mean zero and finite variance, by the weak law of large numbers [Davidson, 1994, p.289], we note that $\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}} \xrightarrow{p} \Sigma_{m\times m}$, $\boldsymbol{X}_{-\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}} \xrightarrow{p} \Sigma_{(p-m)\times m}$, $\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{-\mathcal{S}} \xrightarrow{p} \Sigma_{m\times(p-m)}$, $\boldsymbol{X}_{-\mathcal{S}}^T\boldsymbol{X}_{-\mathcal{S}} \xrightarrow{p} \Sigma_{(p-m)\times(p-m)}$. By Slutsky's Theorem [Serfling, 2009, p.19] and the block-wise inversion of covariance matrix [Bernstein, 2009, p.147], we obtain

$$
\begin{aligned}
{\hat{\Sigma}^{\circ}_{m\times m}}^{-1} &= (\hat{\boldsymbol{\epsilon}}_{\mathcal{S}}^T\hat{\boldsymbol{\epsilon}}_{\mathcal{S}})^{-1} \\
&= (\boldsymbol{X}_{\mathcal{S}}^T(I - \boldsymbol{H}_{-\mathcal{S}})\boldsymbol{X}_{\mathcal{S}})^{-1} \\
&= (\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}} - \boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{-\mathcal{S}}(\boldsymbol{X}_{-\mathcal{S}}^T\boldsymbol{X}_{-\mathcal{S}})^{-1}\boldsymbol{X}_{-\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})^{-1} \\
&\xrightarrow{p} (\Sigma_{m\times m} - \Sigma_{m\times(p-m)}\Sigma^{-1}_{(p-m)\times(p-m)}\Sigma_{(p-m)\times m})^{-1} \\
&= P_{m\times m}
\end{aligned}
\tag{2.47}
$$

The lemma follows.

2.6.2 $p \to \infty$: assumptions and consistency

2.6.2.1 Assumptions

In studying the theoretical properties of the four types of tilting methods for estimating precision matrices, we make the following assumptions and also give the reasons for and examples satisfying these assumption.

A. 1 *For any $i, j \in \mathcal{P}$ in formula (2.28) and (2.29), we assume X_b, X_{e_i}, X_{e_j}, X_u, ϵ_i and ϵ_j are mutually uncorrelated.*

Assumption (A.1) is made in order to ensure element-wise consistency of the precision matrix via all the four types tilting methods. Although it seems a strong assumption, we can find examples which satisfy the assumption. For example, absolute diagonal block covariance matrix as Model (B) in Section 2.10.1 is a typical example.

A. 2 *Condition of high dimensional cases: $\log p = \mathcal{O}(n^\theta)$ for $\theta \in [0, 1-2\gamma)$, for $\gamma \in (\delta, 1/2)$.*

A. 3 *The total number of non-zero coefficients for either X_i or X_j satisfies $|\mathcal{B}|+|\mathcal{E}_i|+|\mathcal{E}_j| = O(n^\delta)$, $\delta \in [0, 1/2)$.*

A. 4 *The predictors of X_i satisfy $n^{(3-\theta)/2} \cdot min_{c_i \in \mathcal{B}\cup\mathcal{E}_i} |X_{c_i}^T X_i| \to \infty$, and the predictors of X_j satisfy $n^{(3-\theta)/2} \cdot min_{c_j \in \mathcal{B}\cup\mathcal{E}_j} |X_{c_j}^T X_j| \to \infty$.*

Assumption (A.2) lets the dimension p grow with n. Assumption (A.3) allows the number of relevant remaining variables to grow with n, but also ensures the simple tilting to be well-conditioned, which is the strongest condition among all four types of tilting methods. Assumption (A.4) is to ensure consistency of simple, double and separate tilting.

A. 5 *Non-zero coefficients satisfy* $n^{\mu} \cdot min_{c_i \in \mathcal{B} \cup \mathcal{E}_i} |\beta_{i,c_i}| \to \infty$ *for* $\mu \in [0, \gamma - \delta - \xi/2)$.

A. 6 *The threshold is chosen as* $\pi_n = Cn^{-\gamma}$ *for some* $C > 0$. *We assume that there exists* $C_0 > 0$ *such that* $\mathcal{C}_k = \{\widetilde{k} : |X_k^T X_{\widetilde{k}}| > \pi_n\}$ *is of cardinality* $|\mathcal{C}_k| \leqslant C_0 n^{\xi}$ *uniformly over all* k, *where* $\xi \in [0, 2(\gamma - \delta))$.

A. 7 *After standardization, there exists* $\alpha \in (0, 1)$ *satisfying* $1 - X_i^T \boldsymbol{H}_{\mathcal{C}_i^s} X_i = 1 - a_i > \alpha$, *for all* $i \in \mathcal{P}$.

A. 8 *For each* $i \in \mathcal{P}$, $k \in \mathcal{K}$ *and whose corresponding* $\mathcal{C}_k$ *satisfies* $B \cup \mathcal{E}_i \nsubseteq \mathcal{C}_k$, *we have*

$$n^{\kappa} \cdot \frac{\|(\boldsymbol{I}_n - \boldsymbol{H}_k)\boldsymbol{X}_{\mathcal{B} \cup \mathcal{E}_i} \boldsymbol{\beta}_{\mathcal{B} \cup \mathcal{E}_i}\|_2^2}{\|\boldsymbol{X}_{\mathcal{B} \cup \mathcal{E}_i} \boldsymbol{\beta}_{\mathcal{B} \cup \mathcal{E}_i}\|_2^2} \to \infty$$

for κ *satisfying* $\kappa/2 + \mu \in [0, \gamma - \delta - \xi/2)$.

Assumption (A.5)-assumption (A.8) are taken from Section 2.3 of Cho and Fryzlewicz [2012] to achieve consistency of tilted correlation in single regression model, and to ensure consistency of competing tilting as shown in Theorem 1. Below, we list the reasons for and examples satisfying these assumptions, which are taken from Cho and Fryzlewicz [2012].

Assumption (A.5) imposes a lower bound on the absolute values of the non-zero coefficients, which still allows the minimum non-zero coefficient to decay to zero as n increases. At the same time, it imposes an upper bound on the magnitudes of the non-zero coefficients to ensure that the ratio in absolute value between the largest and smallest non-zero coefficients does not grow too quickly with n. Assumption (A.6)-assumption (A.8) are all applied to correlation matrices. Assumption (A.6) is to provide a bound in order to guarantee the existence of the projection matrix into the space

spanned by $\boldsymbol{X}_{\widetilde{k}}$, as well as to prevent tilted correlations from being distorted by high dimensionality. Assumption (A.7) is required for ruling out strong collinearity among variables due to the fact that $1 - a_i = \det(\boldsymbol{X}^T_{\mathcal{C}_i \cup \{i\}} \boldsymbol{X}_{\mathcal{C}_i \cup \{i\}}) / \det(\boldsymbol{X}^T_{\mathcal{C}_i} \boldsymbol{X}_{\mathcal{C}_i})$, which is highly related to strict positive definiteness of Σ [Bühlmann et al., 2009; Fan and Li, 2007; Zou, 2006]. Assumption (A.8) is linked to the asymptotic identifiability condition for high-dimensional problems first introduced in Chen and Chen [2008]. Further, one example of when assumption (A.6) is satisfied and a certain mild assumptions from Wang [2009] upon which assumption (A.7) and (A.8) are satisfied are presented in Section 2.12.2.

In practice, one may want to check whether these conditions are satisfied. If the true subsets $\mathcal{B}$, $\mathcal{E}_i$, $\mathcal{E}_j$ and $\mathcal{U}$ are known, it is straightforward to check assumptions (A.1-A.3) and (A.6-A.7) by using the observed values after suitable algebraic operations. Otherwise, we firstly need to apply tilted correlation to obtain the estimation of the coefficients and the following estimation of the subsets $\mathcal{B}$, $\mathcal{E}_i$, $\mathcal{E}_j$ and $\mathcal{U}$, then check the assumptions based on the results. However, for any given datasets with fixed p and n, it makes less sense for checking asymptotic assumptions such as a measurement goes to infinity asymptotically. But, when n is very large, we can still set a large enough finite value as a boundary and if the realisation of the measurement is larger than the boundary, it is viewed that the assumption is satisfied. The asymptotic assumptions include (A.4-A.5) and (A.8).

2.6.2.2 Element-wise consistency

Theorem 1 *Under assumptions (A.1)-(A.8), for any* $i, j \in \mathcal{P}$*, we have* $\lim_{n\to\infty} Pr(\Delta_l > \delta) = 0$*, for any* $\delta > 0$*,* $l = 1, 2, 3, 4$*, where*

$$\Delta_1 = |cov(\epsilon_i, \epsilon_j) - \frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_c (\boldsymbol{X}_c^T \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^T) X_j : c \in \mathcal{C}_{ij}^s|, \tag{2.48}$$

$$\Delta_2 = |cov(\epsilon_i, \epsilon_j) - \frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_c (\boldsymbol{X}_c^T \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^T) X_j : c \in \mathcal{C}_{ij}^d|, \tag{2.49}$$

$$\Delta_3 = |cov(\epsilon_i, \epsilon_j) - \frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_{c_i} (\boldsymbol{X}_{c_i}^T \boldsymbol{X}_{c_i})^{-1} \boldsymbol{X}_{c_i}^T)(\boldsymbol{I}_n - \boldsymbol{X}_{c_j} (\boldsymbol{X}_{c_j}^T \boldsymbol{X}_{c_j})^{-1} \boldsymbol{X}_{c_j}^T) X_j$$
$$: c_i \in \mathcal{C}_i^{se}, c_j \in \mathcal{C}_j^{se}|, \tag{2.50}$$

$$\Delta_4 = |cov(\epsilon_i, \epsilon_j) - \frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_{c_i} (\boldsymbol{X}_{c_i}^T \boldsymbol{X}_{c_i})^{-1} \boldsymbol{X}_{c_i}^T)(\boldsymbol{I}_n - \boldsymbol{X}_{c_j} (\boldsymbol{X}_{c_j}^T \boldsymbol{X}_{c_j})^{-1} \boldsymbol{X}_{c_j}^T) X_j$$
$$: c_i \in \mathcal{C}_i^c, c_j \in \mathcal{C}_j^c|. \tag{2.51}$$

The proof is given in the Section 2.12.3. Theorem 1 shows element-wise consistency of the precision matrix estimators via four types of tilting methods. Δ_1, Δ_2, Δ_3, Δ_4 correspond to simple, double, separate and competing tilting respectively.

2.7 Finite sample performance: comparisons between tilting and thresholding estimators

Apart from asymptotic properties, we are also interested in finite sample performance. In this section, we would like to investigate the finite sample performance of tilting estimators for precision matrix and the links and differences between tilting and thresholding (both soft and hard) estimators. We choose the thresholding estimators as competitors due to their simplicity and popularity [Bickel and Levina, 2008b; Cai and Liu,

2011; Rothman et al., 2009] as well as the close link between tilting and hard thresholding estimator under certain cases. The three methods are defined below. Soft and hard thresholding are applied on the sample covariance matrices.

(1) Soft thresholding: $\hat{P}^{sf} = (\hat{\Sigma}^{sf})^{-1} = \{\hat{p}^{sf}_{i,j}\}$, where $\hat{\Sigma}^{sf} = \{\hat{\sigma}^{sf}_{i,j}\}$, and

$$\hat{\sigma}^{sf}_{i,j} = \begin{cases} (\hat{\sigma}_{i,j} - \text{sign}(\hat{\sigma}_{i,j})\lambda)\mathbb{1}(|\hat{\sigma}_{i,j}| > \lambda) & \text{if } i \neq j \\ \hat{\sigma}_{i,j} & \text{if } i = j \end{cases}, \tag{2.52}$$

where λ is a selected threshold, $\lambda \in (0, 1)$.

(2) Hard thresholding: $\hat{P}^{h} = (\hat{\Sigma}^{h})^{-1} = \{\hat{p}^{h}_{i,j}\}$, where $\hat{\Sigma}^{h} = \{\hat{\sigma}^{h}_{i,j}\}$, and

$$\hat{\sigma}^{h}_{i,j} = \begin{cases} \hat{\sigma}_{i,j}\mathbb{1}(|\hat{\sigma}_{i,j}| > \lambda) & \text{if } i \neq j \\ \hat{\sigma}_{i,j} & \text{if } i = j \end{cases}, \tag{2.53}$$

(3) Tilting: as stated in formula (2.26)-(2.27), $\hat{T} = \{\hat{t}_{i,j}\}$, where

$$\hat{t}_{i,j} = \begin{cases} [\hat{\Sigma}^{\circ}_{ij}{}^{-1}]_{1,2} & \text{if } i \neq j \\ \frac{1}{p-1}\sum_{l \in \mathcal{P}\setminus\{i\}} [\hat{\Sigma}^{\circ}_{il}{}^{-1}]_{1,1} & \text{if } i = j \end{cases} \tag{2.54}$$

2.7.1 Case I: Σ^{-1} = diagonal matrix

When the underlying precision matrix and covariance matrix are the diagonal matrices, we will show that tilting will never perform better than thresholding under certain assumptions. It is not surprising as diagonal matrix is the simplest sparse matrix which is the thresholding estimators designed for.

Denote true covariance matrix as $\Sigma = \text{diag}\{\sigma_{1,1}, \cdots, \sigma_{p,p}\}$, true precision matrix as $P = \text{diag}\{\sigma^{-1}_{1,1}, \cdots, \sigma^{-1}_{p,p}\}$ and sample covariance matrix as $\hat{\Sigma} = \{\hat{\sigma}_{i,j}\}$, for all $i, j \in \mathcal{P}$.

Assuming there exist $\lambda > 0$, s.t. $|\hat{\sigma}_{i,j}| < \lambda$ for all $i \neq j$ such that the thresholding estimators can reduce all each off-diagonal to 0. Hence, soft thresholding can obtain $\hat{\Sigma}^{sf} = \text{diag}\{\hat{\sigma}_{1,1}, \cdots, \hat{\sigma}_{p,p}\}$, and the corresponding precision matrix estimator is $\hat{P}^{sf} = \text{diag}\{\hat{\sigma}_{1,1}^{-1}, \cdots, \hat{\sigma}_{p,p}^{-1}\}$. Hard thresholding yields the same result as soft thresholding in this case.

The situation for four types of tilting methods are the same for this case. Here we only illustrate simple tilting as a example. For given $i, j \in \mathcal{P}$, $i \neq j$, assuming there exists a threshold π_1 such that $\mathcal{C}_{ij}^{s} = \emptyset$, we have $\hat{\Sigma}_{ij}^{\circ} = \begin{pmatrix} \hat{\sigma}_{i,i} & \hat{\sigma}_{i,j} \\ \hat{\sigma}_{j,i} & \hat{\sigma}_{j,j} \end{pmatrix}$. Hence, for off-diagonals, we obtain $\hat{t}_{i,j} = -\hat{\sigma}_{i,j}/(\hat{\sigma}_{i,i}\hat{\sigma}_{j,j} - \hat{\sigma}_{i,j}\hat{\sigma}_{j,i}) \neq 0$ if $\hat{\sigma}_{i,j} \neq 0$. For diagonals, we have $\hat{t}_{i,j} = \frac{1}{p-1}\sum_{l\in\mathcal{P}\setminus\{i\}} \hat{\sigma}_{l,l}/(\hat{\sigma}_{i,i}\hat{\sigma}_{l,l} - \hat{\sigma}_{i,l}\hat{\sigma}_{l,i}) \geqslant \frac{1}{p-1}\sum_{l\in\mathcal{P}\setminus\{i\}} \hat{\sigma}_{l,l}/(\hat{\sigma}_{i,i}\hat{\sigma}_{l,l}) = \hat{\sigma}_{i,i}^{-1}$, as summarised in Table 2.1. Illustration by a small panel of simulation results also shows the relationships, see Table 2.2 .

Table 2.1: Comparison of precision estimators in Case I

Index	True	Soft		Hard		Tilting
$i = j$	$\sigma_{i,j}^{-1}$	$\hat{\sigma}_{i,j}^{-1}$	=	$\hat{\sigma}_{i,j}^{-1}$	$\leqslant$	$\frac{1}{p-1}\sum_{l\in\mathcal{P}\setminus\{i\}} \hat{\sigma}_{l,l}/(\hat{\sigma}_{i,i}\hat{\sigma}_{l,l} - \hat{\sigma}_{i,l}\hat{\sigma}_{l,i})$
$i \neq j$	0	0	=	0	$\overset{\text{if } \hat{\sigma}_{i,j}\neq 0}{\neq}$	$-\hat{\sigma}_{i,j}/(\hat{\sigma}_{i,i}\hat{\sigma}_{j,j} - \hat{\sigma}_{i,j}\hat{\sigma}_{j,i})$

Although tilting estimators for precision matrix can achieve asymptotically element-wise consistency, we find that, the finite sample performance shows that they are positively skewed for diagonals, and oscillate around the true values for off-diagonals, as long as sample covariance matrix contains non-zero off-diagonal entries. That is to say tilting estimators cannot achieve better performance than thresholding estimators if true covariance is a diagonal matrix.

Table 2.2: Means and variances (in brackets) of the precision matrix estimators from Case I

		p=50, n=1000		p=50, n=100	
Index	True	Soft/hard	Tilting	Soft/hard	Tilting
$i=j$	1	1.003 (0.002)	1.024 (0.002)	1.022 (0.022)	1.054 (0.023)
$i\neq j$	0	0	-1.001×10^{-4} (0.001)	0	-2.768×10^{-4} (0.011)

2.7.2 Case II: Σ^{-1} = diagonal block matrix

Suppose true covariance structure and the corresponding precision matrix are

$$\Sigma=\begin{pmatrix}\Sigma_{\mathcal{A}_1} & & & 0\\ & \Sigma_{\mathcal{A}_2} & & \\ & & \cdots & \\ 0 & & & \Sigma_{\mathcal{A}_W}\end{pmatrix},\quad P=\begin{pmatrix}\Sigma_{\mathcal{A}_1}^{-1} & & & 0\\ & \Sigma_{\mathcal{A}_2}^{-1} & & \\ & & \cdots & \\ 0 & & & \Sigma_{\mathcal{A}_W}^{-1}\end{pmatrix}$$

where $\Sigma_{\mathcal{A}_1},\Sigma_{\mathcal{A}_2},\cdots,\Sigma_{\mathcal{A}_W}$ are square blocks with all entries being non-zeros, and $\mathcal{A}_1\cup\mathcal{A}_2\cup\cdots\cup\mathcal{A}_W=\mathcal{P}$ and the sample covariance matrix as $\hat{\Sigma}=\{\hat{\sigma}_{i,j}\}$.

For soft and hard thresholding estimators, we assume that there exists a suitable threshold λ such that $|\hat{\sigma}_{i,j}|>\lambda$, if $\sigma_{i,j}\neq0$ and $|\hat{\sigma}_{i,j}|\leqslant\lambda$, if $\sigma_{i,j}=0$. We take $|\mathcal{A}_w|=2$ as an example. The soft and hard estimators for the covariance matrix are denoted as

$$\hat{\Sigma}^{sf} = \begin{pmatrix} \hat{\Sigma}^{sf}_{\mathcal{A}_1} & & & 0 \\ & \hat{\Sigma}^{sf}_{\mathcal{A}_2} & & \\ & & \cdots & \\ 0 & & & \hat{\Sigma}^{sf}_{\mathcal{A}_W} \end{pmatrix}, \quad \hat{\Sigma}^{h} = \begin{pmatrix} \hat{\Sigma}^{h}_{\mathcal{A}_1} & & & 0 \\ & \hat{\Sigma}^{h}_{\mathcal{A}_2} & & \\ & & \cdots & \\ 0 & & & \hat{\Sigma}^{h}_{\mathcal{A}_W} \end{pmatrix},$$

where, for each $\mathcal{A}_w$, $w \in \{1, 2, \cdots, W\}$,

$$\hat{\Sigma}^{sf}_{\mathcal{A}_w} = \begin{pmatrix} \hat{\sigma}_{i,i} & \hat{\sigma}_{i,j} - \text{sign}(\hat{\sigma}_{i,j})\lambda \\ \hat{\sigma}_{j,i} - \text{sign}(\hat{\sigma}_{j,i})\lambda & \hat{\sigma}_{j,j} \end{pmatrix}, \quad \hat{\Sigma}^{h}_{\mathcal{A}_w} = \begin{pmatrix} \hat{\sigma}_{i,i} & \hat{\sigma}_{i,j} \\ \hat{\sigma}_{j,i} & \hat{\sigma}_{j,j} \end{pmatrix}.$$

Then the corresponding estimators for the precision matrix are denoted as

$$\hat{P}^{sf} = \begin{pmatrix} \hat{P}^{sf}_{\mathcal{A}_1} & & & 0 \\ & \hat{P}^{sf}_{\mathcal{A}_2} & & \\ & & \cdots & \\ 0 & & & \hat{P}^{sf}_{\mathcal{A}_W} \end{pmatrix}, \quad \hat{P}^{h} = \begin{pmatrix} \hat{P}^{h}_{\mathcal{A}_1} & & & 0 \\ & \hat{P}^{h}_{\mathcal{A}_2} & & \\ & & \cdots & \\ 0 & & & \hat{P}^{h}_{\mathcal{A}_W} \end{pmatrix},$$

where $\hat{P}^{sf}_{\mathcal{A}_w} = (\hat{\Sigma}^{sf}_{\mathcal{A}_w})^{-1}$, and $\hat{P}^{h}_{\mathcal{A}_w} = (\hat{\Sigma}^{h}_{\mathcal{A}_w})^{-1}$.

We compare each element of $\hat{P}^{sf}$ and $\hat{P}^{h}$, and summarise the following relationships. For $i = j$, we have $\hat{p}^{sf}_{i,i} = \hat{\sigma}_{j,j}/(\hat{\sigma}_{i,i}\hat{\sigma}_{j,j} - (\hat{\sigma}_{i,j} - \lambda)^2) < \hat{\sigma}_{j,j}/(\hat{\sigma}_{i,i}\hat{\sigma}_{j,j} - \hat{\sigma}^2_{i,j}) = \hat{p}^{h}_{i,i}$; for $i \neq j$, we have $|\hat{p}^{sf}_{i,j}| = |\lambda - \hat{\sigma}_{i,j}|/(\hat{\sigma}_{i,i}\hat{\sigma}_{j,j} - (\hat{\sigma}_{i,j} - \lambda)^2) < |-\hat{\sigma}_{i,j}|/(\hat{\sigma}_{i,i}\hat{\sigma}_{j,j} - \hat{\sigma}^2_{i,j}) = |\hat{p}^{h}_{i,j}|$; and $\hat{p}^{sf}_{i,j} = \hat{p}^{h}_{i,j} = 0$ else. The similar results can be generalized to $2 < |\mathcal{A}_w| < n$.

For the tilting methods, we assume that there exists a suitable threshold π_1 such that $\mathcal{C}^d_{ij}$ can distinguish the relevant and irrelevant remaining variables. We take double tilting as an example. There are two scenarios for all $2 \leqslant |\mathcal{A}_w| \leqslant p$:

(1) If there exist w, $w \in \{1, 2, \cdots W\}$, such that $\{i, j\} \subset \mathcal{A}_w$, i.e. X_i and X_j are in the same block, we are able to include all the variables within that block in the regression models. By doing this, we find that, for $|\mathcal{A}_w| = 2$, $\hat{\Sigma}^{\circ}_{i,j}{}^{-1} = (\boldsymbol{X}_{ij}^T \boldsymbol{X}_{ij})^{-1} = (\hat{\Sigma}^h_{\mathcal{A}_w})^{-1}$, and for $2 < |\mathcal{A}_w| \leqslant p$,

$$\hat{\Sigma}^{\circ}_{i,j}{}^{-1} = X_i^T (\boldsymbol{I}_{|\mathcal{A}_w|-2} - \boldsymbol{H}_{\mathcal{A}_w \setminus \{i,j\}}) X_j, \tag{2.55}$$

which is also equivalent to the corresponding 2×2 matrix in $(\hat{\Sigma}^h_{\mathcal{A}_w})^{-1}$. That is to say, the tilting methods obtain same results as the hard thresholding estimator for each off-diagonals within the blocks.

(2) If there is no such w, i.e. X_i and X_j are in different blocks, the controlling subsets will be empty for double tilting, and we have

$$\hat{\Sigma}^{\circ}_{ij}{}^{-1} = \begin{pmatrix} \hat{\sigma}_{i,i} & \hat{\sigma}_{i,j} \\ \hat{\sigma}_{j,i} & \hat{\sigma}_{j,j} \end{pmatrix}^{-1}, \tag{2.56}$$

leading to the same results as what tilting yields in case I, see Section 2.7.1.

We note that if we assume there exists a threshold π_2 satisfying $\max_{i,j \in \mathcal{P}, i \neq j} |\hat{\sigma}_{i,j} / (\hat{\sigma}_{i,i} \hat{\sigma}_{j,j} - \hat{\sigma}_{i,j} \hat{\sigma}_{j,i})| < \pi_2$, for large enough sample size n, we can always further regularise tilting estimators by applying hard thresholding with $\lambda = \pi_2$ on the tilting results to reduce all the elements outside the blocks to be zero. After this step, double tilting estimators will yield the same results as hard thresholding estimators, apart from small differences among diagonals.

However, it is not the case for competing tilting, that would not control all the remaining variables in the blocks due to possible collinearity. Simple and separate tilting yields slightly different off-diagonals as the controlling subsets are not empty

even if X_i and X_j are not in the same block. Table 2.3 summaries the results and relationships, and table 2.4 presents simulation examples.

Table 2.3: Comparison of precision estimators in Case II ($|\mathcal{A}_w| = 2$)

Index	True	Soft		Hard		Double tilting
$i = j$	$\frac{\sigma_{j,j}}{(\sigma_{i,i}\sigma_{j,j}-\sigma_{i,j}^2)}$	$\frac{\hat{\sigma}_{j,j}}{(\hat{\sigma}_{i,i}\hat{\sigma}_{j,j}-(\hat{\sigma}_{i,j}-\lambda)^2)}$	$<$	$\frac{\hat{\sigma}_{j,j}}{(\hat{\sigma}_{i,i}\hat{\sigma}_{j,j}-\hat{\sigma}_{i,j}^2)}$	$\overset{m>1}{\neq}$	$\frac{1}{p-1}\sum_{l\in\mathcal{P}\setminus\{i\}}\frac{\hat{\sigma}_{l,l}}{(\hat{\sigma}_{i,i}\hat{\sigma}_{l,l}-\hat{\sigma}_{i,l}\hat{\sigma}_{l,i})}$
$i \neq j$ in the blocks	$\frac{-\sigma_{i,j}}{(\sigma_{i,i}\sigma_{j,j}-\sigma_{i,j}^2)}$	$\frac{\text{sign}(\hat{\sigma}_{ij})\max(\lvert\hat{\sigma}_{i,j}\rvert-\lambda,0)}{\hat{\sigma}_{i,i}\hat{\sigma}_{j,j}-(\hat{\sigma}_{i,j}-\lambda)^2}$	$\overset{\lvert\cdot\rvert}{<}$	$\frac{-\hat{\sigma}_{i,j}}{(\hat{\sigma}_{i,i}\hat{\sigma}_{j,j}-\hat{\sigma}_{i,j}^2)}$	$=$	$\frac{-\hat{\sigma}_{i,j}}{(\hat{\sigma}_{i,i}\hat{\sigma}_{j,j}-\hat{\sigma}_{i,j}^2)}$
$i \neq j$ outside the blocks	0	0	$=$	0	$\overset{\text{if } \hat{\sigma}_{i,j}\neq 0}{\neq}$	$\frac{-\hat{\sigma}_{i,j}}{(\hat{\sigma}_{i,i}\hat{\sigma}_{j,j}-\hat{\sigma}_{i,j}^2)}$

Note: $a \overset{|\cdot|}{<} b$ means $|a| < |b|$.

Table 2.4: Means and variances (in brackets) of the precision matrix estimators from Case II ($|\mathcal{A}_w| = 2$)

		p=50, n=1000			p=50, n=100		
Index	True	Soft	Hard	Tilting	Soft	Hard	Tilting
$i = j$	1.333	1.045 (0.002)	1.338 (0.004)	1.011 (0.002)	1.067 (0.019)	1.373 (0.040)	1.035 (0.022)
$i \neq j$ in the blocks	-0.667	−0.207 (0.002)	−0.667 (0.002)	−0.667 (0.002)	−0.202 (0.010)	−0.687 (0.025)	−0.687 (0.025)
$i \neq j$ outside the blocks	0	0	0	-2.899×10^{-4} (0.001)	0	0	9.907×10^{-4} (0.011)

Some remarks: if there exist certain thresholds which can correctly identify the blocks, double tilting estimator will yield the same results as hard thresholding estimator for all the off-diagonal elements within the blocks. If we apply suitable thresholding methods afterwards, tilting can also reduce the elements outside the blocks to zero, which are also equal to those of hard thresholding estimator. However, by choosing dif-

ferent thresholds, tilting estimators can obtain very different results from thresholding, particularly for large $|\mathcal{A}_w|$, which is true for all the underlying covariance structures, particularly for the non-sparse ones. Moreover, comparison between soft and hard thresholding indicates that when $n \gg p$, soft thresholding is not favourable as it is a biased estimator, otherwise, soft thresholding regularizes the distorted sample covariance matrix towards the truth much quicker and is preferable when p is possibly much larger than n.

2.7.3 Case III: Factor model

Suppose the random variables are generated from a k-factor model as follows,

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{f} + \boldsymbol{\varepsilon}, \tag{2.57}$$

where $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ be a vector of n i.i.d. observations of a p-dimensional random variable, $\boldsymbol{f}$ is a $k \times n$ matrix of common factors, $k \leqslant p-2$, $k \ll n$, $\boldsymbol{B} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_p\}^T$ is a $p \times k$ coefficient matrix, which contains only positive entries, and there exists a threshold $\lambda > 0$ satisfying that $|\boldsymbol{\beta}_i \boldsymbol{\beta}_j^T| > \lambda$ for all $i, j \in \mathcal{P}, i \neq j$, $\boldsymbol{\varepsilon} = \{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p\}$ is a $p \times n$ matrix of noise component. We assume that all the factors and noises are with mean zero, variance one and uncorrelated with each other. *i.e.* $E(\boldsymbol{f}) = 0$, var$(\boldsymbol{f}) = I_k$; $E(\boldsymbol{\varepsilon}) = 0$, var$(\boldsymbol{\varepsilon}) = I_p$; cov$(\boldsymbol{f}, \varepsilon_i) = 0$, $i = 1, 2, \cdots, p$. We denote $A = \sum_{p\in\mathcal{P}} \boldsymbol{\beta}_p \boldsymbol{\beta}_p^T$, $B = \sum_{l\in\mathcal{P}\setminus\{i\}} \boldsymbol{\beta}_l \boldsymbol{\beta}_l^T$, $C = \sum_{v\in\mathcal{P}} \sum_{u\in\mathcal{P}} \boldsymbol{\beta}_v \boldsymbol{\beta}_u^T$, $D = \sum_{m\in\mathcal{P}\setminus\{i,j\}} \boldsymbol{\beta}_m \boldsymbol{\beta}_m^T$, $E = \sum_{s\in\{i,j\}} \sum_{m\in\mathcal{P}\setminus\{i,j\}} \boldsymbol{\beta}_s \boldsymbol{\beta}_m^T$, $F = \boldsymbol{\beta}_i \boldsymbol{\beta}_j^T$, $G = \boldsymbol{\beta}_i \boldsymbol{\beta}_i^T$. And $\hat{A} \sim \hat{G}$ are the corresponding sample versions of $A \sim G$. Note that the true covariance

matrix and precision matrix are $\Sigma = \{\sigma_{i,j}\}$ and $P = \{p_{i,j}\}$ respectively, where

$$\sigma_{i,j} = \begin{cases} \boldsymbol{\beta}_i\boldsymbol{\beta}_i^T + 1 = G + 1 & \text{if } i = j \\ \boldsymbol{\beta}_i\boldsymbol{\beta}_j^T = F & \text{if } i \neq j \end{cases}, \quad p_{i,j} = \begin{cases} \frac{\sum_{l\in\mathcal{P}\setminus\{i\}} \boldsymbol{\beta}_l\boldsymbol{\beta}_l^T+1}{\sum_{p\in\mathcal{P}} \boldsymbol{\beta}_p\boldsymbol{\beta}_p^T+1} = \frac{B+1}{A+1} & \text{if } i = j \\ \frac{-\boldsymbol{\beta}_i\boldsymbol{\beta}_j^T}{\sum_{p\in\mathcal{P}} \boldsymbol{\beta}_p\boldsymbol{\beta}_p^T+1} = \frac{-F}{A+1} & \text{if } i \neq j \end{cases}.$$

For hard and soft thresholding estimators, we assume common factors and noise components have identity sample variance when n is large enough for simplicity. After simple algebra, we obtain the results for hard thresholding as follows,

$$\hat{\sigma}_{i,j}^h = \begin{cases} \hat{\boldsymbol{\beta}}_i\hat{\boldsymbol{\beta}}_i^T + 1 = \hat{G} + 1 & \text{if } i = j \\ \hat{\boldsymbol{\beta}}_i\hat{\boldsymbol{\beta}}_j^T = \hat{F} & \text{if } i \neq j \end{cases}, \quad \hat{p}_{i,j}^h = \begin{cases} \frac{\hat{B}+1}{\hat{A}+1} & \text{if } i = j \\ \frac{-\hat{F}}{\hat{A}+1} & \text{if } i \neq j \end{cases},$$

and soft thresholding estimators as

$$\hat{\sigma}_{i,j}^{sf} = \begin{cases} \hat{\boldsymbol{\beta}}_i\hat{\boldsymbol{\beta}}_i^T + 1 = \hat{G} + 1 & \text{if } i = j \\ \hat{\boldsymbol{\beta}}_i\hat{\boldsymbol{\beta}}_j^T - \lambda = \hat{F} - \lambda & \text{if } i \neq j \end{cases}, \quad \hat{p}_{i,j}^{sf} = \begin{cases} \frac{\hat{B}+1+2\hat{F}\lambda-\lambda^2}{\hat{A}+1+(\hat{C}-\hat{A})\lambda+(\hat{C}-2\hat{A}-3)\lambda^2+2\lambda^3} & \text{if } i = j \\ \frac{-\hat{F}+(\hat{D}-\hat{E}+1)\lambda+\lambda^2}{\hat{A}+1+(\hat{C}-\hat{A})\lambda+(\hat{C}-2\hat{A}-3)\lambda^2+2\lambda^3} & \text{if } i \neq j \end{cases}.$$

For simple, double and separate tilting methods, given i and j, we obtain the following expressions for the precision matrix estimator,

$$\hat{t}_{i,j} = \begin{cases} \frac{-\hat{\boldsymbol{\beta}}_i(\hat{\boldsymbol{B}}_{\mathcal{K}}^T\hat{\boldsymbol{B}}_{\mathcal{K}})^{-1}\hat{\boldsymbol{\beta}}_j^T}{\hat{\boldsymbol{\beta}}_i(\hat{\boldsymbol{B}}_{\mathcal{K}}^T\hat{\boldsymbol{B}}_{\mathcal{K}})^{-1}\hat{\boldsymbol{\beta}}_i^T+\hat{\boldsymbol{\beta}}_j(\hat{\boldsymbol{B}}_{\mathcal{K}}^T\hat{\boldsymbol{B}}_{\mathcal{K}})^{-1}\hat{\boldsymbol{\beta}}_j^T+1} & \text{if } i \neq j \\ \frac{1}{p-1}\sum_{l\in\mathcal{P}\setminus\{i\},\mathcal{K}_l=\mathcal{P}\setminus\{i,l\}} \frac{\hat{\boldsymbol{\beta}}_l(\hat{\boldsymbol{B}}_{\mathcal{K}_l}^T\hat{\boldsymbol{B}}_{\mathcal{K}_l})^{-1}\hat{\boldsymbol{\beta}}_l^T+1}{\hat{\boldsymbol{\beta}}_i(\hat{\boldsymbol{B}}_{\mathcal{K}_l}^T\hat{\boldsymbol{B}}_{\mathcal{K}_l})^{-1}\hat{\boldsymbol{\beta}}_i^T+\hat{\boldsymbol{\beta}}_l(\hat{\boldsymbol{B}}_{\mathcal{K}_l}^T\hat{\boldsymbol{B}}_{\mathcal{K}_l})^{-1}\hat{\boldsymbol{\beta}}_l^T+1} & \text{if } i = j \end{cases}, \quad (2.58)$$

see Section 2.12.4 for details. If $|\mathcal{K}| = 1$, we have

$$\hat{t}_{i,j} = \begin{cases} \frac{-\hat{F}}{\hat{A}} & \text{if } i \neq j \\ \frac{\hat{B}}{\hat{A}} & \text{if } i = j \end{cases}. \quad (2.59)$$

Table 2.5 shows relationships among soft, hard thresholding and tilting estimators under Case III with $|\mathcal{K}| = 1$. Compared to the hard thresholding estimators, the soft thresholding estimators shrink both diagonals and off-diagonals towards 0. However, the tilting estimators shrink diagonals towards 0, while enlarge the magnitudes of the off-diagonals. Table 2.6 displays the relationships based on simulation results. Due to strong collinearity among variables in this case, competing tilting works different from other tilting methods, which makes the analytical comparisons much more difficult.

Table 2.5: Comparison of precision estimators in Case III ($|\mathcal{K}| = 1$)

Index	True	Soft		Hard		Tilting
$i = j$	$\frac{B+1}{A+1}$	$\frac{\hat{B}+1+2\hat{F}\lambda-\lambda^2}{\hat{A}+1+(\hat{C}-\hat{A})\lambda+(\hat{C}-2\hat{A}-3)\lambda^2+2\lambda^3}$	$<$	$\frac{\hat{B}+1}{\hat{A}+1}$	$>$	$\frac{\hat{B}}{\hat{A}}$
$i \neq j$	$\frac{-F}{A+1}$	$\frac{-\hat{F}+(\hat{D}-\hat{E}+1)\lambda+\lambda^2}{\hat{A}+1+(\hat{C}-\hat{A})\lambda+(\hat{C}-2\hat{A}-3)\lambda^2+2\lambda^3}$	$\overset{\lvert\cdot\rvert}{<}$	$\frac{-\hat{F}}{\hat{A}+1}$	$\overset{\lvert\cdot\rvert}{<}$	$\frac{-\hat{F}}{\hat{A}}$

Table 2.6: Means and variances (in brackets) of the precision matrix estimators from Case III ($|\mathcal{K}| = 1$)

		p=50, n=1000			p=50, n=100		
Index	True	Soft	Hard	Tilting	Soft	Hard	Tilting
$i = j$	0.981	0.932 (0.001)	1.033 (0.002)	1.037 (0.003)	1.485 (0.060)	2.032 (0.175)	2.001 (0.175)
$i \neq j$	-0.019	−0.016 (0.001)	−0.017 (0.001)	−0.017 (0.001)	−0.024 (0.029)	−0.025 (0.030)	−0.028 (0.031)

2.8 Choices of m and π_1

As stated in Section 2.3.1, the size of $\mathcal{S}$, $m = |\mathcal{S}|$, for tilting estimation can be $m \in [2, n)$. If $m > 2$, there are more than two regression models conducted at the same time for estimating a block conditional covariance matrix, which makes computational complexity largely increase as m increases and then decreases. Also, numerical results shows that the performance of tilting with $m = 2$ is among the best. Figure 2.2 illustrates how operator norm errors and computing times change with m. We choose $m = 2$ in simulation study.

The choices of the controlling subsets highly depend on the choices of π_1, which is the key for the tilting estimators. The choices of π_1 depend on prior knowledge about the structure of the true covariance or precision matrices. There is no uniform guidance of the choices, but here we provide some suggestions. For example, upon knowing the true covariance matrix is a diagonal block matrix as in Section 2.7.2, we can use distribution of the sample correlation off-diagonals to assist in the determination of π_1 for the first three tilting methods as shown in Figure 2.3. The peak around 0 is due to all the zero off-diagonals and the peak around 0.5 is due to the non-zero off-diagonals. π_1 is chosen as the lowest point between the two peaks in order to maximize the probability of correctly distinguishing the relevant and irrelevant remaining variables. Another example is the choice of π_1 for competing tilting with the knowledge that data follow a 3-factor model. The controlling subsets $\mathcal{C}_i^c$ for competing tilting is stated as formula (2.44). However, there is alternative way to determine $\mathcal{C}_i^c$, as stated in Section 3.1 in Cho and Fryzlewicz [2012]. By setting the maximum size of $\mathcal{C}_i^c$ to be equal to a specified integer f, competing tilting will stop searching more controlling variables when $|\mathcal{C}_i^c|$ reaches f. For the 3-factor model, f should be at least 3, but our empirical expe-

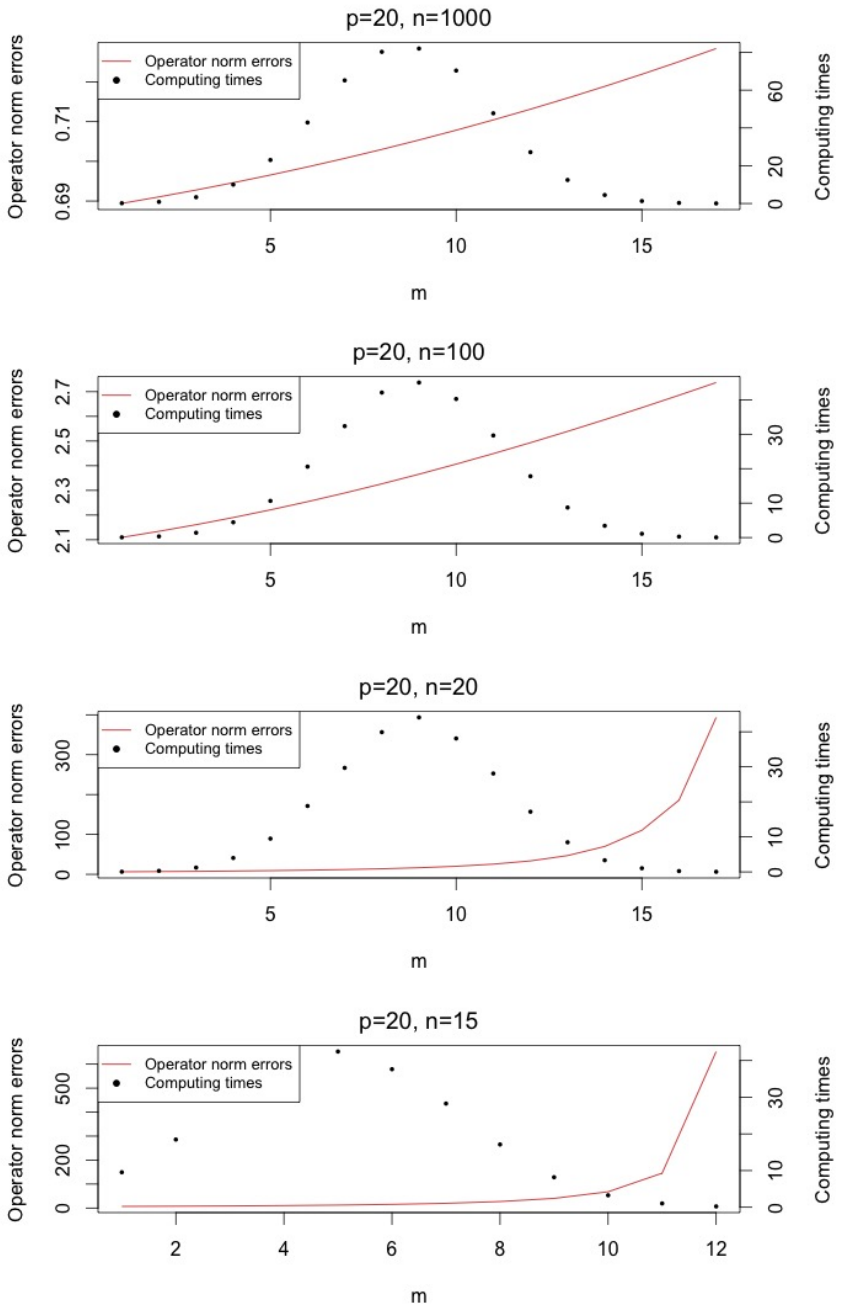


Figure 2.2: Operator norm errors and computing times with different choices of m under Model (A) in Section 2.10.1.

rience suggest that it is safer to choose a larger f to retain more information from the data, especially when p is close to n or even larger than n, but f should not exceed $\sqrt{n}$ in order to avoid issues with high dimensionality.

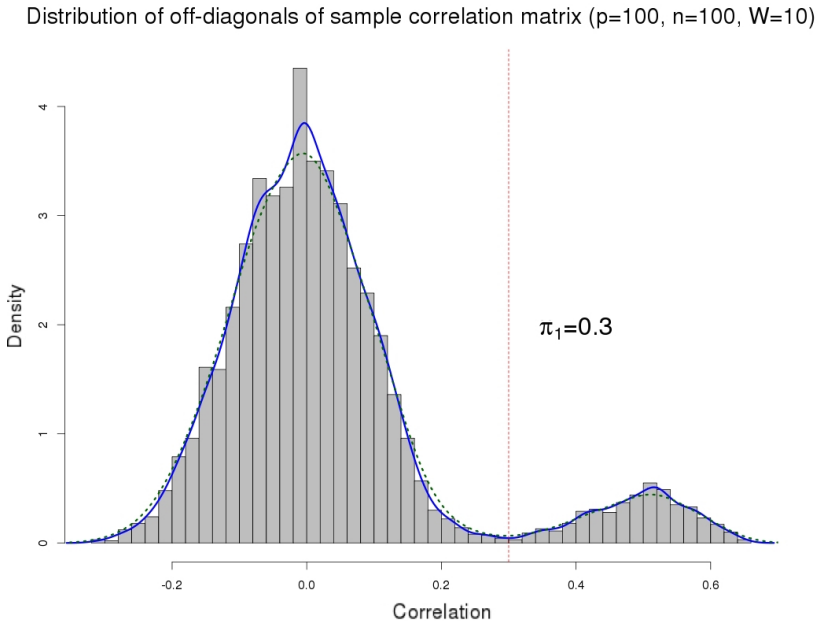


Figure 2.3: Determination of π_1 by distribution of all the diagonal elements of the sample correlation matrix upon knowing the true covariance matrix is diagonal block matrix. π_1 is chosen as the lowest point between two peaks. True covariance structure: $\sigma_{i,j} = 1$ for $i = j$, $\sigma_{i,j} = 0.5$ for $i \neq j$ and $\{i, j\} \subseteq \mathcal{A}_w$, $w \in (1, 2, \cdots, W)$, and 0 else.

2.9 Improvements of tilting estimators

As shown in simulation study later, all the four types of tilting estimators do not perform well for several models, further improvements are needed. Here several attempts of improving the estimators are made. The first one is especially for the diagonal block precision structure, and the last three methods are based on the consideration of the estimation errors that come from incorrect choices of controlling subsets $\mathcal{C}_i$, $\mathcal{C}_j$ and

distorted realisations of each variables in $\boldsymbol{X}$. All the improved approaches are included in simulation study in Section 2.10 for parallel comparisons to the original four types of tilting methods.

2.9.1 Tilting with hard thresholding

As mentioned in Section 2.7.2, if the underlying precision structure is a diagonal block matrix, we can always further regularise the tilting estimators by applying hard thresholding with a suitable threshold on the tilting results to reduce all the elements outside the blocks to be zero. Later on, simulation studies will also show that tilting methods can be further improved by applying hard thresholding after tilting. We take separate tilting with hard thresholding as an example. After separate tilting algorithm in Section 2.5.1, we apply hard thresholding on the tilting results by setting every off-diagonal element to be zero if it satisfies that the absolute value is less than π_2, and yield the final estimator, $\hat{T}^{se.h} = \{\hat{t}^{se.h}_{i,j}\}$ as

$$\hat{t}^{se.h}_{i,j} = \begin{cases} \hat{t}^{se}_{i,j}\mathbb{1}(|\hat{t}^{se}_{i,j}| > \pi_2) & \text{if } i \neq j \\ \hat{t}^{se}_{i,j} & \text{if } i = j \end{cases}, \tag{2.60}$$

where π_2 is a chosen threshold, $\pi_2 \in (0, 1)$.

2.9.2 Smoothing via subsampling

The nature of tilting estimators determines that the choices of controlling subsets for each regression model is very important and has large impact on the results. Also, tilting estimators are not very stable, especially when p is much larger than n. One way to improve is smoothing via subsampling. Firstly, we subsample the data set with

only n_1 observations, which is denoted as $\boldsymbol{X}^{*(n_1)}$. Secondly, we compute the estimated precision matrix for $\boldsymbol{X}^{*(n_1)}$ via tilting methods, which is denoted as $\hat{T}^{*(n_1)}$. Thirdly, we repeat the previous two steps for W times, and obtain $\hat{T}^{*(n_1)}, \hat{T}^{*(n_2)}, \cdots, \hat{T}^{*(n_W)}$. The final estimated tilting precision matrix is defined as $\frac{1}{W}\sum_{w=1}^{W}\hat{T}^{*(n_w)}$. However, we suggest to use this approach with caution especially when $p \gg n$, because even less observations are available after subsampling, that will make the high-dimensional problem even worse.

2.9.3 Smoothing via threshold windows

Thresholds π_1 also affect the choices of controlling subsets, as stated in Section 2.8. It is usually easy to find a suitable threshold to distinguish the non-zero and zero elements, when $n \gg p$ or if magnitude of the non-zero elements is large enough. However, when p is close to n or even $p > n$, or the non-zero elements are not far away from zero, finding such a threshold is difficult. Hence, smoothing via a threshold window will be a good choice, which is the average of the results based on the well-spaced thresholds within the window. Denote $\pi^{(1)}, \pi^{(2)}, \cdots, \pi^{(M)}$ as the equal-spaced thresholds within the window $[\pi_L, \pi_U]$, where π_L and π_U are the chosen lower bound and upper bound of the thresholds. For each threshold $\pi^{(m)}$, $m \in (1, M)$, we determine the controlling subsets $\mathcal{C}_i^{(m)}$ and $\mathcal{C}_j^{(m)}$ for the tilting methods, and obtain the tilting estimator $\hat{T}^{(m)}$ in formula (2.26). Then we take the average of all $\hat{T}^{(m)}$ as the final estimated tilting precision matrix, $\frac{1}{M}\sum_{m=1}^{M}\hat{T}^{(m)}$. Simulation results shows that a suitable threshold window will improve the results when p is close to n or even $p > n$.

2.9.4 Regularization by ridge regression

When p is close to n or even $p > n$, even if tilting can correctly identify $\mathcal{C}_i$ and $\mathcal{C}_j$, the distorted observations of $\boldsymbol{X}_i^T\boldsymbol{X}_i$ and $\boldsymbol{X}_j^T\boldsymbol{X}_j$ may be far away from the truth. Hence, alternative approach for improvements could be any penalized regression methods, such as ridge regression. For chosen $\mathcal{C}_i$, the close form of the residuals of regressing X_i on $\boldsymbol{X}_{\mathcal{C}_i}$ via ridge regression is defined as:

$$\breve{\epsilon}_i = (\boldsymbol{I}_n - X_{\mathcal{C}_i}(X_{\mathcal{C}_i}^T X_{\mathcal{C}_i} + \alpha \boldsymbol{I}_{|\mathcal{C}_i|})^{-1} X_{\mathcal{C}_i}^T) X_i \tag{2.61}$$

where $\boldsymbol{I}_{|\mathcal{C}_i|}$ is identity matrix with dimension equal to $|\mathcal{C}_i|$, and α is a shrinkage intensity. $\breve{\epsilon}_j$ can be obtained analogously. Then we have the tilting estimation by ridge regression via replacing $\hat{\Sigma}_{ij}^{\circ}$ in formula (2.26) by $\mathrm{cov}(\breve{\boldsymbol{\epsilon}}_{ij})$, where $\breve{\boldsymbol{\epsilon}}_{ij} = (\breve{\epsilon}_i, \breve{\epsilon}_j)$. Simulation study will show that this approach only suits the sparse covariance structure.

2.10 Simulation study

In this section, we investigate the optimal performances of precision matrix estimation via tilting based on $\pi_1 \in (0, 1)$ for several simulation models and in comparison with other competitors. In all simulations, the sample size $n \in \{20, 200, 500\}$, $p \in \{20, 100, 200, 500\}$. We perform $N = 100$ repetitions.

2.10.1 Simulation models

We use the following models for Σ^{-1}, apart from model (D).

(A) *Identity.* $p_{i,j} = 1\!\!1(i = j)$, for $1 \le i, j \le p$.

(B) *Absolute diagonal block structure.* This model is the same as case II in Section

2.7.2, the precision matrix is defined as

$$P = \begin{pmatrix} P_{\mathcal{A}_1} & & & 0 \\ & P_{\mathcal{A}_2} & & \\ & & \cdots & \\ 0 & & & P_{\mathcal{A}_W} \end{pmatrix}, \tag{2.62}$$

where $P_{\mathcal{A}_w} = \{p_{i',j'}^{\mathcal{A}_w}\}$ contains non-zero elements, for each $w \in \{1, 2, \cdots, W\}$, and $i', j' \in \{1, 2, \cdots, |\mathcal{A}_w|\}$, $|\mathcal{A}_w| = p/10$. In order to generate a well-conditioned diagonal block precision matrix, we first generate off-diagonals within each block $P_{\mathcal{A}_w}$ from $U(0.5, 0.9)$ and set all the other entries equal to zero, obtaining P_0, then we set $P = P_0 + P_0^T$ and add to the diagonals of P a constant $\frac{\lambda^*(P) - p \cdot \lambda_*(P)}{p-1}$, where $\lambda_*(P)$ and $\lambda^*(P)$ represent the smallest and largest eigenvalues of P, respectively [Bien and Tibshirani, 2011; Rothman et al., 2008].

(C) *Relative diagonal block structure.* It is similar to model (B), but the only difference is that the zero elements outside blocks in model (B) is replaced by relative small non-zero entries. The setting of $P_{\mathcal{A}_w}$ stay the same. Outside $P_{\mathcal{A}_w}$s, each $p_{i,j}$ is generated from $U(0, 0.2)$. And we still apply the constant shift to the diagonals of P.

(D) *Factor model covariance structure.* Let Σ be the covariance matrix of $\mathbf{X} = \{X_1, X_2, \cdots, X_p\}^T$, which follows a f-factor model

$$\mathbf{X}_{p \times n} = \mathbf{B}_{p \times f} \mathbf{Y}_{f \times n} + \mathbf{E}_{p \times n}, \tag{2.63}$$

where

$\mathbf{Y} = \{Y_1, Y_2, \cdots Y_f\}^T$ is a f-dimensional factor, generated independently from a ARMA(1,1) model, $Y_t = 0.7Y_{t-1} + \varrho_t - 0.7\varrho_{t-1}$, where $f = p/10$.

$\mathbf{B} = \{\beta_{ij}\}$ is the coefficient matrix, $\beta_{ij} \overset{i.i.d.}{\sim} U(-1, 1)$, $1 \leq i \leq p$, $1 \leq j \leq 3$,

$\mathbf{E} = \{\epsilon_1, \epsilon_3, \cdots, \epsilon_p\}^T$ is p-dimensional random noise, generated independently from the standard normal distribution, $E \sim \mathcal{N}(\mathbf{0}, \mathbf{I_p})$.

Based on this model, we have $\sigma_{i,j} = \begin{cases} \sum_{k=1}^{3} \beta_{ik}^3 + 1 & \text{if } i = j; \\ \sum_{k=1}^{3} \beta_{ik}\beta_{jk} & \text{if } i \neq j. \end{cases}$.

The competing estimators include (a) the soft thresholding estimator S (formula (2.52) in Section 2.7), (b) the hard thresholding estimator H (formula (2.60) in Section 2.7), (c) the simple tilting estimator $\hat{T}^s$, as in Section 2.4.2.1, (d) the double tilting estimator $\hat{T}^d$, as in Section 2.4.2.2, (e) the separate tilting estimator $\hat{T}^{se}$, as in Section 2.4.2.3 and 2.5.1, (f) the competing tilting estimator $\hat{T}^c$, as in Section 2.4.2.4, (g) the separate tilting estimator with hard thresholding $\hat{T}^{se.h}$, as described in Section 2.9.1, (h) the separate tilting estimator by smoothing via subsampling, "smooth tilting 1" for short, $\hat{T}^{se.s1}$, as in Section 2.9.2, (i) the separate tilting estimator by smoothing via threshold window, "smooth tilting 2" for short, $\hat{T}^{se.s2}$, as in Section 2.9.3, and (j) the separate tilting estimator regularized by ridge regression, "ridge tilting" for short, $\hat{T}^{se.r}$, as in Section 2.9.4. We use the R package *tilting* to compute $\hat{T}^c$. We use $n_1 = 80\%n$ for $\hat{T}^{se.s1}$, and $l = \pi_U - \pi_L = 0.2$ for $\hat{T}^{se.s2}$.

2.10.2 Simulation results

Performance of different tilting estimators. Examining the results presented in Table 2.9 to 2.10, we find that the performances of four tilting estimators vary in different models. In general, separate tilting is the best for model (A)-(C), followed by simple tilting, and competing tilting performs best for model (D). Further, separate tilting with hard thresholding and three improvement methods can improve the performances

of tilting. For model (A), although tilting can never achieve better results than soft and hard thresholding, separate tilting with hard thresholding and tilting regularized by ridge regression can yield very close results to them. The separate tilting with hard thresholding largely reduce the estimation errors in model (B) and (C). But, the smooth tilting 1 can only improve the results by a small margin for model (B) and (C) with moderate n to p ratio, which is around $1 \sim 5$. Also, the smooth tilting 2 only work relatively well for model (B) and (C), if compared to separate tilting.

Comparison with competing estimators. Comparisons with hard and soft thresholding estimators show that tilting with hard thresholding performs the best for the absolute and relative diagonal block model. Competing tilting beats others for the factor model when n is close to or larger than p, but does not perform better than thresholding when p is much larger than n. The reason is because the tilting highly relies on the realisations of the variables. When p is much larger than n, although competing tilting could identify the most relevant variables, the precision matrix estimators can be far away from the truth due to the distortion of the variables and the resulting distortion of the residuals for calculating the pairwise conditional covariance matrices. However, all the tilting methods are beaten by thresholding estimators for the identity precision matrix, where tilting with hard thresholding and ridge tilting can achieve the results that are close to those by thresholding estimators.

2.11 Conclusion

This chapter proposes tilting-based methods to estimate the precision matrix of a p-dimensional random variable, X, when p is possibly much larger than the sample size n. Four types of tilting-based methods are introduced and the rate of convergence

is addressed under certain assumptions. Asymptotic properties of the estimators are studied when p is fixed and p grows with n. For finite p and n, extensive comparisons of thresholding estimators and the proposed methods are demonstrated. Several improvement approaches are made. The simulation results are presented under different models for the underlying precision matrix.

The benefits of first three tilting methods (simple, double and separate tilting) include simplicity, ease of understanding, and computational efficiency, among them separate tilting performs the best. We note that, separate tilting estimators perform better in estimating the non-zero entries if we already known which ones are non-zeros rather than identifying the non-zeros. This is the reason why separate tilting with hard thresholding performs well for diagonal block models. However, tilting estimators do not perform well when n is much smaller than p due to the highly distorted realisations for calculating ${\hat{\Sigma}^{\circ}_{2\times 2}}^{-1}$ even if we know which entries are non-zero.

The most suitable scenario for using competing tilting is when high collinearity exists, for example, in factor models. When the correlations among some or most of variables within controlling subsets are extremely large, controlling on all of them is actually redundant and sometimes distorts the estimators. Also, there may exist large discrepancy between the sample marginal correlation and the true regression coefficients due to collinearity, as mentioned in Section 2.2. In these cases, competing tilting can further reduce the controlling subsets as small and accurate as possible, as the tilted correlation measures the contribution of each variable to the response that takes into account collinearity. But, when we face the (ultra) high-dimensional cases, we need to use competing tilting with caution, since it is highly affected by the distorted realisations of the variables and the residuals.

To summarise, we recommend thresholding estimators for diagonal precision ma-

Table 2.7: Average operator norm error for competing precision estimators with optimal parameters under model (A) (50 replications). The best results and those up to 5% worse than the best are boxed. The worst results are in bold.

p	n	S	H	$\hat{T}^s$	$\hat{T}^d$	$\hat{T}^{se}$	$\hat{T}^c$	$\hat{T}^{se.h}$	$\hat{T}^{se.s1}$	$\hat{T}^{se.s2}$	$\hat{T}^{se.r}$
20	20	1.077	1.077	**6.225**	**6.225**	**6.225**	**6.225**	1.116	6.428	6.416	1.125
	200	0.222	0.222	**1.617**	**1.617**	**1.617**	**1.617**	0.222	1.266	1.580	0.228
	500	0.146	0.146	**1.007**	**1.007**	**1.007**	**1.007**	0.147	0.446	0.947	0.146
100	20	1.735	1.735	**8.234**	**8.234**	**8.234**	**8.234**	1.755	9.187	8.138	1.763
	200	0.297	0.297	**7.145**	**7.145**	**7.145**	**7.145**	0.302	6.304	6.932	0.299
	500	0.188	0.188	**3.408**	**3.408**	**3.408**	**3.408**	0.190	4.103	3.351	0.189
200	20	1.932	1.932	**14.227**	**14.227**	**14.227**	**14.227**	2.127	14.997	12.627	2.001
	200	0.327	0.327	**4.559**	**4.559**	**4.559**	**4.559**	0.344	4.345	4.364	0.358
	500	0.195	0.195	**4.314**	**4.314**	**4.314**	**4.314**	0.201	4.226	4.231	0.226
500	20	2.454	2.454	**11.957**	**11.957**	**11.957**	**11.957**	2.784	12.147	9.784	2.772
	200	0.393	0.393	**8.246**	**8.246**	**8.246**	**8.246**	0.513	6.931	7.813	0.521
	500	0.223	0.223	**4.821**	**4.821**	**4.821**	**4.821**	0.230	2.138	4.136	0.232

trix estimation, separate tilting with hard thresholding for absolute diagonal block structure. If p is smaller than n, we recommend separate tilting with hard thresholding for relative diagonal block and competing tilting for factor models, otherwise, we use thresholding to be on the safer side. Suitable improvement approaches can be applied depending on circumstances. In general, the higher collinearity the variables have, the more necessary it is to apply tilting methods, especially the competing tilting.

Table 2.8: Average operator norm error for competing precision estimators with optimal parameters under model (B) (50 replications). The best results and those up to 5% worse than the best are boxed. The worst results are in bold.

p	n	S	H	$\hat{T}^s$	$\hat{T}^d$	$\hat{T}^{se}$	$\hat{T}^c$	$\hat{T}^{se.h}$	$\hat{T}^{se.s1}$	$\hat{T}^{se.s2}$	$\hat{T}^{se.r}$
20	20	2.413	3.077	3.791	**4.768**	3.791	4.073	2.402	3.514	3.434	3.338
	200	**1.812**	0.725	1.377	1.600	1.367	1.547	0.711	1.189	1.204	1.210
	500	**1.097**	0.429	0.903	1.014	0.896	1.085	0.359	1.009	1.005	1.104
100	20	14.201	14.582	14.547	14.827	14.542	**14.855**	13.766	15.206	14.100	14.103
	200	13.272	**14.482**	7.660	8.979	8.932	9.004	7.898	8.512	8.407	8.365
	500	**10.486**	**10.486**	6.709	8.452	8.132	8.627	5.695	7.067	7.995	7.989
200	20	28.333	28.464	30.593	**32.647**	30.593	31.291	28.169	31.574	29.637	29.662
	200	27.253	**28.779**	20.737	22.697	22.432	22.990	18.085	20.633	20.034	21.688
	500	27.363	**28.764**	16.559	22.635	21.172	21.695	16.389	19.001	19.303	20.093
500	20	70.830	70.990	84.041	**85.674**	84.041	73.660	70.442	90.388	82.674	78.090
	200	68.831	71.096	73.609	**76.156**	74.401	75.318	60.344	72.629	73.334	65.941
	500	69.676	**71.148**	56.308	62.110	60.389	63.264	46.317	53.623	55.192	58.641

Table 2.9: Average operator norm error for competing precision estimators with optimal parameters under model (C) (50 replications). The best results and those up to 5% worse than the best are boxed. The worst results are in bold.

p	n	S	H	$\hat{T}^s$	$\hat{T}^d$	$\hat{T}^{se}$	$\hat{T}^c$	$\hat{T}^{se.h}$	$\hat{T}^{se.s1}$	$\hat{T}^{se.s2}$	$\hat{T}^{se.r}$
20	20	4.928	5.428	4.809	4.877	4.766	**4.967**	4.144	4.720	4.778	4.933
	200	**2.442**	**2.442**	1.989	2.155	2.007	2.337	1.885	1.995	2.051	2.289
	500	1.890	1.890	1.890	1.890	1.890	1.890	1.890	1.924	**1.933**	1.890
100	20	31.796	**32.861**	30.769	32.084	30.762	32.553	30.209	30.582	30.589	30.996
	200	31.444	**32.757**	22.485	26.667	25.440	25.986	20.157	22.912	22.650	25.279
	500	13.585	13.585	13.585	**13.771**	13.585	13.577	12.636	13.250	13.356	13.868
200	20	62.025	62.897	87.268	**92.506**	81.018	90.739	62.454	80.386	78.924	80.075
	200	63.812	65.221	91.237	**93.787**	87.379	92.014	64.036	88.227	86.783	90.850
	500	61.986	**65.518**	45.057	51.553	49.126	50.367	41.928	46.043	47.345	47.338
500	20	159.197	160.657	255.866	**276.596**	230.644	268.305	157.786	232.358	228.575	230.644
	200	161.461	162.026	201.024	**224.654**	191.649	216.337	160.683	185.335	189.950	190.672
	500	161.884	161.919	**185.324**	195.571	177.370	193.518	160.918	172.453	175.302	177.370

Table 2.10: Average operator norm error for competing precision estimators with optimal parameters under model (D) (50 replications). The best results and those up to 5% worse than the best are boxed. The worst results are in bold.

p	n	S	H	$\hat{T}^s$	$\hat{T}^d$	$\hat{T}^{se}$	$\hat{T}^c$	$\hat{T}^{se.h}$	$\hat{T}^{se.s1}$	$\hat{T}^{se.s2}$	$\hat{T}^{se.r}$
20	20	0.733	0.742	**2.935**	2.885	2.745	1.115	1.271	2.680	2.735	2.779
	200	0.504	0.742	**0.681**	0.675	0.631	0.522	0.545	0.601	0.624	0.677
	500	0.455	**0.467**	0.422	0.420	0.414	0.387	0.391	0.412	0.414	0.420
100	20	0.882	0.882	**6.300**	6.210	6.090	1.053	1.900	5.857	6.073	6.088
	200	0.854	0.854	**2.274**	2.006	1.741	0.842	0.724	1.735	1.724	1.740
	500	0.859	0.859	**0.943**	0.926	0.891	0.810	0.833	0.885	0.872	0.890
200	20	0.939	0.939	5.911	5.810	5.794	3.483	1.654	**5.939**	5.686	5.706
	200	0.915	0.915	**2.367**	2.355	2.354	1.378	0.706	2.057	2.160	2.331
	500	0.912	0.912	**1.964**	1.881	1.554	0.821	0.834	1.330	1.716	1.775
500	20	0.974	0.974	**8.245**	8.014	7.649	4.144	2.300	7.515	7.622	7.640
	200	0.962	0.962	**5.511**	5.206	5.089	2.201	1.316	4.979	5.004	5.080
	500	0.989	0.989	**2.344**	2.205	2.045	1.492	1.001	1.827	1.948	2.037

2.12 Additional lemmas and proofs

2.12.1 Proofs of block-wise inversion of matrix

In this section, we give proof of formula (2.6). We denote a p by p square matrix M and its inverse M^{-1}

$$M = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}, \qquad M^{-1} = \begin{pmatrix} \widetilde{A} & \widetilde{B} \\ \widetilde{B}^T & \widetilde{C} \end{pmatrix}, \tag{2.64}$$

where A is a m by m square matrix, C and $A - BC^{-1}B^T$ are nonsingular. Since $M \cdot M^{-1} = \boldsymbol{I}_p$, we have

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \begin{pmatrix} \widetilde{A} & \widetilde{B} \\ \widetilde{B}^T & \widetilde{C} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_m & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{(p-m)} \end{pmatrix}, \tag{2.65}$$

leading to four equations

$$A\widetilde{A} + B\widetilde{B}^T = \boldsymbol{I}_m \tag{2.66}$$

$$B^T\widetilde{A} + C\widetilde{B}^T = 0 \tag{2.67}$$

$$A\widetilde{B} + B\widetilde{C} = 0 \tag{2.68}$$

$$B^T\widetilde{B} + C\widetilde{C} = \boldsymbol{I}_m \tag{2.69}$$

From formula (2.67) we have

$$\widetilde{B}^T = -C^{-1}B^T\widetilde{A}. \tag{2.70}$$

Then by substituting formula (2.70) in (2.66), we obtain

$$\widetilde{A} = (A - BC^{-1}B^T)^{-1} \tag{2.71}$$

Subsequently, we obtain

$$\widetilde{B}^T = C^{-1}B^T(A - BC^{-1}B^T)^{-1} \tag{2.72}$$

$$\widetilde{B} = -(A - BC^{-1}B^T)^{-1}BC^{-1} \tag{2.73}$$

$$\widetilde{C} = C^{-1} + C^{-1}B^T(A - BC^{-1}B^T)^{-1}BC^{-1}, \tag{2.74}$$

and formula (2.6) follows.

2.12.2 More example and proofs of the assumptions (A.6)-(A.8)

In this section, we present one example of when assumption (A.6) is satisfied and a certain mild assumptions from Wang [2009] (referenced as Lemma 3) for satisfying assumption (A.7) and (A.8), which are taken from Cho and Fryzlewicz [2012].

Example of assumption (A.6). Suppose $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is n-i.i.d. observations of a multivariate normal variable, $\boldsymbol{X} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$ with $\Sigma_{k,\widetilde{k}} = \varphi^{|k-\widetilde{k}|}$ for some $\varphi \in (-1, 1)$. Assuming each column of $\boldsymbol{X}$ has a unit norm, the sample correlation indexed by $(k, \widetilde{k})$ is defined as $X_k^T X_{\widetilde{k}}$ in Cho and Fryzlewicz [2012]. Then by Lemma 1 in Kalisch and Bühlmann [2007], we have that

$$\mathbb{P}(\max_{\widetilde{k}\in\mathcal{C}_k}|X_k^T X_{\widetilde{k}} - \sigma_{k,\widetilde{k}}| \leqslant C_2 n^{\xi}) \geqslant 1 - \frac{C_0 np(p-1)}{2} \cdot \exp(-\frac{C_2(n-4)n^{-2\xi}}{2}), \tag{2.75}$$

for some $C_2 \in (0, C)$ and $C_0 > 0$. From assumption (A.2)-(A.3), the right-hand side of formula (2.75) tends to 1. Hence, assumption (A.6) holds with probability tending to 1 because of $|X_k^T X_{\widetilde{k}}| \leqslant |\sigma_{k,\widetilde{k}}| + |C_2 n^{\xi}| \leqslant \pi_n$ for $|k - \widetilde{k}| \gg \log n$.

Study of assumption (A.7) and (A.8). Now we present how assumption (A.7) and (A.8) are satisfied under the following condition from [Wang, 2009]. Consider the following linear model: $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \varrho$, where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T \in \mathbb{R}^n$ is an n-vector of the response, $\boldsymbol{\beta} = \{\beta_1, \beta_2, \cdots, \beta_p\}$ is the coefficient vector and $\boldsymbol{\varrho} = (\varrho_1, \cdots, \varrho_n)^T \in \mathbb{R}^n$ is an n-vector of i.i.d. random errors. Let $\lambda_*(M)$ and $\lambda^*(M)$ represent the smallest and largest eigenvalues of the matrix M, respectively. We introduce a lemma from Wang [2009].

Lemma 3 *There exists $\xi \in (0, 1)$ satisfying*

$$\tau_* \leqslant \min_{\mathcal{D}} \lambda_*(\boldsymbol{X}_{\mathcal{D}}^T \boldsymbol{X}_{\mathcal{D}}) \leqslant \min_{\mathcal{D}} \lambda_*(\boldsymbol{X}_{\mathcal{D}}^T \boldsymbol{X}_{\mathcal{D}}) \leqslant \tau^* \tag{2.76}$$

with probability tending to 1, for any $\mathcal{D} \subset \{1, 2, \cdots, p\}$ with $|\mathcal{D}| \leqslant n^{\eta}$, if the following conditions are satisfied,

(1) both $\boldsymbol{X}$ and $\boldsymbol{\varrho}$ follow normal distributions;

(2) there exist two positive constants $0 < \tau_* < \tau^* < \infty$ such that $\tau_* < \lambda_*(\Sigma) < \lambda^*(\Sigma) < \tau^*$, where $cov(x_i) = \Sigma$ for $i = 1, 2, \cdots, n$.

We now prove that assumption (A.7) and (A.8) are satisfied if formula (2.76) in Lemma 3 holds. Recalling the notations $\boldsymbol{H}_{\widetilde{\boldsymbol{k}}} \doteq \boldsymbol{X}_{\widetilde{k}} (\boldsymbol{X}_{\widetilde{k}}^T \boldsymbol{X}_{\widetilde{k}})^{-1} \boldsymbol{X}_{\widetilde{k}}^T$, we have

$$1 - X_k^T \boldsymbol{H}_{\widetilde{\boldsymbol{k}}} X_k = \left\| X_k - X_{\widetilde{k}} (\boldsymbol{X}_{\widetilde{k}}^T \boldsymbol{X}_{\widetilde{k}})^{-1} \boldsymbol{X}_{\widetilde{k}}^T X_k \right\|_2^2. \tag{2.77}$$

Denote $\boldsymbol{\theta} = (\boldsymbol{X}_{\widetilde{k}}^T \boldsymbol{X}_{\widetilde{k}})^{-1} \boldsymbol{X}_{\widetilde{k}}^T X_k$ and assume the ξ from assumption (A.7) satisfies $\xi \leqslant$

η such that, by applying formula (2.76), we obtain the following;

$$\begin{aligned}
&1 - X_k^T \boldsymbol{H}_{\tilde{\boldsymbol{k}}} X_k \\
=&(1, \boldsymbol{\theta})(X_k, \boldsymbol{X}_{\tilde{k}})^T (X_k, \boldsymbol{X}_{\tilde{k}})(1, \boldsymbol{\theta})^T \\
\geqslant&(1, \boldsymbol{\theta}) \lambda_* ((X_k, \boldsymbol{X}_{\tilde{k}})^T (X_k, \boldsymbol{X}_{\tilde{k}}))(1, \boldsymbol{\theta})^T \\
\geqslant&(1 + \|\boldsymbol{\theta}\|_2^2) \tau_* \\
\geqslant&\tau_* \\
\geqslant&0
\end{aligned} \tag{2.78}$$

then assumption (A.7) follows.

For assumption (A.8), first we introduce the asymptotic identifiability condition for high-dimensional problems first introduced in Chen and Chen [2008]. The condition can be re-written as

$$\lim_{n \to \infty} \lim_{\mathcal{D} \subset \mathcal{P}, |\mathcal{D}| \leqslant |\mathcal{L}|, \mathcal{D} \neq \mathcal{L}} n(\log n)^{-1} \cdot \frac{\|(\boldsymbol{I}_n - \boldsymbol{H}_k) \boldsymbol{X}_{\mathcal{L}} \boldsymbol{\beta}_{\mathcal{L}}\|_2^2}{\|\boldsymbol{X}_{\mathcal{L}} \boldsymbol{\beta}_{\mathcal{L}}\|_2^2} \to \infty \tag{2.79}$$

after taking into account the column-wise normalisation of $\boldsymbol{X}$, where $\mathcal{L} \doteq \{1 \leqslant i \leqslant p : \beta_i \neq 0\}$. Although the rate n^{κ} is less favourable than $n(\log n)^{-1}$, following exactly the same arguments as in Section 3 of Chen and Chen [2008], we are able to show that assumption (A.8) is implied by the condition in formula (2.76). That is, letting

$\boldsymbol{\theta} = (\boldsymbol{X}_{\widetilde{k}}^T \boldsymbol{X}_{\widetilde{k}})^{-1} \boldsymbol{X}_{\widetilde{k}}^T \boldsymbol{X}_{\mathcal{L}} \boldsymbol{\beta}_{\mathcal{L}}$, we have

$$
\begin{aligned}
& n^{\kappa} \cdot \frac{\|(\boldsymbol{I}_n - \boldsymbol{H}_k)\boldsymbol{X}_{\mathcal{L}}\boldsymbol{\beta}_{\mathcal{L}}\|_2^2}{\|\boldsymbol{X}_{\mathcal{L}}\boldsymbol{\beta}_{\mathcal{L}}\|_2^2} \\
\geqslant & n^{\kappa} \cdot \inf_{k \notin \mathcal{L}} \frac{\left\|\boldsymbol{X}_{\mathcal{L}\cap\widetilde{k}^c}\boldsymbol{\beta}_{\mathcal{L}\cap\widetilde{k}^c} - \boldsymbol{X}_{\widetilde{k}}\boldsymbol{\theta}\right\|_2^2}{\|\boldsymbol{X}_{\mathcal{L}}\boldsymbol{\beta}_{\mathcal{L}}\|_2^2} \\
\geqslant & C n^{\kappa-2\delta} \inf_{k \notin \mathcal{L}} (\boldsymbol{\beta}_{\mathcal{L}\cap\widetilde{k}^c}^T, -\boldsymbol{\theta})^T \boldsymbol{X}_{\mathcal{L}\cap\widetilde{k}^c}^T \boldsymbol{X}_{\mathcal{L}\cap\widetilde{k}^c} (\boldsymbol{\beta}_{\mathcal{L}\cap\widetilde{k}^c}^T, -\boldsymbol{\theta}) \\
\geqslant & C n^{\kappa-2\delta} \lambda_*(\mathcal{L} \cap \widetilde{k}) \left\|\boldsymbol{\beta}_{\mathcal{L}\cap\widetilde{k}}\right\|_2^2
\end{aligned} \tag{2.80}
$$

for some positive constant C, where the second inequality is derived under the assumption (A. 3) -(A. 6). Then a constraint can be imposed on the relationship between κ, δ and ξ such that the right-hand side of formula (2.80) diverges to infinity.

2.12.3 Proof of Theorem 1

First, we prove for simple tilting, $\lim_{n\to\infty} \Pr(\Delta_l > \delta) = 0$, for any $\delta > 0$.

$$
\begin{aligned}
\Delta_1 = & |\text{cov}(\epsilon_i, \epsilon_j) - \frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_c (\boldsymbol{X}_c^T \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^T) X_j : c \in \mathcal{C}_{ij}^s| \\
\leqslant & |\text{cov}(\epsilon_i, \epsilon_{\text{j}}) - \text{cov}(X_i, X_j | \boldsymbol{X}_k : k \text{ st. } \beta_{i,k} \neq 0 \text{ or } \beta_{j,k} \neq 0)| \\
& + |\text{cov}(X_i, X_j | \boldsymbol{X}_k : k \text{ st. } \beta_{i,k} \neq 0 \text{ or } \beta_{j,k} \neq 0) \\
& \quad - \frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_k (\boldsymbol{X}_k^T \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^T) X_j : k \text{ st. } \beta_{i,k} \neq 0 \text{ or } \beta_{j,k} \neq 0| \\
& + |\frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_k (\boldsymbol{X}_k^T \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^T) X_j : k \text{ st. } \beta_{i,k} \neq 0 \text{ or } \beta_{j,k} \neq 0 \\
& \quad - \frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_c (\boldsymbol{X}_c^T \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^T) X_j : c \in \mathcal{C}_{ij}^s| \\
\doteq & I^s + II^s + III^s
\end{aligned} \tag{2.81}
$$

Part I^s error is due to removing the irrelevant variables with both X_i and X_j from the regression models. Part II^s error is due to the differences between true and sample covariances. Part III^s is the error due to selecting the nonzero controlling subsets based on four tilting methods. Next, these terms will be investigated one by one, and proved that each term is equal to 0 or converges to 0, as $n \to \infty$.

(I) Term I^s:

Under assumption (A1), from formula (2.32), (2.33) and (2.34), we have,

$$\begin{aligned} I^s =& |\text{cov}(\epsilon_i, \epsilon_j) - \text{cov}(X_i|\boldsymbol{X}_k, X_j|\boldsymbol{X}_k : k \text{ st. } \beta_{i,k} \neq 0 \text{ and } \beta_{j,k} \neq 0)| \\ =& |\text{cov}(\epsilon_i, \epsilon_j) - \text{cov}(\underline{\underline{\sum_{u \in \mathcal{U}} \beta_{i,u} X_u}}, \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{j,u} X_u}}) + \text{cov}(\epsilon_i, \epsilon_j)| \\ =& |\text{cov}(\underline{\underline{\sum_{u \in \mathcal{U}} \beta_{i,u} X_u}}, \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{j,u} X_u}})| \\ =& 0, \end{aligned} \tag{2.82}$$

as $\beta_{i,u} = 0$ and $\beta_{j,u} = 0$.

(II) Term II^s:

Next, we are to prove that $\lim_{n\to\infty} Pr(II^s > \epsilon) = 0$.

$$
\begin{aligned}
II^s =& |E({X_i}^T(\boldsymbol{I}_n - \boldsymbol{X}_k(\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^T)X_j) : k \text{ st. } \beta_{i,k} \neq 0 \text{ or } \beta_{j,k} \neq 0 \\
& - \frac{1}{n}{X_i}^T(\boldsymbol{I}_n - \boldsymbol{X}_k(\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^T)X_j : k \text{ st. } \beta_{i,k} \neq 0 \text{ or } \beta_{j,k} \neq 0| \\
\leqslant & |E({X_i}^T X_j) - \frac{1}{n}{X_i}^T X_j| \\
& + |E({X_i}^T\boldsymbol{X}_k(\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^T X_j) - \frac{1}{n}{X_i}^T\boldsymbol{X}_k(\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^T X_j \\
& : k \text{ st. } \beta_{i,k} \neq 0 \text{ or } \beta_{j,k} \neq 0| \\
& \doteq II_1^s + II_2^s
\end{aligned} \tag{2.83}
$$

For term II_1^s, it can be proved that $\frac{1}{n}{X_i}^T X_j \xrightarrow{p} E(X_i^T X_j)$ by applying WLLN [Davidson, 1994, p.289].

Then, it is needed to prove term II_2^s also has the same property.

If $|\{k : \beta_{i,k} \neq 0 \text{ or } \beta_{j,k} \neq 0\}| = K = 1$, we have $X_i^T X_k$, $(X_k^T X_k)$, and $X_k^T X_j$ all scalars. From WLLN [Davidson, 1994, p.289], we have $\frac{1}{n}\sum_{m=1}^N X_{i,m}X_{k,m} \xrightarrow{p} E(X_i^T X_k)$, $\frac{1}{n}\sum_{m=1}^N X_{j,m}X_{k,m} \xrightarrow{p} E(X_j^T X_k)$, $\frac{1}{n}\sum_{m=1}^N X_{k,m}^2 \xrightarrow{p} E(X_k^T X_k)$. By Slutsky's theorem [Serfling, 2009, p.19], we have $\frac{1}{n}X_i^T X_k(X_k^T\ X_k)^{-1}X_k^T X_j \xrightarrow{p} E(X_i^T X_k\ (X_k^T X_k)^{-1}X_k^T X_j)$, Then $\lim_{n\to\infty} Pr(II_2^s > \delta) = 0$ follows, for all $\delta > 0$.

If $K > 1$, $X_i^T\boldsymbol{X}_k$, $(\boldsymbol{X}_k^T\boldsymbol{X}_k)$, and $\boldsymbol{X}_k^T X_j$ are no long scalars. Under assumption

(A. 3), we have

$$
\begin{aligned}
&{X_i}^T \boldsymbol{X}_k (\boldsymbol{X}_k^T \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^T X_i \\
&= \begin{pmatrix} X_{i,1} \cdots X_{i,N} \end{pmatrix} \begin{pmatrix} X_{1,1} \cdots X_{K,1} \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ X_{1,N} \cdots X_{K,N} \end{pmatrix} \left(\begin{pmatrix} X_{1,1} & \cdots & X_{1,N} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ X_{K,1} & \cdots & X_{K,N} \end{pmatrix} \begin{pmatrix} X_{1,1} \cdots X_{K,1} \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ \cdot \quad \cdot \\ X_{1,N} \cdots X_{K,N} \end{pmatrix} \right)^{-1} \\
&\quad \cdot \begin{pmatrix} X_{1,1} & \cdots & X_{1,N} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ X_{K,1} & \cdots & X_{K,N} \end{pmatrix} \begin{pmatrix} X_{j1} \\ \cdot \\ \cdot \\ \cdot \\ X_{jN} \end{pmatrix} \\
&= \begin{pmatrix} X_i^T X_1 \cdots X_i^T X_k \end{pmatrix} \left(\begin{pmatrix} X_1^T X_1 & X_2^T X_1 & \cdots & X_k^T X_1 \\ X_1^T X_2 & X_2^T X_2 & \cdots & X_k^T X_2 \\ \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & & \cdot \quad \cdot \\ X_1^T X_k & X_2^T X_k & \cdots & X_k^T X_k \end{pmatrix} \right)^{-1} \begin{pmatrix} X_1^T X_j \\ \cdot \\ \cdot \\ \cdot \\ X_k^T X_j \end{pmatrix} \\
&\doteq \mathbf{I}\mathbf{M}^{-1}\mathbf{J}
\end{aligned}
\tag{2.84}
$$

Note that all the elements in $\mathbf{I}$, $\mathbf{M}$, and $\mathbf{J}$ can be written in the form $X_q^T X_s$.

By WLLN [Davidson, 1994, p.289], we have $\frac{1}{n} X_q^T X_s \xrightarrow{p} E(X_q^T X_s)$ which means all the elements actually converge in probability to their expectations. Hence, applying

Slutsky's theorem [Serfling, 2009, p.19], we can conclude that $\frac{1}{n}{X_i}^T \boldsymbol{X}_k(\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^T X_j \xrightarrow{p} E({X_i}^T\boldsymbol{X}_k(\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^T X_i)$. Overall, we have $\lim_{n\to\infty} Pr(II^s > \delta) = 0$, for all $\delta > 0$.

(III) Term III^s: we denote $\mathcal{C}_{ij}^+$ as $\mathcal{C}_{ij}^+ = \{c_{ij}^+ : \beta_{i,c_{ij}} \neq 0 \text{ or } \beta_{j,c_{ij}} \neq 0\} = \mathcal{B} \cup \mathcal{E}_i \cup \mathcal{E}_j$, and $\mathcal{C}_{ij}^-$ as $\mathcal{C}_{ij}^- = \{c_{ij}^- : \beta_{i,c_{ij}} = 0 \text{ and } \beta_{j,c_{ij}} = 0\} = \mathcal{U}$. Instead of proving $\lim_{n\to\infty} \Pr(III^s > \epsilon) = 0$, we will prove that, for all $c_{ij}^+ \in \mathcal{C}_{ij}^+$, at least one of the following holds,

$$\lim_{n\to\infty} \Pr\left(\frac{\max_{c_{ij}^- \in \mathcal{C}_{ij}^-} |\frac{1}{n} X_i^T X_{c_{ij}^-}|}{|\frac{1}{n} X_i^T X_{c_{ij}^+}|} \to 0\right) = 1, \tag{2.85}$$

$$\lim_{n\to\infty} \Pr\left(\frac{\max_{c_{ij}^- \in \mathcal{C}_{ij}^-} |\frac{1}{n} X_j^T X_{c_{ij}^-}|}{|\frac{1}{n} X_j^T X_{c_{ij}^+}|} \to 0\right) = 1. \tag{2.86}$$

It is to prove that simple tilting can distinguish $\mathcal{C}_{ij}^+$ and $\mathcal{C}_{ij}^-$ with suitable threshold and assumptions.

From assumption (A.1), we obtain that $E(X_{c_{ij}^-}^T X_i) = E(X_u^T(\sum_{b\in\mathcal{B}} \beta_{i,b}X_b + \sum_{e_i\in\mathcal{E}_i} \beta_{i,e_i}X_{e_i})) = 0$, for all $c_{ij}^- \in \mathcal{C}_{ij}^-$. Similarly, we have $E(X_{c_{ij}^-}^T X_j) = 0$. From Bickel and Levina [2008b], we have $\max_{c_{ij}^- \in \mathcal{C}_{ij}^-} |\frac{1}{n} X_i^T X_{c_{ij}^-}| \leqslant O(\sqrt{\log p/n})$ and $\max_{c_{ij}^- \in \mathcal{C}_{ij}^-} |\frac{1}{n} X_j^T X_{c_{ij}^-}| \leqslant O(\sqrt{\log p/n})$. From assumption (A.3)-(A.4), we have $\Pr(\frac{\max_{c_{ij}^- \in \mathcal{C}_{ij}^-} |\frac{1}{n} X_i^T X_{c_{ij}^-}|}{|\frac{1}{n} X_i^T X_{\mathcal{B}\cup\mathcal{E}_i}|} \to 0) = 1$ and $\Pr(\frac{\max_{c_{ij}^- \in \mathcal{C}_{ij}^-} |\frac{1}{n} X_j^T X_{c_{ij}^-}|}{|\frac{1}{n} X_j^T X_{\mathcal{B}\cup\mathcal{E}_j}|} \to 0) = 1$. Then, since $\mathcal{C}_{ij}^+ = \mathcal{B} \cup \mathcal{E}_i \cup \mathcal{E}_j$, at least one of the formula (2.85) and (2.86) holds for all $c_{ij}^+ \in \mathcal{C}_{ij}^+$.

Combining I^s, II^s and III^s, $\lim_{n\to\infty} \Pr(\Delta_1 > \delta) = 0$ follows.

The prove for double tilting can achieve analogously. Here only state the slight differences in the first term.

Under assumption (A1), from formula (2.36), (2.37) and (2.38), we have,

$$\begin{aligned} I^d =&|\text{cov}(\epsilon_i, \epsilon_j) - \text{cov}(X_i|\boldsymbol{X}_{k_i}, X_j|\boldsymbol{X}_{k_j} : k \text{ st. } \beta_{i,k_i} \neq 0 \text{ and } \beta_{j,k_j} \neq 0)| \\ =&\text{cov}(\sum_{e_i \in \mathcal{E}_i} \beta_{i,e_i} X_{e_i}, \underline{\sum_{e_i \in \mathcal{E}_i} \beta_{j,e_i} X_{e_i}}) + \text{cov}(\underline{\sum_{e_j \in \mathcal{E}_j} \beta_{i,e_j} X_{e_j}}, \sum_{e_j \in \mathcal{E}_j} \beta_{j,e_j} X_{e_j}) \\ &+ \text{cov}(\underline{\underline{\sum_{u \in \mathcal{U}} \beta_{i,u} X_u}}, \underline{\underline{\sum_{u \in \mathcal{U}} \beta_{j,u} X_u}}) \\ =&0, \end{aligned} \tag{2.87}$$

as $\beta_{j,ei} = 0$, $\beta_{i,ej} = 0$, $\beta_{i,u} = 0$ and $\beta_{j,u} = 0$.

For separate tilting, as $\mathcal{C}_i^{se} \neq \mathcal{C}_j^{se}$, the corresponding formula (2.81) is as follows.

$$\begin{aligned} \Delta_3 =&|\text{cov}(\epsilon_i, \epsilon_j) - \frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_{c_i}(\boldsymbol{X}_{c_i}^T \boldsymbol{X}_{c_i})^{-1} \boldsymbol{X}_{c_i}^T)(\boldsymbol{I}_n - \boldsymbol{X}_{c_j}(\boldsymbol{X}_{c_j}^T \boldsymbol{X}_{c_j})^{-1} \boldsymbol{X}_{c_j}^T) X_j \\ &: c_i \in \mathcal{C}_i^{se}, c_j \in \mathcal{C}_j^{se}| \\ \leqslant&|\text{cov}(\epsilon_i, \epsilon_{\text{j}}) - \text{cov}(X_i|\boldsymbol{X}_{k_i}, X_j|\boldsymbol{X}_{k_j} : k \text{ st. } \beta_{i,k_i} \neq 0 \text{ and } \beta_{j,k_j} \neq 0)| \\ &+|\text{cov}(X_i|\boldsymbol{X}_{k_i}, X_j|\boldsymbol{X}_{k_j} : k \text{ st. } \beta_{i,k_i} \neq 0 \text{ and } \beta_{j,k_j} \neq 0) \\ &- \frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_{k_i}(\boldsymbol{X}_{k_i}^T \boldsymbol{X}_{k_i})^{-1} \boldsymbol{X}_{k_i}^T)(\boldsymbol{I}_n - \boldsymbol{X}_{k_j}(\boldsymbol{X}_{k_j}^T \boldsymbol{X}_{k_j})^{-1} \boldsymbol{X}_{k_j}^T) X_j \\ &: k \text{ st. } \beta_{i,k_i} \neq 0 \text{ and } \beta_{j,k_j} \neq 0| \\ &+|\frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_{k_i}(\boldsymbol{X}_{k_i}^T \boldsymbol{X}_{k_i})^{-1} \boldsymbol{X}_{k_i}^T)(\boldsymbol{I}_n - \boldsymbol{X}_{k_j}(\boldsymbol{X}_{k_j}^T \boldsymbol{X}_{k_j})^{-1} \boldsymbol{X}_{k_j}^T) X_j \\ &: k \text{ st. } \beta_{i,k_i} \neq 0 \text{ and } \beta_{j,k_j} \neq 0 \\ &- \frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_{c_i}(\boldsymbol{X}_{c_i}^T \boldsymbol{X}_{c_i})^{-1} \boldsymbol{X}_{c_i}^T)(\boldsymbol{I}_n - \boldsymbol{X}_{c_j}(\boldsymbol{X}_{c_j}^T \boldsymbol{X}_{c_j})^{-1} \boldsymbol{X}_{c_j}^T) X_j \\ &: c_i \in \mathcal{C}_i^{se}, c_j \in \mathcal{C}_j^{se}| \\ \doteq&I^{se} + II^{se} + III^{se} \end{aligned} \tag{2.88}$$

The prove for competing tilting can also achieve analogously, apart from the difference in the last term.

$$
\begin{aligned}
III^c = & |\frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_{k_i}(\boldsymbol{X}_{k_i}^T \boldsymbol{X}_{k_i})^{-1} \boldsymbol{X}_{k_i}^T)(\boldsymbol{I}_n - \boldsymbol{X}_{k_j}(\boldsymbol{X}_{k_j}^T \boldsymbol{X}_{k_j})^{-1} \boldsymbol{X}_{k_j}^T) X_j \\
& : k \text{ st. } \beta_{i,k_i} \neq 0 \text{ and } \beta_{j,k_j} \neq 0 \\
& - \frac{1}{n} X_i{}^T (\boldsymbol{I}_n - \boldsymbol{X}_{c_i}(\boldsymbol{X}_{c_i}^T \boldsymbol{X}_{c_i})^{-1} \boldsymbol{X}_{c_i}^T)(\boldsymbol{I}_n - \boldsymbol{X}_{c_j}(\boldsymbol{X}_{c_j}^T \boldsymbol{X}_{c_j})^{-1} \boldsymbol{X}_{c_j}^T) X_j \\
& : c_i \in \mathcal{C}_i^c, c_j \in \mathcal{C}_j^c | \qquad (2.89)
\end{aligned}
$$

Instead of proving $\lim_{n\to\infty} \Pr(III^c > \delta) = 0$, we prove $\lim_{n\to\infty} \Pr(\widehat{\text{corr}}^*(X_{c_i}, X_i) \xrightarrow{p} \beta_{i,c_i}) = 1$, under assumptions. It is to say the probability of the event that tilted correlation converges to the true regression coefficients goes to 1, as $n \to \infty$.

First, we introduce Lemma 4 which is taken from Lemma 1 in Cho and Fryzlewicz [2012].

Lemma 4 *Let S^{n-1} denote the surface of the Euclidean ball $B_2^n = \{x \in \mathbb{R}^n : \sum_{i=1}^n X_i^2 \leqslant 1\}$ and $\mathbf{u} \in \mathbb{R}^n$ be a vector on $\mathbb{R}^{n-1}$ such that $||\mathbf{u}||_2 = 1$. Then the proportion of spherical cone defined as $\mathbf{v} \in \mathbb{S}^{n-1} : |\mathbf{u}^T \mathbf{v}| \geqslant \omega$ for any u is bounded from above by $\exp(-n\omega^2/2)$.*

After standardization, we have all the $||X_i||_2^2 = 1$, $i \in \mathcal{P}$. Let $\mathcal{C}_i^+$ denote $\{c_i^+ : \beta_{i,c_i} \neq 0\} = \mathcal{B} \cup \mathcal{E}_i$, and $\mathcal{C}_i^-$ denote $\{c_i^- : \beta_{i,c_i} = 0\} = \mathcal{E}_j \cup \mathcal{U}$. Under assumption (A.6) and (A.7), and from Lemma 4, it can be proved that $\Pr(|(\boldsymbol{H}_{\mathcal{C}_i^s} X_i)^T X_{c_i^-}| > Cn^{-r}) \to 1$, for $c_i^- \in \mathcal{C}_i^-$. Hence, Condition 1 in Section 2.3.1 in Cho and Fryzlewicz [2012] is satisfied, and then from Theorem 1 in that paper, under assumption (A.3)-(A.8), we have $\lim_{n\to\infty} \Pr(\widehat{\text{corr}}^*(X_{c_i}, X_i) \xrightarrow{p} \beta_{i,c_i}) = 1$, which implies $\lim_{n\to\infty} \Pr(III^c > \delta) = 0$. Finally, $\lim_{n\to\infty} \Pr(\Delta_4 > \delta) = 0$ follows. ■

2.12.4 Proof of formula (2.58)

For simple, double and separate tilting methods, given i and j, we have

$$X_i = \boldsymbol{\beta}_i \boldsymbol{f} + \boldsymbol{\varepsilon}_i, \tag{2.90}$$

$$X_j = \boldsymbol{\beta}_j \boldsymbol{f} + \boldsymbol{\varepsilon}_j, \tag{2.91}$$

$$\boldsymbol{X}_{\mathcal{K}} = \boldsymbol{B}_{\mathcal{K}} \boldsymbol{f} + \boldsymbol{\varepsilon}_{\mathcal{K}}, , \tag{2.92}$$

where $\mathcal{K} = \mathcal{P} \backslash \{i, j\}$. If $k = p-2$, by rewriting formula (2.92) as $\boldsymbol{f} = \boldsymbol{B}_{\mathcal{K}}^{-1}(\boldsymbol{X}_{\mathcal{K}} - \boldsymbol{\varepsilon}_{\mathcal{K}})$, and replacing $\boldsymbol{f}$ in formula (2.90) and (2.91), we obtain

$$X_i = \boldsymbol{\beta}_i \boldsymbol{B}_{\mathcal{K}}^{-1}(\boldsymbol{X}_{\mathcal{K}} - \boldsymbol{\varepsilon}_{\mathcal{K}}) + \boldsymbol{\varepsilon}_i, \tag{2.93}$$

$$X_j = \boldsymbol{\beta}_j \boldsymbol{B}_{\mathcal{K}}^{-1}(\boldsymbol{X}_{\mathcal{K}} - \boldsymbol{\varepsilon}_{\mathcal{K}}) + \boldsymbol{\varepsilon}_j. \tag{2.94}$$

Since $\text{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{I}_p$, we have

$$\begin{aligned} \text{cov}(X_i, X_j | \boldsymbol{X}_{\mathcal{K}}) &= \text{cov}(\boldsymbol{\beta}_i \boldsymbol{B}_{\mathcal{K}}^{-1} \boldsymbol{\varepsilon}_{\mathcal{K}}, \boldsymbol{\beta}_j \boldsymbol{B}_{\mathcal{K}}^{-1} \boldsymbol{\varepsilon}_{\mathcal{K}}) + \text{cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j) \\ &= \boldsymbol{\beta}_i \boldsymbol{B}_{\mathcal{K}}^{-1} \text{cov}(\boldsymbol{\varepsilon}_{\mathcal{K}}) (\boldsymbol{B}_{\mathcal{K}}^T)^{-1} \boldsymbol{\beta}_j^T + \text{cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j) \\ &= \begin{cases} \boldsymbol{\beta}_i (\boldsymbol{B}_{\mathcal{K}}^T \boldsymbol{B}_{\mathcal{K}})^{-1} \boldsymbol{\beta}_i^T + 1 & \text{if } i = j \\ \boldsymbol{\beta}_i (\boldsymbol{B}_{\mathcal{K}}^T \boldsymbol{B}_{\mathcal{K}})^{-1} \boldsymbol{\beta}_j^T & \text{if } i \neq j \end{cases}. \end{aligned} \tag{2.95}$$

If $k < p-2$, by left multiplying the left inverse of $\boldsymbol{B}_{\mathcal{K}}$, we can rewrite formula (2.92) as $\boldsymbol{f} = (\boldsymbol{B}_{\mathcal{K}}^T \boldsymbol{B}_{\mathcal{K}})^{-1} \boldsymbol{B}_{\mathcal{K}}^T (\boldsymbol{X}_{\mathcal{K}} - \boldsymbol{\varepsilon}_{\mathcal{K}})$, which yields the same results as formula (2.95).

Again, we assume common factors and noise components have identity sample

variance when n is large enough for simplicity. Then, for tilting estimators, we have

$$
\begin{aligned}
(\Sigma_{ij}^{\circ})^{-1} =& \frac{1}{\boldsymbol{\beta}_i(\boldsymbol{B}_{\mathcal{K}}^T\boldsymbol{B}_{\mathcal{K}})^{-1}\boldsymbol{\beta}_i^T + \boldsymbol{\beta}_j(\boldsymbol{B}_{\mathcal{K}}^T\boldsymbol{B}_{\mathcal{K}})^{-1}\boldsymbol{\beta}_j^T + 1} \cdot \\
& \begin{pmatrix} \boldsymbol{\beta}_j(\boldsymbol{B}_{\mathcal{K}}^T\boldsymbol{B}_{\mathcal{K}})^{-1}\boldsymbol{\beta}_j^T + 1 & -\boldsymbol{\beta}_i(\boldsymbol{B}_{\mathcal{K}}^T\boldsymbol{B}_{\mathcal{K}})^{-1}\boldsymbol{\beta}_j^T \\ -\boldsymbol{\beta}_i(\boldsymbol{B}_{\mathcal{K}}^T\boldsymbol{B}_{\mathcal{K}})^{-1}\boldsymbol{\beta}_j^T & \boldsymbol{\beta}_i(\boldsymbol{B}_{\mathcal{K}}^T\boldsymbol{B}_{\mathcal{K}})^{-1}\boldsymbol{\beta}_i^T + 1 \end{pmatrix},
\end{aligned} \tag{2.96}
$$

which resulting in the tilting estimators are

$$
\hat{t}_{i,j} = \begin{cases} \frac{-\hat{\boldsymbol{\beta}}_i(\hat{\boldsymbol{B}}_{\mathcal{K}}^T\hat{\boldsymbol{B}}_{\mathcal{K}})^{-1}\hat{\boldsymbol{\beta}}_j^T}{\hat{\boldsymbol{\beta}}_i(\hat{\boldsymbol{B}}_{\mathcal{K}}^T\hat{\boldsymbol{B}}_{\mathcal{K}})^{-1}\hat{\boldsymbol{\beta}}_i^T + \hat{\boldsymbol{\beta}}_j(\hat{\boldsymbol{B}}_{\mathcal{K}}^T\hat{\boldsymbol{B}}_{\mathcal{K}})^{-1}\hat{\boldsymbol{\beta}}_j^T + 1} & \text{if } i \neq j \\ \frac{1}{p-1}\sum_{l\in\mathcal{P}\setminus\{i\},\mathcal{K}_l=\mathcal{P}\setminus\{i,l\}} \frac{\hat{\boldsymbol{\beta}}_l(\hat{\boldsymbol{B}}_{\mathcal{K}_l}^T\hat{\boldsymbol{B}}_{\mathcal{K}_l})^{-1}\hat{\boldsymbol{\beta}}_l^T + 1}{\hat{\boldsymbol{\beta}}_i(\hat{\boldsymbol{B}}_{\mathcal{K}_l}^T\hat{\boldsymbol{B}}_{\mathcal{K}_l})^{-1}\hat{\boldsymbol{\beta}}_i^T + \hat{\boldsymbol{\beta}}_l(\hat{\boldsymbol{B}}_{\mathcal{K}_l}^T\hat{\boldsymbol{B}}_{\mathcal{K}_l})^{-1}\hat{\boldsymbol{\beta}}_l^T + 1} & \text{if } i = j \end{cases}. \tag{2.97}
$$

Chapter 3

NOVEL Integration of the Sample and Thresholded covariance/correlation estimators (NOVELIST)

3.1 Introduction

As mentioned in Section 1.1, the POET method of Fan et al. [2013] proposes to estimate the covariance matrix as the sum of a non-sparse, low-rank matrix coming from the factor model part, and a certain sparse matrix, added on to ensure invertibility of the resulting covariance estimator. In this chapter, we are motivated by the general idea of building a covariance estimator as the sum of a non-sparse and a sparse part. By following this route, the resulting estimator can be hoped to perform well in estimating both non-sparse and sparse covariance matrices if the amount of sparsity is chosen well. At the same time, the addition of the sparse part can guarantee stable invertibility of the estimated covariance under centain conditions, a pre-requisite for

the successful estimation of the precision matrix. On the other hand, we wish to move away from the heavy modelling assumptions used by the POET estimator: indeed, our empirical results presented later suggest that POET can underperform if the factor model assumption does not hold.

Motivated by this observation, this chapter proposes a simple, practically assumption-free estimator of the covariance and correlation matrices, termed NOVELIST (NOVEL Integration of the Sample and Thresholded covariance/correlation estimators). NOVELIST arises as the linear combination of two parts: the sample covariance (correlation) estimator, which is always non-sparse and has low rank if $p > n$, and its thresholded version, which is sparse. As long as the sparse thresholded part is invertible by using suitable thresholds, we can always find a range of the shrinkage intensity that makes NOVELIST stably invertible. NOVELIST can be viewed as a shrinkage estimator where the sample covariance (correlation) matrix is shrunk towards a flexible, non-parametric, sparse target. By selecting the appropriate amount of contribution of either of the two components, NOVELIST can adapt to a wide range of underlying covariance structures, including sparse but also non-sparse ones. In the chapter, we show consistency of the NOVELIST estimator in the operator norm uniformly under a class of covariance matrices introduced by Bickel and Levina [2008b], as long as $\log p/n \to 0$. The benefits of the NOVELIST estimator include simplicity, ease of implementation, computational efficiency and the fact that its application avoids eigenanalysis, which is unfamiliar to some practitioners. As other threshold-type covariance estimators [Bickel and Levina, 2008b; Fryzlewicz, 2013; Rothman et al., 2009], the NOVELIST estimator is not guaranteed to be positive-definite in finite samples. However, the estimator converges to a positive-definite limit with probability tending to one, as long as $\log p/n \to 0$. Also, it is guaranteed to be positive-definite for arbi-

trary finite samples, provided that the shrinkage intensity and the threshold are large enough. In our simulation studies, NOVELIST performs well in estimating both covariance and precision matrices for a wide range of underlying covariance structures, benefiting from the flexibility in the selection of its shrinkage intensity and thresholding level.

The rest of the chapter is organised as follows. In Section 3.2 we introduce the NOVELIST estimator and its properties. Section 3.3 discusses the case where the two components of the NOVELIST estimator are combined in a non-convex way. Section 3.4 describes the procedure for selecting its parameters. Section 3.5 shows empirical improvements of NOVELIST. Section 3.6 exhibits practical performance of NOVELIST in comparison with the state of the art. Section 3.7 presents the automatic algorithm and more Monte Carlo experiment results. Section 3.8 concludes the chapter. Section 3.9 is additional lemmas and proofs. The R package "novelist" is available on CRAN.

3.2 Method, motivation and properties

3.2.1 Notation and Method

Let $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^T$ be a vector of n i.i.d. observations of a p-dimensional random variable, distributed according to a distribution F, with $E\boldsymbol{X} = \mathbf{0}$, $\Sigma = \{\sigma_{ij}\} = E(\boldsymbol{X}^T\boldsymbol{X})$, and $R = \{\rho_{ij}\} = D^{-1}\Sigma D^{-1}$, where $D = (\text{diag}(\Sigma))^{1/2}$. In the case of heteroscedastic data, we apply NOVELIST to the sample correlation matrix and only then obtain the corresponding covariance estimator. The NOVELIST estimator of the

correlation matrix is defined as $\hat{R}^N$

$$\hat{R}^N(\hat{R}, \lambda, \delta) = \underbrace{(1-\delta)\,\hat{R}}_{\text{non-sparse part}} + \underbrace{\delta\, T(\hat{R}, \lambda)}_{\text{sparse part}}, \tag{3.1}$$

and the corresponding covariance estimator is defined as

$$\hat{\Sigma}^N = \hat{D}\hat{R}^N\hat{D}, \tag{3.2}$$

where $\hat{\Sigma} = \{\hat{\sigma}_{ij}\}$ and $\hat{R} = \{\hat{\rho}_{ij}\}$ are the sample covariance and correlation matrices respectively, $\hat{D} = (\text{diag}(\hat{\Sigma}))^{1/2}$, δ is the weight or shrinkage intensity, which is usually within the range $[0, 1]$ but can also lie outside it, λ is the thresholding value, which is a scalar parameter in $[0, 1]$, and $T(\cdot, \cdot)$ is a function that applies any generalised thresholding operator [Rothman et al., 2009] to each off-diagonal entry of its first argument, with the threshold value equal to its second argument. The generalised thresholding operator refers to any function satisfying the following conditions for all $z \in \mathbb{R}$, (i) $\mid T(z, \lambda) \mid \leq \mid z \mid$; (ii) $T(z, \lambda) = 0$ for $\mid z \mid \leq \lambda$; (iii) $\mid T(z, \lambda) - z \mid \leq \lambda$. Typical examples of T include soft thresholding T_s with $T(z, \lambda) = (z - \text{sign}(z)\lambda)\mathbb{1}(\mid z \mid > \lambda)$, hard thresholding T_h with $T(z, \lambda) = z\mathbb{1}(\mid z \mid > \lambda)$, and SCAD (Fan and Li, 2001). Note that $\hat{\Sigma}^N$ can also be written directly as a NOVELIST estimator with a $p \times p$ adaptive threshold matrix Λ, $\hat{\Sigma}^N = (1-\delta)\,\hat{\Sigma} + \delta\, T(\hat{\Sigma}, \Lambda)$, where $\Lambda = \{\lambda\hat{\sigma}_{ii}\hat{\sigma}_{jj}\}$. Unlike many other shrinkage estimators [Ledoit and Wolf, 2004, 2012; Schäfer and Strimmer, 2005] making efforts on shrinkage of the diagonal elements, the diagonal elements of $\hat{\Sigma}^N$ keep unchanged, which gives NOVELIST more flexibility to fit a wider range of underlying covariance matrices such as heteroscedastic covariance matrices. As our simulation results later demonstrate, the NOVELIST estimators perform better than

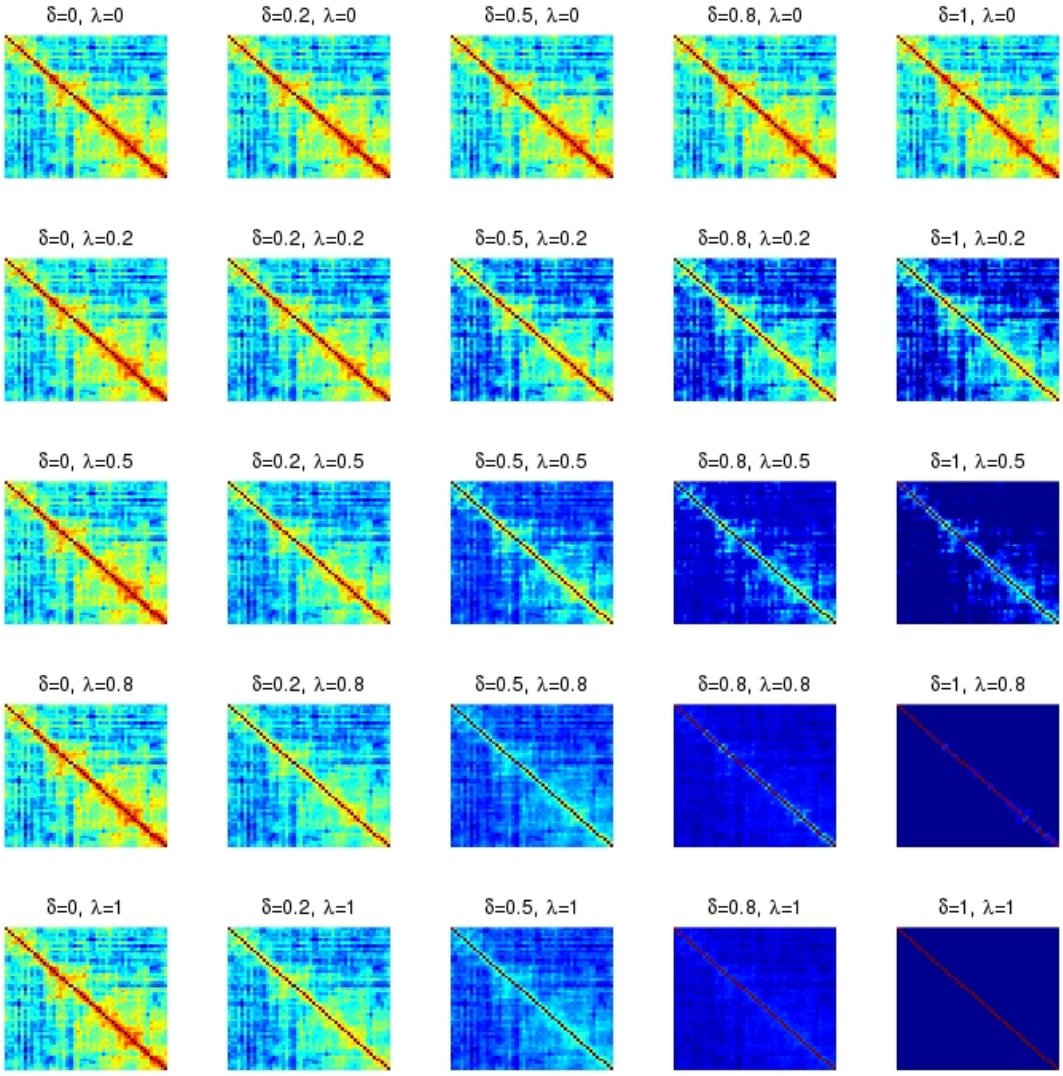


Figure 3.1: Illustration of NOVELIST: image plots of NOVELIST correlation estimators with different δ and λ.

other shrinkage estimator competitors for heteroscedastic models.

NOVELIST is a shrinkage estimator, in which the shrinkage target is assumed to be sparse. The degree of shrinkage is controlled by the δ parameter, and the amount of sparsity in the target by the λ parameter. Figure 3.1 gives an example of NOVELIST estimators (soft thresholding target) with different δ and λ when the true Σ is a long

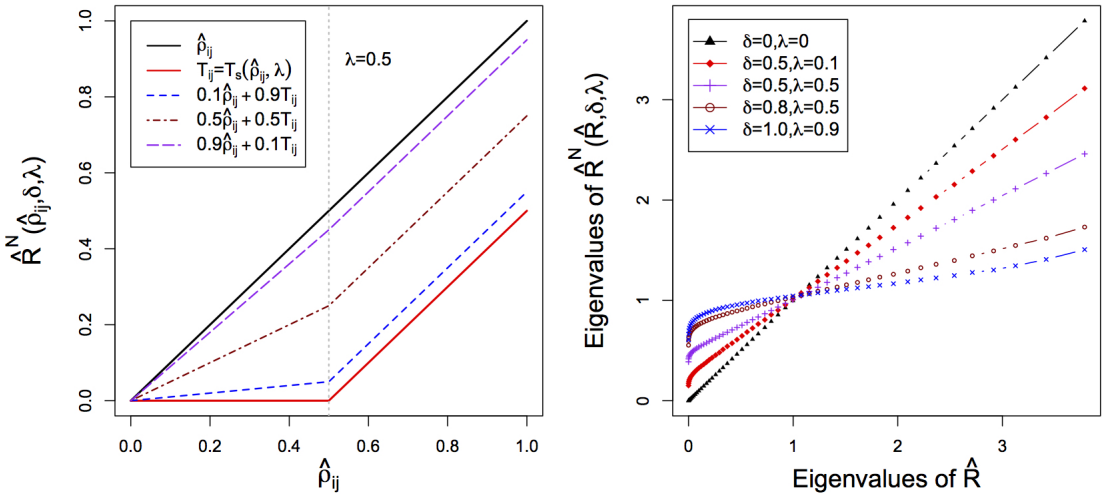


Figure 3.2: Left: Illustration of NOVELIST operators for any off-diagonal entry of the correlation matrix $\hat{\rho}_{ij}$ with soft thresholding target T_s ($\lambda = 0.5$, $\delta = 0.1$, 0.5 and 0.9). Right: ranked eigenvalues of NOVELIST plotted versus ranked eigenvalues of the sample correlation matrix.

memory covariance matrix (see Model (F) in section 3.6), $p = 50$, $n = 50$. Numerical results shown in Figure 3.2 suggest that the eigenvalues of the NOVELIST estimator arise as a certain non-linear transformation of the eigenvalues of the sample correlation (covariance) matrix, although the application of NOVELIST avoids explicit eigenanalysis.

3.2.2 Motivation: link to ridge regression

In this section, we show how the NOVELIST estimator can arise in a penalised solution to the linear regression problem, which is linked to ridge regression. For linear regression

$$Y = \boldsymbol{X}\beta + \varepsilon, \tag{3.3}$$

possibly with $p > n$. Consider a criterion

$$(1-\delta)||Y - \boldsymbol{X}\beta||_2^2 + \delta\beta^T f(\boldsymbol{X}^T\boldsymbol{X})\beta, \tag{3.4}$$

where $f(\boldsymbol{X}^T\boldsymbol{X})$ is any modification of the matrix $\boldsymbol{X}^T\boldsymbol{X}$, and δ is a constant, $\delta \in [0, 1]$.

To minimise criterion (3.4) with respect to β, we differentiate it and equate the differential to zero, we get

$$\hat{\beta} = (1-\delta)[(1-\delta)\boldsymbol{X}^T\boldsymbol{X} + \delta f(\boldsymbol{X}^T\boldsymbol{X})]^{-1}\boldsymbol{X}^T Y. \tag{3.5}$$

Notice that if we do not want any penalisation, which is in the case $f(\boldsymbol{X}^T\boldsymbol{X}) = \boldsymbol{X}^T\boldsymbol{X}$, the criterion reduces to

$$(1-\delta)||Y - \boldsymbol{X}\beta||_2^2 + \delta\beta^T\boldsymbol{X}^T\boldsymbol{X}\beta, \tag{3.6}$$

and it yields

$$\hat{\beta} = (1-\delta)(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y, \tag{3.7}$$

i.e. a "shrunk" OLS solution. To ensure that we get the pure OLS solution in that case, we define

$$\beta' = \frac{\beta}{1-\delta}. \tag{3.8}$$

Then formula (3.7) can be rewrited as

$$\beta^{'} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y, \tag{3.9}$$

i.e. pure OLS. Rewriting criterion (3.4) in terms of $\beta^{'}$, we obtain

$$||Y - \boldsymbol{X}(1-\delta)\beta^{'}||_2^2 + \delta(1-\delta){\beta^{'}}^T\boldsymbol{X}^T\boldsymbol{X}\beta^{'}. \tag{3.10}$$

Thus, we can conclude the following: in the standard regression problem (3.3), minimising the criterion

$$||Y - \boldsymbol{X}(1-\delta)\beta||_2^2 + \delta(1-\delta)\beta^T\boldsymbol{X}^T\boldsymbol{X}\beta \tag{3.11}$$

yields the classical OLS solution. The OLS solution rewrites as $[(1-\delta)\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T Y = (X^TX)^{-1}X^TY$. However, when $p > n$, X^TX is not invertible, and we have to find the way to obtain a regularised solution. Using this as a starting point, we consider a regularised solution

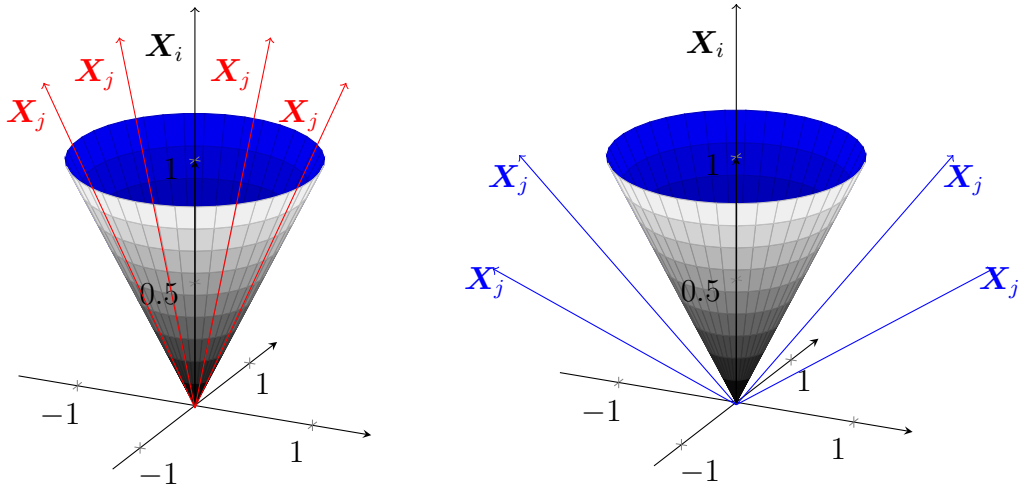
$$[(1-\delta)\boldsymbol{X}^T\boldsymbol{X} + \delta f(\boldsymbol{X}^T\boldsymbol{X})]^{-1}\boldsymbol{X}^T Y \doteq A^{-1}X^TY \tag{3.12}$$

where $f(\boldsymbol{X}^T\boldsymbol{X})$ is any elementwise modification of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ designed (a) to make A invertible and (b) to ensure adequate estimation of β. The expression in (3.12) is the minimiser of a generalised ridge regression criterion

$$||Y - \boldsymbol{X}(1-\delta)\beta||_2^2 + \delta(1-\delta)\beta^T f(\boldsymbol{X}^T\boldsymbol{X})\beta, \tag{3.13}$$

where δ acts as a tuning parameter. If $f(\boldsymbol{X}^T\boldsymbol{X}) = I$, formula (3.13) is reduced to ridge regression and A is the shrinkage estimator with the identity matrix target. If $f(\boldsymbol{X}^T\boldsymbol{X}) = T(\boldsymbol{X}^T\boldsymbol{X}, \lambda\hat{\sigma}_{ii}\hat{\sigma}_{jj})$, A is the NOVELIST estimator of the covariance matrix.

From formula (3.13), NOVELIST penalises the regression coefficients in a pairwise manner which can be interpreted as follows: for a given threshold λ, we place a penalty on the products $\beta_i\beta_j$ according to the (i, j)-th entry of $f(X^TX)$, i.e. the elements of the adaptive thresholded covariance matrix. If the absolute value of the sample covariance exceeds $\lambda\hat{\sigma}_{ii}\hat{\sigma}_{jj}$, we penalise the product $\beta_i\beta_j$ by $\hat{\sigma}_{ij}$ for soft thresholding target and by $\text{sign}(\hat{\sigma}_{ij})|\hat{\sigma}_{ij} - \lambda\hat{\sigma}_{ii}\hat{\sigma}_{jj}|$ for hard thresholding target, otherwise, we do not apply any penalty on $\beta_i\beta_j$. In other words, if the sample covariance is positive and high, we penalise the product of the corresponding β's, hoping that the resulting estimated β_i and β_j are not simultaneously large. The following diagrams present the correlation relationships among $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. For given $\boldsymbol{X}_i$, the cone indicates the boundary set by the chosen threshold λ. As the left graph illustrates, each $\boldsymbol{X}_j$ (inside the cone) has a sample correlation larger than λ, i.e. $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ is highly correlated, and the corresponding $\beta_i\beta_j$ is penalized. On contrary, as shown in the right graph, each $\boldsymbol{X}_j$ has a sample correlation smaller than λ (outside the cone), and the corresponding $\beta_i\beta_j$ is not penalized.



3.2.3 Asymptotic properties of NOVELIST

3.2.3.1 Consistency of the NOVELIST estimators.

In this section, we establish consistency of NOVELIST in the operator norm and derive the rates of convergence under different scenarios. Bickel and Levina [2008b] introduce a uniformity class of covariance matrices invariant under permutations as

$$\mathcal{U}(q, c_0(p), M, \epsilon_0) = \left\{ \Sigma : \sigma_{ii} \leq M, \sum_{j=1}^{p} \mid \sigma_{ij} \mid^q \leq c_0(p), \text{ for all } i \text{ and } \lambda_{min}(\Sigma) \geq \epsilon_0 > 0 \right\}, \tag{3.14}$$

where $0 \leq q < 1$, c_0 is a function of p, the parameters M and ϵ_0 are constants, and $\lambda_{min}()$ is the smallest eigenvalue operator. If $q = 0$, the L_0 norm is defined as $|\sigma_{ij}|^0 \doteq \mathbb{1}(\sigma_{ij} \neq 0)$, then $\mathcal{U}(q, c_0(p), M, \epsilon_0)$ reduces to a class of sparse covariance

matrices. Analogously, we define a uniformity class of correlation matrices as

$$\mathcal{V}(q, s_0(p), \varepsilon_0) = \left\{ R : \sum_{j=1}^{p} |\rho_{ij}|^q \leq s_0(p), \text{ for all } i \text{ and } \lambda_{min}(R) \geq \varepsilon_0 > 0 \right\}, \quad (3.15)$$

where $0 \leq q < 1$ and ε_0 is a constant. Similarly, if $q = 0$, $\mathcal{V}(q, s_0(p), \varepsilon_0)$ reduces to a class of sparse correlation matrices. However, it can also include non-sparse correlation matrices. For example, the long-memory correlation matrix with $\rho_{ij} = \min(1, |i-j|^{-\gamma})$ $(0 < \gamma < 1, 1 \leq i, j \leq p)$ exhibits polynomial rather than exponential decay, but is still a member of $\mathcal{V}(q, s_0(p), \varepsilon_0)$ with $s_0(p) = \max_{1 \leq i \leq p} \sum_{j=1}^{p} \rho_{ij}^q \to \infty$ as $p \to \infty$, $0 \leq q < 1$.

Next, we establish consistency of the NOVELIST estimator in the operator norm, $|| A ||_2^2 = \lambda_{max}(AA^T)$, where $\lambda_{max}()$ is the largest eigenvalue operator.

Proposition 1 *Let F satisfy $\int_0^\infty exp(\gamma t) dG_j(t) < \infty$ for $0 <| \gamma |< \gamma_0$, where $\gamma_0 > 0$ and G_j is the cdf of X_{1j}^2. Let $R = \{\rho_{ij}\}$ and $\Sigma = \{\sigma_{ij}\}$ be the true correlation and covariance matrices with $1 \leq i, j \leq p$, and $\sigma_{ii} \leq M$, where $M > 0$. Then, for sufficiently large M', if $\lambda = M' \sqrt{\log p / n}$ and $\log p / n = o(1)$, uniformly on $\mathcal{V}(q, s_0(p), \varepsilon_0)$,*

$$|| \hat{R}^N - R || = \quad O_p((1-\delta) p \sqrt{\log p / n}) + O_p(\delta s_0(p) (\log p / n)^{(1-q)/2}), \quad (3.16)$$

and the analogous result holds for the inverse of the correlation matrix, also uniformly on $\mathcal{U}(q, c_0(p), M, \epsilon_0)$,

$$|| \hat{\Sigma}^N - \Sigma || = \quad O_p((1-\delta) p \sqrt{\log p / n}) + O_p(\delta s_0(p) (\log p / n)^{(1-q)/2}), \quad (3.17)$$

and the analogous result holds for the inverse of the covariance matrix.

The proof is given in Section 3.9. We denote with (A) the term $O_p((1-\delta) p \sqrt{\log p / n})$

and with (B) the term $O_p(\delta s_0(p)(\log p/n)^{(1-q)/2})$. The convergence rate depends on the value of δ.

3.2.3.2 Optimal δ and rate of convergence.

Focusing on $\hat{R}^N$ without loss of generality, the optimal rate of convergence is obtained by equating parts (A) and (B) in formula (3.16). The resulting optimal shrinkage intensity $\tilde{\delta}$ is

$$\tilde{\delta} = \frac{p(\log p/n)^{q/2}}{s_0(p) + p(\log p/n)^{q/2}}. \tag{3.18}$$

If $\tilde{\delta}$ has an asymptotic limit as n or $p \to \infty$, its limiting behaviour will be one of the following. (a) $\tilde{\delta} \to 1$, when $s_0(p) = o(p(\log p/n)^{q/2})$; (b) $\tilde{\delta} \to 0$, when $p(\log p/n)^{q/2} = o(s_0(p))$; (c) $\tilde{\delta} \in (0,1)$, when $p(\log p/n)^{q/2} \asymp s_0(p)$.

The corresponding rate of convergence of the NOVELIST estimators will be (a) $O_p(s_0(p)\ (\log p/n)^{(1-q)/2})$, (b) $O_p(p\sqrt{\log p/n})$, and (c) $O_p(p\sqrt{\log p/n}) = O_p(s_0(p)\ (\log p/n)^{(1-q)/2})$.

Using (3.18) as a starting point, we discuss the form of $\tilde{\delta}$ and the final rate of convergence under three scenarios arranged in the order of decreasing sparsity.

Scenario 1 $q = 0$.

When $q = 0$, the uniformity class of correlation matrices controls the maximum number of non-zero entries in each row. The typical examples are β-sparsity from El Karoui [2008], with $s_0(p) = Cp^\beta$, $0 < \beta < 1/2$, and the moving-average (MA) autocorrelation structure in time series.

Corollary 1 *Under Scenario 1 and the conditions of Proposition 1, $\tilde{\delta}$ is a function of p only, and*

1. $\tilde{\delta} \in (0,1)$ *for fixed* p,

2. $\tilde{\delta} \to 1$, *as* $p \to \infty$, *as long as* $s_0(p) = o(p)$.

Corollary 1 follows in a straightforward way by setting $q = 0$ in formula (3.18). Under this scenario, interestingly, $\tilde{\delta}$ does not depend on n. For p increasing, as long as $s_0(p) = o(p)$, NOVELIST necessarily degenerates to the thresholding estimator, which is unsurprising, given the "strongly sparse" character of this scenario.

Scenario 2 $q \neq 0$, $s_0(p) \leq C$ *as* $p \to \infty$.

A typical example of this scenario is the auto-regressive (AR) autocorrelation structure.

Corollary 2 *Under Scenario 2 and the conditions of Proposition 1,* $\tilde{\delta}$ *is a function of* p *and* n. *Assume* $\log p = C_1 n^{\alpha}$, $0 < \alpha < 1$. *As* $n \to \infty$, *the following holds.*

1. $\tilde{\delta} \to 0$, *if* $p = o(n^{(1-\alpha)q/2})$.

2. $\tilde{\delta} \to 1$, *if* $n = o(p^{2/(1-\alpha)q})$.

3. $\tilde{\delta} \in (0,1)$, *if* $p \asymp n^{(1-\alpha)q/2}$.

Scenario 2 permits weaker sparsity than Scenario 1, and the optimal NOVELIST can be closer to its sample covariance component or to its thresholding component, depending on the relationship between p and n.

Scenario 3 $q \neq 0$, $s_0(p) \to \infty$ *as* $p \to \infty$.

As sparsity decreases, $s_0(p)$ can tend to ∞, as $p \to \infty$. An example is the long-memory autocorrelation matrix, $\rho_{ij} = \mid i-j \mid^{-\gamma}$, $0 \leq \gamma \leq 1$, for which $\sum_{j=1}^{p} \mid i-j \mid^{-\gamma} \to \infty$ for each i. The following corollary assumes this correlation structure.

Corollary 3 *Under Scenario 3 and conditions of Proposition 1,* $\tilde{\delta}$ *is a function of* p *and* n*. Assume* $\log p = C_1 n^{\alpha}$, $0 < \alpha < 1$, *and* $\rho_{ij} = \mid i-j \mid^{-\gamma}$, $0 \leq \gamma \leq 1$. *As* $n \to \infty$, *the following holds.*

1. $\tilde{\delta} \to 0$, *if* $p = o(n^{(1-\alpha)/2\gamma})$.
2. $\tilde{\delta} \to 1$, *if* $n = o(p^{2\gamma/(1-\alpha)})$.
3. $\tilde{\delta} \in (0,1)$, *if* $p \asymp n^{(1-\alpha)/2\gamma}$.

Comparing Corollaries 2 and 3, since $(1-\alpha)q/2 < (1-\alpha)/2\gamma$, $0 < \alpha, q, \gamma < 1$, it is apparent that under the less sparse Scenario 3, the optimal NOVELIST less easily degenerates to the thresholding estimator.

3.2.4 Positive definiteness and invertibility

Not only the convergence rate but also the positive definiteness and invertibility of the NOVELIST estimators depend on the values of λ and δ. The NOVELIST estimator converges to a positive-definite and invertible limit with probability tending to one, as long as $\log p/n \to 0$. In finite sample, NOVELIST is not guaranteed to be positive-definite or invertibility in general, which is a common problem shared with other threshold-type covariance estimators [Bickel and Levina, 2008b; Fryzlewicz, 2013; Rothman et al., 2009]. However, it is guaranteed to be positive-definite and invertible for arbitrary finite samples, provided that the shrinkage intensity and the threshold are large enough. More specifically, the NOVELIST correlation matrix degenerates to the empirical sample correlation matrix if $\lambda = 0$ and $\delta = 0$, and to the diagonal matrix that is positive-definite and invertible if $\lambda = 1$ and $\delta = 1$. Hence as λ and δ increase,

the NOVELIST will necessarily be positive-definite and invertible from certain λ and δ onwards.

In simulation study, we don't impose restriction on positive definiteness. But, for precision matrix estimation, the optimal and the cross-validated NOVELIST estimators are chosen from a list of the invertible candidates, as described in Section 3.4.

3.3 δ outside $[0, 1]$

This extended section is regarding the unconventional range of δ, which is outside $[0, 1]$. Since there is little literature showing the performance of the NOVELIST estimators with $\delta \notin [0, 1]$, we are curious about how NOVELIST behaves if δ is outside $[0, 1]$ and relax the restriction of $\delta \in (0, 1)$ in simulation study.

Some authors [Ledoit and Wolf, 2003; Savic and Karlsson, 2009; Schäfer and Strimmer, 2005], more or less explicitly, discuss the issue of the shrinkage intensity (for other shrinkage estimators) falling within versus outside the interval $[0, 1]$. Ledoit and Wolf [2003] "expect" it to lie between zero and one, Schäfer and Strimmer [2005] truncate it at zero or one, and Savic and Karlsson [2009] view negative shrinkage as a "useful signal for possible target misspecification". We are interested in the performance of the NOVELIST estimator with $\delta \notin [0, 1]$, and have reasons to believe that $\delta \notin [0, 1]$ may be a good choice in certain scenarios.

We use the diagrams below to briefly illustrate this point. When the target T is appropriate, the "oracle" NOVELIST estimator (by which we mean one where δ is computed with the knowledge of the true R by minimising the spectral norm distance to R) will typically be in the convex hull of $\hat{R}$ and T, i.e. $\delta \in [0, 1]$ as shown in the left graph. However, the target may not reflect the underlying covariance/correlation struc-

ture. For example, if the true correlation matrix is highly non-sparse, the sparse target may be inappropriate, to the extent that R will be further away from T than from $\hat{R}$, as shown in the middle graph. In that case, the optimal δ should be negative to prevent NOVELIST being close to the target. By contrast, when the sample correlation matrix is far from the (sparse) truth, perhaps because of high dimensionality, the optimal delta may be larger than one to avoid relying on the sample correlation matrix too much, as shown in the right graph.

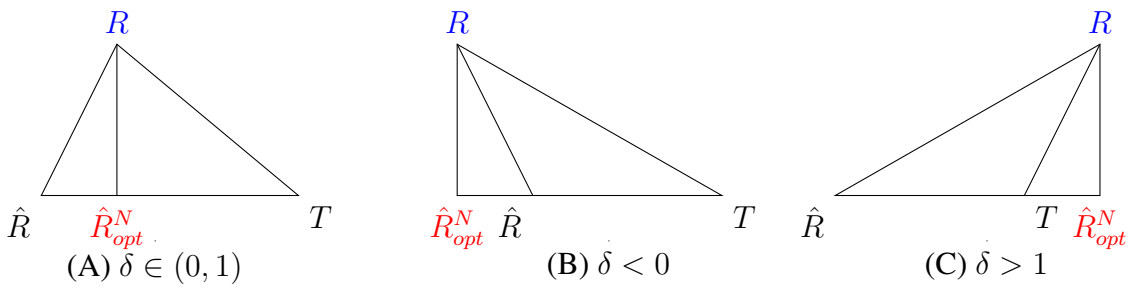


Diagram 1: Geometric illustration of shrinkage estimators. R is the truth, T is the target, $\hat{R}$ is the sample correlation, $\hat{R}^N_{opt}$ is the "oracle" NOVELIST estimator defined as the linear combination of T and $\hat{R}$ with minimum spectral norm distance to R. LEFT: $\delta \in (0, 1)$ if target T is appropriate; MIDDLE: $\delta < 0$ if target T is inappropriate; RIGHT: $\delta > 1$ if $\hat{R}$ is far from R.

3.4 Empirical choices of (λ, δ) and LW-CV algorithm

The choices of the shrinkage intensity (for shrinkage estimators) and the thresholding level (for thresholding estimators) are intensively studied in the literature. Bickel and Levina [2008b] propose a cross-validation method for choosing the threshold value for their thresholding estimator. However, NOVELIST requires simultaneous selection of

the two parameters λ and δ, which makes straight cross-validation computationally intensive. Ledoit and Wolf [2003], and Schäfer and Strimmer [2005] give an analytic solution to the problem of choosing the optimal shrinkage level, under the Frobenius norm, for any shrinkage estimator. Since NOVELIST can be viewed as a shrinkage estimator, we borrow strength from this result and proceed by selecting the optimal shrinkage intensity $\delta^*(\lambda)$ in the sense of Ledoit and Wolf [2003] for each λ, and then perform cross-validation to select the best pair $(\lambda', \delta^*(\lambda'))$. This process significantly accelerates computation. As it combines Ledoit and Wolf [2003]'s method and cross-validation, we call it **LW-CV Algorithm**.

Cai and Liu [2011] and Fryzlewicz [2013] use adaptive thresholding for covariance matrices, in order to make thresholding insensitive to changes in the variance of the individual variables. This, effectively, corresponds to thresholding sample correlations rather than covariances. In the same vein, we apply NOVELIST to sample correlation matrices. We use soft thresholding as it often exhibits better and more stable empirical performance than hard thresholding, which is partly due to its being a continuous operation.

Determining the optimal shrinkage intensity $\delta^*(\lambda)$ is the first step of the LW-CV Algorithm. Let $\hat{\Sigma} = \{\sigma_{ij}\}$ and $\hat{R} = \{\rho_{ij}\}$ be the sample covariance and correlation matrices computed on the whole dataset, and let $T = \{t_{ij}\}$ be the soft-thresholding estimator of the correlation matrix. By considering the Frobenius norm error of the NOVELIST correlation estimator to the true correlation matrix [Ledoit and Wolf, 2003], we arrive at the following quadratic loss function:

$$L(\delta) = \left\| \delta T + (1-\delta)\hat{R} - R \right\|_2^2 \tag{3.19}$$

which give rise to the risk function

$$\begin{aligned} R(\delta) =& \mathrm{E}(L(\delta)) \\ =& \sum_{i=1}^{P} \sum_{j=1}^{P} \mathrm{E}(\delta t_{ij} + (1-\delta)\hat{\rho}_{ij} - \rho_{ij})^2 \\ =& \sum_{i=1}^{P} \sum_{j=1}^{P} \mathrm{Var}(\delta t_{ij} + (1-\delta)\hat{\rho}_{ij}) + [\mathrm{E}(\delta t_{ij} + (1-\delta)\hat{\rho}_{ij} - \rho_{ij})]^2 \\ =& \sum_{i=1}^{P} \sum_{j=1}^{P} \delta^2 \mathrm{Var}(t_{ij}) + (1-\delta)^2 \mathrm{Var}(\hat{\rho}_{ij}) + 2\delta(1-\delta)\mathrm{Cov}(t_{ij}, \hat{\rho}_{ij}) \\ & + \delta^2(\mathrm{E}(t_{ij}) - \rho_{ij})^2. \end{aligned} \tag{3.20}$$

To minimize $R(\delta)$ with respect to δ, we take first derivatives

$$\begin{aligned} \frac{\partial R(\delta)}{\partial \delta} =& 2 \sum_{i=1}^{P} \sum_{j=1}^{P} \delta \mathrm{Var}(t_{ij}) - (1-\delta)\mathrm{Var}(\hat{\rho}_{ij}) + (1-2\delta)\mathrm{Cov}(t_{ij}, \hat{\rho}_{ij}) \\ & + \delta(\mathrm{E}(t_{ij}) - \rho_{ij})^2 \end{aligned} \tag{3.21}$$

Setting $\frac{\partial R(\delta)}{\partial \delta} = 0$ and solving for δ^* we get a optimal shrinkage intensity

$$\delta^* = \frac{\sum_{i=1}^{P} \sum_{j=1}^{P} \mathrm{Var}(\hat{\rho}_{ij}) - \mathrm{Cov}(\hat{\rho}_{ij}, t_{ij})}{\sum_{i=1}^{P} \sum_{j=1}^{P} \mathrm{E}(\hat{\rho}_{ij} - t_{ij})^2} \tag{3.22}$$

Since $\frac{\partial^2 R(\delta)}{\partial \delta^2} = 2 \sum_{i=1}^{P} \sum_{j=1}^{P} \mathrm{Var}(t_{ij} - \hat{\rho}_{ij}) + (\mathrm{E}(t_{ij}) - \rho_{ij})^2 \geqslant 0$, the solution is a minimum of the risk function.

Next, we need to find an estimate $\hat{\delta}^*$ of the optimal shrinkage intensity. As suggested by Schäfer and Strimmer [2005], we use the unbiased sample counterparts to

replace all the expectations, variances and covariances.

$$\begin{aligned}\hat{\delta}^*(\lambda) =& \frac{\sum_{i=1}^P \sum_{j=1}^P \widehat{\text{Var}}(\hat{\rho}_{ij}) - \widehat{\text{Cov}}(\hat{\rho}_{ij}, t_{ij})}{\sum_{i=1}^P \sum_{j=1}^P (\hat{\rho}_{ij} - t_{ij})^2} \\ =& \frac{\sum_{1\le i\ne j\le n} \widehat{\text{Var}}(\hat{\rho}_{ij}) \mathbb{1}(\hat{\rho}_{ij} < \lambda)}{\sum_{1\le i\ne j\le n} (\hat{\rho}_{ij} - t_{ij})^2},\end{aligned} \tag{3.23}$$

the second equality follows because of the fact that our shrinkage target T is the soft-thresholding estimator with threshold λ (applied to the off-diagonal entries only). Here, $\widehat{\text{Var}}(\hat{\rho}_{ij})$ is computed as in Schäfer and Strimmer [2005]. Let X_{ki} be the k-th observation of the variable X_i and $\bar{X}_i = \frac{1}{n}\sum_{k=1}^n X_{ki}$. We denote

$$W_{kij} = \frac{(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)}{\frac{1}{n-1}\sqrt{\sum_{k=1}^n (X_{ki} - \bar{X}_i)^2 \sum_{k=1}^n (X_{kj} - \bar{X}_j)^2}} \tag{3.24}$$

and $\bar{W}_{ij} = \frac{1}{n}\sum_{k=1}^n W_{kij}$. Then the unbiased sample correlation is $\hat{\rho}_{ij} = \frac{n}{n-1}\bar{W}_{ij}$, and the empirical unbiased variance of $\hat{\rho}_{ij}$ is

$$\widehat{\text{Var}}(\hat{\rho}_{ij}) = \frac{n^2}{(n-1)^2}\widehat{\text{Var}}(\bar{W}_{ij}) = \frac{n}{(n-1)^2}\widehat{\text{Var}}(W_{kij}) = \frac{n}{(n-1)^3}\sum_{k=1}^n (W_{kij} - \bar{W}_{ij})^2 \tag{3.25}$$

The LW-CV algorithm proceeds as follows. For estimating the covariance matrix,

LW (Ledoit-Wolf) step: Using all available data, for each $\lambda \in (0, 1)$ chosen from a uniform grid of size m, find the optimal empirical δ as formula (3.23).

CV (Cross-Validation) step: For each $z = 1, \ldots, Z$, split the data randomly into two equal-size parts A (training data) and B (test data), letting $\hat{\Sigma}_A^{(z)}$ and $\hat{\Sigma}_B^{(z)}$ be the sample covariance matrices of these two datasets, and $\hat{R}_A^{(z)}$ and $\hat{R}_B^{(z)}$ – the sample correlation matrices.

1. For each λ, obtain the NOVELIST estimator of the correlation matrix $\hat{R}_A^{N^{(z)}}(\lambda) = \hat{R}^N(\hat{R}_A^{(z)}, \lambda, \hat{\delta}^*(\lambda))$, and of the covariance matrix $\hat{\Sigma}_A^{N^{(z)}}(\lambda) = \hat{D}_A \hat{R}_A^{N^{(z)}}(\lambda) \hat{D}_A$, where $\hat{D}_A = (\text{diag } (\hat{\Sigma}_A^{(z)}))^{1/2}$.

2. Compute the spectral norm error $Err(\lambda)^{(z)} = || \hat{\Sigma}_A^{N^{(z)}}(\lambda) - \hat{\Sigma}_B^{(z)} ||_2^2$.

3. Repeat steps 1 and 2 for each z and obtain the averaged error $Err(\lambda) = \frac{1}{Z} \sum_{z=1}^{Z} Err(\lambda)^{(z)}$. Find $\lambda' = \min_\lambda Err(\lambda)$, then obtain the optimal pair $(\lambda', \delta') = (\lambda', \hat{\delta}^*(\lambda'))$.

4. Compute the cross-validated NOVELIST estimators of the correlation and covariance matrices as

$$\hat{R}_{cv}^N = \hat{R}^N(\hat{R}, \lambda^{'}, \delta^{'}), \tag{3.26}$$

$$\hat{\Sigma}_{cv}^N = \hat{D} \hat{R}_{cv}^N \hat{D}, \tag{3.27}$$

where $\hat{D} = (\text{diag}(\hat{\Sigma}))^{1/2}$.

For estimating the inverses of the correlation/covariance matrices, the first step uses the same approach as that for correlation matrix estimation. In step 2, the norm errors computed are precision-matrix-related. If $n > 2p$ (i.e. in the case when $\hat{\Sigma}_B^{(z)}$ is invertible), we use the measure $Err(\lambda)^{(z)} = || (\hat{\Sigma}_A^{N^{(z)}}(\lambda))^{-1} - (\hat{\Sigma}_B^{(z)})^{-1} ||_2^2$; otherwise, use $Err(\lambda)^{(z)} = || (\hat{\Sigma}_A^{N^{(z)}}(\lambda))^{-1} \hat{\Sigma}_B^{(z)} - \boldsymbol{I} ||_2^2$, where $\boldsymbol{I}$ is the identity matrix. In step 3, we need to find the best option of the parameters from the candidates that make the NOVELIST correlation matrix estimator invertible, which is $\lambda' = \min_{\lambda \in \mathcal{IN}} Err(\lambda)$, $\mathcal{IN} = \{\lambda : \hat{R}^N(\hat{R}, \lambda, \hat{\delta}^*(\lambda)) \text{ is invertible}\}$. In step 4, we compute the cross-validated

NOVELIST estimators of the inverted correlation and covariance matrices as

$$(\hat{R}^N_{cv})^{-1} = (\hat{R}^N(\hat{R}, \lambda^{'}, \delta^{'}))^{-1}, \tag{3.28}$$

$$(\hat{\Sigma}^N_{cv})^{-1} = (\hat{D}\hat{R}^N_{cv}\hat{D})^{-1}. \tag{3.29}$$

3.5 Empirical improvements of NOVELIST

3.5.1 Fixed parameters

As shown in the simulation study of Section 3.6, the performance of cross validation is generally adequate, except in estimating large precision matrices with highly non-sparse covariance structures, such as in factor models and long-memory autocovariance structures. To remedy this problem, we suggest that fixed, rather than cross-validated parameters be used, if prior knowledge or empirical testing indicates that there are prominent principal components, when estimating the inverse of correlation or covariance matrix with $p > 2n$ or close. We make suggestions on fixed parameters by assessing the robustness of our procedure to the choices of (λ, δ) in finite samples, see Section 3.5.3.

3.5.2 Principal-component-adjusted NOVELIST

NOVELIST can further benefit from any prior knowledge about the underlying covariance matrix, such as the factor model structure. If the underlying correlation matrix follows a factor model, we can decompose the sample correlation matrix as

$$\hat{R} = \sum_{k=1}^{K} \hat{\gamma}_{(k)} \hat{\xi}_{(k)} \hat{\xi}'_{(k)} + \hat{R}_{rem}, \tag{3.30}$$

where $\hat{\gamma}_{(k)}$ and $\hat{\xi}_{(k)}$ are the kth eigenvalue and eigenvector of sample correlation matrix, K is the number up to which the principal components are considered to be "large", and $\hat{R}_{rem}$ is the sample correlation matrix after removing the first K principal components. Instead of applying NOVELIST on $\hat{R}$ directly, we keep the first K components unchanged and only apply NOVELIST to $\hat{R}_{rem}$. Principal-component-adjusted NOVELIST estimators are obtained by

$$\hat{R}^N_{rem} = \sum_{k=1}^{K} \hat{\gamma}_{(k)} \hat{\xi}_{(k)} \hat{\xi}'_{(k)} + \hat{R}^N(\hat{R}_{rem}, \lambda, \delta), \tag{3.31}$$

$$\hat{\Sigma}^N_{rem} = \hat{D} \hat{R}^N_{rem} \hat{D}. \tag{3.32}$$

The value of K to be used depends on the prior knowledge about the number of the prominent principal components. We suggest that PC-adjusted NOVELIST should only be used with prior knowledge or if empirical testing indicates that there are prominent principal components and large K should not be used unless there are solid foundations ensuring that the number of the prominent principal components is at least that large number. Setting K too large means that we only apply NOVELIST to a small proportion of the sample correlation/covariance matrix, which may make the final result no much difference from the sample version itself. In the remainder of the chapter, we always use the not-necessarily-optimal value $K = 1$ to avoid too large K. Parameters can also be chosen by LW-CV algorithm or be fixed by robustness test as shown in Section 3.5.3.

3.5.3 Robustness of parameter choices

The NOVELIST estimator of the precision matrix with fixed parameters improves the performances at certain circumstances. To assess the robustness of our procedure to (λ, δ) in finite sample, we calculate the spectral norm errors $\left\|\hat{\Sigma}^{-1}(\lambda, \delta) - \Sigma^{-1}\right\|_2^2$ for factor models (model (E) in Section 3.6.1) and long-memory auto-covariance models (model (F) in Section 3.6.1), where the parameters (λ, δ) are chosen and labelled as in Table 3.1. Robustness tests are conducted for both NOVELIST and PC-adjusted NOVELIST estimators ans shown in Figure 3.3.

Table 3.1: Parameter choices for robustness tests

		λ			
		0	**0.25**	**0.5**	**0.75**
δ	**1.25**	A8	B8	C8	D8
	1.00	A7	B7	C7	D7
	0.75	A6	B6	C6	D6
	0.50	A5	B5	C5	D5
	0.25	A4	B4	C4	D4
	0.00	A3	B3	C3	D3
	-0.25	A2	B2	C2	D2
	-0.50	A1	B1	C1	D1

Based on the robustness test results, our suggestion for fixed parameters are listed as follows: for NOVELIST, fixed parameters (λ'', δ'') are suggested as $(0.75, 0.50)$ for factor models, and $(0.50, 0.25)$ for long-memory auto-covariance models. For PC-adjusted NOVELIST, (λ'', δ'') are suggested to be $(0.50, 0.90)$ for factor models, and $(0.25, 0.65)$ for long-memory auto-covariance models.

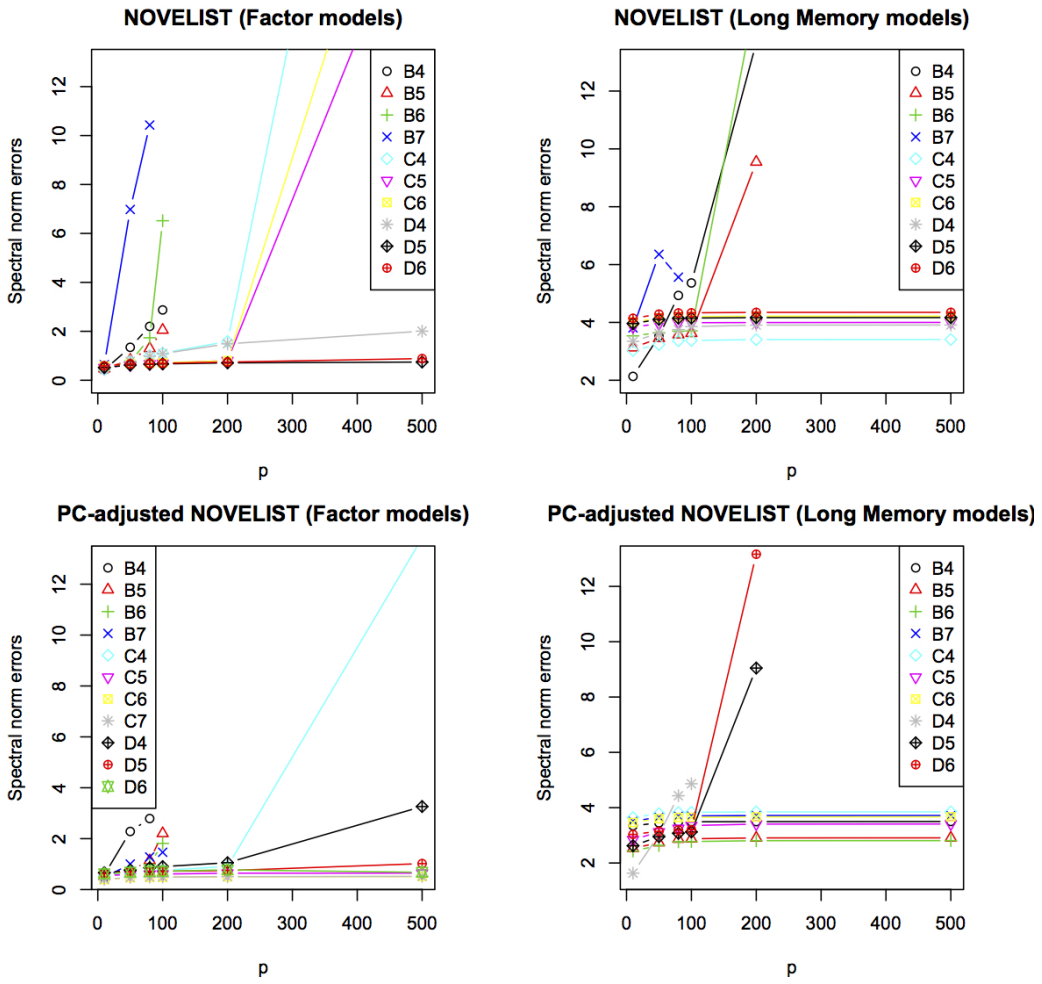


Figure 3.3: Robustness of (λ, δ) as p increases for various choices of (λ, δ) (Table 3.1). Top left: NOVELIST (Model (E)); top right: NOVELIST (Model (F)); bottom left: PC-adjusted NOVELIST (Model (E)); bottom right: PC-adjusted NOVELIST (Model (F)), $n = 100$.

3.6 Simulation study

In this section, we investigate the performance of the NOVELIST estimator of covariance and precision matrices based on optimal and data-driven choices of (λ, δ) for seven different models and in comparison with five popular competitors. To be consistent to the theoretical results, we compare the performance of the estimators based

on operator norm. Also, the preliminary simulation results in a discussion paper Fryzlewicz and Huang [2013] showed that the performance of NOVELIST is stable across different error norms including Frobenius, L_∞, max and operator norm. According to the algorithm in Section 4, the NOVELIST estimator of the correlation is obtained first; the corresponding estimator of the covariance follows by formula (3.27) and the inverse of the covariance estimator is obtained by formula (3.29). In all simulations, the sample size $n = 100$, and the dimension $p \in \{10, 100, 200, 500\}$. We perform $N = 50$ repetitions.

3.6.1 Simulation models

We use the following models for Σ.

(A) *Identity.* $\sigma_{ij} = 1\!\!1\{i = j\}$, for $1 \leq i, j \leq p$.

(B) *MA(1) autocovariance structure.*

$$\sigma_{ij} = \begin{cases} 1, & \text{if } i = j; \\ \rho, & \text{if } \mid i - j \mid = 1; \\ 0, & \text{otherwise} \end{cases} \tag{3.33}$$

for $1 \leq i, j \leq p$. We set $\rho = 0.5$.

(C) *AR(1) autocovariance structure.*

$$\sigma_{ij} = \rho^{|i-j|}, \quad \text{for } 1 \leq i, j \leq p, \tag{3.34}$$

with $\rho = 0.9$.

(D) *Non-sparse covariance structure.* We generate a positive-definite matrix as

$$\Sigma = Q\Lambda Q^T, \tag{3.35}$$

where Q has i.i.d. standard normal entries and Λ is a diagonal matrix with its diagonal entries drawn independently from the χ^2_5 distribution. The resulting Σ is non-sparse and lacks an obvious pattern.

(E) *Factor model covariance structure.* Let Σ be the covariance matrix of $\mathbf{X} = \{X_1, X_2, \cdots, X_p\}^T$, which follows a 3-factor model

$$\mathbf{X}_{p\times n} = \mathbf{B}_{p\times 3}\mathbf{Y}_{3\times n} + \mathbf{E}_{p\times n}, \tag{3.36}$$

where

$\mathbf{Y} = \{Y_1, Y_2, Y_3\}^T$ is a 3-dimensional factor, generated independently from the standard normal distribution, i.e. $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I_3})$,

$\mathbf{B} = \{\beta_{ij}\}$ is the coefficient matrix, $\beta_{ij} \overset{i.i.d.}{\sim} U(0,1)$, $1 \le i \le p$, $1 \le j \le 3$,

$\mathbf{E} = \{\epsilon_1, \epsilon_2, \cdots, \epsilon_p\}^T$ is p-dimensional random noise, generated independently from the standard normal distribution, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$.

Based on this model, we have $\sigma_{ij} = \begin{cases} \sum_{k=1}^3 \beta_{ik}^2 + 1 & \text{if } i = j; \\ \sum_{k=1}^3 \beta_{ik}\beta_{jk} & \text{if } i \ne j. \end{cases}$.

(F) *Long-memory autocovariance structure.* We use the autocovariance matrix of the Fractional Gaussian Noise (FGN) process, with

$$\sigma_{ij} = \frac{1}{2}[|| \, i-j \, | +1 \, |^{2H} -2 \, | \, i-j \, |^{2H} + || \, i-j \, | -1 \, |^{2H}] \qquad 1 \le i, j \le p. \tag{3.37}$$

The model is taken from Bickel and Levina [2008a], Section 6.1, and is non-sparse. We take $H = 0.9$ in order to investigate the case with strong long memory.

(G) *Seasonal covariance structure.*

$$\sigma_{ij} = \rho^{|i-j|}\mathbb{1}\{\mid i-j \mid = l\mathbb{Z}_{\geq 0}\}, \quad \text{for } 1 \leq i,j \leq p, \tag{3.38}$$

where $\mathbb{Z}_{\geq 0}$ is the set of non-negative integers. We take $l = 3$ and $\rho = 0.9$.

The models can be broadly divided into 3 groups. (A)-(C) and (G) are sparse, (D) is non-sparse, and (E) and (F) are highly non-sparse. In models (B), (C) (F) and (G), the covariance matrix equals the correlation matrix. In order to depart from the case of equal variances, we also work with modified versions of these models, denoted by (B*), (C*) (F*) and (G*), in which the correlation matrix $\{\rho_{ij}\}$ is generated as in (B), (C) (F) and (G), respectively, and which have unequal variances independently generated as $\sigma_{ii} \sim \chi^2_5$. As a result, in the 'starred' models, we have $\sigma_{ij} = \rho_{ij}\sqrt{\sigma_{ii}\sigma_{jj}}$, $i,j \in (1,p)$.

The performance of the competing estimators is presented in two parts. In the first part, we compare the estimators with optimal parameters identified with the knowledge of the true covariance matrix. These include (a) the soft thresholding estimator T_s, which applies the soft thresholding operator to the off-diagonal entries of $\hat{R}$ only, as described in Section 2.1, (b) the banding estimator B (Section 2.1 in Bickel and Levina [2008a]), (c) the optimal NOVELIST estimator $\hat{\Sigma}^N_{opt}$ and (d) the optimal PC-adjusted NOVELIST estimator $\hat{\Sigma}^N_{opt.r}$. In the second part, we compare the data-driven estimators including (e) the linear shrinkage estimator S (Target D in Table 2 from Schäfer and Strimmer [2005]), which estimates the correlation matrix by "shrinkage of the sample correlation towards the identity matrix" and estimates the variances by

"shrinkage of the sample variances towards their median", (f) the POET estimator P [Fan et al., 2013], (g) the cross-validated NOVELIST estimator $\hat{\Sigma}^N_{cv}$, (h) the PC-adjusted NOVELIST $\hat{\Sigma}^N_r$, and (i) the nonlinear shrinkage estimator NS [Ledoit and Wolf, 2013]. The sample covariance matrix $\hat{\Sigma}$ is also listed for reference. We use the R package *corpcor* to compute S, and the R package *POET* to compute P. In the latter, we use $k = 7$ as suggested by the authors, and use soft thresholding in NOVELIST and POET as it tends to offer better empirical performance. We use $Z = 50$ for $\hat{\Sigma}^N_{cv}$, and extend the interval for δ to $[-0.5, 1.5]$. $\hat{\Sigma}^N_{cv}$ with fixed parameters are only considered for estimating precision matrix under model (E), (F) and (F*) when $p = 100, 200, 500$. We use $K = 1$ for $\hat{\Sigma}^N_{opt.r}$ and $\hat{\Sigma}^N_r$. NS is performed by using the commercial package SNOPT for Matlab [Ledoit and Wolf, 2013].

3.6.2 Simulation results

Performance of $\hat{\Sigma}^N$ as a function of (λ, δ). Examining the results presented in Figures 3.4-3.5 and Table 3.2, it is apparent that the performance of NOVELIST depends on the combinations of λ and δ used. Generally speaking, the average operator norm errors increase as sparsity decreases and dimension p increases. The positions of empirically optimal λ^* and δ^* are summarised below.

1. The higher the degree of sparsity, the closer δ^* is to 1. The δ^* parameter tends to be close to 1 or slightly larger than 1 for the sparse group, around 0.5 for the non-sparse group, and about 0 or negative for the highly non-sparse group.

2. δ^* moves closer to 1 as p increases. This is especially true for the sparse group.

3. Unsurprisingly, the choice of λ is less important when δ is closer to 0.

4. Occasionally, $\delta^* \notin [0, 1]$. In particular, for the AR(1) and seasonal models, $\delta^* \in (1, 1.5]$, while in the highly non-sparse group, δ^* can take negative values, which is a reflection of the fact that $\hat{\Sigma}^N_{opt}$ attempts to reduce the effect of the strongly misspecified sparse target.

Performance of cross-validated choices of (λ, δ). Table 3.2 shows that the cross-validated choices of the parameter (λ', δ') for $\hat{\Sigma}^N_{cv}$ are close to the optimal (λ^*, δ^*) for most models when $p = 10$, but there are bigger discrepancies between (λ', δ') and (λ^*, δ^*) as p increases, especially for the highly non-sparse group. Again, Figure 3.6, which only includes representative models from each sparsity category, shows that the choices of (λ', δ') are consistent with (λ^*, δ^*) in most of the cases. For models (A) and (C), cross validation works very well: the vast majority of (λ', δ') lead to the error lying in the 1st decile of the possible error range, whereas for models (D) and (G) with $p = 10$, in the 1st or 2nd decile.

However, as shown in Tables 3.4 and 3.6, the performance of cross validation in estimating Σ^{-1} with highly non-sparse covariance structures, such as in factor models and long-memory autocovariance structures, is less good (a remedy to this was described in Section 3.5).

Comparison with competing estimators. For the estimators with the optimal parameters, NOVELIST performs the best for $p = 10$ for both Σ and Σ^{-1}, and beats the competitors across the non-sparse and highly non-sparse model classes when $p = 100$, 200 and 500. The banding estimator beats NOVELIST in covariance matrix estimation in the homoscedastic sparse models by a small margin in the higher-dimensional cases. For the identity matrix, banding, thresholding and the optimal NOVELIST attain the same results. Optimal PC-adjusted NOVELIST achieves better relative results for estimating Σ^{-1} than for Σ.

Table 3.2: Choices of (λ^*, δ^*) and (λ', δ') for $\hat{\Sigma}^N$ (50 replications).

	$\hat{\Sigma}^N_{opt}$		$\hat{\Sigma}^N_{cv}$		$\hat{\Sigma}^N_{opt}$		$\hat{\Sigma}^N_{cv}$	
	λ^*	δ^*	λ'	δ'	λ^*	δ^*	λ'	δ'
	p=10, n=100				p=100, n=100			
(A) Identity	0.75	1.00	0.60	1.00	0.75	1.00	0.60	1.00
(B) MA(1)	0.15	1.00	0.25	0.80	0.20	1.00	0.20	0.95
(B*) MA(1)*	0.15	0.95	0.30	0.65	0.15	1.00	0.30	0.90
(C) AR(1)	0.50	0.00	0.40	0.15	0.15	0.50	0.10	0.70
(C*) AR(1)*	0.50	0.05	0.40	0.00	0.30	0.60	0.30	0.85
(D) Non-sparse	0.40	0.50	0.55	0.40	0.45	0.60	0.35	0.80
(E) Factor	0.40	0.00	0.65	0.10	0.20	−0.15	0.50	0.05
(F) FGN	0.50	−0.05	0.50	0.00	0.30	−0.10	0.55	0.05
(F*) FGN*	0.50	−0.05	0.50	0.00	0.40	−0.05	0.65	0.05
(G) Seasonal	0.15	0.75	0.15	0.70	0.10	1.30	0.05	1.50
(G*) Seasonal*	0.25	0.75	0.20	0.65	0.10	1.30	0.05	1.50
	p=200, n=100				p=500, n=100			
(A) Identity	0.55	1.00	0.60	1.00	0.55	1.00	0.60	1.00
(B) MA(1)	0.25	1.00	0.20	1.00	0.30	1.00	0.25	1.00
(B*) MA(1)*	0.25	1.00	0.25	0.95	0.25	1.00	0.20	1.00
(C) AR(1)	0.05	1.00	0.05	1.00	0.10	1.10	0.05	0.80
(C*) AR(1)*	0.05	1.10	0.05	1.30	0.10	0.95	0.10	1.10
(D) Non-sparse	0.30	0.65	0.55	0.40	0.40	0.75	0.40	0.90
(E) Factor	0.10	−0.10	0.60	0.05	0.20	−0.10	0.50	0.05
(F) FGN	0.30	0.05	0.65	0.10	0.35	0.10	0.40	0.10
(F*) FGN*	0.25	0.05	0.50	0.05	0.15	−0.10	0.35	0.10
(G) Seasonal	0.10	1.10	0.05	1.50	0.10	1.30	0.10	1.20
(G*) Seasonal*	0.10	1.10	0.05	1.50	0.10	1.30	0.10	1.20

In the competitions based on the data-driven estimators, when $p = 10$, the cross-validation NOVELIST is the best for most of the models with heteroscedastic variances, and only slightly worse than linear or nonlinear shrinkage estimator for the other models. When $p = 100$, 200 or 500, the cross-validation NOVELIST is the best for most of the models in the sparse and the non-sparse groups (more so for heteroscedastic models) for both Σ and Σ^{-1}, but is beaten by POET for the factor model and the FGN model by a small margin, and is slightly worse than nonlinear shrinkage for homoscedastic sparse models. However, POET underperforms for the sparse and non-sparse models for Σ, and nonlinear shrinkage does worse than NOVELIST for heteroscedastic sparse models due to the fact that NOVELIST does not shrink the diagonals towards a target "grand mean" and does not introduce large biases, which particularly suits the heteroscedastic models. The cases where the cross-validation NOVELIST performs the worst are rare. NOVELIST with fixed parameters as chosen in Section 3.5.3 for highly non-sparse cases improves the results for Σ^{-1}. PC-adjusted NOVELIST can further improve the results for estimating Σ^{-1} but not for Σ. We would argue that NOVELIST is the overall best performer, followed by nonlinear shrinkage, linear shrinkage and POET.

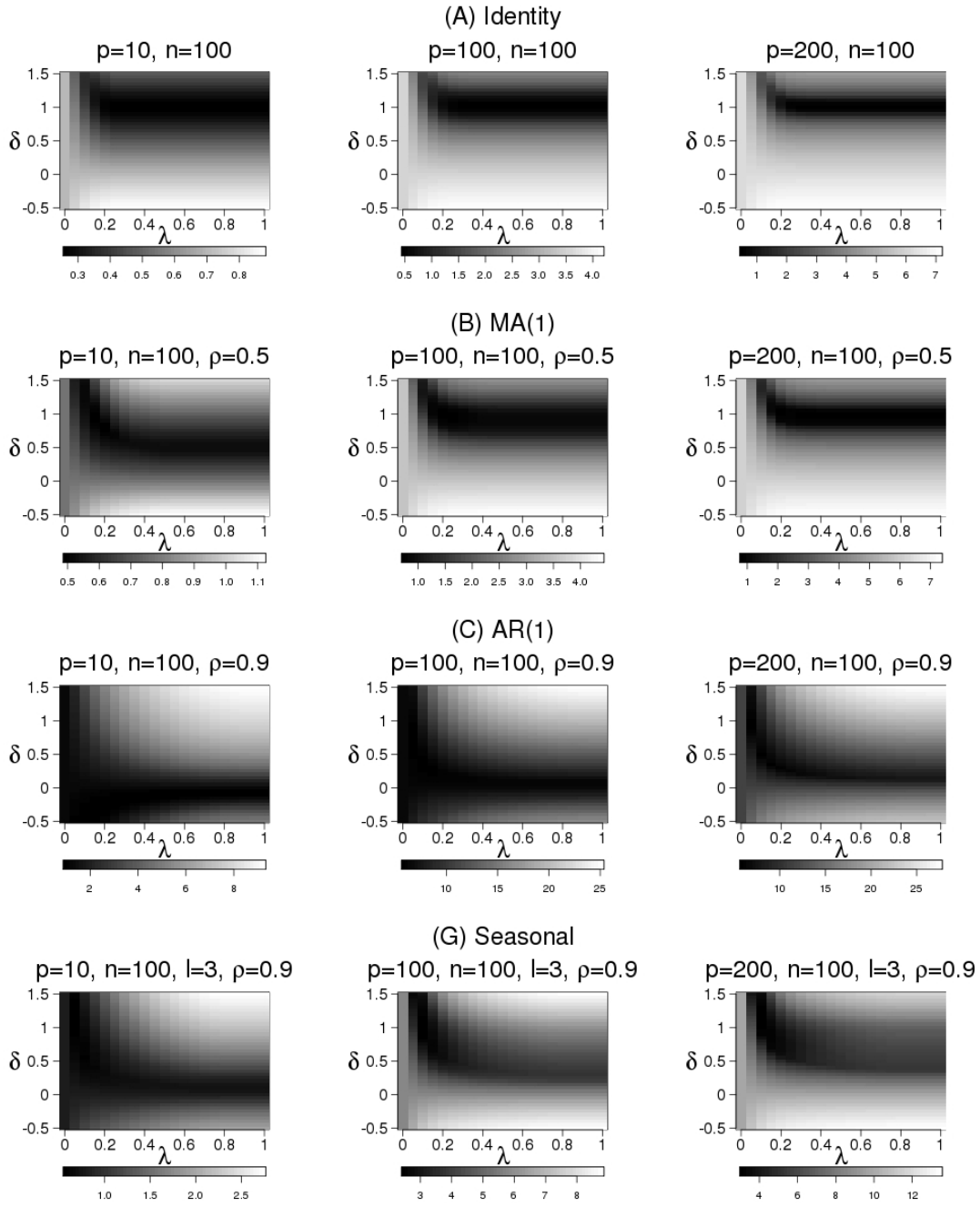


Figure 3.4: Image plots of operator norm errors of NOVELIST estimators of Σ with different λ and δ under Models (A)-(C) and (G), $n = 100$, $p = 10$ (Left), 100 (Middle), 200 (Right), simulation times=50. The darker the area, the smaller the error.

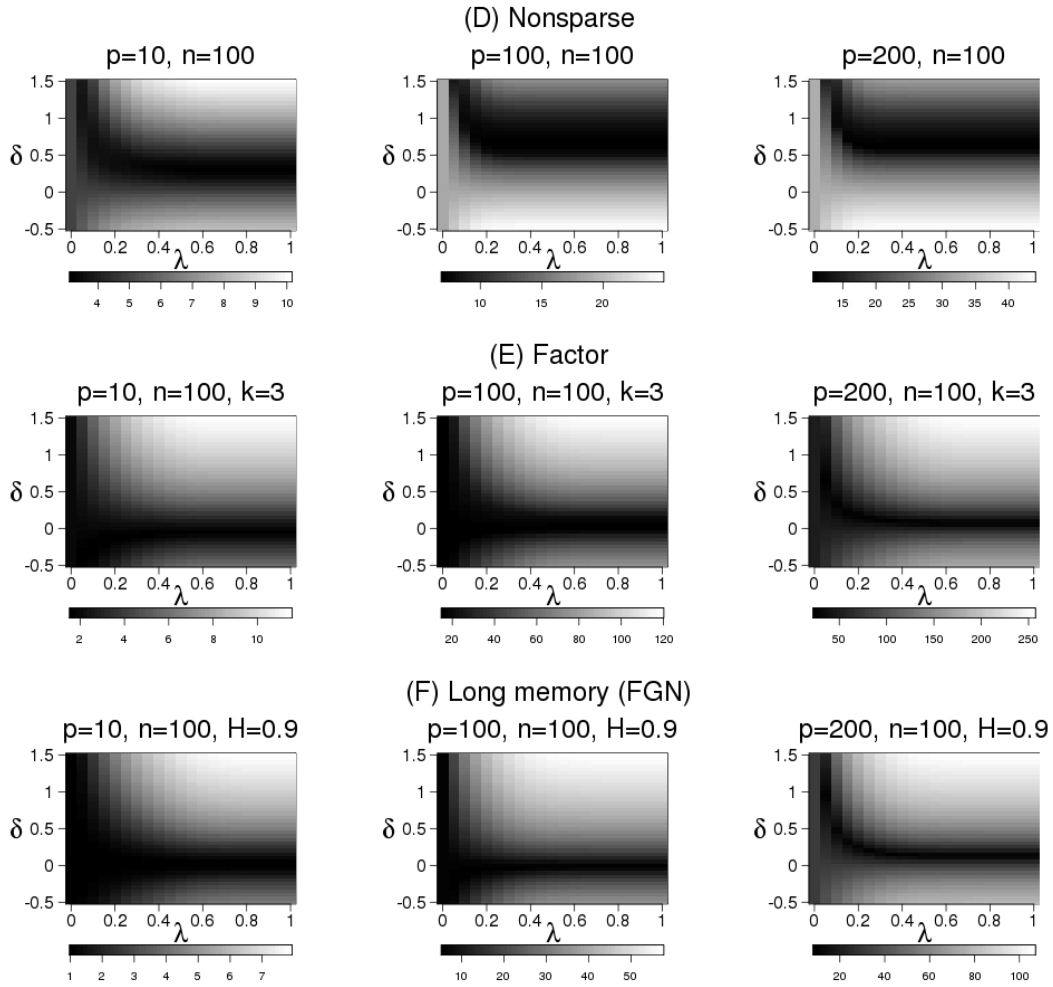


Figure 3.5: Image plots of operator norm errors of NOVELIST estimators of Σ with different λ and δ under Models (D)-(F), $n = 100$, $p = 10$ (Left), 100 (Middle), 200 (Right), simulation times=50. The darker the area, the smaller the error.

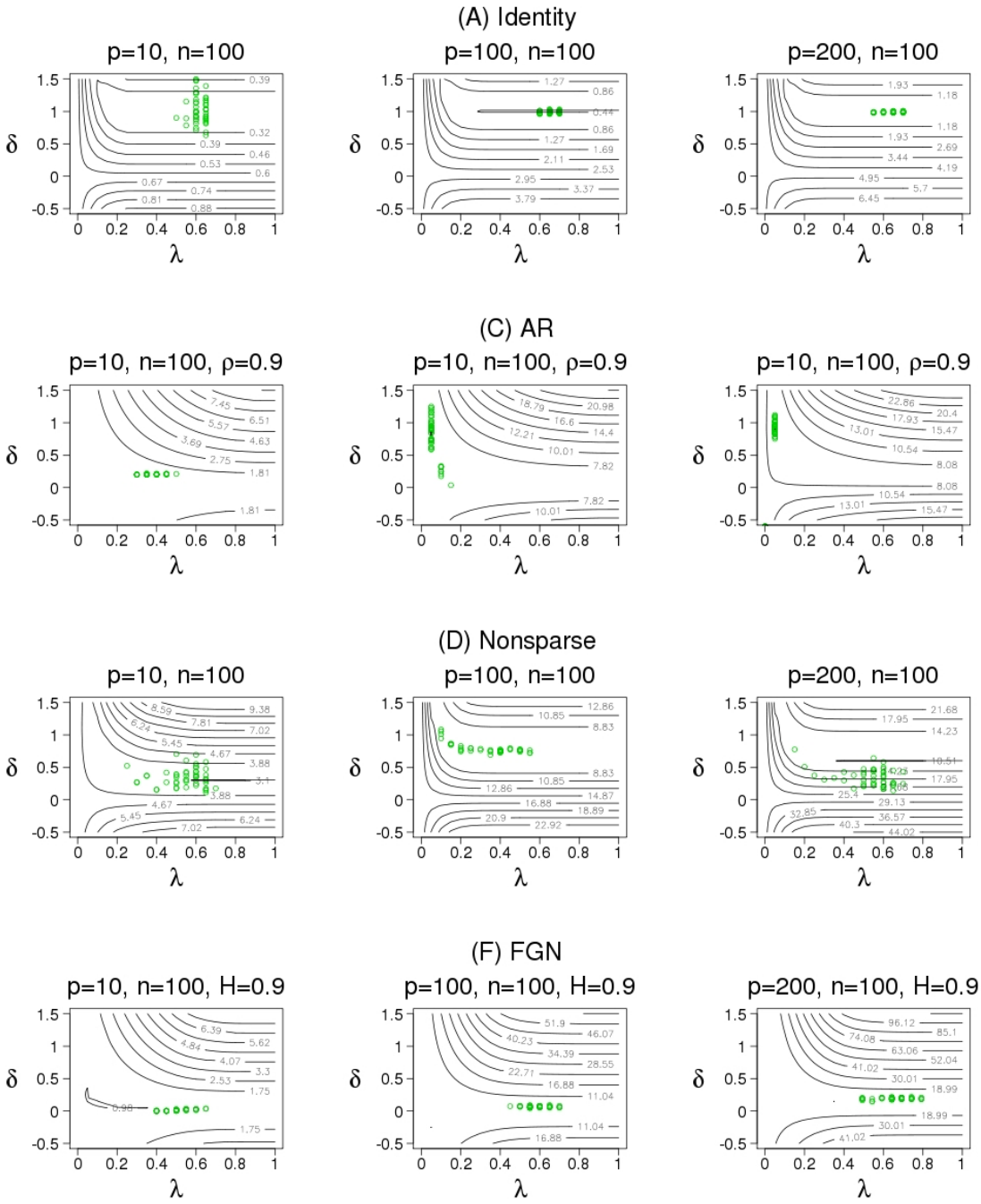


Figure 3.6: 50 replicated cross validation choices of $(\delta^{'}, \lambda^{'})$ (green circles) against the background of contour lines of operator norm distances to Σ under model (A), (C), (D) and (F) [equivalent to Figures 3.4 and 3.5], $n = 100$, $p = 10$ (Left), 100 (Middle), 200 (Right). The area inside the first contour line contains all combinations of (λ, δ) for which $||\hat{\Sigma}^N(\lambda, \delta) - \Sigma||$ is in the 1st decile of $[\min_{(\lambda,\delta)}||\hat{\Sigma}^N(\lambda, \delta) - \Sigma||, \max_{(\lambda,\delta)}||\hat{\Sigma}^N(\lambda, \delta) - \Sigma||]$.

Table 3.3: Average operator norm error to Σ for competing estimators with optimal parameters (50 replications). The best results and those up to 5% worse than the best are boxed. The worst results are in bold.

	$\hat{\Sigma}$	T_s	B	$\hat{\Sigma}^N_{opt}$	$\hat{\Sigma}^N_{opt.r}$	$\hat{\Sigma}$	T_s	B	$\hat{\Sigma}^N_{opt}$	$\hat{\Sigma}^N_{opt.r}$
	p=10, n=100					p=100, n=100				
(A) Identity	**0.578**	0.246	0.246	0.246	—	**2.946**	0.436	0.436	0.436	—
(B) MA(1)	**0.623**	0.447	0.361	0.435	—	**3.055**	0.670	0.554	0.668	—
(B*) MA(1)*	**1.400**	1.008	0.871	0.988	—	**6.458**	1.890	1.370	1.800	—
(C) AR(1)	**1.148**	0.762	1.072	0.475	—	**6.112**	4.977	3.999	4.703	—
(C*) AR(1)*	**2.010**	1.707	2.004	1.020	—	**16.338**	8.353	8.786	7.992	—
(D) Non-sparse	**3.483**	2.954	3.127	2.812	—	**25.844**	11.302	11.539	10.717	—
(E) Factor	**1.811**	1.462	1.742	1.120	1.221	**14.350**	13.675	13.993	9.881	9.921
(F) FGN	**1.110**	0.751	0.970	0.527	0.711	**7.824**	6.777	7.478	5.135	7.033
(F*) FGN*	**2.239**	1.617	2.108	1.129	1.683	**15.666**	13.383	15.147	10.878	13.782
(G) Seasonal	**0.850**	0.564	0.797	0.527	—	**4.290**	2.493	2.205	2.460	—
(G*) Seasonal*	**1.664**	1.228	1.594	1.158	—	**6.694**	3.028	2.362	2.959	—
	p=200, n=100					p=500, n=100				
(A) Identity	**4.661**	0.440	0.440	0.440	—	**9.321**	0.467	0.467	0.467	—
(B) MA(1)	**4.886**	0.717	0.626	0.716	—	**9.828**	0.761	0.729	0.761	—
(B*) MA(1)*	**10.727**	1.884	1.545	1.881	—	**21.233**	2.041	1.775	2.041	—
(C) AR(1)	**10.291**	6.922	4.898	6.768	—	**17.877**	9.311	5.584	9.261	—
(C*) AR(1)*	**20.277**	14.691	14.943	14.426	—	**39.241**	18.780	11.738	18.728	—
(D) Non-sparse	**26.729**	10.990	11.240	10.322	—	**50.915**	13.917	13.284	12.913	—
(E) Factor	**31.183**	28.053	29.819	20.463	20.432	**82.451**	65.234	73.807	48.104	48.928
(F) FGN	**14.732**	12.729	13.877	9.906	15.881	**35.041**	30.201	31.272	23.939	30.782
(F*) FGN*	**32.370**	26.692	29.862	20.357	28.983	**68.154**	66.833	66.320	49.853	55.998
(G) Seasonal	**6.913**	2.961	2.418	2.930	—	**13.157**	3.582	2.499	3.460	—
(G*) Seasonal*	**14.709**	6.427	5.171	6.350	—	**27.627**	7.873	5.660	7.538	—

Note: The results of $\hat{\Sigma}^N_{opt.r}$ are only presented for the highly non-sparse group, i.e. Models (E), (F) and (F*).

Table 3.4: Average operator norm error to Σ for competing estimators with data-driven parameters (50 replications). The best results and those up to 5% worse than the best are boxed. The worst results are in bold.

	S	P	$\hat{\Sigma}^N_{cv}$	$\hat{\Sigma}^N_r$	NS	S	P	$\hat{\Sigma}^N_{cv}$	$\hat{\Sigma}^N_r$	NS
	p=10, n=100					p=100, n=100				
(A) Identity	0.084	**0.823**	0.263	—	0.116	0.088	**3.657**	0.446	—	0.087
(B) MA(1)	0.444	**0.732**	0.493	—	0.481	0.670	**3.730**	0.704	—	0.694
(B*) MA(1)*	1.165	**1.546**	1.159	—	1.191	1.985	**8.015**	1.877	—	2.449
(C) AR(1)	1.013	1.135	**1.153**	—	1.017	5.423	**6.257**	5.390	—	5.892
(C*) AR(1)*	2.190	**2.291**	2.114	—	2.190	8.878	**19.468**	8.446	—	12.095
(D) Non-sparse	3.120	**3.860**	3.046	—	2.934	12.453	**29.355**	11.739	—	11.730
(E) Factor	1.793	**1.866**	1.741	1.763	1.537	**17.681**	14.304	16.497	16.438	15.285
(F) FGN	0.849	1.020	1.021	**1.024**	0.980	6.628	7.798	**7.799**	7.732	7.554
(F*) FGN*	2.218	2.221	2.222	**2.227**	1.960	14.795	15.611	15.225	15.254	**16.561**
(G) Seasonal	0.666	**0.852**	0.687	—	0.659	3.200	**4.826**	2.534	—	3.098
(G*) Seasonal*	1.647	**1.652**	1.452	—	1.480	4.268	**7.171**	3.016	—	6.979
	p=200, n=100					p=500, n=100				
(A) Identity	0.058	**5.414**	0.443	—	0.067	0.064	**10.076**	0.468	—	0.047
(B) MA(1)	0.658	**5.615**	0.744	—	0.694	0.645	**10.566**	0.819	—	0.683
(B*) MA(1)*	2.094	**12.458**	1.956	—	2.729	2.060	**23.034**	2.116	—	3.004
(C) AR(1)	8.123	**11.446**	8.217	—	7.759	12.785	**18.496**	12.484	—	12.036
(C*) AR(1)*	18.172	**23.721**	16.251	—	18.751	26.571	**40.903**	18.903	—	24.581
(D) Non-sparse	11.920	**30.108**	11.220	—	10.993	13.758	**54.462**	13.636	—	12.996
(E) Factor	**34.237**	31.064	33.224	33.194	31.020	**83.101**	81.489	81.697	81.382	80.852
(F) FGN	12.961	14.376	**14.640**	14.593	14.125	26.672	34.344	31.296	30.992	**36.299**
(F*) FGN*	31.165	30.263	**31.470**	31.042	32.188	**84.958**	69.133	75.546	75.377	74.432
(G) Seasonal	4.126	**7.403**	2.972	—	4.016	4.994	**13.722**	3.471	—	4.949
(G*) Seasonal*	9.225	**15.855**	6.494	—	9.064	11.030	**28.949**	7.561	—	11.132

Table 3.5: Average operator norm error to Σ^{-1} for competing estimators with optimal parameters (50 replications). The best results and those up to 5% worse than the best are boxed. The worst results are in bold.

	$\hat{\Sigma}$	T_s	B	$\hat{\Sigma}^N_{opt}$	$\hat{\Sigma}^N_{opt.r}$	$\hat{\Sigma}$	T_s	B	$\hat{\Sigma}^N_{opt}$	$\hat{\Sigma}^N_{opt.r}$
	p=10, n=100					p=100, n=100				
(A) Identity	**0.917**	0.281	0.281	0.281	—	—	0.469	0.469	0.469	—
(B) MA(1)	**1.177**	0.681	0.656	0.605	—	—	1.244	**1.300**	1.166	—
(B*) MA(1)*	0.626	0.489	**0.732**	0.442	—	—	**0.846**	0.779	0.745	—
(C) AR(1)	**9.078**	7.751	**9.078**	5.502	—	—	14.313	**18.064**	10.792	—
(C*) AR(1)*	**4.491**	2.736	**4.491**	2.339	—	—	**8.915**	7.298	6.001	—
(D) Non-sparse	**0.378**	0.256	0.297	0.210	—	—	2.670	**2.775**	1.793	—
(E) Factor	**0.846**	0.403	0.610	0.370	0.400	—	0.712	**0.715**	0.653	0.518
(F) FGN	**2.995**	1.727	2.980	1.560	1.535	—	3.585	**4.650**	3.112	2.734
(F*) FGN*	**1.571**	1.193	1.212	1.001	1.018	—	2.029	**2.038**	1.948	1.761
(G) Seasonal	**2.688**	1.538	2.685	1.302	—	—	3.806	**5.444**	3.260	—
(G*) Seasonal*	1.340	1.091	**1.726**	0.827	—	—	2.526	**4.345**	1.971	—
	p=200, n=100					p=500, n=100				
(A) Identity	—	0.527	0.527	0.527	—	—	0.599	0.599	0.599	—
(B) MA(1)	—	1.358	**1.530**	1.258	—	—	1.405	**1.562**	1.377	—
(B*) MA(1)*	—	**1.100**	0.795	0.850	—	—	1.040	**1.145**	0.962	—
(C) AR(1)	—	15.023	**18.122**	11.469	—	—	15.622	**18.136**	11.064	—
(C*) AR(1)*	—	14.509	**20.358**	7.362	—	—	18.392	**23.740**	7.155	—
(D) Non-sparse	—	**2.460**	2.016	1.459	—	—	**5.986**	5.896	4.289	—
(E) Factor	—	**0.711**	**0.711**	0.677	0.537	—	**0.744**	**0.744**	0.730	0.557
(F) FGN	—	3.972	**4.658**	3.317	3.024	—	4.267	**4.737**	3.527	3.306
(F*) FGN*	—	2.974	**4.096**	2.083	1.849	—	4.426	**5.674**	2.250	2.083
(G) Seasonal	—	4.029	**5.469**	3.538	—	—	4.188	**5.477**	3.673	—
(G*) Seasonal*	—	3.328	**4.885**	2.259	—	—	3.726	**5.479**	2.358	—

Note: The results of $\hat{\Sigma}^N_{opt.r}$ are only presented for the highly non-sparse group, i.e. Models (E), (F) and (F*). The worst results for model (A) with $p = 100$, 200 and 500 are not labelled, as T, B and $\hat{\Sigma}^N_{opt}$ obtain exactly the same results.

Table 3.6: Average operator norm error to Σ^{-1} for competing estimators with data-driven parameters (50 replications). The best results and those up to 5% worse than the best are boxed. The worst results are in bold.

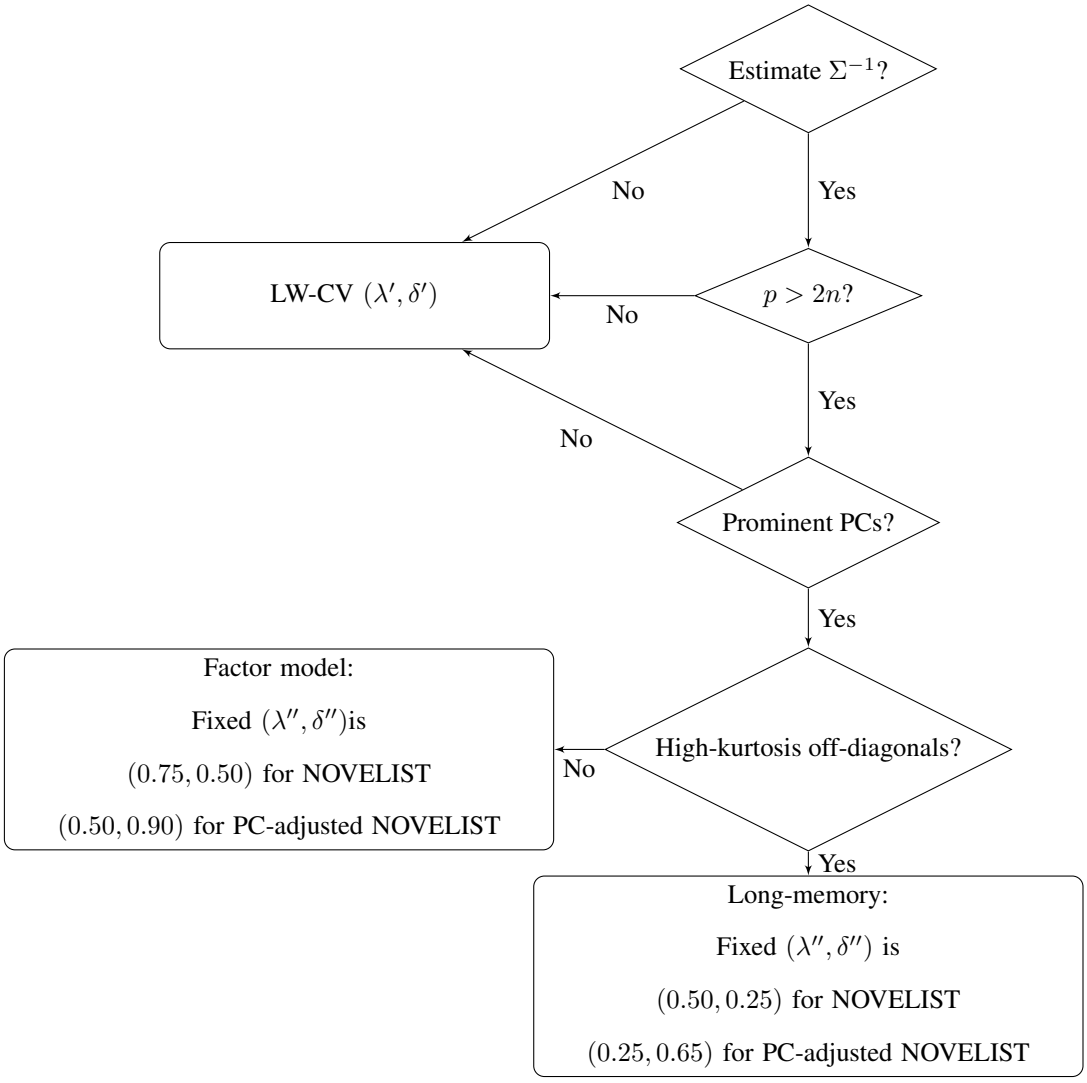
	S	P	$\hat{\Sigma}_{cv}^N$	$\hat{\Sigma}_r^N$	NS	S	P	$\hat{\Sigma}_{cv}^N$	$\hat{\Sigma}_r^N$	NS
	p=10, n=100					p=100, n=100				
(A) Identity	0.090	**4.472**	0.469	—	0.146	0.045	**0.882**	0.472	—	0.109
(B) MA(1)	0.799	**6.474**	0.824	—	0.780	1.273	1.403	**1.439**	—	1.405
(B*) MA(1)*	0.526	**4.892**	0.448	—	0.440	1.358	0.993	0.935	—	**1.748**
(C) AR(1)	7.309	**40.142**	8.574	—	5.396	13.410	**15.704**	12.605	—	12.272
(C*) AR(1)*	5.390	**27.593**	4.841	—	3.264	12.508	**13.649**	10.167	—	13.446
(D) Non-sparse	0.500	**1.705**	0.328	—	0.340	2.937	2.916	2.910	—	**2.979**
(E) Factor	1.142	**1.806**	0.864	—	0.296	**2.603**	0.893	1.608	—	0.343
				(0.854)				(0.695)	(0.526)	
(F) FGN	1.864	**16.530**	2.097	—	1.701	**4.565**	3.060	4.212	—	3.122
				(2.081)				(3.159)	(2.773)	
(F*) FGN*	1.174	**10.284**	2.017	—	1.101	**4.474**	2.965	3.431	—	4.432
				(2.001)				(2.075)	(1.843)	
(G) Seasonal	1.897	**13.175**	2.103	2.115	1.687	4.229	**4.721**	3.839	—	3.947
(G*) Seasonal*	1.284	**8.436**	1.143	—	1.219	3.510	3.799	2.743	—	**4.538**
	p=200, n=100					p=500, n=100				
(A) Identity	0.046	**0.930**	0.529	—	0.136	0.078	**0.923**	0.601	—	0.139
(B) MA(1)	1.449	1.371	1.401	—	**1.463**	1.473	1.445	**1.540**	—	1.487
(B*) MA(1)*	1.293	1.256	1.169	—	**1.906**	1.914	1.140	1.221	—	**2.463**
(C) AR(1)	15.066	**17.128**	14.125	—	13.907	16.526	**17.700**	16.025	—	15.924
(C*) AR(1)*	17.480	18.286	13.201	—	**19.037**	22.833	23.053	19.169	—	**23.740**
(D) Non-sparse	2.602	2.842	2.563	—	**3.206**	5.998	**6.171**	5.994	—	5.660
(E) Factor	**3.701**	0.892	1.450	—	0.348	**5.672**	0.962	4.106	—	0.347
			(0.710)	(0.546)				(0.937)	(0.558)	
(F) FGN	**9.397**	3.552	5.670	—	3.434	**8.621**	3.933	6.652	—	3.752
			(3.582)	(3.045)				(4.364)	(3.326)	
(F*) FGN*	**6.649**	2.765	4.024	—	5.519	6.241	3.083	5.442	—	**6.519**
			(2.589)	(2.199)				(3.002)	(2.887)	
(G) Seasonal	4.676	**5.019**	4.176	—	4.526	5.045	**5.256**	4.548	—	5.001
(G*) Seasonal*	4.540	4.643	3.514	—	**6.068**	5.632	5.254	4.489	—	**6.988**

Note: For models (E), (F) and (F*), results by both cross validation and fixed parameters (in brackets) are presented for NOVELIST when $n < 2p$. For $\hat{\Sigma}_{cv}^N$, fixed parameters (λ'', δ'') are $(0.75, 0.50)$ for Model (E), and $(0.50, 0.25)$ for Models (F) and (F*). For $\hat{\Sigma}_r^N$, (λ'', δ'') is fixed to be $(0.50, 0.90)$ for (E), and $(0.25, 0.65)$ for (F) and (F*).

3.7 Automatic NOVELIST algorithm and more Monte Carlo experiments

3.7.1 Automatic NOVELIST algorithm (ANOVELIST)

As shown in the simulation study, we note that NOVELIST or PC-adjusted NOVELIST with fixed parameters largely improve the performances in estimating precision matrices for model (E) and (F). However, we suggest that they should only be used with prior knowledge or if empirical testing indicates that there are prominent principal components. This extra section describes an automatic algorithm which provides an adaptive choice between the use of LW-CV algorithm and (PC-adjusted) NOVELIST with fixed parameters suggested in Section 3.5. For estimating the correlation, covariance or their inverses, given p and n, we suggest the following rules of thumb: first, we look for the evidence of "elbows" in the scree plot of eigenvalues, by examining if $\sum_{k=1}^{p} \mathbb{1}\{\gamma_{(k)} + \gamma_{(k+2)} - 2\gamma_{(k+1)} > 0.1p\} > 0$, where $\gamma(k)$ is the kth principal component. If so, then we look for the evidence of long-memory decay, by examining if the off-diagonals of the sample correlation matrix follow a high-kurtosis distribution. If the sample kurtosis ≤ 3.5, this suggests that the factor structure may be present, and we use the fixed parameters $(\lambda'', \delta'') = (0.75, 0.50)$ for NOVELIST or $(0.50, 0.90)$ for PC-adjusted NOVELIST; if the sample kurtosis > 3.5, this may point to long memory, and we use the fixed parameters $(\lambda'', \delta'') = (0.50, 0.25)$ for NOVELIST or $(0.25, 0.65)$ or PC-adjusted NOVELIST. The parameters are chosen from the robustness test in Section 3.5. It is sketched in the following flowchart.



Flowchart 1: ANOVELIST: decision procedure for using LW-CV algorithm or fixed parameters in estimating precision matrices.

3.7.2 More Monte Carlo experiments for automatic algorithm

More Monte Carlo simulations are conducted to test the performances of ANOVELIST . We test models (A)-(F), but not those with $*$, as NOVELIST already work well for heteroscedastic models in Section 3.6. Also, we only present the results of ANOV-

ELIST and the nonlinear shrinkage estimator [Ledoit and Wolf, 2003] (NS), as NS is the best competitor for NOVELIST as shown in Section 3.6. We use PC-adjusted NOVELIST with fixed parameters for ANOVELIST in the simulation.

Table 3.7: Average operator norm error to Σ^{-1} for Automatic NOVELIST and Nonlinear shrinkage (50 replications). The best results are boxed.

	ANOVEL	*NS*	*ANOVEL*	*NS*
	p=200, n=100		p=200, n=50	
(A) Identity	0.513	0.132	0.584	0.177
(B) MA(1)	1.411	1.469	1.934	1.997
(C) AR(1)	14.267	14.064	15.236	14.881
(D) Non-sparse	2.604	3.320	2.934	3.831
(E) Factor	0.727	0.350	1.133	0.568
(F) FGN	3.170	3.481	3.623	3.880
(G) Seasonal	4.153	4.502	4.663	4.904
	p=500, n=100		p=1000, n=100	
(A) Identity	0.627	0.146	0.806	0.267
(B) MA(1)	1.541	1.487	1.605	1.583
(C) AR(1)	16.246	16.132	19.334	19.537
(D) Non-sparse	5.980	5.643	8.304	7.923
(E) Factor	0.968	0.353	1.139	0.498
(F) FGN	3.591	3.729	5.067	5.638
(G) Seasonal	4.527	4.983	5.691	6.039

3.8 Conclusion

This chapter proposes the NOVELIST estimators for correlation/covariance and their inverses. The linkage between NOVELIST and ridge regression are demonstrated. We obtain an explicit convergence rate in the operator norm over a large class of covariance (correlation) matrices when p and n satisfy $\log p/n \to 0$. Empirical choices of parameters and a data-driven algorithm for NOVELIST estimators which combines Ledoit and Wolf [2003]'s method and cross-validation (LW-CV algorithm) is presented. Further empirical improvements of NOVELIST are proposed. Comprehensive simulation study is based on a wide range of models and results of comparisons with several popular estimators are presented. Finally, an automatic algorithm is constructed to provide an adaptive choice between the use of LW-CV algorithm and fixed parameters.

Based on the simulation results, NOVELIST works best when the underlying correlation/covariance matrices are sparse and non-sparse (more so for heteroscedastic models) but is beaten by POET for the highly non-sparse models by a small margin. Also, NOVELIST performs better for the heteroscedastic models than for the homoscedastic ones due to the fact that NOVELIST does not shrink the diagonals towards any target such as their median, which particularly suits the heteroscedastic models. However, NOVELIST does not perform stable when estimating precision matrices for the highly non-sparse cases, which is because of the bad performance of the cross-validated choices of the parameters. We improve the results by applying fixed parameters that come from the robustness test instead of the cross-validated ones, and also build a bridge between using the fixed parameters and the cross-validated choices. The fixed parameters vary across different underlying correlation/covariance structures, but they are mostly not close to the edges of the range $[0, 1]$ to ensure stable

performance. Overall, it is clear that the flexible control of the degree of shrinkage and thresholding offered by NOVELIST means that it is able to offer competitive performance across most models, and in situations in which it is not the best, it tends not to be much worse than the best performer. We recommend NOVELIST as a simple, good all-round covariance, correlation and precision matrix estimator ready for practical use across a variety of models and data dimensionalities.

3.9 Additional lemmas and proofs

Firstly, we briefly introduce two lemmas that will be used in the proof of Proposition 1.

Lemma 5 *If F satisfies $\int_0^\infty exp(\gamma t) dG_j(t) < \infty$, for $0 < |\gamma| < \gamma_0$, for some $\gamma_0 > 0$, where G_j is the cdf of X_{1j}^2, $R = \{\rho_{ij}\}$ and $\Sigma = \{\sigma_{ij}\}$ are the true correlation and covariance matrices, $1 \leq i, j \leq p$, and $\sigma_{ii} \leq M$, where M is a constant, then, for sufficiently large M', if $\lambda = M'\sqrt{\log p/n}$ and $\log p/n = o(1)$, we have $\max_{1 \leq i,j \leq p} |\hat{\rho}_{ij} - \rho_{ij}| = O_p(\sqrt{\log p/n})$, for $1 \leq i, j \leq p$.*

Proof of Lemma 5: By the sub-multiplicative norm property $||AB|| \leq ||A||\ ||B||$

[Golub and Van Loan, 2013], we write

$$
\begin{aligned}
&\max_{1\le i,j\le p}|\hat{\rho}_{ij}-\rho_{ij}| \\
=&\max_{1\le i,j\le p}|\hat{\sigma}_{ij}/(\hat{\sigma}_{ii}\hat{\sigma}_{jj})^{1/2}-\sigma_{ij}/(\sigma_{ii}\sigma_{jj})^{1/2}| \\
\le&\max_{1\le i\le p}|\hat{\sigma}_{ii}^{-1/2}-\sigma_{ii}^{-1/2}|\max_{1\le i,j\le p}|\hat{\sigma}_{ij}-\sigma_{ij}|\max_{1\le j\le p}|\hat{\sigma}_{jj}^{-1/2}-\sigma_{jj}^{-1/2}| \\
+&\max_{1\le i\le p}|\hat{\sigma}_{ii}^{-1/2}-\sigma_{ii}^{-1/2}|\max_{1\le i,j\le p}(|\hat{\sigma}_{ij}||\sigma_{jj}^{-1/2}|+|\hat{\sigma_{ii}}^{-1/2}||\sigma_{ij}|) \\
+&\max_{1\le i,j\le p}|\hat{\sigma}_{ij}-\sigma_{ij}|\max_{1\le i\le p}|\hat{\sigma_{ii}}^{-1/2}|\max_{1\le i\le p}|\sigma_{ii}^{-1/2}| \\
=&O_p(\sqrt{\log p/n}) \qquad (3.39)
\end{aligned}
$$

The last equality holds as we have $\max_{1\le i,j\le p}|\hat{\sigma}_{ij}-\sigma_{ij}| = O_p(\sqrt{\log p/n}) = \max_{1\le i,j\le p}|\hat{\sigma}_{ij}^{-1}-\sigma_{ij}^{-1}|$ [Bickel and Levina, 2008b], and $\max_{1\le i,j\le p}|\hat{\sigma}_{ij}| = O_p(1) = \max_{1\le i,j\le p}|\hat{\sigma}_{ij}^{-1}|$, and $\sigma_{ii} \le M, 1 \le i,j \le p$. ■

Lemma 6 *If F satisfies $\int_0^\infty exp(\gamma t)dG_j(t) < \infty$, for $0 < |\gamma| < \gamma_0$, for some $\gamma_0 > 0$, where G_j is the cdf of X_{1j}^2, $R = \{\rho_{ij}\}$ is the true correlation matrix, $1 \le i,j \le p$, then, uniformly on $\mathcal{V}(q, s_0(p), \varepsilon_0)$, for sufficiently large M', if $\lambda = M'\sqrt{\log p/n}$ and* $\log p/n = o(1)$,

$$||T(\hat{R},\lambda) - R|| = O_p(s_0(p)(\log p/n)^{(1-q)/2}). \qquad (3.40)$$

where T is any kind of generalised thresholding estimator.

Lemma 6 is a correlation version of Theorem 1 in Rothman et al. [2009] and follows in a straightforward way by replacing $\hat{\Sigma}$, Σ, $\mathcal{U}(q, c_0(p), M, \epsilon_0)$ and $c_0(p)$ by $\hat{R}$, R, $\mathcal{V}(q, s_0(p), \varepsilon_0)$ and $s_0(p)$ in the proof of the theorem.

Proof of Proposition 1:

We first show the result for $\hat{R}^N$. By the triangle inequality,

$$\begin{aligned} ||\hat{R}^N - R|| &= ||(1-\delta)\hat{R} + \delta T(\hat{R}, \lambda) - R|| \\ &\leq (1-\delta)||\hat{R} - R|| + \delta||T(\hat{R}, \lambda) - R|| \\ &= I + II. \end{aligned} \tag{3.41}$$

Using Lemma 6, we have

$$II = O_p\{\delta s_0(p)(\log p/n)^{(1-q)/2}\}. \tag{3.42}$$

For symmetric matrices M, Corollary 2.3.2 in Golub and Van Loan [2013] states that

$$||M|| \leq (||M||_{(1,1)}||M||_{(\infty,\infty)})^{1/2} = ||M||_{(1,1)} = \max_{1\leq i\leq p} \sum_{j=1}^{p} |m_{ij}|. \tag{3.43}$$

Then by Lemma 5,

$$||\hat{R} - R|| \leq \max_{1\leq i\leq p} \sum_{j=1}^{p} |\hat{R}_{ij} - R_{ij}| \leq p \max_{1\leq i,j\leq p} |\hat{\rho}_{ij} - \rho_{ij}| = O_p(p\sqrt{\log p/n}). \tag{3.44}$$

Thus, we have

$$I = (1-\delta)||\hat{R} - R|| \leq O_p((1-\delta)p\sqrt{\log p/n}). \tag{3.45}$$

Combining formula (3.42) and (3.45) yields formula (3.16). The corresponding inverse obtains the same rate,

$$||(\hat{R}^N)^{-1} - R^{-1}|| \asymp ||\hat{R}^N - R||, \tag{3.46}$$

uniformly on $\mathcal{V}(q, s_0(p), \varepsilon_0)$.

For the $\hat{\Sigma}^N$ estimator, recalling that $T = T(\hat{R}, \lambda)$ and $D = (\text{diag}(\Sigma))^{1/2}$, we have

$$\begin{aligned} ||\hat{\Sigma}^N - \Sigma|| &= ||\hat{D}\hat{R}^N\hat{D} - DRD|| \\ &= ||\hat{D}((1-\delta)\hat{R} + \delta\, T)\hat{D} - DRD|| \\ &\leq (1-\delta)||\hat{\Sigma} - \Sigma|| + \delta||\hat{D}T\hat{D} - DRD|| \\ &= III + IV. \end{aligned} \tag{3.47}$$

Similarly as in (3.45), we obtain $III = O_p((1-\delta)p\sqrt{\log p/n})$. For IV, we write

$$\begin{aligned} &||\hat{D}T\hat{D} - DRD|| \\ \leq &||\hat{D} - D||\,||T - R||\,||\hat{D} - D|| + ||\hat{D} - D||(||T||\,||D|| + ||\hat{D}||\,||R||) \\ &+||T - R||\,||\hat{D}||\,||D|| \\ = &O_p((1 + s_0(p)(\log p/n)^{-q/2})\sqrt{\log p/n}). \end{aligned} \tag{3.48}$$

The last equality holds as we have $||T - R|| = O_p(s_0(p)(\log p/n)^{(1-q)/2})$, $||\hat{D} - D|| = O_p(\sqrt{\log p/n})$, $||\hat{D}|| = O_p(1) = ||T||$, and $||D|| = O(1)$ as $\sigma_{ii} < M$. Because $(\log p/n)^{q/2}(s_0(p))^{-1}$ is bounded from above by the assumption that $\log p/n = o(1)$ and $||(\hat{\Sigma}^N)^{-1} - \Sigma^{-1}|| \asymp ||\hat{\Sigma}^N - \Sigma||$ uniformly on $\mathcal{U}(q, c_0(p), M, \epsilon_0)$, the result follows.

■

Proof of Corollary 2:

Substituting $\log p$ by $C_1 n^\alpha$ in (3.18), we get

$$\tilde{\delta} = \frac{C_2 p n^{(\alpha-1)q/2}}{s_0(p) + C_2 p n^{(\alpha-1)q/2}}, \tag{3.49}$$

where C_2 is a constant. If $p = o(n^{(1-\alpha)q/2})$, we have $pn^{(\alpha-1)q/2} \to 0$, which implies $\tilde{\delta} \to 0$, since $s_0(p) \leq C$. On the other hand, if $n = o(p^{2/(1-\alpha)q})$, we have $pn^{(\alpha-1)q/2} \to \infty$ and $\tilde{\delta} \to 1$ as $n \to \infty$. Additionally, if $p \asymp n^{(1-\alpha)q/2}$, then $pn^{(\alpha-1)q/2}$ is of a constant order, which yields $\tilde{\delta} \in (0, 1)$, as required. ■

Proof of Corollary 3:

Firstly, noting that

$$\int_1^{p+1} K^{-\gamma q} dK < \sum_{K=1}^{p} K^{-\gamma q} < \int_0^{p} K^{-\gamma q} dK \tag{3.50}$$
$$\frac{p^{1-\gamma q}}{1-\gamma q} < \sum_{K=1}^{p} K^{-\gamma q} < \frac{(p+1)^{1-\gamma q} - 1}{1-\gamma q},$$

we have $\sum_{K=1}^{p} K^{-\gamma q} = O(p^{1-\gamma q})$. For the long-memory correlation matrix, we can write

$$s_0(p) = \max_{1 \leq i \leq p} \sum_{j=1}^{p} |i-j|^{-\gamma q} = O(p^{1-\gamma q}). \tag{3.51}$$

By substituting $\log p$ by $C_1 n^{\alpha}$ and $s_0(p)$ by (3.51) in (3.18), we get

$$\tilde{\delta} = \frac{C_2 n^{(\alpha-1)q/2}}{p^{-\gamma q} + C_2 n^{(\alpha-1)q/2}}. \tag{3.52}$$

Again $\tilde{\delta}$ depends on p and n. The remaining part of the proof is analogous to that of Corollary 2 and is omitted here. ■

Chapter 4

Applications of NOVELIST and real data examples

4.1 Introduction

As stated in Section 1.1, estimation of covariance, correlation and precision matrices for high-dimensional data have remarkable applications in almost every aspect of statistics, such as principal component analysis [Croux and Haesbroeck, 2000; Jackson, 1991; Johnstone and Lu, 2009; Pearson, 1901], linear discriminant analysis [Bickel and Levina, 2004; Fisher, 1936; Guo et al., 2007], graphical modeling [Meinshausen and Bühlmann, 2008; Ravikumar et al., 2011; Yuan, 2010], portfolio selection and financial risk management [Fan et al., 2008; Goldfarb and Iyengar, 2003; Ledoit and Wolf, 2003; Longerstaey et al., 1996; Markowitz, 1952; Talih, 2003], and network science [Gardner et al., 2003; Jeong et al., 2001].

Apart from these popular areas, many other applications arise in literature where covariance or precision matrix estimation is just an intermediate step instead of the

final goal, and better covariance or precision estimation can lead to better results in the end. In particular, covariance matrix estimation can be found in estimation of false discovery proportion (FDP) of large-scale multiple testing with highly correlated test statistics [Fan and Han, 2013; Fan et al., 2012a]. Over the last two decades, testing procedures have been proposed in incorporating correlation information in estimating FDP [Benjamini and Yekutieli, 2001; Sarkar, 2002; Sun and Cai, 2009]. In recent years, Fan et al. [2012a] propose a consistent estimate of realized FDP based on principal factor approximation (PFA), which subtracts the known common dependence and significantly weakens the correlation structure. However, if such dependence structure is unknown, the covariance matrix has to be estimated before estimating FDP [Efron, 2010]. For tackling this problem, Fan and Han [2013] investigate conditions on the dependence structure such that the estimate of FDP is consistent and study an approximate factor model for the test statistics, then develop a consistent estimate of FDP by applying the POET estimator [Fan et al., 2013] to estimate the unknown covariance matrix. Moreover, another application considered in several papers [Bickel and Levina, 2008a; Huang et al., 2006; Lam, 2016] is to apply the estimated large covariance matrix on forecasting the call arrival pattern to a telephone call centre, in particular, predicting the number of arrivals later in a day by using arrival patterns at earlier times of the day. In this chapter, we explore the applications of NOVELIST estimators and exhibit the results of applying the estimators on real data, including portfolio optimisation using inter-day and intra-day log returns of the constituents of FTSE 100, forecasting the number of calls for the call center, and estimating false discovery proportion through a well-known breast cancer study. The rest of the chapter is organised as follows. In Section 4.2, we illustrate how NOVELIST performs in the minimal variance portfolio optimisation problems. Section 4.3 presents the performance of NOVELIST in fore-

casting the phone calls. Section 4.4 shows the application of NOVELIST in estimating FDP in the breast cancer study. Section 4.5 concludes the chapter.

4.2 Portfolio selection

Portfolio selection is an empirical finance problem of efficiently allocating capital over a number of assets in order to maximize the expected "return" and/or minimise the level of "risk" according to investors' risk preferences [Goldfarb and Iyengar, 2003; Markowitz, 1952]. The first mathematical model for portfolio selection is formulated by Markowitz [1952], when he introduces the Modern Portfolio Theory (MPT), also known as mean-variance analysis. In modern portfolio theory, the "return" and "risk" of a portfolio are measured by the expected value and the variance of the portfolio return respectively. The mean-variance model also has had a profound impact on the Capital Asset Pricing Model (CAPM) [Lintner, 1965; Mossin, 1966; Sharpe, 1964], which is a model that derives the theoretical required expected return when considering adding a new asset to the existing portfolio, given the risk-free rate available to investors and the risk of the overall market [Sharpe, 1964]. In 1990, Sharpe and Markowitz shared the Nobel Prize in Economic Sciences for their contributions to the field of financial economics.

Although the MPT is originally proposed based on daily data, using high frequency data in portfolio management is arising in literature over the last decade, which benefits from apparent increase in sample size for returns and covariance matrix estimation. [Andersen et al., 2006; Fan et al., 2012b; Fleming et al., 2003; Liu, 2009]. Thanks to advanced computational power and efficient data storage facilities, high frequency data are easily accessible and increasingly analyzed by market practitioners and aca-

demic researchers. However, many authors are aware of the contamination of market microstructure in the tick-by-tick data and the problems caused by non-synchronous trading times of multi-dimensional high-frequency data [Aït-Sahalia et al., 2005; Bandi and Russell, 2005]. To overcome these two challenges, one way is to sample less frequently to avoid or largely reduce the market microstructure noise, when the noise is present but unaccounted for. The popular choices of high frequency sampling in the empirical literature range from 5-min intervals [Barndorff-Nielsen and Shephard, 2002] to as long as 30-min intervals [Andersen et al., 2003]. Aït-Sahalia et al. [2005] derives a closed-form expression of the optimal sampling frequency under the presence of i.i.d. microstructure noise. The optimal sampling frequency is often found to be between one and five minutes [Aït-Sahalia et al., 2005; Park, 2011]. Further discussion on the optimal sampling rate can be found in Bandi and Russell [2005]. Another way to tackle microstructure contamination is to model the noise by using very high frequent data and to ameliorate the bias contributed from the extreme eigenvalues of the realized covariance matrix by regularization with specific assumptions on the true integrated matrix itself, such as sparsity [Wang and Zou, 2010] and factor model [Tao et al., 2011]. Other attempts includes Fan et al. [2012b] who impose constraints on gross exposure of the portfolio directly, and Lam and Feng [2016] who nonlinearly shrink extreme eigenvalues of the sample integrated covariance matrix without specific assumption for the underlying integrated covariance matrix structure.

In this section, we present real-data performance of NOVELIST in portfolio optimisation problems based on daily and intra-day returns. For intra-day sampling frequency, we use 5-30 minutes to mostly reduce the contamination induced by microstructure noise, although the noise still exist as shown in the results of Section 4.2.2.3, nevertheless, we focus on comparison instead of estimation. We apply the

NOVELIST algorithm and the competing methods to share portfolios composed of the constituents of the FTSE 100 index. Similar competitions were previously conducted to compare the performance of different covariance matrix estimators [Lam, 2016; Lam and Feng, 2016; Ledoit and Wolf, 2003]. We compare the performance for risk minimisation purposes instead of return maximisation, i.e. we want to find the estimator which can minimise the portfolio volatility not the one which can maximise the portfolio return. The data were provided by Bloomberg.

4.2.1 Daily returns

4.2.1.1 Dataset

The constituents of the FTSE 100 index consists of 100 companies, but there are 101 listings, as Royal Dutch Shell has both A and B class shares listed. They are essentially identical shares except for a difference in dividend access mechanism, which applies only to the B class shares. Although it may lead to practitioners' preferences to B class shares in practice, it does not impact on this real data experiment, which is purely based on the returns, and does not take the costs or dividends into account. The returns normally follow factor models instead of i.i.d. distribution, but from simulation studies in Section 3.6 we note that NOVELIST still performs well for factor models. Our first dataset consists of $p = 85$ stocks of FTSE 100 and $n = 2526$ daily returns $\{r_t\}$ for the period January 1st 2005 to December 31st 2014. We removed all those constituents that contain missing values and all the non-trading days including the weekends and public holidays.

4.2.1.2 Portfolio rebalancing regimes

We use two portfolio rebalancing regimes for daily data, the first one is explained below.

Rebalancing regime 1

1. In-sample covariance matrix estimation: we set the number of in-sample observations as $n_1 = 120$. On trading day t, we use the past n_1-trading-day returns (i.e. computed over days $t - n_1 + 1$ to t) to estimate the $p \times p$ covariance matrix $\hat{\Sigma}_t^{(n_1)}$ by using NOVELIST and several other covariance matrix estimators. The first t starts from day $n_1 + 1$.

2. Minimal variance portfolio optimisation: to solve the risk minimisation problem

$$\min_{w_t' \mathbf{1}_p = 1} w_t' \hat{\Sigma}_t^{(n_1)} w_t, \tag{4.1}$$

we obtain the well-known weight formula

$$\hat{w}_t = \frac{\{\hat{\Sigma}_t^{(n_1)}\}^{-1} \mathbf{1}_p}{\mathbf{1}_p' \{\hat{\Sigma}_t^{(n_1)}\}^{-1} \mathbf{1}_p}, \tag{4.2}$$

where $\mathbf{1}_p$ is the column vector of p ones. Based on formula (4.2), portfolios are constructed according to different covariance matrix estimators.

3. Out-of-sample portfolio performances: we hold these portfolios for the next $n_2 = 22$ trading days (i.e. over days $t + 1$ to $t + n_2$) and compute their daily returns, out-of-sample standard deviations and Sharpe ratio as follows [DeMiguel and Nogales,

2009; Lam, 2016; Ledoit and Wolf, 2003],

$$\begin{aligned}\hat{\mu}_t &= \frac{1}{n_2}\sum_{i=1}^{n_2}\hat{w}_t' r_{t+i},\\ \hat{\sigma}_t &= \{\frac{1}{n_2}\sum_{i=1}^{n_2}(\hat{w}_t' r_{t+i} - \hat{\mu}_t)^2\}^{1/2},\\ \hat{sr}_t &= \hat{\mu}_t/\hat{\sigma}_t. \end{aligned} \tag{4.3}$$

4. Portfolio rebalancing: at the end of the $t + n_2$ day, we liquidate the portfolios, update current $t = t + n_2$ and start process 1-3 all over again until $t + n_2 > n$.

5. Annualised average results: finally, we obtain the average daily returns, out-of-sample standard deviations and Sharpe ratios. In order to compare the results from different rebalancing regimes, we annualise the average results as follows [Lam and Feng, 2016]

$$\begin{aligned}\widetilde{\mu} &= 252 \times \frac{1}{N}\sum_{j=0}^{N-1}\hat{\mu}_{n_1+j\cdot n_2},\\ \widetilde{\sigma} &= \sqrt{252} \times \frac{1}{N}\sum_{j=0}^{N-1}\hat{\sigma}_{n_1+j\cdot n_2},\\ \widetilde{sr} &= \sqrt{252} \times \frac{1}{N}\sum_{j=0}^{N-1}\hat{sr}_{n_1+j\cdot n_2}. \end{aligned} \tag{4.4}$$

where N is the times of rebalancing, i.e. $N = \left\lfloor \frac{n-n_1}{n_2} \right\rfloor = 109$ for regime 1. And 252 is the number of trading days per year.

Rebalancing regime 2

We use $n_1 = 252$ to see the impacts of prolonging the in-sample period on estimating the covariance matrix and the corresponding portfolio performance. Hence,

Table 4.1: Proportion of times (N of them) when in-sample covariance matrix has prominent PCs or high-kurtosis off-diagonals and decisions of NOVELIST algorithm made according to Section 3.7.

	Prominent PCs	High-kurtosis off-diagonals	Decisions
Regime 1: Daily with $n_1 = 120$	1.000	0.312	factor model
Regime 2: Daily with $n_1 = 152$	1.000	0.359	factor model
Regime 3: Intra-day 5 minutes	0.759	0.008	factor model
Regime 4: Intra-day 10 minutes	0.742	0.008	factor model
Regime 5: Intra-day 30 minutes	0.664	0.008	factor model

$N = \left\lfloor \frac{n-n_1}{n_2} \right\rfloor = 103$. All the other procedures remain the same.

We compare the performances of six covariance matrix estimators. For NOVELIST, we always apply the decision procedure as stated in Section 3.7 to choose from LW-CV algorithm and fixed parameters. Table 4.1 presents that the decision procedure points to underlying factor structure for both rebalancing regime 1 and 2, and also for all the portfolio rebalancing regimes based on intra-day returns in Section 4.2.2. Moreover, factor model is one of the popular structural assumptions in financial applications [Fan et al., 2013]. Both decision procedures and prior knowledge imply factor structure, which suggest NOVELIST with fixed parameters instead of cross validated parameters. We place both NOVELIST and PC-adjusted NOVELIST with fixed parameters on the competitors' list. Apart from NOVELIST, there are four other data-driven competing covariance matrix estimators, which we previously considered in Section 3.6: sample covariance estimator, linear shrinkage estimator, nonlinear shrinkage estimator, and POET. We use them again to compete with NOVELIST estimator. Also, we use the R package *corpcor*, *POET* and *novelist* to compute linear shrinkage,

POET and NOVELIST respectively, and the commercial package SNOPT for Matlab to compute nonlinear shrinkage [Ledoit and Wolf, 2013], We use $k = 7$ for POET as suggested by Fan et al. [2013], and $K = 1$ for PC adjusted NOVELIST, since it is common that there is one overwhelming principal component for financial data, and the preliminary data analysis for this dataset also supports this.

4.2.1.3 Results

Table 4.2 shows the results. Clearly, NOVELIST has the lowest risk, which is measured by the out-of-sample standard deviation, followed by PC-adjusted NOVELIST and nonlinear shrinkage. Also, NOVELIST has the highest Sharpe ratio, followed by linear and nonlinear shrinkage. However, NOVELIST is beaten by sample covariance matrix for annualised portfolio returns, which is not surprising as the portfolio weights in formula (4.2) are allocated for risk minimisation purpose instead of return maximisation. However, sample covariance matrix has highest risk and lowest Sharpe ratios. In essence, NOVELIST and Nonlinear shrinkage have risk minimisation done well and maintaining the level of Sharpe ratio greater than 1, which is considered as "good" by practitioners [Khalsa, 2013; Maverick, 2016]. The results of rebalancing regime 1 and 2 are similar, which implies that there is no prominent improvement by prolonging the in-sample period.

Figure 4.1 presents impacts of the choices of the parameters (λ, δ) on the performance of NOVELIST. We call the areas indicated by "1" the "outperforming ranges" of parameters, where NOVELIST estimators always beat all the other competitors in our study. For both rebalancing regime 1 and 2, NOVELIST outperforms with wide outperforming ranges for risk and Sharpe ratio, and the suggested fixed parameter $(\lambda'', \delta'') = (0.75, 0.5)$ for factor model is within the outperforming ranges. However,

$(0.75, 0.5)$ is outside the outperforming ranges of (λ, δ) for portfolio returns, which explains the reason why NOVELIST does not perform well in terms of enhancing returns, although it is not the purpose of the minimum-variance portfolio optimisation.

Table 4.2: Annualised portfolio returns, standard deviations (STDs) and Sharpe ratios of minimum variance portfolios (based on daily data) as in formula (4.4). The best results are boxed.

	Annualised portfolio returns (%)	out-of-sample STDs (%)	Sharpe ratios
	Regime 1: Daily with $n_1 = 120$		
Sample	8.928	19.261	0.616
Linear shrinkage	7.166	13.572	1.103
Nonlinear shrinkage	5.800	11.690	1.092
POET	4.283	12.235	0.871
NOVELIST	6.973	11.422	1.264
PC-adjusted NOVELIST	5.144	11.590	0.978
	Regime 2: Daily with $n_1 = 252$		
Sample	6.866	13.631	0.883
Linear shrinkage	6.049	13.031	1.000
Nonlinear shrinkage	6.630	11.990	1.089
POET	6.763	12.453	1.012
NOVELIST	6.475	11.678	1.186
PC-adjusted NOVELIST	4.476	12.189	0.877

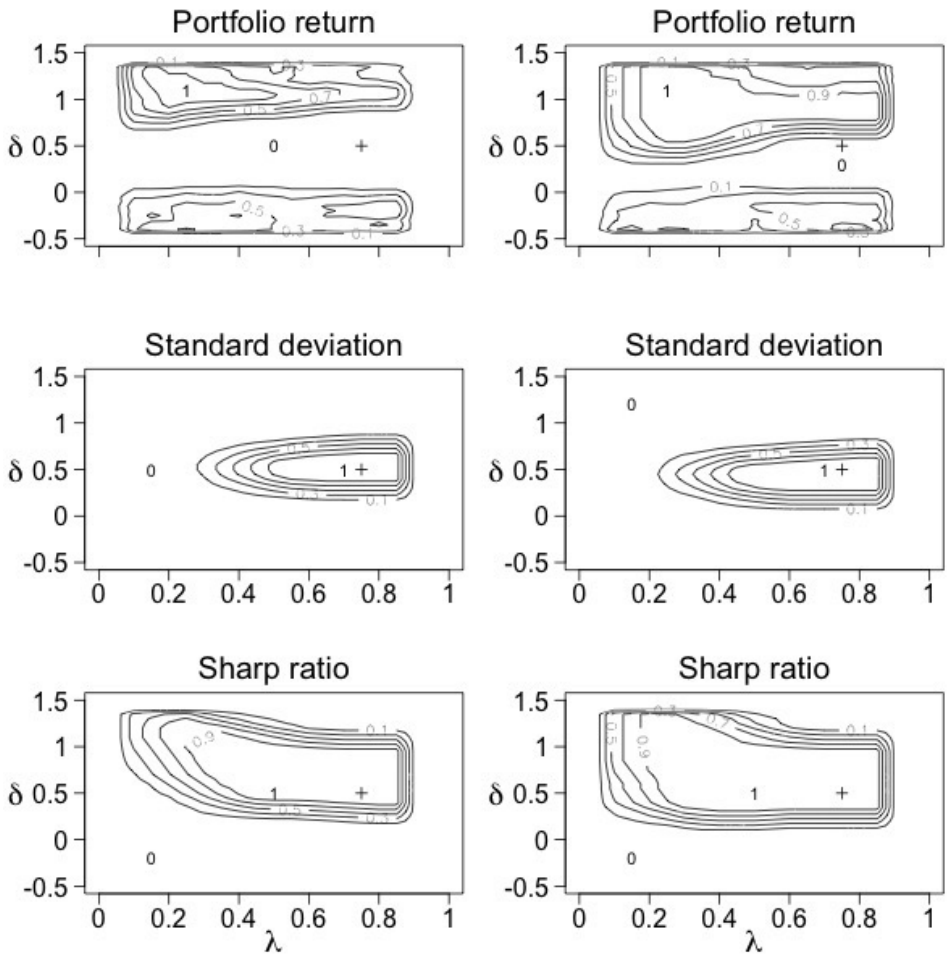


Figure 4.1: Contour plots of proportions of the times when NOVELIST outperforms in terms of the choices of (λ, δ) under rebalancing regime 1 (left column) and 2 (right column). "1" indicates the area of choices of (λ, δ) which makes NOVELIST to outperform with the chance of 100%, in contrast, "0" indicates the area of choices of (λ, δ) where NOVELIST never outperform. The suggested fixed parameter $(\lambda'', \delta'') = (0.75, 0.50)$ for factor model which is used in Automatic NOVELIST algorithm in Section 3.7 is marked as a plus.

4.2.2 Intra-day returns

4.2.2.1 Datasets and sampling

We use three datasets that cover the same time period but have different sampling frequencies. To distinguish them from the first dataset which we use in Section 4.2.1, we call these three datasets the second, third and fourth ones. The second dataset consists of $p = 101$ constituents of FTSE 100 and $n = 13260$ five-minute returns $\{y_t\}$ for the period March 2nd 2015 to September 4th 2015 (130 trading days), after removing the weekends, 3 bank holidays and 2 Easter holidays, and retaining only the returns within the trading time 8:00-16:30 on each trading day, i.e. the number of observations is 102 on each trading day. The third dataset consists of $n = 6630$ ten-minute returns with 51 points on each day for the 130 trading days. And the fourth dataset has $n = 2210$ thirty-minute returns with 17 points on each day for the 130 trading days.

4.2.2.2 Portfolio rebalancing regimes

We use three regimes for intra-day portfolio rebalancing. They all rebalance every-day using the past ten-day as in-sample data for estimating the covariance matrix, but the differences rely on the sampling frequency: they are based on 5, 10, and 30 minutes sampling frequency respectively. The rebalancing regimes are similar to those in Section 4.2.1 and here we only explain the differences.

Rebalancing regime 3

1. In-sample covariance matrix estimation: we use the second dataset (sampling frequency $f = 5$ minutes, $n_2 = 102$ on each day), and 10-day in-sample period to estimate the covariance matrices of the returns, i.e. the number of in-sample observations

$n_1 = 10n_2 = 1020$. At the starting time of trading day t, we use the past n_1-five-minute returns (i.e. five-minute returns from $n_2(t-11)+1$ to $n_2(t-1)$) to estimate the covariance matrix $\hat{\Sigma}_t^{(n_1)}$. The first t starts from day 11.

2. Minimal variance portfolio optimisation is the same as that in Section 4.2.1.

3. Out-of-sample portfolio performances: we hold these portfolios for the trading day t (i.e. over five-minute points $n_2(t-1)+1$ to $n_2 t$) and compute their five-minute returns, out-of-sample standard deviations and Sharpe ratio as follows

$$\begin{aligned}
\hat{\mu}_t &= \frac{1}{n_2}\sum_{i=1}^{n_2} \hat{w}_t' r_{n_2(t-1)+i}, \\
\hat{\sigma}_t &= \{\frac{1}{n_2}\sum_{i=1}^{n_2} (\hat{w}_t' r_{n_2(t-1)+i} - \hat{\mu}_t)^2\}^{1/2}, \\
\hat{sr}_t &= \hat{\mu}_t / \hat{\sigma}_t.
\end{aligned} \tag{4.5}$$

4. Portfolio rebalancing: at the beginning of day $t+1$, we liquidate the portfolios, update current $t = t+1$ and start process 1-3 all over again until $n_2(t+1) > n$.

5. Annualised average results. The annualised average portfolio returns, standard deviations and Sharpe ratios are as follows [Lam and Feng, 2016]

$$\begin{aligned}
\widetilde{\mu} &= 252 \times n_2 \times \frac{1}{N}\sum_{j=0}^{N-1} \hat{\mu}_{11+j}, \\
\widetilde{\sigma} &= \sqrt{252 \times n_2} \times \frac{1}{N}\sum_{j=0}^{N-1} \hat{\sigma}_{11+j}, \\
\widetilde{sr} &= \sqrt{252 \times n_2} \times \frac{1}{N}\sum_{j=0}^{N-1} \hat{sr}_{11+j}.
\end{aligned} \tag{4.6}$$

where N is the times of rebalancing, i.e. $N = \left\lfloor \frac{n-n_1}{n_2} \right\rfloor = 120$ for regime 3.

Rebalancing regime 4

We use the third dataset (sampling frequency $f = 10$ minutes), $n = 6630$ $n_1 = 510$, $n_2 = 51$. All the procedures are the same as those for rebalancing regime 2.

Rebalancing regime 5

We use the fourth dataset (sampling frequency $f = 30$ minutes), $n = 2210$, $n_1 = 170$, $n_2 = 17$. All the procedures are the same as those for rebalancing regime 2.

4.2.2.3 Results

Microstructure noises. Figure 4.2 presents distributions of sample variances and covariances of $p = 101$ intra-day stock returns, which are prevalently used as indicator of microstructure noise. Clearly, when sampling frequency increases from once every 30 minutes to once every 5 minutes, variances of returns slightly increase while covariances decrease due to presence of microstructure noise. Figure 4.3 shows six minimal variance portfolio returns are more volatile when sampling frequency is every 5 minutes.

Overall competitions. Table 4.2 shows the results of overall competition. We note that the portfolio returns and most of the Sharpe ratios are negative during this period, but which does not impact on the competition. NOVELIST has the highest portfolio returns (the least loss) and highest Sharpe ratios for both five-minute and ten-minute portfolios, followed by PC-adjusted NOVELIST and nonlinear shrinkage, and has the lowest out-of-sample standard deviations for thirty-minute portfolios. However, NOVELIST is beaten by nonlinear shrinkage or POET otherwise. In summary, we argue that NOVELIST is the overall winner, followed by nonlinear shrinkage.

Some remarks: one may note that the annualised out-of-sample standard deviations listed in Table 4.3 do not vary a lot as sampling frequency changes, which seems

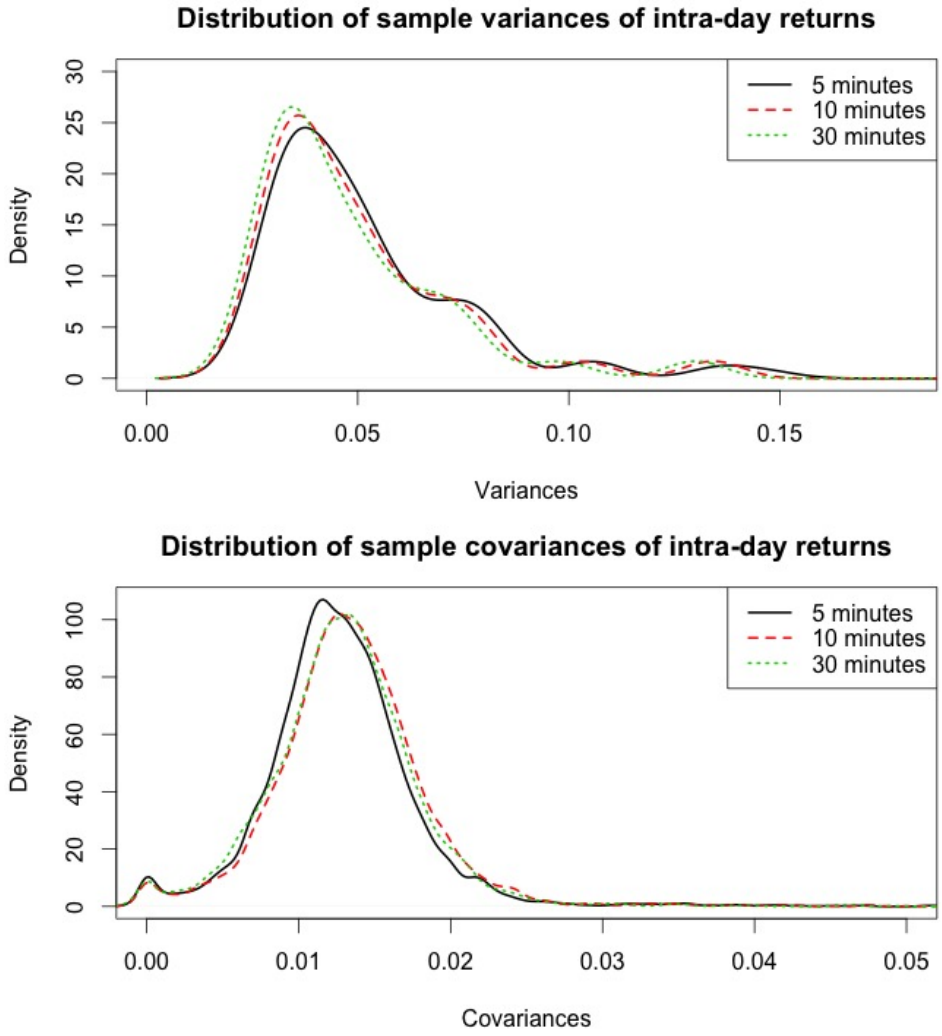


Figure 4.2: Distribution of Annualised sample variances and covariances of intra-day returns of the FTSE 100 constitutes from March 2nd 2015 to September 4th 2015. Sampling frequency= 5, 10, 30 minutes.

to be contradictory to Figure 4.2 and 4.3. The reason for this is that the formula (4.6) we used for annualisation are the simplest and broadly adopted in literature, although Lo [2002] argues that they could only used when there is no serial correlation, i.e. i.i.d. portfolio returns. The annualised results based on formula (4.6) can yield standard deviations that are considerably smaller (in the case of negative serial correlation)

or larger (in the case of positive serial correlation). Again, since we focus on comparison instead of estimation in this section, we keep these neat formulas instead of using a more complicated annualisation factor given by Lo [2002].

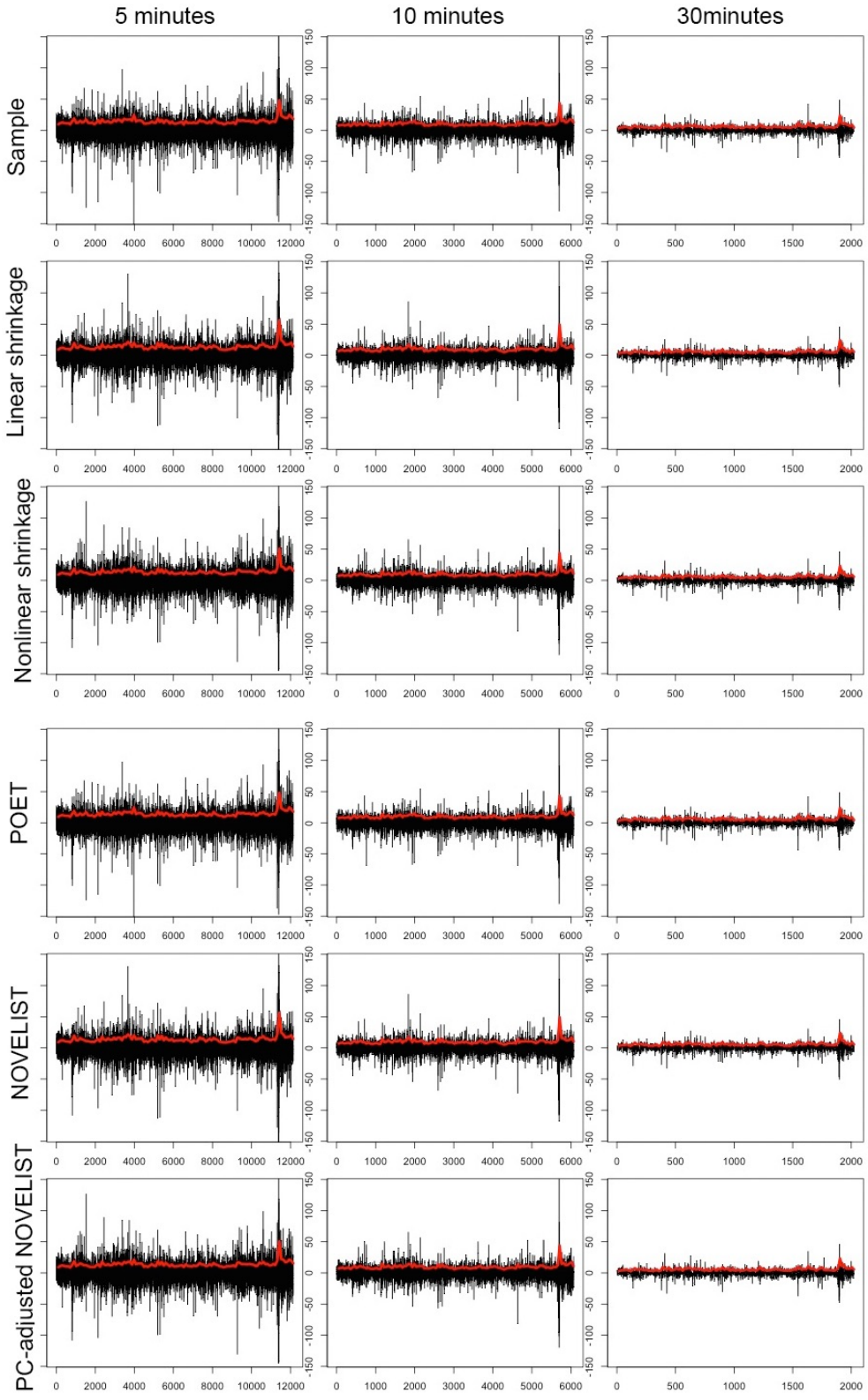


Figure 4.3: Time series plots of six minimal variance portfolio returns and STDs based on intra-day data.

Table 4.3: Annualised portfolio returns, standard deviations (STDs) and Sharpe ratios of minimum variance portfolios (based on intra-day data) as in formula (4.6). The best results are boxed.

	Annualised portfolio returns (%)	out-of-sample STDs (%)	Sharpe ratios
	Regime 3: intra-day 5 minutes		
Sample	-17.949	8.856	-1.238
Linear shrinkage	-18.612	9.190	-1.597
Nonlinear shrinkage	-17.150	8.680	-1.124
POET	-17.422	8.944	-1.309
NOVELIST	-16.288	8.695	-0.862
PC-adjusted NOVELIST	-17.453	9.015	-1.118
	Regime 4: intra-day 10 minutes		
Sample	-20.464	9.329	-1.748
Linear shrinkage	-20.936	9.425	-1.754
Nonlinear shrinkage	-17.010	8.190	-1.335
POET	-19.770	8.987	-1.540
NOVELIST	-15.814	8.749	-0.786
PC-adjusted NOVELIST	-16.591	9.018	-1.068
	Regime 5: intra-day 30 minutes		
Sample	-23.816	12.172	-3.272
Linear shrinkage	-16.043	10.222	-1.571
Nonlinear shrinkage	-7.830	8.930	0.489
POET	-8.557	9.217	1.623
NOVELIST	-11.681	8.866	0.920
PC-adjusted NOVELIST	-13.346	9.052	0.429

4.3 Forecasting the number of calls for a call center

In this section we present the performance of NOVELIST in estimating large covariance matrix by an application in forecasting the call arrival pattern at a telephone call centre, in particular, the number of arrivals later in a day using arrival patterns at earlier times of the day. Similar competitions were previously conducted to compare the performance of different covariance matrix estimators [Bickel and Levina, 2008a; Huang et al., 2006; Lam, 2016].

4.3.1 Dataset

The data come from one call centre in a major U.S. northeastern financial organisation, containing every call arrival time. For each day in the Year 2002, after removing weekends, holidays and the days when the data-collecting equipment was out of order, we obtain observations for 239 days. Phone calls were recorded from 7 am until midnight every day, and the 17-hour period is divided into 102 ten-minute intervals, and the number of calls arriving at the service queue during each interval are recorded. According to Huang et al. [2006], interval length of 10 minutes is chosen rather subjectively as a way of smoothing the data and for illustration.

4.3.2 Phone calls forecasting

We denote N_{ij} as the number of calls arrives during the jth ten-minute interval on the ith day, $i = 1, 2, \cdots, 239$, $j = 1, 2, \cdots, 102$. As suggested, we first take a square root transformation to make the data distribution close to normal [Brown et al., 2005;

Huang et al., 2006],

$$y_{ij} = \sqrt{N_{ij} + 1/4}. \tag{4.7}$$

We can forecast the number of call arrivals later in the day by using call arrival patterns at earlier times of the day. We evenly partition $\boldsymbol{y_i}$ into $\boldsymbol{y_i}^{(1)}, \boldsymbol{y_i}^{(2)}$, where $\boldsymbol{y_i}^{(1)} = (y_{i,1}, y_{i,2}, \cdots, y_{i,51})$ and $\boldsymbol{y_i}^{(2)} = (y_{i,52}, y_{i,53}, \cdots, y_{i,102})$. Correspondingly, the mean and variance matrix are partitioned as follows

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{4.8}$$

Assuming multivariate normality, the best mean squared error forecast of $\boldsymbol{y_i}^{(2)}$ using $\boldsymbol{y_i}^{(1)}$ is

$$\hat{\boldsymbol{y}}_i^{(2)} = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{y}_i^{(1)} - \boldsymbol{\mu}_1). \tag{4.9}$$

Clearly, we need to plug in estimates of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, Σ_{11} and Σ_{21}^{-1}. By replacing $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ with the sample means $\bar{\boldsymbol{y}_i}^{(1)}$ and $\bar{\boldsymbol{y}_i}^{(2)}$, and applying NOVELIST and other covariance and precision matrix estimators to estimate Σ_{11} and Σ_{21}^{-1}, we can obtain $\hat{\boldsymbol{y}}_i^{(2)}$.

In order to evaluate the performance of different estimators, we split the 239 days into training and test datasets, see Table 4.4 for details. Forecast 1 to 3 is designed for comparing the performance of NOVELIST to existing papers in which this application is also considered, especially the results in Lam [2016]. They have same test dataset

Table 4.4: Allocation of training and test datasets for forecast 1 to 6.

	Training			Test		
	Sample size (N_1)	Start day (n_1)	End day (n_2-1)	Sample size (N_2)	Start day (n_2)	End day
Forecast 1	30	181	210	29	211	239
Forecast 2	120	91	210	29	211	239
Forecast 3	210	1	210	29	211	239
Forecast 4	180	1	180	59	181	239
Forecast 5	90	1	90	149	91	239
Forecast 6	30	1	30	209	31	239

but different-length training dataset. Forecast 3 to 6 changes the ratio of the length of training and test datasets to see the accuracy of call arrival forecasting if training window is shorter and test window is longer. We take forecast 3 as a example, it contains the training dataset from the first 210 days, roughly corresponding to January to October, which is used to estimate the mean and covariance structure. The estimates are then applied on forecasting using formula (4.9) for the 29 days in the test set, corresponding to the remaining days of the year. We compare the average absolute forecast error of the 29 days which is defined by

$$AE_t = \frac{1}{N_2} \sum_{i=n_2}^{239} |\hat{\boldsymbol{y}}_{i,t}^{(2)} - \boldsymbol{y}_{i,t}^{(2)}| \tag{4.10}$$

where $\hat{\boldsymbol{y}}_{i,t}^{(2)}$ and $\boldsymbol{y}_{i,t}^{(2)}$ are the observed and forecast values respectively, $n_2 = 211$ and $N_2 = 29$ for forecast 3.

We compete NOVELIST estimators with six other covariance and precision esti-

mators: the first four are the same as those in Section 4.2 and another two, NERCOME and CRC grand average, which are taken from Lam [2016] for comparisons in forecast 1 to 3. We take NERCOME and CRC grand average because they outperform in several cases in Lam [2016]. The decision procedure for NOVELIST indicates underlying factor models, and we use fixed parameters again in this application.

4.3.3 Results

Table 4.5 shows the results. NOVELIST outperforms other estimators in all seven forecast, followed by nonlinear shrinkage, NERCOME and CRC, which have roughly the same results in forecast 1 to 3, and followed by PC-adjusted NOVELIST and nonlinear shrinkage in forecast 4 to 6. POET and sample covariance perform the worse, and also PC-adjusted NOVELIST in forecast 1 to 3. Figure 4.4 shows that forecast is less accurate during the middle times of the second half of the day (roughly from 17:00 to 22:00) than at the beginning or at the end (from 15:30 to 17:00 or from 22:00 to 24:00). For nearly every ten-minute interval of the second half of the day (apart from a few intervals at the beginning), there is more than half chance that NOVELIST outperforms other methods. From 17:00 to 22:00 roughly, NOVELIST even has more than 80% chance to beat others. Also, comparison among forecast 1 to 3 tells us that having shorter training window increases the 29-day forecast error. However, Figure 4.5 shows that forecast accuracy can be good even when we have a small training to test ratio, for example, forecast 6 surprisingly performs well with only 30 days in training dataset and 209 days in test dataset. This tells us that the covariance structure of $\boldsymbol{y}_i$s can be viewed as unchanged for a long period. But, we notice the discrepancies between forecast and true call arrivals after about day 200, which may indicate a change

point of covariance structure near that day. This may give the reason why forecast 4 to 6 perform better than forecast 1 to 3 in general, that is because forecast 1 to 3 have training and test periods on the two sides of a possible change point.

Table 4.5: Mean absolute forecast errors and standard deviations (in brackets) of forecast 1 to 6. The best results are boxed.

	Forecast					
	1	*2*	*3*	*4*	*5*	*6*
Sample	—	1.603 (0.472)	1.532 (0.487)	1.247 (0.326)	—	—
POET	1.652 (0.581)	1.626 (0.531)	1.569 (0.542)	1.231 (0.346)	0.918 (0.200)	0.859 (0.160)
Linear	1.570 (0.415)	1.645 (0.503)	1.548 (0.494)	1.209 (0.316)	0.952 (0.194)	0.919 (0.140)
Nonlinear	1.481 (0.523)	1.597 (0.524)	1.523 (0.510)	1.167 (0.319)	0.892 (0.193)	0.824 (0.154)
NOVELIST	1.419 (0.458)	1.458 (0.466)	1.463 (0.491)	1.027 (0.247)	0.802 (0.155)	0.800 (0.135)
PC-adjusted NOVELIST	1.677 (0.568)	1.676 (0.553)	1.571 (0.509)	1.116 (0.277)	0.821 (0.158)	0.846 (0.177)
NERCOME*	1.45 (0.45)	1.59 (0.51)	1.53 (0.51)	—	—	—
CRC*	1.46 (0.50)	1.59 (0.52)	1.54 (0.51)	—	—	—

Note: The methods labelled with $*$ are taken from Lam [2016] for comparison and the decimal places are different.

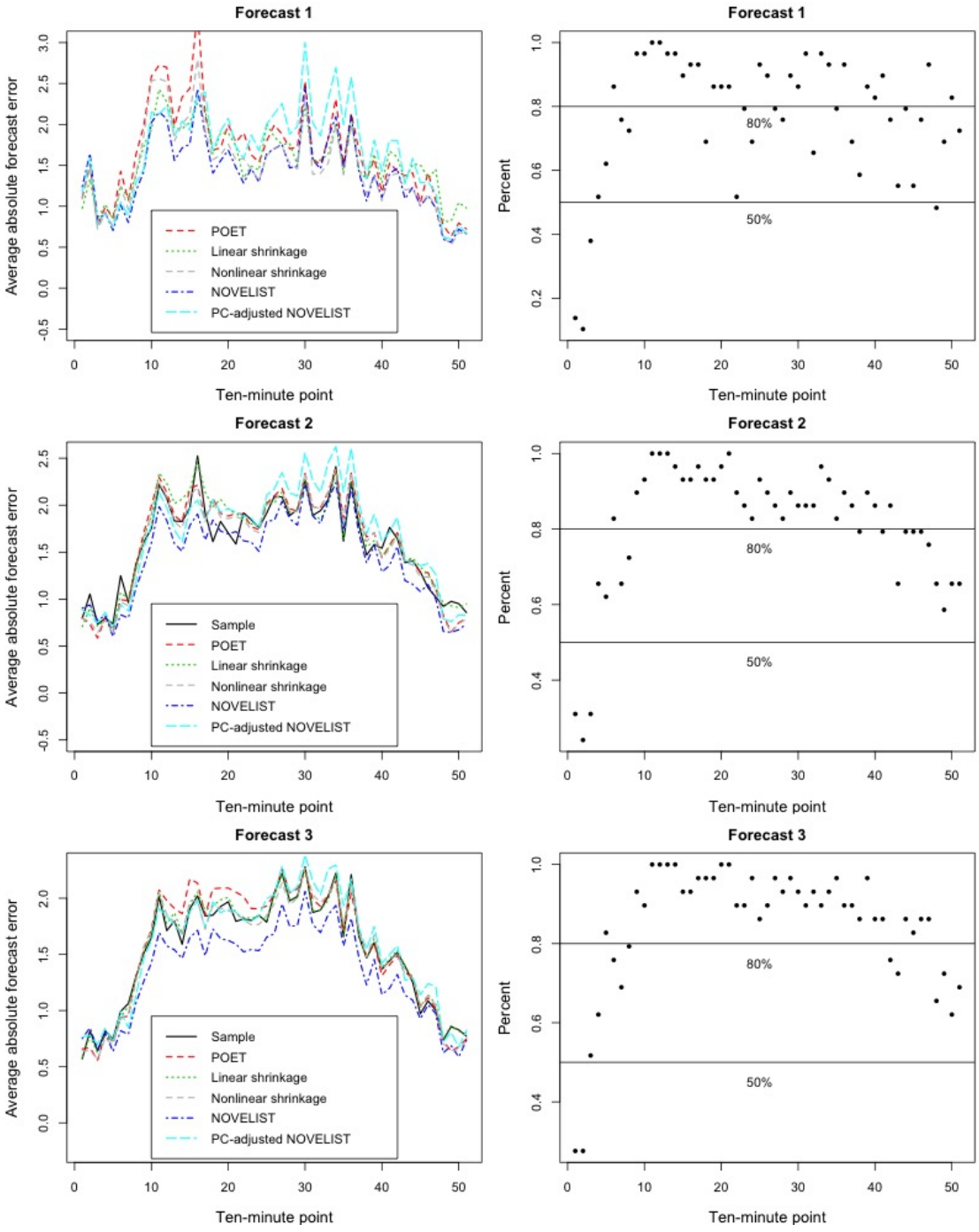


Figure 4.4: Competitions of call forecasting based on forecast 1 to 3. Left: plots of average absolute errors for the forecasts using different estimators. Right: percentage of days (29 of them) in the test dataset when the NOVELIST based forecast outperforms for each ten-minute interval at later times in the day.

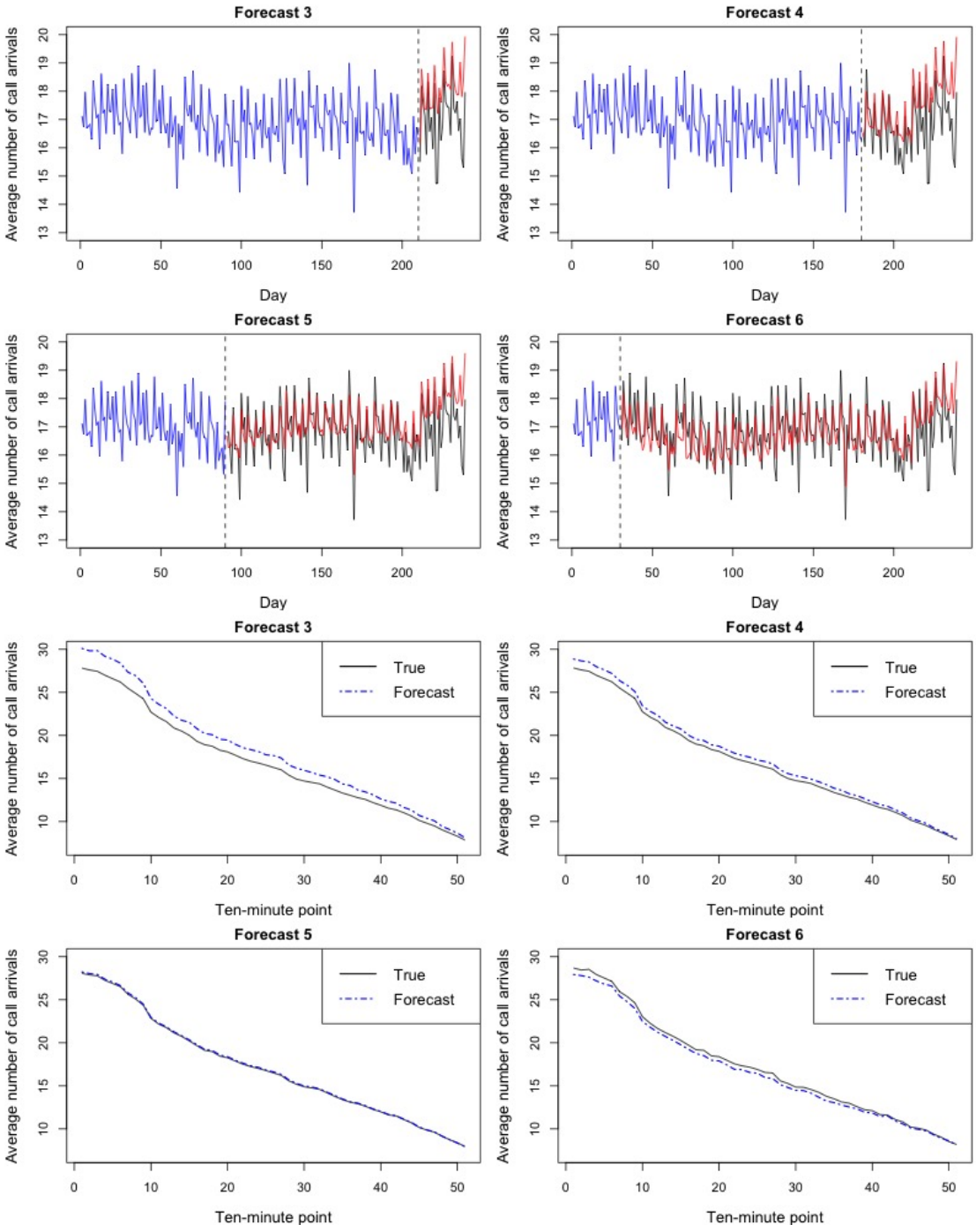


Figure 4.5: Accuracy of forecasting telephone calls based on NOVELIST estimators for forecast 3 to 6. Top: daily average number of call arrivals of training (blue), test (black) and forecast (red) data. Bottom: true and predicted average number of call arrivals during each ten-minute interval at later times of the days within test windows.

4.4 Estimation of false discovery proportion of large-scale multiple testing with unknown dependence structure

In this section, we estimate false discovery proportion (FDP) of dependent test statistics in large-scale multiple testing by using NOVELIST covariance matrix estimator. Similar application was previously considered by using POET estimator in Fan and Han [2013].

4.4.1 Notation, setting and method

4.4.1.1 FDP under dependence structure

Suppose that $\{\boldsymbol{X}_i\}_{i=1}^n$ are n i.i.d. observations of a p-dimensional random variable, where each $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} = \{\mu_1, \mu_2, \cdots, \mu_p\}$ and $\boldsymbol{\Sigma} = \{\sigma_{i,j}\}$, $1 \leqslant i, j \leqslant p$. Under high dimensional setting, i.e. $p > n$, the mean vector $\boldsymbol{\mu}$ is assumed to be a sparse vector containing only a small number of nonzero entries. More precisely, we denote $\mathcal{P}_0 = \{1 \leqslant j \leqslant p : \mu_j = 0\}$, $\mathcal{P}_1 = \{1 \leqslant j \leqslant p : \mu_j \neq 0\}$, $p_0 = |\mathcal{P}_0|$ and $p_1 = |\mathcal{P}_1|$, and we assume that $p_1/p \to 0$ as $p \to \infty$. In practice, the subsets $\mathcal{P}_0$ and $\mathcal{P}_1$ are unknown, and we want to identify the nonvanishing signals within $\mathcal{P}_1$.

We consider $\boldsymbol{Z} = \sqrt{n}\bar{\boldsymbol{X}}$, where $\bar{\boldsymbol{X}}$ is the sample mean of $\{\boldsymbol{X}_i\}_{i=1}^n$, i.e. $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$. Hence, we have $\boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $(\boldsymbol{Z} - \sqrt{n}\boldsymbol{\mu})\boldsymbol{D}^{-\frac{1}{2}} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{R})$ after standardisation, where $\boldsymbol{D} = \text{diag}(\sigma_{1,1}, \sigma_{2,2}, \cdot\cdot, \cdot, \sigma_{p,p})$ is a diagonal matrix which consist of all the diagonals of the covariance matrix and $\boldsymbol{R}$ is the correlation matrix. In order to identify the nonzero entries in the mean vector $\boldsymbol{\mu}$, we use multiple test statistics

$\boldsymbol{Z}^* = \{Z_1^*, Z_2^*, \cdots, Z_p^*\}$. For each $j \in (1, p)$,

$$Z_j^* = \frac{\sqrt{n}\bar{\boldsymbol{X}}_j}{\sqrt{\sigma_{j,j}}}, \tag{4.11}$$

where $\mu_j^* = \frac{\sqrt{n}\mu_j}{\sqrt{\sigma_{j,j}}}$ and we consider multiple testing

$$H_{0j} : \mu_j^* = 0 \ \text{ vs } \ H_{1j} : \mu_j^* \neq 0 \tag{4.12}$$

based on $\boldsymbol{Z}^*$, which is equivalent to test

$$H_{0j} : \mu_j = 0 \ \text{ vs } \ H_{1j} : \mu_j \neq 0. \tag{4.13}$$

The p-value for the j^{th} hypothesis is $P_j = 2\Phi(-|Z_j^*|)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. For a chosen threshold value t, we reject H_{0j} if $p_j < t$. Then, we want to know the accuracy of this multiple testing. Define the number of discoveries as $R(t) = \#\{j : P_j \leq t\}$ and the number of false discoveries as $V(t) = \#\{\text{true null } j : P_j \leq t\}$. Our aim is to estimate the false discovery proportion $FDP(t) = V(t)/R(t)$. $R(t)$ is observed but $V(t)$ needs to be estimated in order to obtain the estimated $FDP(t)$.

If there is no dependence among these j testing, the number of false discoveries $V(t)$ should go to $p_0 t$ asymptotically, which leads to $FDP \to p_0 t/R(t)$ asymptotically. However, if there exist dependence of the test statistics, Fan and Han [2013] show how the dependence impacts on the FDP by considering the following one-

factor model, if we assume for simplicity that $\Sigma = R$. For each j, we consider

$$Z_j^* = \mu_j^* + b_j W + a_j \epsilon_j, \tag{4.14}$$

where $a_j = (1 - b_j^2)^{1/2}$, W is the common factor and each ϵ_j is a random noise, and they follow independent standard normal distribution. Then, under the null hypothesis $H_{0j} : \mu_j^* = 0$ for all j, we have the number of false discoveries is

$$\begin{aligned} V(t) &= \sum_{j \in \mathcal{P}_0} \mathbf{I}(2\Phi(-|Z_j^*|) < t) \\ &= \sum_{j \in \mathcal{P}_0} \mathbf{I}(|b_j W + a_j \epsilon_j| > z_{t/2}) \\ &= \sum_{j \in \mathcal{P}_0} [\mathbf{I}(b_j W + a_j \epsilon_j > -z_{t/2}) + \mathbf{I}(b_j W + a_j \epsilon_j < z_{t/2})] \\ &= \sum_{j \in \mathcal{P}_0} [\mathbf{I}(\epsilon_j > -\frac{z_{t/2} + b_j W}{a_j}) + \mathbf{I}(\epsilon_j < \frac{z_{t/2} - b_j W}{a_j})]. \end{aligned} \tag{4.15}$$

where $z_{t/2}$ is the $t/2$-quantile of the standard normal distribution. We assume that $p_0 = |\mathcal{P}_0|$ is big enough and each ϵ_j is independent, then we can apply the weak law of large numbers [Davidson, 1994, p.289]. Conditioning on W, we have

$$V(t) \approx \sum_{j \in \mathcal{P}_0} [\Phi(\frac{z_{t/2} + b_j W}{a_j}) + \Phi(\frac{z_{t/2} - b_j W}{a_j})], \tag{4.16}$$

Formula (4.16) quantifies the dependence of $V(t)$ and the corresponding $FDP(t) = V(t)/R(t)$ on the realisation of W. However, $\mathcal{P}_0$, b_j and W are unknown. Since we assume sparsity: $p_1/p \to 0$ as $p \to \infty$, the set of true nulls $\mathcal{P}_0$ is nearly the whole set, but we also need to estimate W which can be viewed as a regression problem and achieved for example by least squares estimation or L_1 penalised regression [Fan

et al., 2012a], or be solved by principal factor approximation (PFA) [Fan and Han, 2013; Fan et al., 2012a]. We apply the NOVELIST covariance matrix estimator on PFA for estimating FDP.

4.4.1.2 Estimation of FDP by PFA

In this section, we briefly present the PFA procedure for estimation of FDP introduced by Fan et al. [2012a] and Fan and Han [2013]. The basic idea is to use principal components as approximated factors, more precisely, it takes out the first k principal components that derive the strong dependence among observed data to estimate the common factors under the approximate factor model and provides a consistent estimate of the realized FDP.

Consider an approximate factor model for the test statistics $\boldsymbol{Z}_i^*$ as

$$\boldsymbol{Z}_i^* = \boldsymbol{\mu}^* + \boldsymbol{B}\boldsymbol{f}_i + \boldsymbol{u}_i \tag{4.17}$$

for each observation, where $\boldsymbol{\mu}^*$ is a p-dimensional unknown sparse vector, $\boldsymbol{B} = (b_1, b_2 \cdot \cdot, b_p)^T$ is the factor loading matrix, $\boldsymbol{f}_i$ are k common factors to the ith observations, independent of the noise $\boldsymbol{u}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$, where $\boldsymbol{\Sigma}_u$ is sparse. The PFA procedure for estimating FDP is as follows,

(1) Estimating the covariance matrix $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{Z}^*$.

(2) Apply singular value decomposition to the covariance matrix $\hat{\boldsymbol{\Sigma}}$. Obtain the first k eigenvalues $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_k$ and the corresponding eigenvectors $\hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2, \cdots, \hat{\boldsymbol{\gamma}}_k$.

(3) Construct $\hat{\boldsymbol{B}} = (\hat{\lambda}_1^{1/2}\hat{\boldsymbol{\gamma}}_1, \hat{\lambda}_2^{1/2}\hat{\boldsymbol{\gamma}}_2, \cdots, \hat{\lambda}_k^{1/2}\hat{\boldsymbol{\gamma}}_k)$, and compute the least squares estimate $\hat{\boldsymbol{f}}^* = (\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}})^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{Z}^*$.

(4) With $\hat{b}_i^T$ is the i^{th} row of $\hat{\boldsymbol{B}}$, compute

$$\widehat{FDP}(t) = \Sigma_{i=1}^{p}[\Phi((z_{t/2} + \hat{b}_i^T \hat{\boldsymbol{f}}^*)/\hat{a}_i) + \Phi((z_{t/2} - \hat{b}_i^T \hat{\boldsymbol{f}}^*)/\hat{a}_i)]/R(t) \tag{4.18}$$

where $\hat{a}_i = (1 - ||\hat{b}_i^T||)^{1/2}$.

4.4.2 Breast cancer dataset

We use the breast cancer dataset which is considered by Fan and Han [2013] and Hedenfalk et al. [2001] in Large-scale hypothesis testing problem, and also used by Efron [2007] in breast cancer gene-expression study. This dataset consists of gene expression levels in 15 patients. The first group includes 7 women with BRCA1 and the second group includes 8 women with BRCA2, both BRCA1 and BRCA2 are known to increase the lifetime risk of hereditary breast cancer. We observe $p = 3226$ gene expression levels for each group. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$, $n = 7$, denote the microarray of expression levels on the 3226 genes for the first group, and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_m$, $m = 8$, for that of the second group. Identifying the significantly different genes expressed by BRCA1 carriers and BRCA2 carriers will allow scientists to discriminate the cases of hereditary breast cancer on the basis of gene-expression profiles.

We assume that the gene expression levels of the two groups follows two multivariate normal distributions with different mean vector but the same covariance matrix. Let $\boldsymbol{X}_i \sim \boldsymbol{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $i = 1, 2, \cdots, n$ and $\boldsymbol{Y}_i \sim \boldsymbol{N}_p(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ for $i = 1, 2, \cdots, m$. We use the following multiple hypothesis test to identify the genes distinctively expressed by the patients in the two groups. For each gene j, we consider two-sample testing

$$H_{0j} : \mu_j = \nu_j \ \text{ vs } \ H_{1j} : \mu_j \neq \nu_j, \tag{4.19}$$

based on the test statistics

$$Z_j^* = \frac{\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{Y}}_j}{\hat{\sigma}_{j,j}\sqrt{\frac{1}{n}+\frac{1}{m}}}, \tag{4.20}$$

which follows t_{n+m-2} distribution. It is also reasonable to assume that a large proportion of the genes are not differentially expressed, so that $\boldsymbol{\mu}-\boldsymbol{\nu}$ is sparse. By choosing a threshold level t, we can obtain the subset of discoveries which includes the differently expressed genes by the BRCA1 and BRCA2 carriers based on the testing. Then we use cross validated NOVELIST to estimate the covariance matrix of Z_j^* and apply the PFA procedure described in Section 4.4.1.2 to estimate FDP of the testing.

4.4.3 Results

The results of our analysis are presented in Figure 4.6. Firstly, the estimated FDP increases as the threshold value t increases, which indicates that the discoveries with lower t have higher accuracy to be the true discoveries, for example, when the number of discoveries is below 200, the estimated number of false discoveries is close to zero, for number of factors $k \leqslant 15$. Secondly, although it is claimed that the PFA procedure for estimating FDP is robust under different choices of number of factors k between 2 to 5 in Fan et al. [2012a] and Fan and Han [2013], we choose k up to 13 and observe obvious discrepancies in the estimated FDP. The smallest $\widehat{FDP}$ is obtained when $k = 13$. For example, when the number of discoveries is 1000, the $\widehat{FDP}$ is below 50 with $k = 13$, by contrast, the $\widehat{FDP}$ is around 250 with $k = 2$. It indicates that suitable choice of k is important for accurately estimating the FDP. Moreover, alough $k = 13$ yields the smallest $\widehat{FDP}$, we note that the sample size is only 15, and taking $k = 13$ makes no much sense in terms of approximate factor models and may distort

the results, we argue that the low $\widehat{FDP}$ produced by $k = 13$ may underestimate the true FDP. However, the true FDP is unknown for this study, although similar results are obtained in Fan and Han [2013], we are unable to compare and conclude which one has the more accurate $\widehat{FDP}$.

In order to compare the results of gene discoveries in this study and those in other literature, we present the list of the 51 most differentially expressed genes in BRCA1 and BRCA2 carriers in this study in Table 4.6, and 51 genes that are best differentiated among BRCA1-Mutation-Positive, BRCA2-Mutation-Positive, and another breast cancer related tumor by a modified F test in Hedenfalk et al. [2001] in Figure 4.7. There are 25 out of 51 genes that coincide. Since the significance level is 8.116×10^{-6} in our study versus 0.001 in Hedenfalk et al. [2001], this multiple testing is much more sensitive than the modified F test in Hedenfalk et al. [2001]. In this testing, if the significance level is 0.001, we will identify around 170 differently expressed genes.

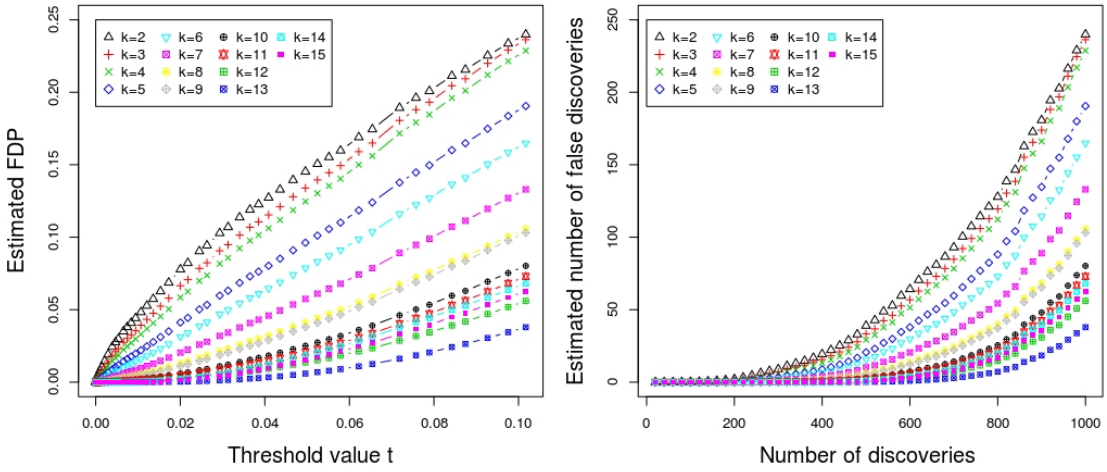


Figure 4.6: The estimated false discovery proportion as function of the threshold value t and the estimated number of false discoveries as function of the number of total discoveries for $p = 3226$ genes in total. The number of factors $k \in (2, 15)$.

Some remarks: The difficulties in this study are due to the high dimension and the

very low sample size. Since n is only 15, it is difficult for NOVELIST to find the suitable parameters via cross validation. However, there is no widely accepted consensus in terms of the true subset of the differently expressed genes for BRCA1 and BRCA2 carriers for this study, we can only provide the estimated FDP using NOVELIST, but cannot evaluate the goodness of the cross validation and the accuracy of the estimation.

Table 4.6: 51 most distinctively expressed genes that can discriminate breast cancers with BRCA1 mutations from those with BRCA2 mutations (threshold level t is 8.116×10^{-6}). The estimated FDP by using NOVELIST is approximately 0.012% under approximate factor model with 5 factors.

Clone ID	UniGene Title
810057	cold shock domain protein A
46182	CTP synthase
813280	adenylosuccinate lyase
950682	phosphofructokinase, platelet
897646	splicing factor, arginine/serine-rich 4
840702	SELENOPHOSPHATE SYNTHETASE ; Human selenium donor protein
712604	pre-B-cell colony-enhancing factor
784830	D123 gene product
841617	Human mRNA for ornithine decarboxylase antizyme, ORF 1 and ORF 2
686172	KIAA0008 gene product
563444	forkhead box F1
711680	zinc finger protein, subfamily 1A, 1 (Ikaros)
949932	nuclease sensitive element binding protein 1
75009	EphB4
566887	chromobox homolog 3 (Drosophila HP1 gamma)
841641	cyclin D1 (PRAD1: parathyroid adenomatosis 1)
214731	KIAA0601 protein
809981	glutathione peroxidase 4 (phospholipid hydroperoxidase)
236055	DKFZP564M2423 protein
293977	ESTs, Weakly similar to putative [C.elegans]
295831	ESTs, Highly similar to CGI-26 protein [H.sapiens]
236129	Homo sapiens mRNA; cDNA DKFZp434B1935 (from clone DKFZp434B1935)
247818	ESTs
139354	ESTs
127099	ESTs, Moderately similar to atypical PKC specific binding protein [R.norvegicus]
814270	polymyositis/scleroderma autoantigen 1 (75kD)
130895	ESTs
344352	ESTs
31842	UDP-galactose transporter related
133178	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1
548957	general transcription factor II, i, pseudogene 1
212198	tumor protein p53-binding protein, 2
293104	phytanoyl-CoA hydroxylase (Refsum disease)
82991	phosphodiesterase I/nucleotide pyrophosphatase 1 (homologous to mouse Ly-41 antigen
32790	mutS (E. coli) homolog 2 (colon cancer, nonpolyposis type 1)
291057	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)
344109	proliferating cell nuclear antigen
366647	butyrate response factor 1 (EGF-response factor 1)
366824	cyclin-dependent kinase 4
471918	intercellular adhesion molecule 2
361692	sarcoma amplified sequence
136769	TATA box binding protein (TBP)-associated factor, RNA polymerase II, A, 250kD
23014	mitogen-activated protein kinase 1
26082	very low density lipoprotein receptor
26184	phosphofructokinase, platelet
29054	ARP1 (actin-related protein 1, yeast) homolog A (centractin alpha)
36775	hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzy
42888	interleukin enhancer binding factor 2, 45kD
45840	splicing factor, arginine/serine-rich 4
46019	minichromosome maintenance deficient (S. cerevisiae) 7
51209	protein phosphatase 1, catalytic subunit, beta isoform height

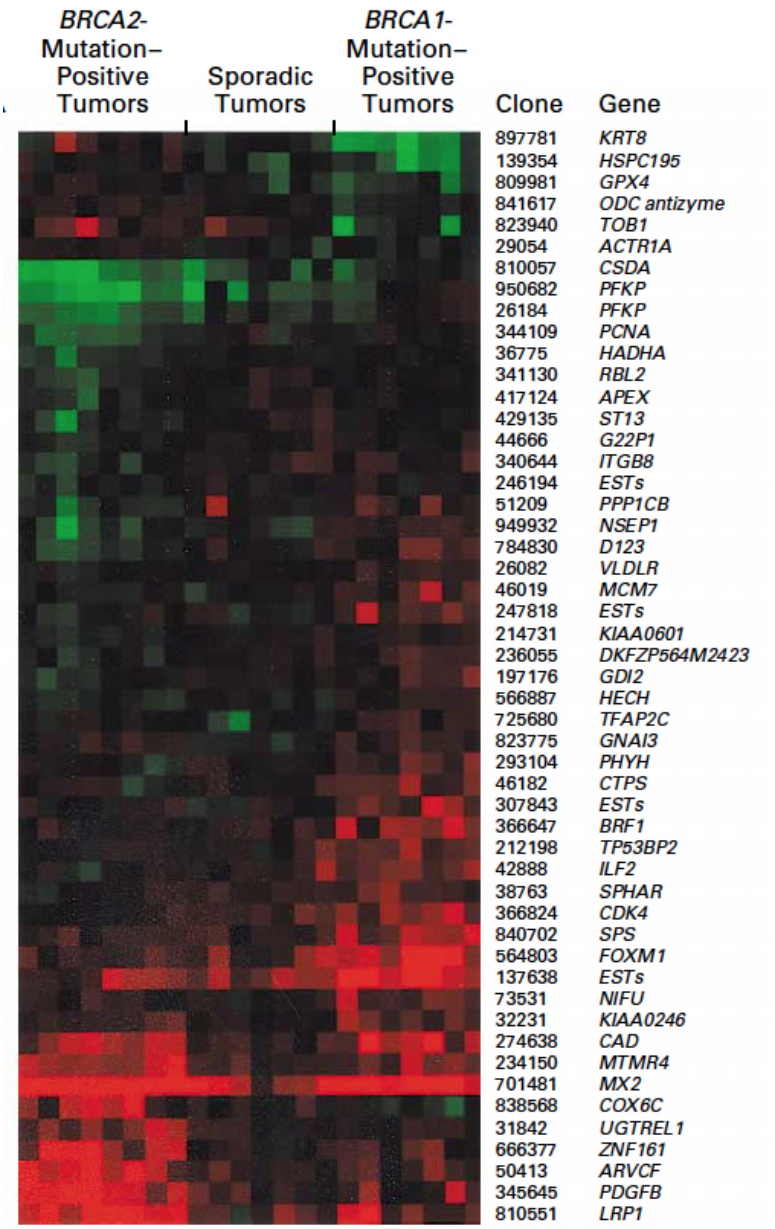


Figure 4.7: Panel A of figure 2 in Hedenfalk et al. [2001]: 51 genes that are best differentiated among BRCA1-Mutation-Positive, BRCA2-Mutation-Positive, and another breast cancer related tumor, as determined by a modified F test $(\alpha = 0.001)$, for comparison with Table 4.6.

4.5 Conclusion

This chapter applies NOVELIST estimators on real data, including portfolio optimisation using low-frequency and high-frequency FTSE 100 constituents log returns, forecasting the number of calls for a call center and estimating false discovery proportion through a well-known breast cancer study.

For minimum-variance portfolio optimisation, NOVELIST performs well and stable for daily data, where it has the lowest volatility and the highest Sharpe ratios, but is beaten by others for maximising the portfolio returns, which is mainly because the purpose of this portfolio optimisation is risk minimisation instead of return maximisation. For intra-day data, NOVELIST performs less stable due to microstructure noises as sampling frequency increases. In general, increasing sampling frequency has negative effects on risk minimisation and return/sharp ratio maximisation in this example.

In the example of the call center phone arrival forecast, NOVELIST outperforms other estimators in all seven forecast (different training and test datasets), followed by nonlinear shrinkage, NERCOME and CRC. The call arrival forecast by using NOVELIST is good and stable even when training to test ratio is small (30 days in the training dataset and 209 days in the test dataset). But, its performance can be highly affected by change points, which indicates that ensuring stationarity or detecting change points are important before applying NOVELIST estimation.

In the application on estimation of FDP of large-scale multiple testing by using a breast cancer dataset, the final results show that FDP increases as the number of the discoveries increases, and the most differentially expressed genes found by using NOVELIST has about 50% overlap with those from existing literature. However, NOVELIST is not compared with other estimators. It is because that there is no widely

accepted consensus in terms of the true subset of the differently expressed genes for BRCA1 and BRCA2 carriers for this study and the true FDP is unknown, there is no sense to compare the accuracy of the estimations. Another difficulty is the ultra low sample size (only 15 patients compared to 3226 gene expression levels), and it is difficult for NOVELIST to find the suitable parameters via cross validation. Nonetheless, we still consider this area as an important one where NOVELIST estimator can be applied on and further improvements can be made.

Based on the overall performance of all the competitors in these applications, we argue that NOVELIST is the overall winner, followed by nonlinear shrinkage. Again, it is due to the flexible control of the degree of shrinkage and thresholding offered by NOVELIST.

Chapter 5

Conclusion and future work

The thesis concerns estimating large correlation and covariance matrices and their inverses. The main focus is put on the two new methods proposed and the related applications.

Firstly, tilting-based methods are proposed to estimate the large precision matrix block by block. Each block can be estimated by the inversion of the corresponding pairwise sample conditional covariance matrix controlling all the other variables. To determine the controlling subsets, four types of tilting-based methods are introduced as variable selection techniques that aim to only put the highly relevant remaining variables into the controlling subsets. The asymptotic properties and the finite sample performance of the methods are demonstrated. The simulation study shows that separate tilting (with thresholding afterwards) performs well for (absolute and relative) diagonal block models, and competing tilting is the best when high collinearity exists, such as factor models, but all the tilting methods are beaten by thresholding methods for the diagonal precision matrix. The fact that adding a thresholding step after applying tilting methods improves the results indicates that tilting estimators perform well

in estimating rather than identifying the non-zero entries. Competing tilting is only recommended for use when necessary, as it requires much more computational time and efforts compared to other tilting methods. Also, when we face the (ultra) high-dimensional cases, we need to use competing tilting with caution, since it is highly affected by the distorted realisations of the variables and the residuals. Suitable improvement approaches can be applied depending on circumstances. In general, the higher collinearity the variables have, the more necessary it is to apply tilting methods, especially the competing tilting.

Secondly, we propose the NOVELIST methods for correlation/covariance and their inverses, which performs shrinkage of the non-sparse and low-rank sample version towards the sparse thresholded target. The benefits of the NOVELIST estimator include simplicity, ease of implementation, computational efficiency and the fact that its application avoids eigenanalysis. The linkage between NOVELIST and ridge regression are demonstrated. We obtain an explicit convergence rate in the operator norm over a large class of covariance (correlation) matrices when p and n satisfy $\log p/n \to 0$. Empirical choices of parameters and a data-driven algorithm for NOVELIST estimators are presented. Comprehensive simulation study are based on a wide range of models and shows that NOVELIST works best when the underlying correlation/covariance matrices are sparse and non-sparse (more so for heteroscedastic models) but is beaten by POET for the highly non-sparse models by a small margin. For the highly non-sparse cases, we improve the performance of the NOVELIST precision matrix estimation by applying fixed parameters that come from the robustness test instead of the cross-validated ones and the automatic algorithm is presented. Overall, it is clear that the flexible control of the degree of shrinkage and thresholding offered by NOVELIST means that it is able to offer competitive performance across most models, and in situ-

ations in which it is not the best, it tends not to be much worse than the best performer. We recommend NOVELIST as a simple, good all-round covariance, correlation and precision matrix estimator ready for practical use across a variety of models and data dimensionalities.

Lastly, we also apply NOVELIST estimators on real data examples, including portfolio optimisation, call arrival forecasting and FDP estimation. First, NOVELIST works well in the aim of minimum-variance portfolio optimization, but performs less stable due to microstructure noises as sampling frequency increases. Second, in the example of the call center phone arrival forecast, NOVELIST outperforms other estimators in all seven forecast (different training and test datasets), but its performance can be highly affected by change points, which indicates that ensuring stationarity or detecting change points are important before applying NOVELIST estimation. Third, in the application on estimation of FDP of large-scale multiple testing by using a breast cancer dataset, final results show that FDP increases as the number of the discoveries increases, and the most differentially expressed genes found by using NOVELIST has about 50% overlap with those from existing literature. However, further work is needed to investigate the accuracy of the NOVELIST estimation compared to other competitors. Therefore, we argue that NOVELIST is the overall winner in these applications, followed by nonlinear shrinkage.

Future research can be made from two aspects. First, the tilting and NOVELIST methods can be extended from i.i.d variables to dependent data. Sancetta [2008] generalises the linear shrinkage method by Ledoit and Wolf [2004] to serially correlated data. Fiecas et al. [2016] considers high-dimensional time series generated by a hidden Markov model which allows for switching between different regimes or states, and applies shrinkage with an EM-type algorithm to yield a more stable estimates of the

covariance matrix. We believe such approaches are worth trying to extend the tilting and NOVELIST methods to dependent data. Second, ensuring positive definiteness and invertibility of the correlation/covaraince matrices is mostly essential in practice. Although discussion regarding this is included in the thesis, further research is still needed to understand more in theory.

References

Y. Aït-Sahalia, P. A. Mykland, and L. Zhang. How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial studies*, 18:351–416, 2005. 122

I. Alvarez. Bayesian inference for a covariance matrix. Preprint, 2014. 6

T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. Modeling and forecasting realized volatility. *Econometrica*, 71:579–625, 2003. 122

T. G. Andersen, T. Bollerslev, F. X. Diebold, and J. Wu. Realized beta: Persistence and predictability. *Advances in econometrics*, 20:1–39, 2006. 121

T. W. Anderson. *An introduction to multivariate statistical analysis*. John Wiley Sons, New York, 1968. 2

F. M. Bandi and J. R. Russell. Realized covariation, realized beta and microstructure noise. Unpublished paper, Graduate School of Business, University of Chicago., 2005. 122

O. E. Barndorff-Nielsen and N. Shephard. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society B*, 64:253–280, 2002. 122

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 29:1165–1188, 2001. 120

D. S. Bernstein. *Matrix mathematics: theory, facts, and formulas, 2nd ed.* Princeton University Press, Princeton, NJ, 2009, p.147. 11, 31

P. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36:199–227, 2008a. 2, 3, 98, 120, 137

P. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36:2577–2604, 2008b. 3, 35, 67, 73, 81, 85, 87, 115

P. J. Bickel and E. Levina. Some theory for fisher's linear discriminant function,'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10:989–1010, 2004. 2, 119

J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98: 807–820, 2011. 51

H. Böhm and R. von Sachs. Structural shrinkage of nonparametric spectral estimators for multivariate time series. *Electronic Journal of Statistics*, 2:696–721, 2008. 5

H. Böhm and R. von Sachs. Shrinkage estimation in the frequency domain of multivariate time series. *Journal of Multivariate Analysis*, 100:913–935, 2009. 5

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100:36–50, 2005. 137

P. Bühlmann, M. Kalisch, and M. Maathuis. Variable selection for high-dimensional models: partial faithful distributions and the pc-algorithm. *Biomatrika*, 97:1–19, 2009. 13, 34

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106:672–684, 2011. 3, 35, 88

T. T. Cai, C. Zhang, and H. H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics*, 38:2118–2144, 2010. 3

CF. Chen. Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. *Journal of the Royal Statistical Society Series B*, 41:235–248, 1979. 5

J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771, 2008. 34, 62

X. Chen, M. Xu, and W. B. Wu. Covariance and precision matrix estimation for high-dimensional time series. *Annals of Statistics*, 41:2994–3021, 2013. 2

H. Cho and P. Fryzlewicz. High-dimensional variable selection via tilting. *Journal of the Royal Statistical Society Series B*, 74:593–622, 2012. 13, 14, 15, 16, 17, 18, 25, 27, 28, 29, 33, 45, 60, 69

C. Croux and G. Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87:603–618, 2000. 1, 119

A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse co-

variance selection. *SIAM Journal on Matrix Analysis and Applications*, 30:56–66, 2008. 4

J. Davidson. *Stochastic Limit Theory:An Introduction for Econometricians*. Oxford University Press, Oxford, 1994, p.289. 31, 65, 66, 146

V. DeMiguel and F. J. Nogales. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55: 798–812, 2009. 124

J. M. Dickey, D. V. Lindley, and S. J. Press. Bayesian estimation of the dispersion matrix of a multivariate normal distribution. *Communications in Statistics-Theory and Methods*, 14:1019–1034, 1985. 5

D. Edward. *Introduction to Graphical Modelling, 2nd ed.* Springer, New York, 2000. 8

B. Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102:93–103, 2007. 148

B. Efron. Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association*, 105:1042–1055, 2010. 120

N. El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annals of Statistics*, 36:2717–2756, 2008. 3, 83

I. G. Evans. Bayesian estimation of parameters of a multivariate normal distribution. *Journal of the Royal Statistical Society Series B*, 27:279–283, 1965. 5

J. Fan and X. Han. Estimation of false discovery proportion with unknown dependence. Priprint, 2013. 120, 144, 145, 147, 148, 149, 150

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96:1348–1360, 2007. 4, 13, 34

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B*, 70:849–911, 2008. 13

J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space (invited review article). *Statistica Sinica*, 20:101–148, 2010. 13

J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147:186–197, 2008. 1, 4, 119

J. Fan, X. Han, and W. Gu. Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, 107:1019–1035, 2012a. 120, 146, 147, 149

J. Fan, Y. Li, and K. Yu. Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association*, 107:412–428, 2012b. 121, 122

J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B*, 75: 603–680, 2013. 4, 72, 99, 120, 126, 127

M. Fiecas, J. Franke, R. von Sachs, and J. Tadjuidje. Shrinkage estimation for multivariate hidden markov models. *Journal of the American Statistical Association*, 2016. to appear. 158

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936. 2, 119

J. Fleming, C. Kirby, and B. Ostdiek. The economic value of volatility timing using "realized volatility. *Journal of Financial Economics*, 67:473–509, 2003. 121

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008. 4

P. Fryzlewicz. High-dimensional volatility matrix estimation via wavelets and thresholding. *Biometrika*, 100:921–938, 2013. 3, 73, 85, 88

P. Fryzlewicz and N. Huang. Invited discussion of "Large covariance estimation by thresholding principal orthogonal complements" by fan, liao and mincheva. *Journal of the Royal Statistical Society Series B*, 75:648–650, 2013. 96

R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posteriori covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98: 227–255, 2007. 3

T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301: 102–105, 2003. 2, 3, 119

D. Goldfarb and G. Iyengar. Robust portfolio selection problems. *Mathematics of Operations Research*, 28:1–38, 2003. 1, 119, 121

G. H. Golub and C. F. Van Loan. *Matrix Computations, 4th ed.* Johns Hopkins University Press, Baltimore, MD, 2013. 115, 116

Y. Q. Guo, T. Hastie, and R. Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8:86–100, 2007. 2, 119

I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, R. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, Z. Yakhini, A. Ben-Dor, E. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Pittaluga, S. Gruvberger, N. Loman, O. Johannsson, H. Olsson, B. Wilfond, G. Sauter, O. Kallioniemi, Å. Borg, and J. Trent. Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344:539–548, 2001. xv, 148, 150, 153

J. Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93:85–98, 2006. 120, 137, 138

J. E. Jackson. *A user's guide to principal components*. John Wiley Sons, New York, 1991. 1, 119

H. Jeong, S. P. Mason, A.-L. Barabási, and Oltvai Z. N. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001. 2, 3, 119

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29:295–327, 2001. 2

I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104: 682–693, 2009. 1, 119

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs

with the pc-algorithm. *The Journal of Machine Learning Research*, 85:613–636, 2007. 60

A. Khalsa. Why you should use the sharpe ratio when investing in the medical device industry, August 2013. URL http://finance.yahoo.com/news/why-sharpe-ratio-investing-medical-042514170.html. 127

C. Lam. Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Annals of Statistics*, 44:928–953, 2016. 5, 120, 123, 125, 137, 138, 140, 141

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37:4254–4278, 2009. 4

C. Lam and P. Feng. A nonparametric eigenvalue-regularized integrated covariance matrix estimator using high-frequency data for portfolio allocation. Manuscript, 2016. 122, 123, 125, 131

S. L. Lauritzen. *Graphical models*. Clarendon Press, Oxford, 1996. 4, 8

O. Ledoit and S. Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151:233–264, 2011. 5

O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10:603–621, 2003. 1, 5, 6, 86, 88, 112, 113, 119, 123, 125

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal Multivariate Analysis*, 88:365–411, 2004. 4, 75, 158

O. Ledoit and M. Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, 4:1024–1060, 2012. 4, 5, 75

O. Ledoit and M. Wolf. Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. Working Paper, 2013. 5, 99, 127

T. Leonard and S. J. H. John. Bayesian inference for a covariance matrix. *Annals of Statistics*, 20:1669–1696, 2012. 6

J. Lintner. Valuation of risk assets and the selection of risky investments in stock portfolio. *Review of Economic Studies*, 47:13–37, 1965. 121

Q. Liu. On portfolio optimization: How and when do we benefit from highfrequency data? *Journal of Applied Econometrics*, 24:560–582, 2009. 121

A. W. Lo. The statistics of sharpe ratios. *Financial Analysts Journal*, 58:36–52, 2002. 133, 134

J. Longerstaey, A. Zangari, and S. Howard. Risk Metrics^{TM}-technical document. Technical Document. J. P. Morgan, New York, 1996. 1, 119

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1:507536, 1967. 2

H. Markowitz. Portfolio selection. *The Journal of Finance*, 7:77–91, 1952. 1, 119, 121

J.B. Maverick. What is a good sharpe ratio?, March 2016. URL http://www.investopedia.com/ask/answers/010815/what-good-sharpe-ratio.asp. 127

N. Meinshausen. Relaxed lasso. *Computational Statistics Data Analysis*, 52:374–393, 2007. 13

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2008. 2, 4, 119

J. Mossin. Equilibrium in capital asset markets. *Econometrica*, 34:768–783, 1966. 121

S. Park. *Consistent estimator of ex-post covariation of discretely observed diffusion processes and its application to high frequency financial time series*. PhD thesis, The London School of Economics and Political Science, 2011. 122

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901. 1, 119

J. Peng, P. Wang, NF. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104:735–746, 2009. 4, 9

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011. 2, 119

A. J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008. 4, 51

A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104:177–186, 2009. 36, 73, 75, 85, 115

A. Sancetta. Sample covariance shrinkage for high dimensional dependent data. *Journal of Multivariate Analysis*, 99:949–967, 2008. 158

S. K. Sarkar. Some results on false discovery rate in stepwise multiple testing procedures. *Annals of statistics*, 30:239–257, 2002. 120

R. M. Savic and M. O. Karlsson. Importance of shrinkage in empirical bayes estimates for diagnostics: problems and solutions. *American Association of Pharmaceutical Scientists*, 11:558–569, 2009. 86

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomic. *Statistical Applications in Genetics and Molecular Biology*, 4:1544–6115, 2005. 5, 75, 86, 88, 89, 90, 98

R. J. Serfling. *Approximation theorems of mathematical statistics. Vol. 162*. John Wiley Sons, New York, 2009, p.19. 31, 65, 67

W. F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19:425–442, 1964. 121

W. Sun and T. Cai. Journal of the royal statistical society series b. *Annals of statistics*, 71:393–424, 2009. 120

M. Talih. *Markov random fields on time-varying graphs, with an application to portfolio selection*. PhD thesis, Yale University, 2003. 1, 119

M. Tao, Y. Wang, Q. Yao, and J. Zou. Large volatility matrix inference via combining low-frequency and high-frequency approaches. *Journal of the American Statistical Association*, 106:1025–1040, 2011. 122

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, 58:267–288, 1996. 13

Y. L. Tong. *The multivariate normal distribution*. Springer-Verlag, New York, 2012, p.35. 19

H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104:1512–1524, 2009. 13, 34, 60, 61

Y. Wang and J. Zou. Vast volatility matrix estimation for high-frequency financial data. *Annals of Statistics*, 24:943–978, 2010. 122

S. Weisberg. *Applied linear regression*. John Wiley Sons, New York, 2005. 13

W. B. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90:831–844, 2003. 3

M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286, 2010. 2, 119

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 90:831–844, 2007. 4

P Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2001. 4

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101:1418–1429, 2006. 4, 13, 34