THE EFFECTS OF HIGH DIMENSIONAL COVARIANCE MATRIX ESTIMATION ON ASSET PRICING AND GENERALIZED LEAST SQUARES

A Thesis
Presented to
The Academic Faculty

by

Soo-Hyun Kim

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2010

THE EFFECTS OF HIGH DIMENSIONAL COVARIANCE MATRIX ESTIMATION ON ASSET PRICING AND GENERALIZED LEAST SQUARES

Approved by:

Dr. Ming Yuan, Advisor,
Committee Chair
H. Milton Stewart School of Industrial
and Systems Engineering
Georgia Institute of Technology

Dr. Shijie Deng
H. Milton Stewart School of Industrial
and Systems Engineering
Georgia Institute of Technology

Dr. Seong-Hee Kim
H. Milton Stewart School of Industrial
and Systems Engineering
Georgia Institute of Technology

Dr. Nicoleta Serban
H. Milton Stewart School of Industrial
and Systems Engineering
Georgia Institute of Technology

Dr. Yixin Fang
Department of Mathematics and
Statistics
Georgia State University

Date Approved: 10 June 2010

To lovely wife,

Jee Min

ACKNOWLEDGEMENTS

First of all, I would like to thank my wife, Jee Min Lee. I could not have finished the PhD study without her devotion. She has been always with me whenever I faced any kinds of challenges and she walked me through it. For me, PhD study has been all about intellectual struggling, which I could not overcome without my wife's love and her ceaseless effort to cheer me up.

I also deeply appreciate the support from my parents, parents-in-law, sisters and brother-in-law. With their endless support, I could enjoy the journey of PhD program.

I would like to thank my advisor, Dr. Ming Yuan. Although I was not always successful in doing research, he shared his knowledge and experience with me so that I could see through the problem and keep trying. I sincerely appreciate his treating me as his student and as a friend. I am lucky to have Dr. Yuan as my advisor.

I also want to express my gratitude to my dissertation committee members, Dr. Deng, Dr. Kim, Dr. Serban, and Dr. Fang. Their valuable comments improved the quality of my thesis. I learned a lot from their sharp intuition and research experience, by which I could see different aspects of the research.

PhD program of ISyE at Georgia Institute of Technology was very special. Its research environment is very helpful in that it has a lot of top class faculty members to discuss problems with. Smart colleague students have me to try harder to improve myself. I am very proud that I was educated in such a privileged institute.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Covariance matrix estimation is the essence of measuring risks in multivariate statistics. Existing research efforts are mostly devoted to asymptotic behaviors as sample size increases or to modeling covariance matrices with structural assumptions. In this thesis we investigate alternative methods that do not depend on such restrictions.

High dimensional covariance matrix estimation is considered in the context of empirical asset pricing. In asset pricing models covariance matrices are used more intensively and potentially make significant difference in estimating or testing errors because the nature of asset pricing models is far more complicated. In order to see the effects of covariance matrix estimation on asset pricing, parameter estimation, model specification test, and misspecification problems are explored. Along with existing techniques, which is not yet tested in applications, diagonal variance matrix is simulated to evaluate the performances in these problems. We found that modified Stein type estimator outperforms all the other methods in all three cases. In addition, it turned out that heuristic method of diagonal variance matrix works far better than existing methods in Hansen-Jagannathan distance test.

High dimensional covariance matrix as a transformation matrix in generalized least squares is also studied. Since the feasible generalized least squares estimator requires ex ante knowledge of the covariance structure, it is not applicable in general cases. We propose fully banding strategy for the new estimation technique. Apart from analytical efforts to examine the behaviors of our estimation, guided simulations are provided to support our claim that more spread-out diagonals of covariance matrix lead to better relative outperformance of GLS estimation over OLS estimation. First we look into the sparsity of covariance matrix and the performances of GLS. Then we move onto the discussion of diagonals of covariance matrix and column summation of inverse of covariance matrix to see the effects on GLS estimation. In addition, factor analysis is employed to model the covariance matrix and it turned out that communality truly matters in efficiency of GLS estimation.

CHAPTER I

MOTIVATION AND OUTLINE

1.1 Motivation

Measuring risks is one of the important statistical tasks both in theoretical and practical perspective. Recent outburst of financial crisis rooted in sub-prime mortgage was also considered as an example of mis-judged investment risks. Among others, variance is not only the traditional statistical methods but also the most widely used measure. For univariate random variables variance makes a solid standpoint in practice since it is fairly well-defined and straightforward to compute.

However, in modern statistical research, we often encounter multivariate problems and covariance has to be entertained in addition to variance. Covariance matrices, multivariate counterpart of variance in univariate case, are natural choice for risk measure in the multivariate case. Apart from the simple risk measure of random variables, it is extended to many other usages in various applications. For example, covariance matrices are commonly used in generalized least squares as the transformation matrices, or in generalized method of moments as the weighting matrices.

Despite the importance of covariance matrices, it is not an easy job to estimate them precisely especially when high dimensional covariance matrices are considered. Even a small universe of ten assets, for instance, requires 55 parameters to be estimated. Additionally, in many applications from asset pricing, the inverse covariance matrices are needed rather than covariance matrices. The small sample properties get even worse if we take inverse of the covariance matrix estimates.

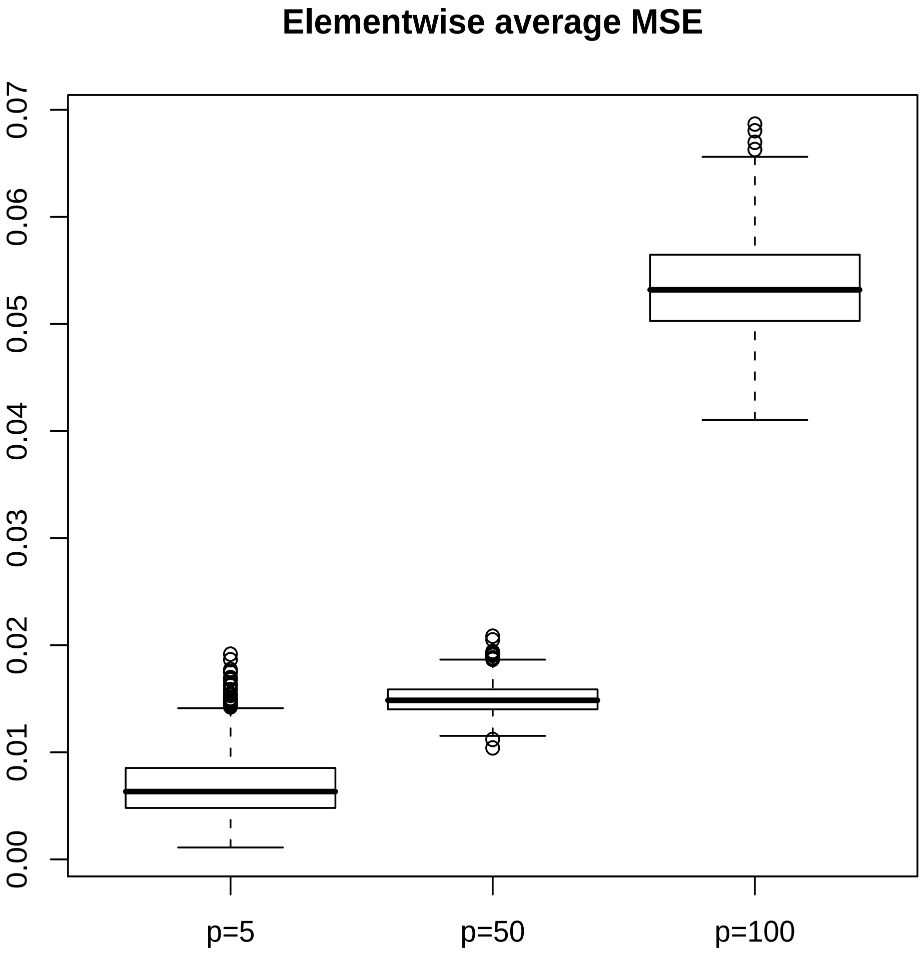


Figure 1: Average MSE of precision matrix estimate

An example with simple simulation will make it clearer. Let p be the dimensionality, and Σ be a $p \times p$ covariance matrix. Draw 200 random samples from $N(0_p, \Sigma)$. Here we only take very simple case of Σ=diag(1,...,1). With inverse sample covariance matrix, $\hat{\Sigma}^{-1}$, let error matrix be $ER = \Sigma^{-1} - \hat{\Sigma}^{-1}$. We compute the element wise average squared error as

$$\frac{1}{p^2}||ER||_F^2 = \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p} er_{ij}^2$$

where er_{ij} is the element of ER in row i and column j. The simulation results with 1,000 iterations with three cases of p=5, 50, 100 are given as the box plots in figure 1. As the dimensionality p becomes larger, the figure shows that the error gets bigger exponentially. Even a very simple case of covariance matrix with zero off-diagonal elements makes significant differences between low and high dimensional cases. Since we often cope with far more complex covariance structure, the problem is be expected to be much worse.

As noted earlier, multivariate applications arise in many applications, and its dimensionality becomes higher nowadays. The problems of high dimensional covariance matrix estimation are noted by many researchers but the optimal solutions for such problems have not been identified. Previous considerations are mainly focused on model based structural covariance matrices.(See [36], [20] and [42]). Although it improves our understanding of high dimensional covariance matrices, it imposes too strong assumptions structurally, which limits their practical value. We will review some existing results in next chapters.

Mathematically rigorous theories usually require parametric approach with prior assumptions in models. But this may limit our understanding about behaviors of high dimensional covariance matrix estimation in general. Therefore, we would like to take a different view in this thesis. Although it is almost impossible to set up a fully theoretical approach, it would be very beneficial to have a simulation kicked-in to understand the problem. There are already a lot of model-free covariance estimation methods developed but they are not yet tested in high dimensional asset pricing applications. Concentrating on asset pricing model in the context of empirical financial problems, we would like to show the performance of existing methods and propose a new estimation method for covariance matrix. In addition, generalized least squares will also be considered. Generalized least squares rely heavily on a good estimate of covariance matrices. We try to understand the effects of covariance matrix estimation from a theoretical point of view and supplement our proposal with guided simulations.

1.2 Thesis Outline

The rest of thesis is organized as follows. Chapter 2 focuses on the effects of high dimensional covariance matrix estimation on empirical asset pricing models. The chapter starts with an introduction to covariance matrix distribution and previous development of covariance matrix estimation techniques with the comparison to sample covariance matrix. Section 3 of the chapter applies various estimation techniques in parameter estimation problems, especially the popular two-pass procedure. Through simulation studies, we will illustrate the performances of each estimation method. Section 4 discusses model specification testing with an example of Hansen-Jagannathan distance. In hypothetical testing, covariance matrices are frequently used as weighting matrices. We also take a close look at the model misspecification to evaluate the covariance matrix estimation performance in terms of type-2 error.

Chapter 3 covers covariance estimation in relation with generalized least squares in three settings corresponding to sparse, diagonal and factor covariance matrices. We argue that the banding strategy is useful for high dimensional covariance matrices estimation. In order to support the idea, simulations with different degree of sparsity are provided in section 3. Inspired by the simulation results, both analytical and simulation studies focusing on diagonal covariance matrices are explored in the following section. We derived that spread-out diagonals make more difference between OLS and GLS estimation in terms of efficiency. In section 5, we consider factor covariance matrices. Fully banded strategy is investigated by means of analytical calculations for single factor models. Three types of matrix norms are used to measure the distance between inverse of specific covariance matrices and inverse of covariance matrices. Communality and specific variance ratios turn out to be one of the crucial elements of high dimensional cases.

We summarize and conclude the thesis in chapter 4. Potential development on both the academic side and practical applications are discussed as well. Supplementary plots pertaining to the simulation in chapter 2 and 3 are also given in appendix.

CHAPTER II

HIGH DIMENSIONAL COVARIANCE MATRIX ESTIMATION AND ASSET PRICING MODEL

2.1 Introduction

Covariance matrices play a key role in finance. Markowitz portfolio theory followed by Capital Asset Pricing Model (CAPM) of Sharpe are considered fundamental basis of modern finance theory, and both rely heavily upon covariance structure among the returns of risky assets. Covariance matrices are a crucial part of any asset pricing model because prices of any asset depend not on idiosyncratic risk but only on systematic risk which is measured through covariance structure of the market. Apart from the theory, there are many empirical techniques that take advantage of covariance matrices: factor analysis, principle component analysis(PCA), generalized method of moments(GMM), and generalized least squares(GLS), etc.

In finance, it is well known that the volatility of financial assets is not constant over time. Historical evidences suggest that large volatility tends to cluster together. ARCH and GARCH models are in part motivated by such observations(see [5] and [16]). An example of Microsoft stock return movement is shown in figure 2. The volatility becomes large in Oct. 2008 and continues to be large until Jan. 2009, while the other periods have smaller and stable volatility. As the example illustrates, it is not always plausible to utilize large sample in estimating parameters because statistical characteristic may change over time in certain cases. Although small sample properties of covariance matrix estimation are very important practically, it seems that not enough research attention has been given.

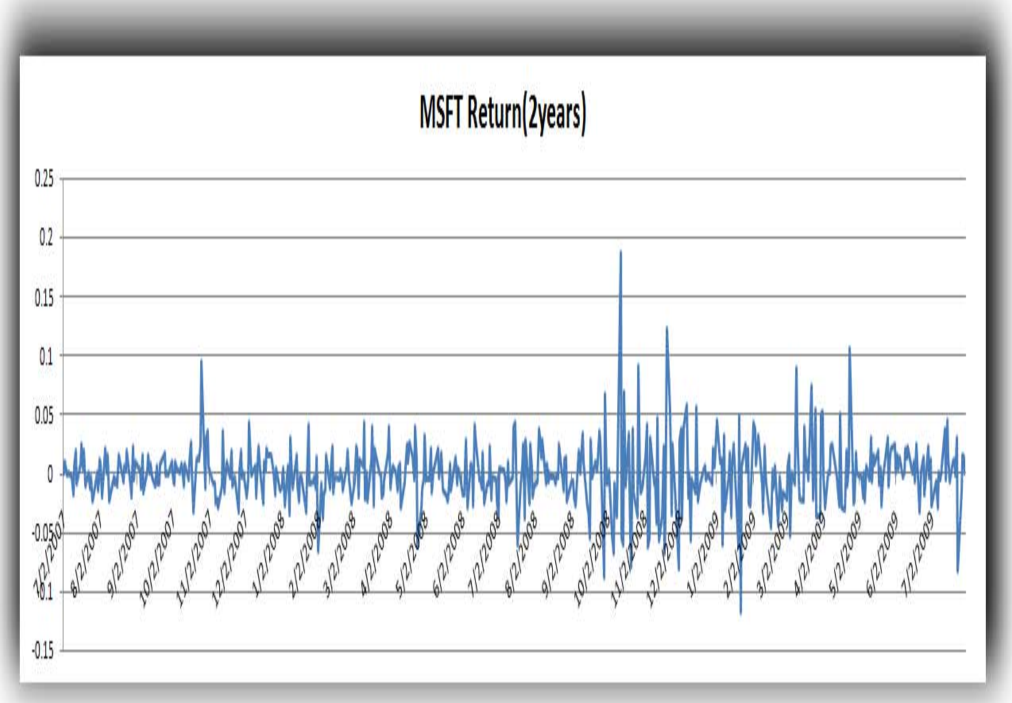


Figure 2: MSFT stock return movement

Among the few work, [36] focuses on the role of covariance matrix in portfolio selection and shows that shrinkage method improves the performance. [1] reports that sample covariance matrix fails to test model specification based on Hansen-Jagannathan distance [32]. It also observes that the test overrejects the true model extremely often especially when the number of time-series data is relatively small comparing to the number of assets. [42] suggests shrinkage method in estimating covariance matrix and shows that small sample properties are much improved when applying to Hansen-Jagannathan distance.

In this chapter, we examine the impacts of various covariance matrix estimators when applying to the asset pricing with simulation studies. Section 2.3 explores the impact of covariance estimation on parameter estimation, while Section 2.4 studies model specification testing. In many cases, inverse of covariance matrices, or precision matrices are used as weighting matrices. [47] introduces several estimation methods of precision matrices: modified adjusted(MAU), modified Perron-type(MPR), modified Stein(MST), the usual estimator(US), Efron-Morris-type(EM), and Dey(DY) estimators. We will employ seven of them and compare their performances with sample covariance matrices or structural true covariance matrix. A brief summary of these estimators is given in section 2.2.

2.2 Various precision matrix estimators

This section briefly reviews several estimators given in [47]. Overall discussion on the estimators from different perspectives are given as well. The precision matrix estimation has been studied from three aspects: adjusting eigenvalue estimation, Bayesian approach and shrinkage method. Each approach will be introduced as follows.

2.2.1 Adjusting eigenvalues

Given true covariance matrix Σ and sample covariance estimation S we have orthogonal decomposition as,

$$\hat{\Sigma}_{UB}^{-1} = (n-p-1)S^{-1} = (n-p-1)R\phi(L)R',$$

where $L = diag(l_1, l_2, \cdots, l_p)$ with eigenvalue l_i of S. The usual form of ϕ is, of course, inverse. Implausible small sample property stems from the fact that the eigenvalue estimations of S^{-1} are more spread out than the eigenvalues of Σ^{-1}. Therefore correction of eigenvalue estimation is one of possible alternative estimating methods. Moreover, it is shown that this class of estimators is better than the usual unbiased estimator with trace loss function $L(\hat{\Sigma}^{-1}, \Sigma^{-1}) = tr(\hat{\Sigma}^{-1} - \Sigma^{-1})^2$, under the following conditions:

(i) $\frac{\partial \delta_i(L)}{\partial l_i} \geq 0, \forall i$

(ii) $n-p-5 \leq \delta_p(L) \leq \cdots \leq \delta_1(L) \leq n-p-1.$

In this chapter, we employ three types of estimators from this class: Adjusted, Perron-type, and Stein-type. The formula of these estimators are as follow.

2.2.1.1 Adjusted estimation

$$\hat{\Sigma}^{-1}_{AU} = R\phi_{AU}(L)R' \tag{1}$$

where $\phi_{AU}(L) = diag(\delta^{AU}_1(L)/l_1, ..., \delta^{AU}_p(L)/l_p)$ with $\delta^{AU}_i(L) = n-p-1-4(i-1)/(p-1)$.

Modified AU estimator is given by

$$\hat{\Sigma}^{-1}_{MAU} = \hat{\Sigma}^{-1}_{AU} + cS/tr(S^2) \tag{2}$$

where $c = p^2 + p - 4$.

2.2.1.2 Perron-type estimator

Let $h_i = h(1/l_{p+1-i})$, $d_i = n-p-5+4(i-1)/(p-1)$, and $H = diag(h_1, ..., h_p)$ with positive valued nondecreasing function $h(\cdot)$. Here we used $h(x) = \sqrt{x}$.

Let $W(H)$ be $p \times p$ matrix with its component w_{ik},

$$w_{ik}(H) = \frac{tr_{k-1}(H_i)}{tr_{k-1}(H)} - \frac{tr_k(H_i)}{tr_k(H)}$$

where, $H_i = diag(h_1, ..., h_{i-1}, 0, h_{i+1}, ..., h_p)$.

$$tr_k(H) = \begin{cases} 1 & \text{if } k = 0 \\ \sum_{1 \leq i_1 \leq i_2, ..., i_k \leq p} \prod_{j=1}^{k} h_{i_j} & k = 1, 2, .., p \\ 0 & \text{otherwise} \end{cases}$$

Then the Perron type estimator is given by,

$$\hat{\Sigma}_{PR}^{-1} = R\phi_{PR}(L)R' \tag{3}$$

where $\phi_i^{PR}(L) = \delta_i^{PR}(L)/l_i$ with $\delta_{p+1-i}^{PR}(L) = \sum_{k=1}^{p} w_{ik}(H)d_k$.

Modified PR estimator is given by

$$\hat{\Sigma}_{MPR}^{-1} = \hat{\Sigma}_{PR}^{-1} + cS/tr(S^2) \tag{4}$$

where $c = p^2 + p - 4$.

2.2.1.3 Stein-type estimator

$$\hat{\Sigma}_{ST}^{-1} = R\phi_{ST}(L)R' \tag{5}$$

where $\phi_i^{ST}(L) = \frac{1}{l_i}(n - p - 3 + \sum_{j \neq i} \frac{l_i}{l_i - l_j})$. Since it is not monotone, we apply isotonic regression to use the fitted values of $\bar{\phi}_i^{ST}$.

Modified PR estimator is given by

$$\hat{\Sigma}_{MST}^{-1} = \hat{\Sigma}_{ST}^{-1} + cS/tr(S^2) \tag{6}$$

where $c = p^2 + p - 4$.

2.2.2 Bayesian Approach

Now consider the following class of estimators.

$$\hat{\Sigma}_G^{-1} = aS^{-1} + G$$

where G is a $p \times p$ symmetric matrix with the elements of G being functions of S. It has been proven that this class of estimators is better than the usual unbiased estimator under trace loss function, if satisfying the following numerical conditions.

(i) $n - p - 5 \leq a \leq n - p - 1$

(ii) $tr(G^2 - 2(n - p - 1 - a)S^{-1}G - 4D_S G) \leq 0$

where $D_S = (1/2)(1 + \delta_{ij})\partial/\partial S_{ij})$ with δ_{ij} being the Kronecker delta.

Upon the prior given to Σ^{-1} or G, the conditional expectations of posterior can be obtained as Bayesian estimators. The following are the examples of these and will be used in our simulation study.

2.2.2.1 The usual estimator

Suppose the uniform distribution is given as prior of Σ^{-1}, $p(\Sigma^{-1}) \propto 1/|\Sigma^{-1}|^{(n+p+1-a)/2}$ where a is a constant. Then the posterior distribution is derived as Gaussian distribution, and the Bayseian estimator is given as

$$\mathbb{E}(\Sigma^{-1}|S) = \hat{\Sigma}_{US}^{-1} = aS^{-1} \tag{7}$$

where $a = n - p - 3$.

2.2.2.2 Efron-Morris-type estimator

Efron-Morris-type estimator assumes the precision matrix decomposition as $\Sigma^{-1} = \omega\mathbf{I}_p + \xi\xi'$ where ξ is $p \times a$ random matrix. Giving the prior distribution to ξ as $p(\xi|\omega) \propto \omega^{-pa/2}|\mathbf{I}_p + \xi\xi'/\omega|^{-n/2}$, we obtain the Bayseian estimator

$$\mathbb{E}(\Sigma^{-1}|S, \omega) = \omega\mathbf{I}_p + aS^{-1}$$

From the assumption that the marginal distribution of S is Wishart , estimate of ω is computed as $\hat{\omega} = \frac{(n-a)p-2}{tr(S)}$. Finally the EM estimator is given as

$$\hat{\Sigma}_{EM}^{-1} = aS^{-1} + \frac{b(t)}{t}Q(Q'Q)^{-1}Q' \tag{8}$$

where $a = n - p - 4$, $t = tr(S)$, $b(t) = 1$, and Q is a $p \times q$ matrix with rank q.

2.2.2.3 Haff-type estimator

Similarly, Haff-type estimator assumes Wishart distribution $W_p(\lambda\mathbf{I}_p, m - p - 1)$ for the prior of Σ^{-1} with some constant m. Then the Bayesian estimator is given as,

$$\begin{aligned} \hat{\Sigma}_{HF}^{-1} &= a_0(S + ub(u)I_p)^{-1} && (9) \\ &= \hat{\Sigma}_{UB}^{-1} - a_0ub(u)(S^2 + ub(u)S)^{-1} && (10) \end{aligned}$$

where $a_0 = n - p - 1$, $u = \frac{1}{tr(S^{-1})}$, and $b(u) = 2/(n - p - 5)$

2.2.2.4 Dey estimator

Shrinkage method is often employed to improve the performance of the original estimator. Let $\hat{\Sigma}_M^{-1} = \hat{\Sigma}_{SH}^{-1} + M$ where $\hat{\Sigma}_{SH}^{-1}$ satisfies $\hat{\Sigma}_{UB}^{-1} - \hat{\Sigma}_{SH}^{-1}$'s being positive semi-definite. Then $\hat{\Sigma}_M^{-1}$ dominates $\hat{\Sigma}_{SH}^{-1}$ under certain numerical conditions. See [47] for the numerical conditions and related theorems. One example is Dey's estimator which takes adjusted estimator for $\hat{\Sigma}_{SH}^{-1}$ and $\mathbf{I}_p$ for Q.

$$\hat{\Sigma}_{DY}^{-1} = aS^{-1} + \frac{b(S)}{tr(S^2)}S \tag{11}$$

where $a = n - p - 3$ and $b(S) = p^2 + p - 4$.

2.2.3 Discussion on the simulation result

The simulation study of [47] shows that none of the estimators is dominantly better than the others, under the percentage reduction in average loss(PRIAL) relative to the unbiased estimator, which is defined as below.

Let the loss function $L(\cdot)$ be $L(\hat{\Sigma}^{-1}, \Sigma^{-1}) = tr(\hat{\Sigma}^{-1} - \Sigma^{-1})^2$. The risk function, therefore, can be written as $R(\hat{\Sigma}^{-1}, \Sigma^{-1}) = E[L(\hat{\Sigma}^{-1}, \Sigma^{-1})]$. The criteria of PRIAL is computing the relative improvement of risk of each estimators in comparison with the unbiased estimator in percentage sense, i.e.

$$PRIAL = 100 \times (\hat{R}(\hat{\Sigma}_{UB}^{-1}) - \hat{R}(\hat{\Sigma}_{\cdot}^{-1}))/\hat{R}(\hat{\Sigma}_{UB}^{-1})$$

The risk, $R(\cdot)$ is approximated by 10,000 iterations of random draws from multivariate normal distribution.

This experiment presents a couple of useful points to our interest. Although no estimators can be found as universally better over the other candidates, MST looks the best in our application to asset pricing. Especially with the true covariance matrix, $\Sigma = diag(1, 1, 1, 1, 1)$, PRIAL of MST estimator outperforms the others except for case of sample size 12. This can be considered as an extreme case in that the sample size is too small for the dimension. While relative performance of other estimators over unbiased estimator disappears quickly as the sample size gets larger, the speed of reduction in PRIAL of MST is much slower so that its performance outstands in the cases of sample size 30 and 50 comparing to other estimators. When the magnitude of diagonal elements of Σ are far different to each other, PRIAL of MST is not plausible. It turns out that MST is the worst with $\Sigma = diag(4, 4^2, 4^3, 4^4, 4^5)$ case.

MST is not universally the best estimator yet is the most proper to asset pricing application. The reason is that the variances of asset returns are not very different to one another after conditioning the common pricing factors. Let us take CAPM model as an example. One may observe a stock price very volatile than the others. This is usually because beta of the stock is very large in magnitude, not because the variance of the firm specific risk is greater than the others. In this sense, it is a widely accepted notion in finance theory that the idiosyncratic risk does not matter in pricing the asset. Once common pricing factors are specified correctly or conditioning the information correctly, the remaining disturbances or idiosyncratic risks have very similar variances each other. Since this case is similar to the previous simulation with $\Sigma = diag(1, 1, 1, 1, 1)$, we can guess that MST may serve the best precision matrix estimator in asset pricing.

Table 1: PRIAL of precision matrix estimators

Dimension	5	10	15	20	25
MAU	58.42	74.70	84.12	90.36	94.60
MPR	-419.04	-132.14	5.50	62.17	85.28
MST	67.95	80.75	86.67	91.26	94.76
US	45.05	70.74	84.43	91.35	95.11
EM	44.89	70.52	84.25	91.22	90.02
DY	49.80	69.18	80.32	87.67	92.73
MHF	41.60	59.96	73.31	82.86	89.53

Again, a simple numerical example will give clearer intuition on our discussion. Since our main interest lies in behavior of the inverse of the covariance matrix estimators, we fix the number of the sample size and examine the behavior in connection to dimensionality. We may also think of this example as follows. We only have limited number of time-series data on asset returns, and would like to compare the preciseness of several estimators as the number of assets increases. The simulation of this example fixes the number of time-series data to 50. The true covariance matrix is set to be the simplest diagonal case, $\Sigma = diag(1, 1, \cdots, 1)$, with $dim(\Sigma)$ increasing from 5 to 25. The result is provided in table 1.

As expected, the PRIAL's are getting bigger as the dimensionality increases. This means that the estimators provide more precise estimates in higher dimensionality than the inverse of the sample covariance matrix estimator. In case of MPR, its performance is much worse than the performance of sample covariance matrix when the dimension is 5 or 10. As the dimensionality becomes higher, the relative performance of all the seven estimators over the inverse of sample covariance matrix gets greater. First of all, MAU, MST, US and EM estimators show very high PRIAL as 94% to 95% when the ratio of sample size to dimension is extremely small. We could see that sample covariance matrix performs very poor and the other estimators improve the small sample properties significantly. It is also obvious from this simulation example that none of the estimator is universally better over the other estimators.

In extreme case as in dimension of 25, US and EM appear the best, while MST outperforms them up to dimension of 20. Again, here we only have very simple covariance structure. Based on this observation, tt is natural to step forward considering more complex structure. We take this step with asset pricing application, which is highly complicated in covariance structure. In the next two sections, parameter estimation and model specification testing problem will be explored in terms of the effects of the precision matrix estimators.

2.3 Covariance Matrix in Parameter Estimation: two-pass procedure

In this section, the effects of high dimensional covariance matrix estimation on two-pass procedure are discussed. This method is developed as an empirical technique testing CAPM framework. [46] and [39] show that beta of any risky asset in equilibrium can be derived from mean-variance space, and that expected return of an asset is linearly related to its beta.

$$\mathbb{E}[r_i] = \gamma_0 + \gamma_1 \beta_i, \text{ for all assets i} \tag{12}$$

where $\beta_i = \frac{cov(r_i,\gamma_1)}{var(r_i)}$ and γ_1 is the market risk premium. CAPM nicely models the co-movement of financial asset returns with returns on market premium. Naturally, one can think of equation (12) as a regression with slope coefficient β. Under beta-pricing type regression models, Two-pass procedure is a popular empirical statistical method.

2.3.1 two-pass procedure

[39] suggests two-pass procedure in empirical asset pricing with panel data. It consists of two stages: time-series regression as first-pass and cross-section regression(CRS) as second-pass. The first-pass is the stage where the beta of each asset is estimated.

$$r_{it} = \gamma_0 + \gamma_{1t}\beta_i + \epsilon_{it}, \tag{13}$$

where r_{it} is the return of i-th asset at time t, and disturbance ϵ_{it} is normally distributed with mean zero. The number of assets and observable discrete time are assumed to be N and T, respectively, i.e. $i = 1, 2, \cdots, N$ and $t = 1, 2, \cdots, T$. Suppose that ϵ_{it}

is independent across time t. The first pass, therefore, estimates β_i by time-series Ordinary Least Square(OLS) with each asset returns. With the estimated $\hat{\beta}_i$ from regression model (13) and average return of each asset, CRS regression is formed as follows,

$$\bar{r}_i = \gamma_0 + \gamma_1 \hat{\beta}_i + \eta_i. \tag{14}$$

In order to estimate γ_0 and γ_1, Litner used OLS in second-pass CRS. Estimated $\hat{\gamma_0}$ and $\hat{\gamma_1}$ can be interpreted as average risk free rate and average return on market portfolio over the time $t = 1, 2, \cdots, T$. [15] compares $\hat{\gamma_0}$ and $\hat{\gamma_1}$ with the yield to maturity of government bond R_F over the same time period of the empirical model and average excess market return $\bar{R}_M - R_F$, respectively. It is found that estimated $\hat{\gamma_0}$ is far greater than R_F and $\hat{\gamma_1}$, which implies the inconsistency of CAPM model with reality.

Fama and Macbeth adopts "rolling" betas to improve the second-pass regression in [19]. They first estimate β each month using previous historical time-series data. Then CRS is conducted with the beta estimates for that month. They repeat this procedure month by month to obtain a time-series estimates of $\hat{\gamma}_0$ and $\hat{\gamma}_1$, then take the average to compare with R_f and $\bar{R}_M - R_F$.

Although Fama-Macbeth approach improves the errors-in-variable, and is proved to be consistent as the number of time-series data(T) becomes large enough, it still misleads the result because cross-sectional correlation and heteroscedasticity are not taken into account. Portfolio Grouping is employed in [4] and [19], yet significant cross-sectional correlation is remained. [45] suggests GLS to be employed in the second pass of CSR, which is the natural remedy to this problem.

Since GLS procedure requires inverse of covariance matrix as a weighting matrix, it is a good example to see how the high dimensional covariance matrix estimation matters in more complicated situation such as two-pass procedure. We extend the market model to the multiple factor model and apply several precision matrix estimators as GLS weighting matrices.

2.3.2 Data Generating Process

In order to compare several precision matrix estimators thoroughly, four cases of data generating processes are taken into account. First of all, three-factor model is assumed as follows.

$$R_{ti} = \alpha + X_{t1}\beta_{1i} + X_{t2}\beta_{2i} + X_{t3}\beta_{3i} + \epsilon_{ti}, \tag{15}$$

where X_{tj} are randomly drawn from $N(0.0022, 6.944 \times 10^{-5})$, factor loading β's are drawn from $U(0, 2)$. The parameters distributions are selected to be consistent with historical evidences. See [1] for more detail. We will consider four cases depending on the ways of constructing random error e_{ti}. Let Ω be the covariance matrix of random error ϵ_i. As commonly accepted, the disturbance term in return process is assumed to be independent in time-series direction so that $corr(\epsilon_{ti}, \epsilon_{t'i}) = 0$ for all i and $t \neq t'$. We want to examine the behavior of the estimates from two-pass procedure according to the inter-asset correlation structure Ω. Here, different inter-asset correlation $corr(\epsilon_{ti}, \epsilon_{ti'})$ structures are simulated as follows.

• Homoscedastic with zero correlation case

Disturbance in equation (15) ϵ_{ti}'s are independently random-drawn from $N(0, 6.944 \times 10^{-5})$ for all i and t. In other words, the disturbances are assumed to be independent across the assets.

• Homoscedastic with AR(1) correlation case

Here, the term, AR, is abused for convenience. AR(1) refers to the case that the only adjacent disturbances have correlations, i.e. $corr(\epsilon_i, \epsilon_j) \neq 0$ if $|i-j| = 1$, and $corr(\epsilon_i, \epsilon_j) = 0$ otherwise. In our simulation studies, we set $corr(\epsilon_i, \epsilon_j) = 0.5$ for all i, j such that $|i-j| = 1$.

- **Homoscedastic with AR(2) correlation case**

Similarly, AR(2) here means that disturbances are correlated only if $|i-j| \leq 2$. We set $corr(\epsilon_i, \epsilon_j) = 0.5$ if $|i-j| = 1$, and $corr(\epsilon_i, \epsilon_j) = 0.25$ if $|i-j| = 2$. All the others are set to be non-correlated.

- **Heteroscedastic with completely random correlation case**

Let $\Omega = \mathbb{E}[\epsilon_1, \epsilon_2, \cdots, \epsilon_N]'[\epsilon_1, \epsilon_2, \cdots, \epsilon_N]$. The diagonal elements of Ω are randomly chosen from $U(0.00004944, 0.00008944)$, whose mean is the same as homoscedastic case. The off-diagonal elements are drawn from $U(-0.2, 0.2) \times 0.00006944$. From this setting, we can generate heteroscedastic(diagonal elements) random correlation(off-diagonal) case.

2.3.3 Estimation and Simulation results

The expression of GLS estimate is given as

$$\hat{\Gamma} = \bar{R}(\hat{B}'\hat{\Omega}^{-1}\hat{B})^{-1}\hat{B}'\hat{\Omega}^{-1}, \tag{16}$$

where $\Gamma = [X_1, X_2, X_3]'$ and $B = [1_N, \beta_1, \beta_2, \beta_3]$. Note that covariance matrix of disturbance Ω is estimated from the residuals of OLS in first-pass.

In our simulation, we'd like to replace the inverse covariance matrix Ω^{-1} with various estimators. In order to compare the performances of the estimators, PRIAL is used here also as measure for the improvement. Let $m = \Gamma_{True} - \hat{\Gamma}$, or the estimating error of $\hat{\Gamma}$. Taking $L = m'm$ as the loss function, we have the risk

Table 2: PRIAL: Homoscedastic with zero correlation

N \ T	MAU	MPR	MST	US	EM	DY	MHF
25\160	5.67	5.54	10.38	$\approx$0	0.007	5.31	5.27
25\330	1.85	1.44	4.66	$\approx$0	0.002	1.71	1.70
25\700	0.17	-0.02	0.54	$\approx$0	0.001	0.15	0.15
100\160	35.78	35.77	50.68	$\approx$0	0.007	34.69	34.17
100\330	14.79	14.74	24.00	$\approx$0	0.001	14.53	14.46
100\700	4.11	4.04	8.49	$\approx$0	0.001	4.04	4.03

$R = \mathbb{E}(L) = \mathbb{E}(m'm)$. As before, PRIAL is defined as $PRIAL = 100 \times (\hat{R}(\hat{\Sigma}_{UB}^{-1}) - \hat{R}(\hat{\Sigma}_{\cdot}^{-1}))/\hat{R}(\hat{\Sigma}_{UB}^{-1})$, percentage improvement in risk over inverse of sample covariance.

We consider the homoscedastic with zero correlation case first. This is the most simple case because inter-asset correlations are set to be zeros and the idiosyncratic risk of each asset has the same variances. The simulation result is shown in table 2.

The first observation is that all the PRIAL's are positive except for MPR with N=25 and T=700, meaning that the performances of the estimators are all better than that of sample covariance matrix. The relative performance gets greater as the ratio between the number of asset N and sample size T increases. The second observation is that MST dominates in PRIAL. In all the cases, PRIAL of MST is the best, and it improves the risk reductoin as high as 50.7% relative to the inverse of sample covariance matrix in N=100,T=160 case. MAU, MPR, DY and MHF also give a strong evidence of improvement over sample covariance matrix, however, PRIAL of MST is nearly twice of them.

The following are the cases of AR(1), AR(2). We can find interesting result from table 3 for AR(1) case. Most of the PRIAL's are negative indicating that sample covariance matrix outperforms all the other estimators. More interesting part is that the PRIAL's of seven estimators improve as the sample size gets bigger. Among the estimators with relative weak performance, EM is the best yet worse than the inverse of sample covariance matrix. AR(2) case gives completely different evidence. See table 4.

Most of PRIAL's are positive and in some cases it is as high as 27.45% in AR(2) simulation. Recall that AR(2) is constructed with more complex covariance structure than AR(1) in that asset $i-2, i-1, i+1$ and $i+2$ are directly correlated with asset i. As before, MST performs the best. Especially the cases with N=100, MST is significantly more precise estimator than sample covariance matrix. PRIAL's of US estimator are almost zero, so it performs almost the same as sample covariance matrix. This phenomenon is also shown in all the other cases.

In our simple example in table 1, US estimator shows relative advantage over sample covariance matrix, nonetheless, it behaves very similar to sample covariance matrix in more complicated situations. This observation comes clearer through the pair plots provided next. We can check in the plot that EM also shows similar behavior as sample covariance matrix. All the other estimators performs similar to each other, worse than MST but better than EM and US. PRIAL's are almost the same.

Table 3: PRIAL: AR(1)

N \ T	MAU	MPR	MST	US	EM	DY	MHF
25\160	-3.52	-2.48	-0.83	≈0	-0.42	-3.70	-3.60
25\330	-0.98	-0.91	-0.45	≈0	-0.72	-1.01	-0.99
25\700	-0.26	-0.24	-0.22	≈0	-0.81	-0.27	-0.27
100\160	-36.26	-35.53	-8.29	≈0	-0.68	-36.97	-33.90
100\330	-14.40	-13.10	-6.60	≈0	-1.83	-14.40	-14.24
100\700	-1.84	-0.97	-1.37	≈0	-1.28	-1.89	-1.87

Now look at table 5 for completely random correlation case. This simulation constructs the covariance matrix to have different variances and completely random correlations. Hence, this can be seen as the most complex case in two-pass procedure simulations.

Table 4: PRIAL: AR(2)

N \ T	MAU	MPR	MST	US	EM	DY	MHF
25\160	1.85	1.76	2.04	≈0	-3.90	1.8	1.8
25\330	-0.26	-0.22	-0.70	≈0	-0.04	-0.23	-2.22
25\700	0.1	-0.5	-0.03	≈0	0.11	0.11	0.11
100\160	14.75	14.76	27.45	≈0	-0.89	13.95	13.84
100\330	3.78	3.86	5.69	≈0	0.52	3.69	3.70
100\700	1.24	1.22	1.65	≈0	0.34	1.22	1.22

Table 5: PRIAL: Heteroscedastic with random correlation

N \ T	MAU	MPR	MST	US	EM	DY	MHF
25\160	3.41	3.40	3.96	≈0	-0.006	3.29	3.28
25\330	0.89	0.79	1.08	≈0	0.30	0.87	0.87
25\700	0.24	-0.18	0.24	≈0	0.15	0.24	0.23
100\160	11.89	11.91	18.95	≈0	0.59	11.13	11.20
100\330	3.49	3.55	3.78	≈0	0.37	3.44	3.43
100\700	1.63	1.55	2.01	≈0	0.33	1.61	1.61

Almost all PRIAL's are positive and MST performs the best in any combinations of N and T. All the findings from AR(2) are also valid in this case as well. One more thing to be mentioned here is that the improvement of small sample properties become smaller than AR(2). Basically when N=100, the ratio of dimensionality to sample size is bigger than the case with N=25. Therefore, more improvement in PRIAL is expected in the case with N=100, and it actually is shown in both AR(2) and completely random case. But AR(2) case makes bigger difference between N=100 and N=25.

For example, PRIAL of MST in N=25 and T=160 is 10.38 and that of N=100, T=160 is 18.95 when completely random case is considered. On the other hand, the corresponding PRIAL's of AR(2) are 2.04 and 27.45. This can be applied to other combinations of N and T or other estimators. If we take the example of zero correlation case, table 2, this trend becomes clearer. It seems that the more complex covariance structure is, the weaker the improvement of the small sample properties we obtain via new estimators.

In this section, we have seen the application of new precision matrix estimators to two-pass procedure estimating problem. Several simulation experiments give us evidence that the new estimators mostly outperform the sample covariance matrix with the PRIAL criteria. In various simulation settings, we found that the estimators make very significant improvement over sample covariance matrix, especially in extreme cases with high ratio of dimensionality to sample size, such as N=100, T=160. As we expected in section 2.2, MST's relative outperformance to sample covariance matrix is the most significant, and it is recommended to use in GLS of two-pass procedure.

2.4 Covariance Matrix in Model specification test

In this section, effects of covariance matrix estimation are explored in the context of model specification testing problem. In asset pricing model, Hansen statistic based on GMM in [25] is first developed to detect the errors in estimating stochastic discount factor. See [12] for more detail. Although the statistic has very convenient property of asymptotic χ^2 distribution, it has a couple of weaknesses. First of all, Hansen statistic favors highly variable pricing error. This is so because Hansen statistic is minimizing the quadratic error with inverse of consistent estimators of covariance matrix of pricing error as weighting matrices. Second, the statistic is too large in magnitude with finite samples, as pointed out in [6] and [21].

To overcome these issues, Jagannathan and Wang[32] develop a distance measure that enables testing a linear asset pricing model specification. Hansen-Jagannathan distance(HJ-distance) assesses the errors by taking least square of maximum distance between stochastic discount factor(SDF) from the specified model and the family of true SDF's which prices the asset correctly [24]. It is worth emphasizing that HJ-distance uses the covariance matrix of asset returns, not pricing error as the weighting matrix. Therefore it overcomes the problem of favoring the highly variable pricing error in Hansen's statistic.

Additionally, Hansen-Jagannathan distance uses the same weighting matrix regardless of the pricing models specified. This is also because Hansen-Jagannathan distance uses second moment of return, which remains the same whichever model we choose. Although HJ-distance does not have the nice property of asymptotic χ^2 distribution as Hansen statistic, a simulation method for computing empirical p-value is developed in [32].

Despite the improvement in testing asset pricing model with HJ-distance, [1] finds that the HJ-distance test overrejects the true model too frequently if one uses sample covariance matrix for the weighting matrix. Although a few researchers attempt to solve the overrejection problem by adjusting degrees-of-freedom([21]), the Monte Carlo experiment reveals that it is not enough to accept the model specification test with small sample size. Ahn and Gadarowski identify that the problem rises from poor small sample properties of sample covariance matrix in [1]. They show that the rejection rate is nearly 100% for theoretical p-value of 1%, 5% and 10% when the relative number of asset to time-series sample size is large. This result suggests that the statistic is practically useless in testing many assets with small time-series sample size.

[42] constructs shrinkage method to improve the overrejection problem in HJ-distance. Shrinkage method is also introduced in [36] in asset allocation problem. The shrinkage approach used in [42] is a bit different from the usual one. They use the linear combination of sample covariance matrix of asset returns and the estimated structural covariance matrix implied by the specified model. Optimal weights between return covariance matrix and the structural covariance matrix are computed by minimizing the trace loss function.

In spite of the improvement in small sample properties, it has a limitation in that shrinkage method imposes structural covariance matrix in the stage of constructing the test statistics. This method uses the model specification via covariance matrix estimation, and uses the statistic to test the same model specification. It only makes sense only when the model is correctly specified. Therefore, it is difficult to apply this approach to the real world practice because we never able to specify the model very precisely. Moreover, this method is not plausible in that shrinkage method cannot make the advantage of merit of HJ-distance over Hansen statistic. Recall that HJ-distance is better than Hansen statistic because its weighting matrix is from second moment of returns, which is not dependent on model specifications. But shrinkage method uses different weighting matrices whenever different specifications are tested, which does not fit for the spirit of HJ-distance. In the section, we'd like to employ the seven precision matrix estimators which are completely independent from the model specifications, and see the improvement in the small sample properties.

2.4.1 Hansen-Jagannathan distance

Hansen-Jagannathan distance measures how far stochastic discount factor(SDF) implied in the specified model is from SDF of the true model that generates the asset returns. As briefly discussed above, HJ-distance computes the maximum distance between the SDF in the model of our interest and the family of SDF which possibly generates the data. Although HJ-distance can be applied to any asset pricing model, we limit our focus on the linear models.

Historically, most of the asset pricing models suggested are linear. Arbitrage Pricing Theory(APT) ([43]), Fama-French three factor model ([17],[18]), five macro factor model by Chen, Roll and Ross([11]) are the famous examples. Factor analysis and principal component analysis are the important methodologies employed for analyzing the linear asset pricing models([37],[13]). See [8] more discussion on this. Jagannathan and Wang provide convenient form of HJ-distance which enables us to use it in practice. Previous literature mostly utilizes the HJ-distance with linear pricing models. For example, Jagannathan and Wang study conditional CAPM model, cross-sectional regression models, and stochastic discount factor based models([32], [30], [31]); Campbell and Cochrane apply HJ-distance comparing several versions of CAPM models with consumption based models([7]). Hodrick and Zhang consider the specification errors of various empirical asset pricing models([26]), and there are more that make use of linear asset pricing models with HJ-distance(see [29], [34], [38], [48]). In this subsection, the derivation of HJ-distance in linear case is briefly reviewed.

Suppose we have N assets and let R_t be gross return vector at time t. A stochastic discount factor, m_t, is the factor that prices the asset. Therefore if SDF prices the return correctly, we have the condition $\mathbb{E}(m_t R_t') = 1_N$, where R_t and 1_N are $1 \times N$ vectors. Remember that R_t is the gross return, so the price of the asset should be 1, if perfectly priced. Considering K-factor model including intercept, stochastic discount factor can be expressed as $m_t = Y_t'\delta$, with $K \times 1$ vector Y_t. Y_t is the vector of linear pricing factors. This relation is so because we only consider the linear asset pricing model. Linear asset pricing model implies that SDF is the linear combination of factors with weight δ. δ is the SDF parameter. See [24] for more detail. From $\mathbb{E}(R_t Y_t'\delta) = 1_N$, the pricing error can be defined as $w_t(\delta) = R_t Y_t'\delta - 1_N$. We can estimate parameter δ by the least square of pricing error. Hansen and Jagannathan([24]) propose the distance measure as

$$HJ(\delta) = \sqrt{(E[w_t(\delta)]'G^{-1}E[w_t(\delta)]}, \tag{17}$$

where $G = E[R_t R_t']$.

HJ-distance is a measure for the quadratic pricing error weighted by covariance matrix of returns. As seen in equation (17), $G_T^{-1} = E[R_t R_t']$ is used as the weighting matrix, which does not favor the variability of pricing error. Moreover, it remains the same regardless of the model specified. Because of these plausible properties, we can use HJ-distance for comparing different asset pricing models with the same data set.

For practical application, Jagannathan and Wang ([32]) suggests the equation below as the estimate of HJ-distance

$$HJ_T(\delta) = \sqrt{(E[w_T(\delta)]'G_T^{-1}E[w_T(\delta)]}, \tag{18}$$

where $D_T = T^{-1}\sum_{t=1}^{T} R_t Y_t'$, $w_T(\delta) = T^{-1}\sum_{t=1}^{t} w_t(\delta) = D_T\delta - 1_N$ and $G_T = T^{-1}\sum_{t=1}^{T} R_t' R_t$. Moreover, δ_T can be estimated by deriving the first order condition in quadratic minimization problem.

$$\delta_T = argmin[w_T(\delta)G_T^{-1}w_T(\delta)] = (D'_T G_T^{-1} D_T)^{-1} D'_T G_T^{-1} 1_N \qquad (19)$$

Hansen's statistic follows χ^2 distribution since covariance matrix of pricing error is used. In spite of the nice improvement by HJ-distance, it loses the asymptotic property as second moment of returns replaces that of pricing error. Instead, [32] provides the following algorithm computing empirical p-value.

$$p = M^{-1} \sum_{j=1}^{M} I(u_j \geq T[HJ_T(\delta_T)]^2), \qquad (20)$$

where $u_j = \sum_{i=1}^{N-K} \lambda_i v_{ij}$. v_j is $\chi^2(1)$ random draws for M times and λ_i is non-negative eigenvalues of

$$\psi = S^{1/2}G^{-1/2}[I_N - G(^{-1/2})'D(D'G^{-1}D)^{-1}D'G^{-1/2}](G^{-1/2})'(S^{1/2})', \qquad (21)$$

where $S = E[w_t(\delta)w_t(\delta)']$ and $D = E[R'_t Y_t]$. In practice, we replace S and D with the usual consistent estimates $S_T = \frac{1}{T}\sum_{t=1}^{T} w_t(\delta_T)w_t(\delta_T)'$ and D_T.

2.4.2 Simulation Results

Simulation scheme here is similar to the one from Section 2.3. First, asset returns are generated from three factor model, i.e. $K = 4$ including intercept. Three factor linear model is expressed as $R_{ti} = \alpha + X_{t1}\beta_{1i} + X_{t2}\beta_{2i} + X_{t3}\beta_{3i} + e_{ti}$, with X_t's are the factors and e_{ti} is the disturbance or idiosyncratic risk. Factor X_t's are drawn from $N(0.0022, 6.944 \times 10^{-5})$, factor loading β's are drawn from $U(0, 2)$, and e_{ti}'s are from $N(0, 6.944 \times 10^{-5})$. In order to see the difference across various dimensions, the number of asset is set either to N=25 or to N=100, and the sample size is set to T=160, T=330, or T=700. The simulation setting is taken after [1] and [42] for the comparison purpose.

The simulation results are given in table 6 through table 14. When the sample size is large relative to dimension such as N=25 and T=700 case, all the estimators show good performances because the rejection rate is close enough to the p-value. Looking at the high dimensional cases, the sample covariance matrix turns out to have severe overrejection especially in N=100, T=160 case. The rejection rates are 99.9%. Other estimators still have the same problem. Only MST shows improvement in small sample properties (table 8). For instance, MST gives reasonable empirical rejection rates in N=25/T=330 or N=100/700 cases. Since all the other estimators are practically the same as the inverse of sample covariance matrix, MST is the only option that we can replace the sample covariance matrix for the small sample size with high dimensionality.

Table 6: Modified Adjusted estimator

	p-value	T=160	T=330	T=700
N=25	1%	5.1	2.1	1.4
	5%	14.3	9	7.3
	10%	24.1	15.4	13.1
N=100	1%	99.4	48.3	11.2
	5%	99.9	68.1	27.6
	10%	99.9	78.4	38.7

Table 7: Modified Perron-type estimator

	p-value	T=160	T=330	T=700
N=25	1%	5.1	2.1	1.3
	5%	14.2	9	7.2
	10%	24.1	15.4	13
N=100	1%	99.4	48.3	11.7
	5%	99.9	68.1	27.6
	10%	99.9	78.8	39.3

Table 8: Modified Stein estimator

	p-value	T=160	T=330	T=700
N=25	1%	2.6	1.6	0.9
	5%	8.6	6.7	6.2
	10%	16.2	13.3	11.2
N=100	1%	46.1	10.8	3.5
	5%	73	28.6	12.6
	10%	83.3	41	22.1

Table 9: The usual estimator

	p-value	T=160	T=330	T=700
N=25	1%	5.1	2.1	1.4
	5%	14.7	9.1	7.3
	10%	24.5	15.5	13.1
N=100	1%	99.5	48.7	11.3
	5%	99.9	68.5	27.8
	10%	99.9	79	39.4

Table 10: Efron-Morris-type estimator

	p-value	T=160	T=330	T=700
N=25	1%	5.1	2.1	1.4
	5%	14.7	9.1	7.3
	10%	24.5	15.5	13.1
N=100	1%	99.5	48.7	11.3
	5%	99.9	68.5	27.8
	10%	99.9	49	39.4

Table 11: Dey Estimator

	p-value	T=160	T=330	T=700
N=25	1%	5.1	2.1	1.4
	5%	14.7	9.1	7.3
	10%	24.5	15.5	13.1
N=100	1%	99.5	48.7	11.3
	5%	99.9	68.5	27.8
	10%	99.9	49	39.4

Table 12: Adjusted Haff-type estimator

	p-value	T=160	T=330	T=700
N=25	1%	5.1	2.1	1.4
	5%	14.7	9.1	7.3
	10%	24.5	15.5	13.1
N=100	1%	99.5	48.7	11.3
	5%	99.9	68.5	27.8
	10%	99.9	49	39.4

Table 13: Diagonal Variance Matrix

	p-value	T=160	T=330	T=700
N=25	1%	1.2	0.7	0.8
	5%	5.1	5.3	4.3
	10%	11.2	11.5	8.9
N=100	1%	0.2	0.1	1
	5%	2.2	2.6	4.9
	10%	7	7.6	9.6

Table 14: True covariance

	p-value	T=160	T=330	T=700
N=25	1%	1.4	0.5	1.4
	5%	5.2	4.5	4.7
	10%	10.8	9.1	10.5
N=100	1%	1.9	1.8	1.8
	5%	7.5	7.2	6.5
	10%	15.4	14.2	12.8

Furthermore, additional estimator is taken into consideration: diagonal variance matrix. Diagonal variance matrix is computed by simply suppressing off-diagonal elements of sample covariance matrix to zeros, and taking inverse. Of course, this estimator is somewhat unreasonable because we ignore the correlation among the asset returns in the first place.

Recall that the asset returns are generated from three common factors with linear model. Therefore the correlations among the assets exist significantly, and yet our diagonal variance matrix only counts the variance. Surprisingly, empirical rejection rates of HJ-distance using diagonal variance matrix are very close to theoretical *p*-values even in N=100, T=160 case (table 13). Unlike all the other estimators, the empirical rejection rates of diagonal variance matrix indicate underrejection problem in high dimensional case.

However, the deviation of the rejection rate of diagonal variance matrix seems not a big problem in that we still have the overrejection even with the simulation with true covariance matrix(table 14). Hence, we may conclude that we can obtain plausible testing results by suppressing the correlations to zeros. The pair plots of *p*-values across the covariance matrix estimators are provided in figure 3 through 5.

The plots support the same conclusion as the tables. Comparing with the true covariance matrix (upper right corner), all the empirical *p*-values of the estimators look too small to be almost identical to x or y axis. *p*-values of MST is the only one that can be comparable to true covariance matrix, although *p*-values of MST is not perfectly aligned with 45 degree line. Pair plots of all the estimators with true covariance matrix are getting closer to 45 degree line as the sample size increases, but MST is still the closest to true covariance matrix.

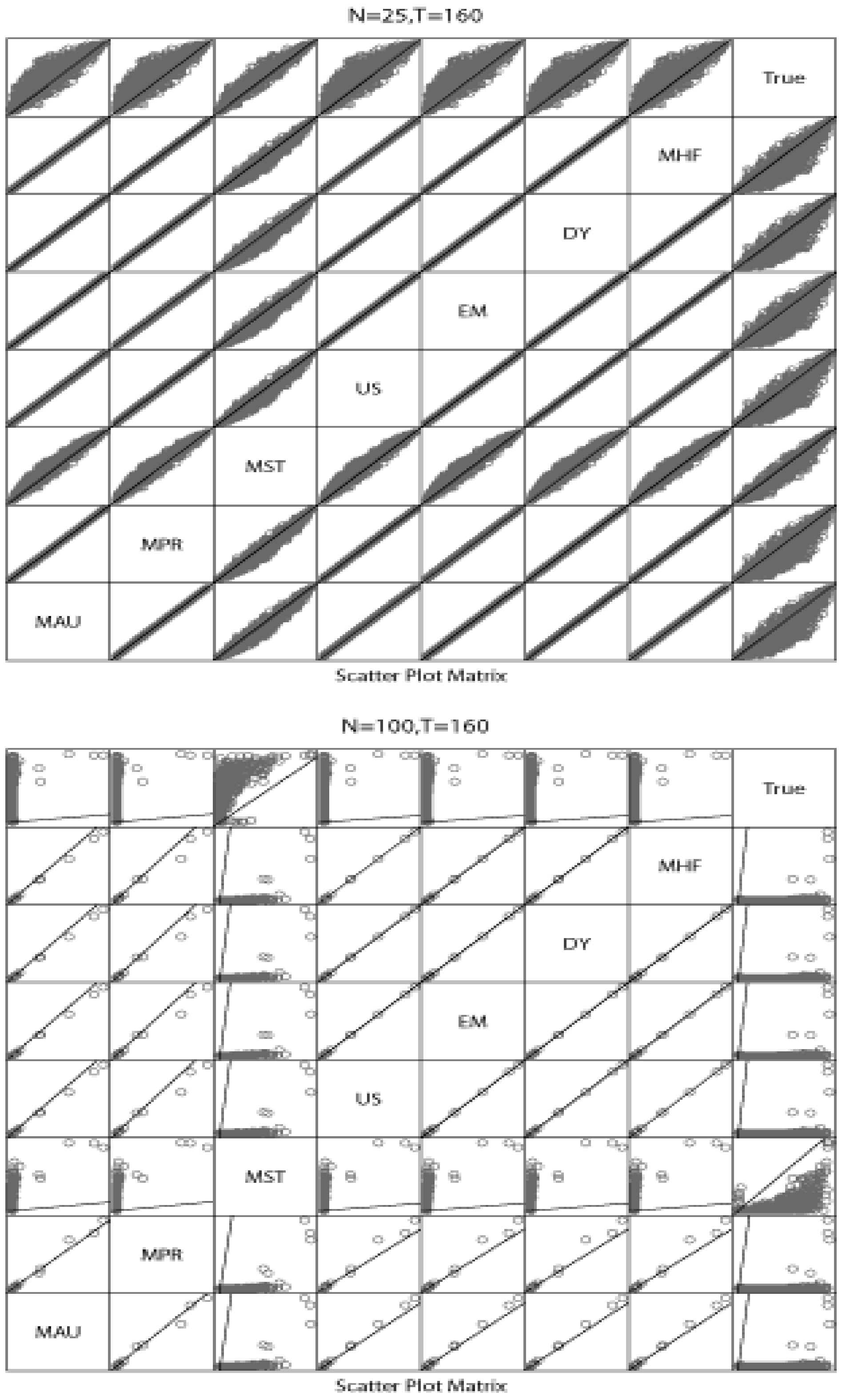
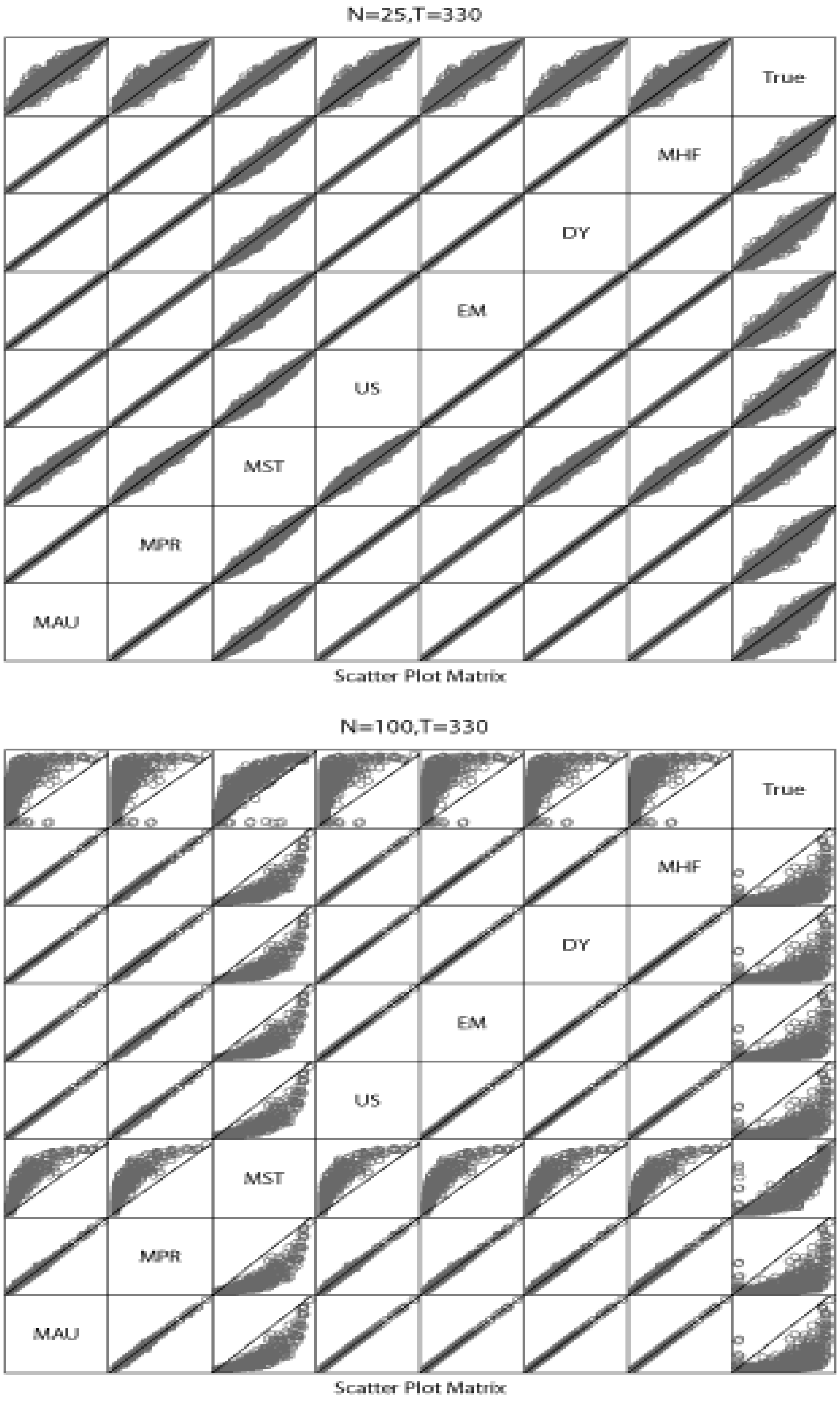


Figure 3: Pair Plot of p-values: T=160



N=25,T=330

N=100,T=330

Figure 4: Pair Plot of p-values: T=330

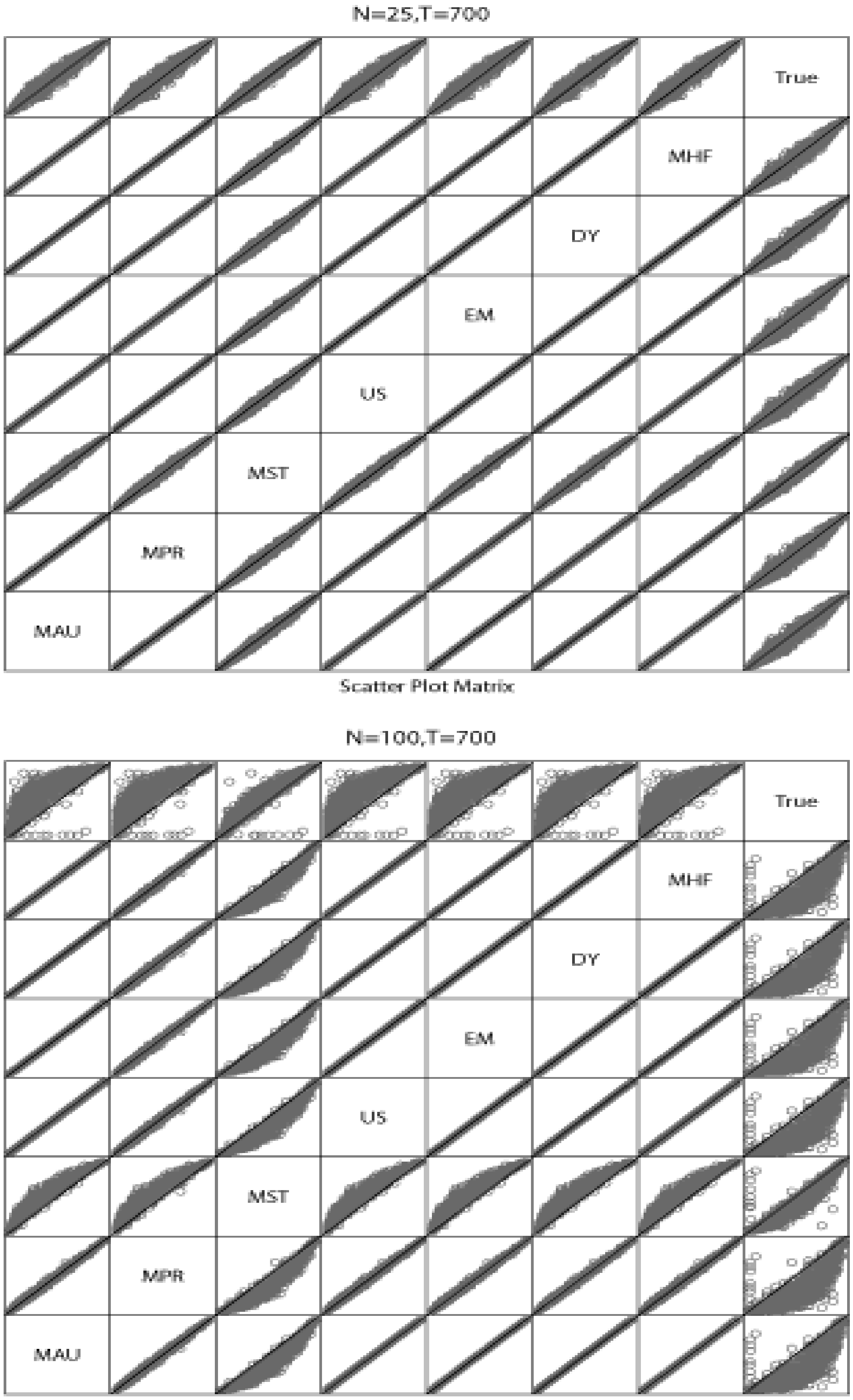


Figure 5: Pair Plot of p-values: T=700

2.4.3 Simulation of Misspecification

Simulation so far portraits the situation where the asset pricing model is correctly specified. Remember the argument about the advantage of HJ-distance over Hansen statistic. Our estimators as weighting matrices are invariant to model specifications in HJ-distance, and we would like to explore the estimators from different aspects in this subsection. We expect that the HJ-distance successfully differentiates various candidate models from the true models with proper distance or degree because our estimators do not change over the different model specifications.

Suppose the true data generating process is a four factors linear model. If one candidate model correctly specifies three factors while the other candidate takes only two of them into its specification, then the HJ-distance of the first model specification should be less than that of the second specification. In other words, our interest lies in how HJ-distance reacts to the model misspecification with different precision matrix estimators.

The simulation setting of model misspecification is as follows. In data generating process, additional factor, X_4 is introduced to linear factor model. Unlike the four-factor true model, the specified models are assumed to be three-factors, i.e. $R_{ti} = \alpha + X_{t1}\beta_{1i} + X_{t2}\beta_{2i} + X_{t3}\beta_{3i} + e_{ti}$. On the other hand, the true model is $R_{ti} = \alpha + X_{t1}\beta_{1i} + X_{t2}\beta_{2i} + X_{t3}\beta_{3i} + X_{t4}\beta_{4i} + e_{ti}$. X_4 is the missing factor in the specified model. By changing the values of coefficients β_{4i} from 0 to 0.5, we examine the HJ-distance distributions with respect to the degree of misspecification. β_{4i} being zero indicates there is no misspecification problem. Bigger β_{4i} means greater degree of misspecification. Figure 6 through 8 are the distribution plots of HJ-distance with several degrees of misspecifications.

As the degree of model misspecification gets larger, we expect that HJ-distance distribution locates farther to the right comparing with the case of $\beta_{4i} = 0$. We can exactly observe this in case of using true covariance matrix and diagonal variance matrix estimator. In addition to this, we have two interesting and intuitive observations.

As sample size T increases, the HJ-distance separates the various degrees misspecification more clearly. HJ-distributions are plotted as results of simulation of 5,000 repetitions for each model specification. Fixing the number of asset N, we can examine how HJ-distance behaves across different sample sizes. We also can observe that the more assets (larger N) we have, the better we can tell the differences between the correctly specified model and misspecified models. This result is quite intuitive because large N means that we have many cross sectional data to test the model with, which should lead to better testing outcome. This justifies the reason why we need to consider high dimensional covariance matrix estimation seriously in asset pricing. The asset pricing models can be precisely tested only with large number of assets, which requires high dimensional covariance matrix estimation.

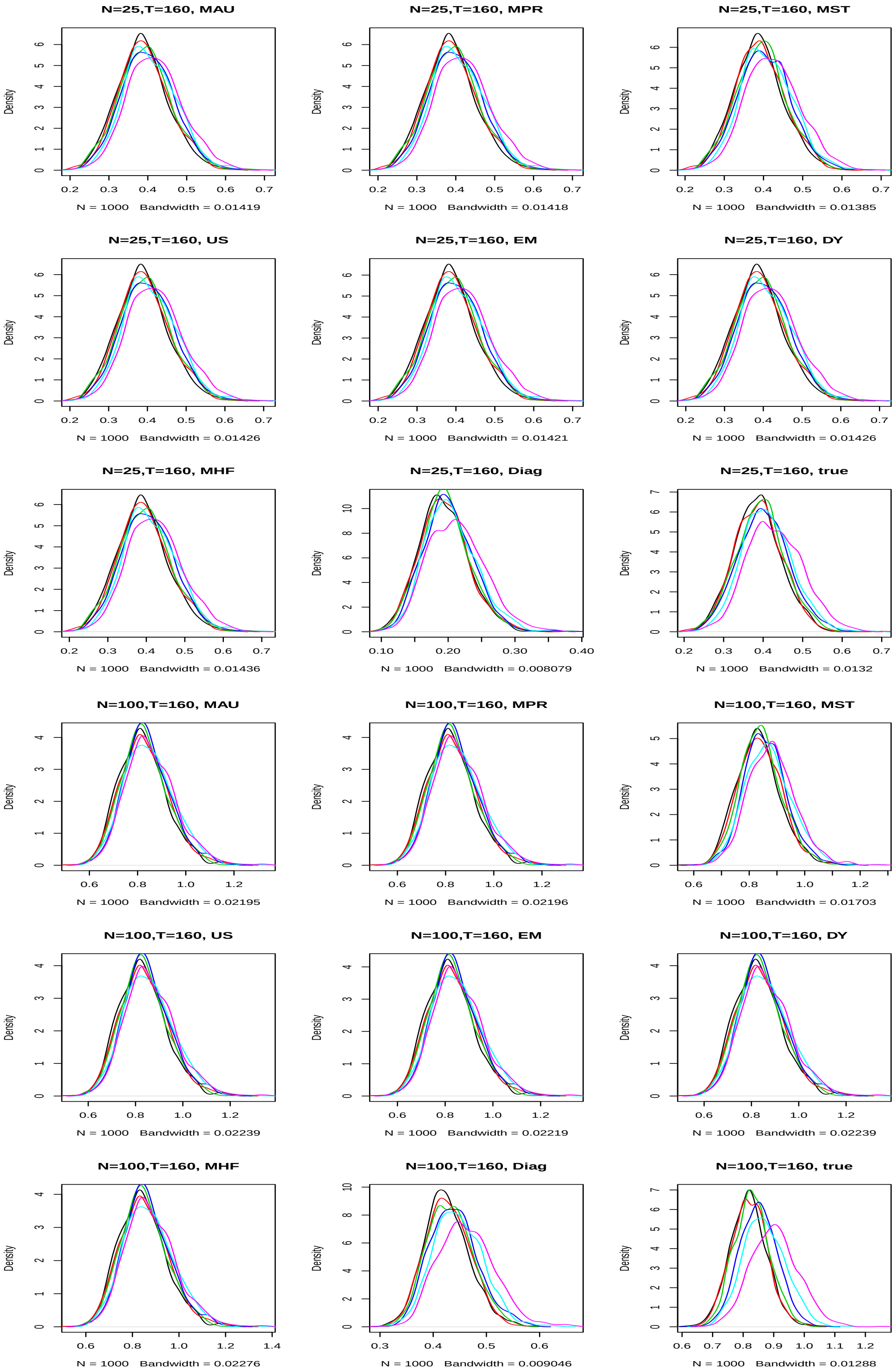


Figure 6: HJ-distance distribution: T=160

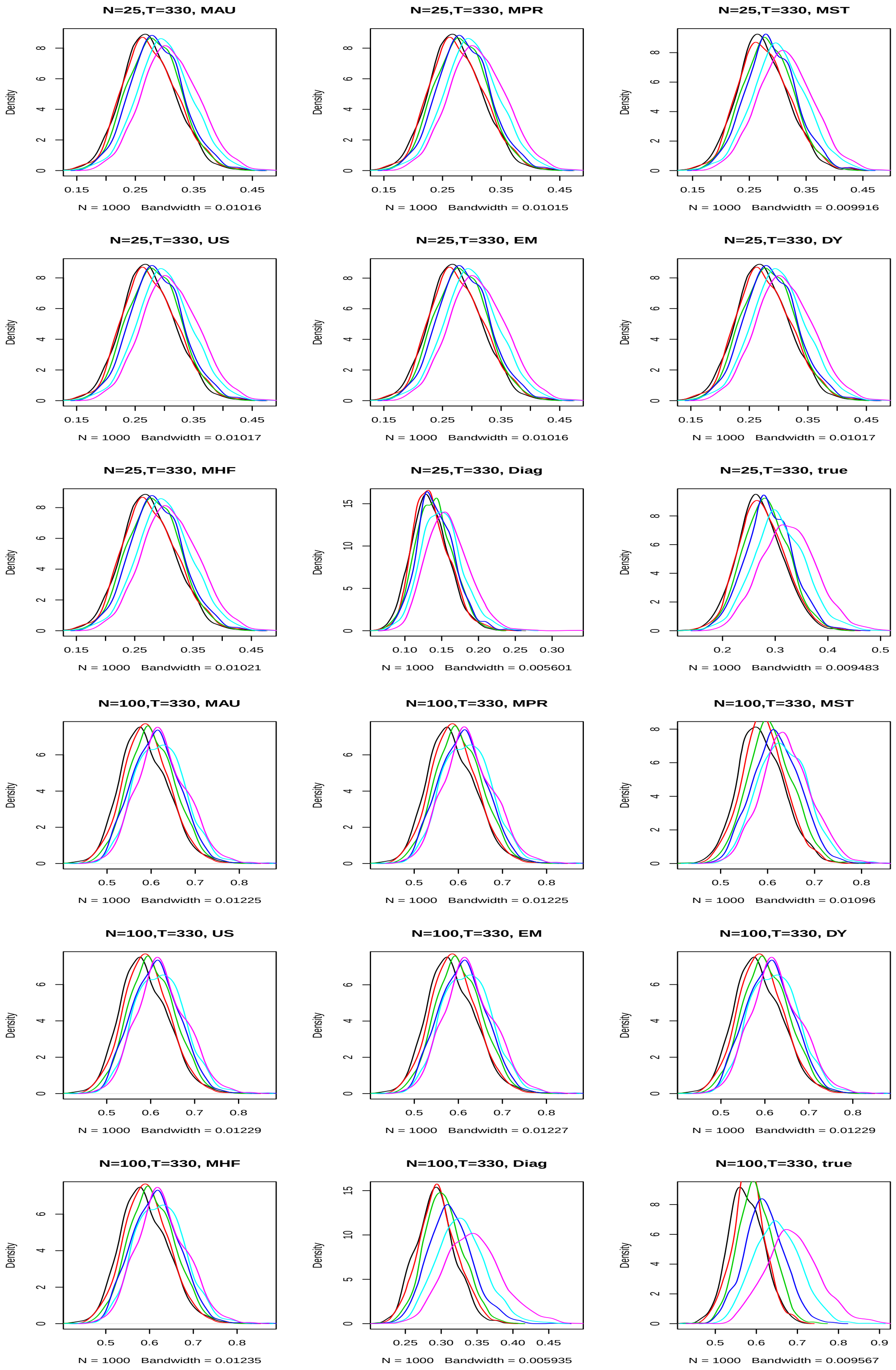


Figure 7: HJ-distance distribution: T=330

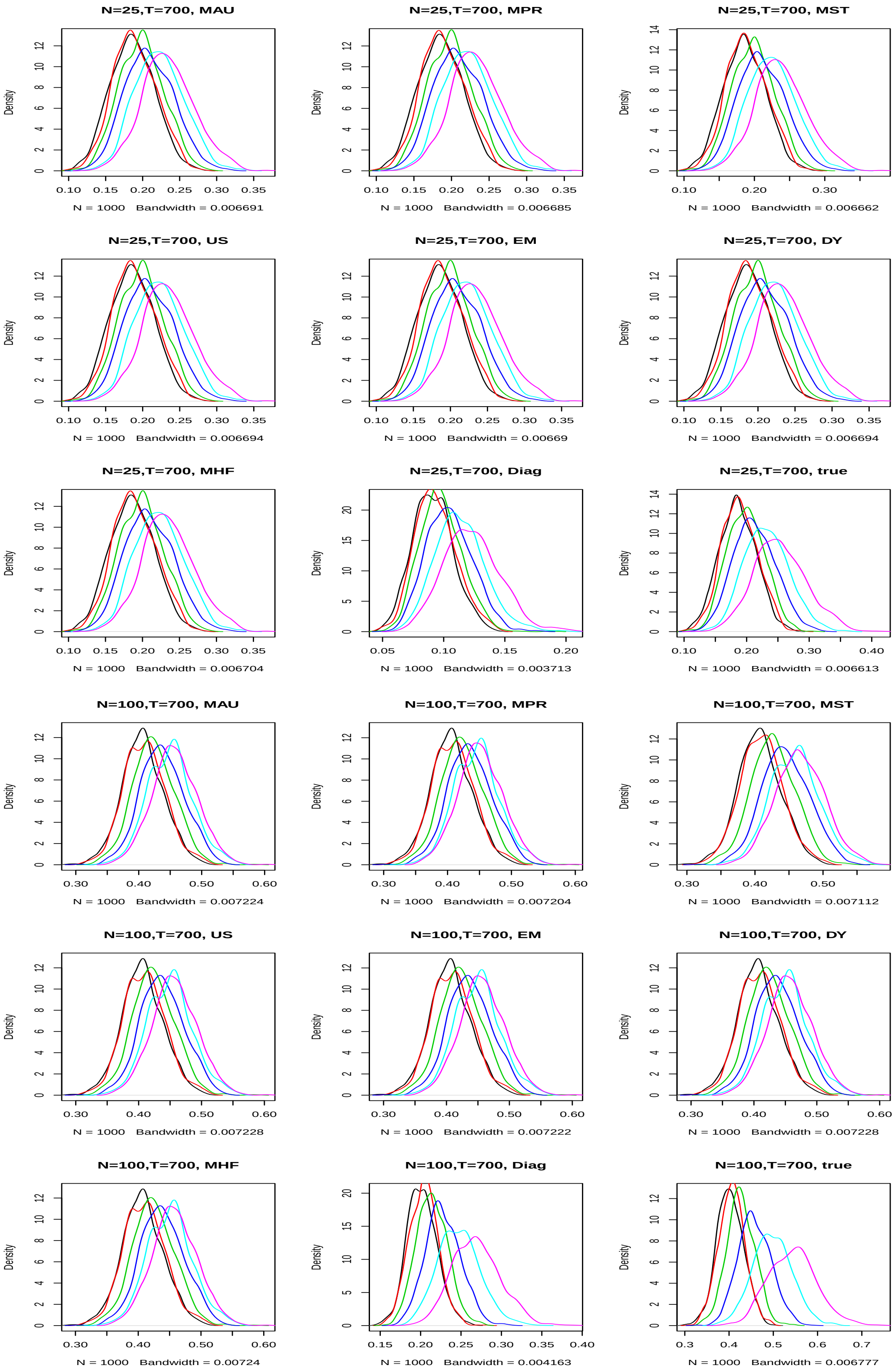


Figure 8: HJ-distance distribution: T=700

All the other estimators including MST do not perform as clearly as diagonal variance matrix estimator or true covariance matrix. Among others, MST does relatively good job in separating out the HJ-distance across different misspecified models. From table 8, we have seen that MST is the best when we look at type-1 error. MST also performs the best when type-2 error is considered. This is so because the more clearly HJ-distance differentiates different levels of model misspecification, the less probable we accept the wrong specified model as the true one.

2.5 Summary

Estimating high dimensional covariance matrices is challenging due to the large number of parameter to be estimated. In this chapter, we explored the effects of high dimensional covariance matrix estimators in the context of financial asset pricing. Empirical financial studies often require covariance matrix estimators for weighting matrices. Estimating covariance matrix for the financial panel data, moreover, faces small sample properties.

By simulation studies, we conducted simulation experiments on the effects of several covariance matrix estimators both in two-pass procedure and in Hansen-Jagannathan distance. We find that MST works pretty well in GLS setting even with complicated correlation structure, while sample covariance matrix produces too much error. On the other hand, in testing environment of HJ-distance, diagonal variance matrix works well. Among the seven precision matrix estimators, MST improves the small sample properties, yet not a match for diagonal variance matrix. Moreover, by the experiment of model misspecification, we have shown that MST is the best candidate in measuring type-2 error among seven estimators. Again, our heuristic estimator of diagonal variance matrix is even better than MST in the case as well. We also discusses that high dimensional covariance matrix is important issue especially in the asset pricing model because the cross-sectional relationship is crucial in testing the model.

Based on the observation that we find in this chapter, we can go forward to interesting research projects. As pointed out in the first section, the volatility of stock returns is not constant over time, so that empirically constructed minimum-variance portfolio performs unstable. We may apply the new precision matrix estimators to make minimum-variance portfolio and see how the portfolio performs. Another possible application is the option pricing. Covariance matrix must play an important role in a basket option with many underlying assets, therefore precise covariance matrix estimation is crucial in pricing. Value-at-Risk of portfolios consisting of complicated securities seems an interesting issue as well. From theoretical point of view, the convergence speed of seven estimators is meaningful. In addition, mathematical study on the relation of ratio of dimension to sample size might enlarge our understanding of high dimensional covariance estimate.

CHAPTER III

COVARIANCE MATRIX ESTIMATION AND GENERALIZED LEAST SQUARES

3.1 Introduction

Among other statistical methodologies, Ordinary Least Square(OLS) has been one of the most important and practical techniques both in theories and applications in various fields. Especially in empirical studies, OLS is the most commonly adopted method because it provides very useful statistical tools such as t-statistics or R^2.

Moreover, OLS estimator has other plausible properties such as unbiasedness and consistency. Efficiency, in particular, is considered the most important benefit of OLS estimates. By Gauss-Markov theorem OLS estimator is proven to be the least variance among linear unbiased estimator. We will briefly review the usual assumptions imposed on OLS analysis.

$$y = X\beta + \epsilon \tag{22}$$

Where y and ϵ are $n \times 1$ vectors while X and β are $n \times k$, $k \times 1$ respectively. The classical assumptions are:

1. Regressor X is non-stochastic.

2. $\mathbb{E}(y) = X\beta$ and $var(Y) = \sigma^2 I_n$ for some $\sigma > 0$

3. y is a random vector following multivariate normal distribution, i.e. $y \sim MN(X\beta, \sigma^2 I_n)$.

Although the useful statistical properties of OLS are derived from these classical conditions, sometimes these should be relaxed when dealing with more realistic models or data set. In the case of violating the second assumption, i.e. the errors being homoscedastic and uncorrelated to each other, Generalized Least Square(GLS) model has to be introduced as a remedy. Portfolio analysis is one of many examples. If we are interested in the risk of the portfolio value with respect to oil price change, a typical approach would be running a regression using historical returns of the portfolio on oil price changes. In this case, it is not reasonable to preserve the second classical assumption because stock returns might be significantly correlated to each other even after conditioning with oil price changes. A portfolio containing Exxon mobile and Shell will be an example. In reality , since it is impossible to introduce all meaningful conditioning variables in the model especially in social sciences, OLS is not always the best option in regression analysis.

In spite of the shortcoming explained above, GLS is not as often used as it should be. The main reason is the covariance matrix. As will be explained in detail in the following section, GLS is basically transforming the original variables with covariance matrix to satisfy the classical conditions of OLS. However, covariance matrix is unknown in most cases. Furthermore, we do not have any guideline when GLS has to be used or under what kind of covariance structure it is even more useful to employ GLS than OLS.

Motivated by importance of GLS, we would like to explore the covariance structure to study the effects of covariance matrix estimation on GLS. Sparse, diagonal and factor covariance matrix are mainly considered as the essential ingredients of the problem. Guided simulations from analytical derivations will be shown. The rest of the chapter is organized as follows. We will start with brief review of GLS theory.

In section 3.4, new covariance estimators will be suggested. Section 3.5 discusses the conditions under which GLS is even more efficient, followed by the section of factor covariance matrix.

3.2 GLS overview

Problems that classical assumptions do not hold are often found in practice. Generalized Least Square(GLS) is a remedy for the breakdown of the second assumption that covariance matrix of y or ϵ, is scalar variance, σI. In this situation, employing OLS estimate does not guarantee the properties of OLS. OLS estimator would not be the most efficient estimator, i.e. no longer Best Linear Unbiased Estimator(BLUE). Since standard hypothesis tests are based on the scalar variance assumption, they are not valid either. Therefore we need a method that allows more general forms of linear model. In this section we briefly review GLS theory. Overall discussion both on OLS and GLS is well described in [41],[23]. For details on GLS see [35].

3.2.1 OLS with general covariance matrix

We have the model as in (22):

$$y = X\beta + \epsilon,$$

where y and ϵ are $n \times 1$ vectors while X and β are $n \times k$, $k \times 1$, respectively. The assumptions are the same as before except that $var(y) = var(\epsilon) = \Sigma$. Σ needs not to be diagonal. The only requirement is positive-semi definite symmetric matrix. Now the OLS estimate becomes:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y.$$

Even under our relaxed assumption, the OLS estimate is still unbias as long as $\mathbb{E}(\epsilon) = 0$. Under general covariance structure, the variance of OLS has a form as,

$$var(\hat{\beta}_{OLS}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}. \tag{23}$$

From this, we can see that statistical inferences are not valid any more because all the inferences such as $t - statistics$ or $F - statistics$ are based on the fact that $var(\hat{\beta}_{OLS}) = \sigma(X'X)^{-1}$. Moreover, OLS estimator is not BLUE. This will be verified by GLS version of Gauss-Markov Theorem. Therefore, under the relaxed assumption on the general covariance structure of y or equivalently ϵ, we need to have different approach from OLS.

3.2.2 Generalized Least Square Estimator

GLS, as a remedy for the violation of the second assumption, is basically transforming the model in order to satisfy the classical OLS conditions. Consequently, the transforming matrix is turned out to be $\Sigma^{-1/2}$. The derivation and properties of GLS estimator, also called as Aitken estimator, will be discussed.

Since covariance matrix Σ is positive semi-definite and symmetric, it can be spectral-decomposed as follows.

$$\begin{aligned} \Sigma =& U\Lambda U' \\ \Sigma^{-1} =& U\Lambda^{-1}U', \end{aligned}$$

with positive diagonal matrix Λ and orthogonal matrix U. If we take $G' = U\Lambda^{-1/2}$ to transform the model (22), it becomes

$$Gy = GX\beta + G\epsilon.$$

By renaming the transformed variables with asterisk mark, we have GLS model as,

$$y_* = X_*\beta + \epsilon_*. \tag{24}$$

Then, the covariance matrix of y is computed as

$$\begin{aligned} var(y) = var(\epsilon) &= \mathbb{E}(\epsilon_*\epsilon_*') \\ &= G\Sigma G' \\ &= \Lambda^{-1/2}U'U\Lambda U'U\Lambda^{-1/2} \\ &= I_n. \end{aligned} \tag{25}$$

With transformed variables, we can apply OLS procedure to get GLS estimator.

$$\begin{aligned} \hat{\beta}_{GLS} &= (X_*'X_*)^{-1}X_*y \\ &= (X'U'UX)^{-1}X'U'Uy \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \end{aligned} \tag{26}$$

And its variance-covariance matrix is computed as follows.

$$\begin{aligned} var(\hat{\beta}_{GLS}) &= var((X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y) \\ &= (X'\Sigma^{-1}X)^{-1} \end{aligned} \tag{27}$$

By Gauss-Markov theorem, we can verify that (27) is, in fact BLUE. In addition, if we impose normality of ϵ, we can construct log-likelihood function as

$$L = -\frac{n}{2}log(2\pi) - \frac{1}{2}log(|\Sigma|) - \frac{1}{2}(y - X\beta)'\Sigma^{-1}(y - X\beta).$$

Then, by taking differentiating L with respect to β for the first order condition, we get $X'\Sigma^{-1}(y - X\beta) = 0$, which leads to the maximum likelihood estimator,

$$\hat{\beta}_{MLE} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y.$$

And this is the same as what we get as GLS estimator (26). Furthermore, we have the information matrix as below.

$$\mathbb{E}[X'\Sigma^{-1}(y - X\beta)(y - X\beta)'\Sigma^{-1}X] = X'\Sigma^{-1}X$$

Applying Cramer-Rao bound, we have the minimum variance $(X\Sigma^{-1}X)^{-1}$, which is simply the variance of GLS as (27). Summing up, imposing normality of ϵ, GLS estimator is equivalent to maximum likelihood estimator and it is the most efficient estimator, or best unbiased estimator(BUE).

3.2.3 Feasible Generalized Least Squares

As seen above, GLS estimator is a reasonable remedy for non-scalar covariance structure. We also saw that GLS is the most efficient estimator. However, it is difficult to obtain GLS estimator in practice because the covariance matrix Σ is not available in most cases. Feasible Generalized Least Squares(FGLS) is the estimator with available covariance estimator $\hat{\Sigma}$ for Σ:

$$\hat{\beta}_{FGLS} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}y.$$

However, in most often situation we encounter in practice, we only have one observation in each y_i. Thus, it is impossible to have an estimate for covariance matrix Σ. Even though we have multiple observations such as in panel data analysis, estimating Σ is not a simple job because $n(n+1)/2$ parameters are to be estimated. The usual parametric approach is to impose assumptions on Σ with simple covariance structure. One example is the case of serial correlation.

Suppose we believe that y_i's are serially correlated as in AR(1) model, i.e.

$$Cov(y_t, y_{t-i}) = \rho^i\sigma.$$

Then, we have the expression for Σ and G as below.

$$
\begin{aligned}
\Sigma &= \sigma^2 \begin{pmatrix} 1 & \cdots & \rho^{n-1} \\ \vdots & \ddots & \\ \rho^{n-1} & \cdots & 1 \end{pmatrix} \\
G &= \Sigma^{-1/2} \\
&= \frac{1}{\sigma^2} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \frac{-\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} & 0 & \cdots \\ 0 & \frac{-\rho}{\sqrt{1-\rho^2}} & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & & \frac{1}{\sqrt{1-\rho^2}} \end{pmatrix}
\end{aligned}
$$

FGLS with this transformation is called Cochrane-Orcutt estimation [10]. The estimation problem of Σ is reduced to estimate one parameter, ρ. Since we have at least n observations to estimate this, GLS is quite feasible. There are other cases in which FGLS is doable. Another simple example would be heteroscedasticity where off-diagonals of covariance matrix are assumed to be zeros, but the diagonals are not necessarily the same. Since we only need to estimate n estimates of diagonals, FGLS is also doable under this assumption.

These parametrical approaches provide useful solutions in a few cases, yet not reasonable to be generalized in practice: they impose too strong assumptions on covariance matrix. As a matter of fact, covariance structure is unlikely to be known at all. Therefore, these parametrical approach to FGLS is not actually feasible in many cases.

In this chapter, we will explore effects of covariance matrix structure from various aspects in order to get a guideline for FGLS estimates. Heuristic approach of banding is studied and sparsity, factor models are considered. Guided simulations by analytical reasoning will be provided as well.

3.3 Sparsity and FGLS estimation

Let's look at the regression model of (22):

$$y = X\beta + \epsilon.$$

Suppose y and ϵ are $n \times 1$. As emphasized previously, since we only have one observation for each y_i, it is impossible to estimate covariance matrix of ϵ_i directly without imposing additional assumptions on ϵ. One possible approach to FGLS, is via OLS as follows:

(1) Run OLS to get residual vector e.

(2) Take ee' to obtain $n \times n$ matrix.

(3) Make some changes of ee' to estimate Σ.

Since e is $n \times 1$ vector, ee' is rank 1 and thus not invertible. Thus it cannot serve as an estimate for covariance matrix Σ. If we want to put an assumption that the matrix is somewhat sparse, or, some of the off-diagonal elements of Σ are zeros, then we have two natural ways of transforming the matrix: truncation and banding. Truncation is a method that suppresses matrix elements to zeros if the elements do not meet the pre-specified criteria. In this case, from our assumption of sparse covariance matrix, the off-diagonal elements whose absolute values are smaller than certain level are set to be zeros.

Banding is suppressing sub- and symmetrically corresponding super-diagonals to zeros. For example band-1 of 5×5 matrix is setting (5,1) and (1,5) zeros and band-2 is to set (4,1), (5,2) and (1,4), (2,5) to zeros in addition to the result of band-1. In other words, it is setting the sub- and super-diagonals zeros inward from the very last sub- and super-diagonal.

To check if these methods work, the invertibility should be taken into consideration. After banding and truncation with different degrees, the number of non-zero singular values are counted. The simulations follow procedure as below.

(1) Generate the data, $y = X\beta + \epsilon$, by random draws of X, β, and ϵ.

(2) Run OLS to get residual e, and compute ee'.

(3) Banding and truncating ee', then count the number of non-zero singular values for each degree.

(4) Numerical threshold of zero is set to 10^{-10}.

The plots for the number of non-singular values using truncation and banding are provided in figure 9 and figure 10, respectively. Note that the dimension n is set to be 25 for truncation simulation. The thresholds of truncation procedure is $4 \times i$ percentage quantiles, $i = 1, 2, \cdots, 25$.

As seen in figure 9, the numbers of non-zero singular values are not 25 in most cases. Only the most truncated case, or diagonal matrix case, gives the invertible matrix estimation. Therefore, the other truncation strategies are not usable since

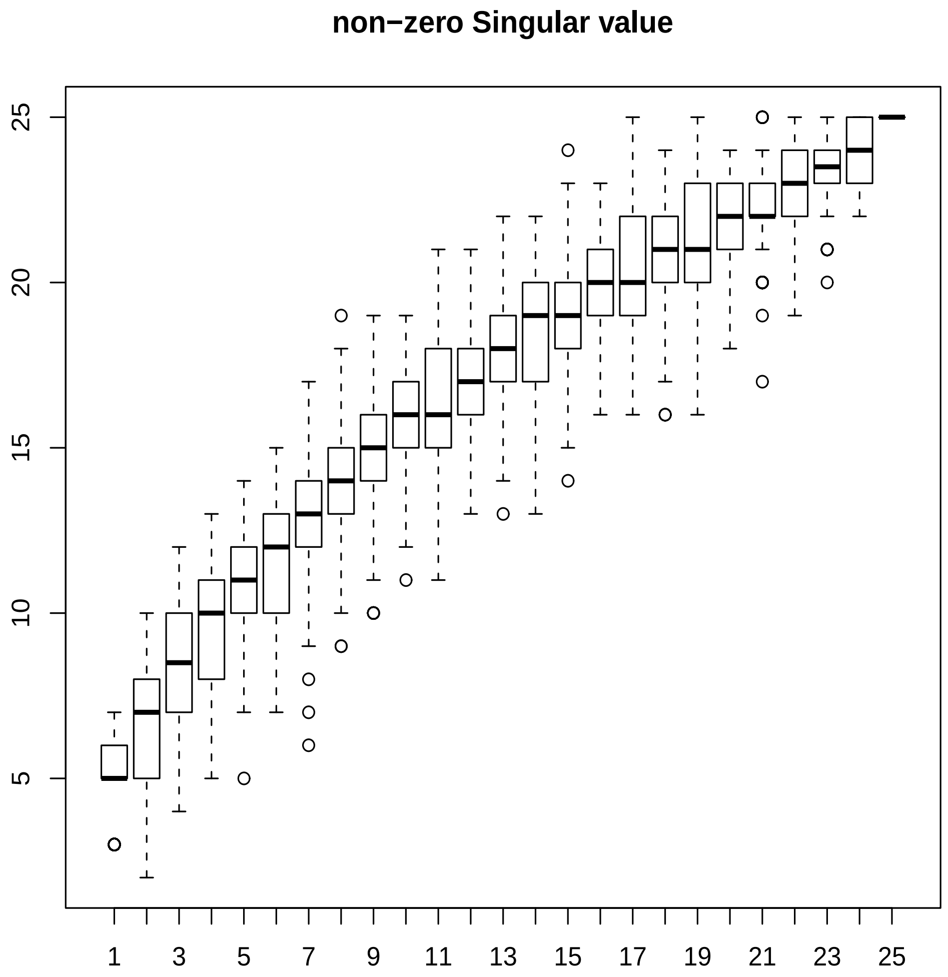
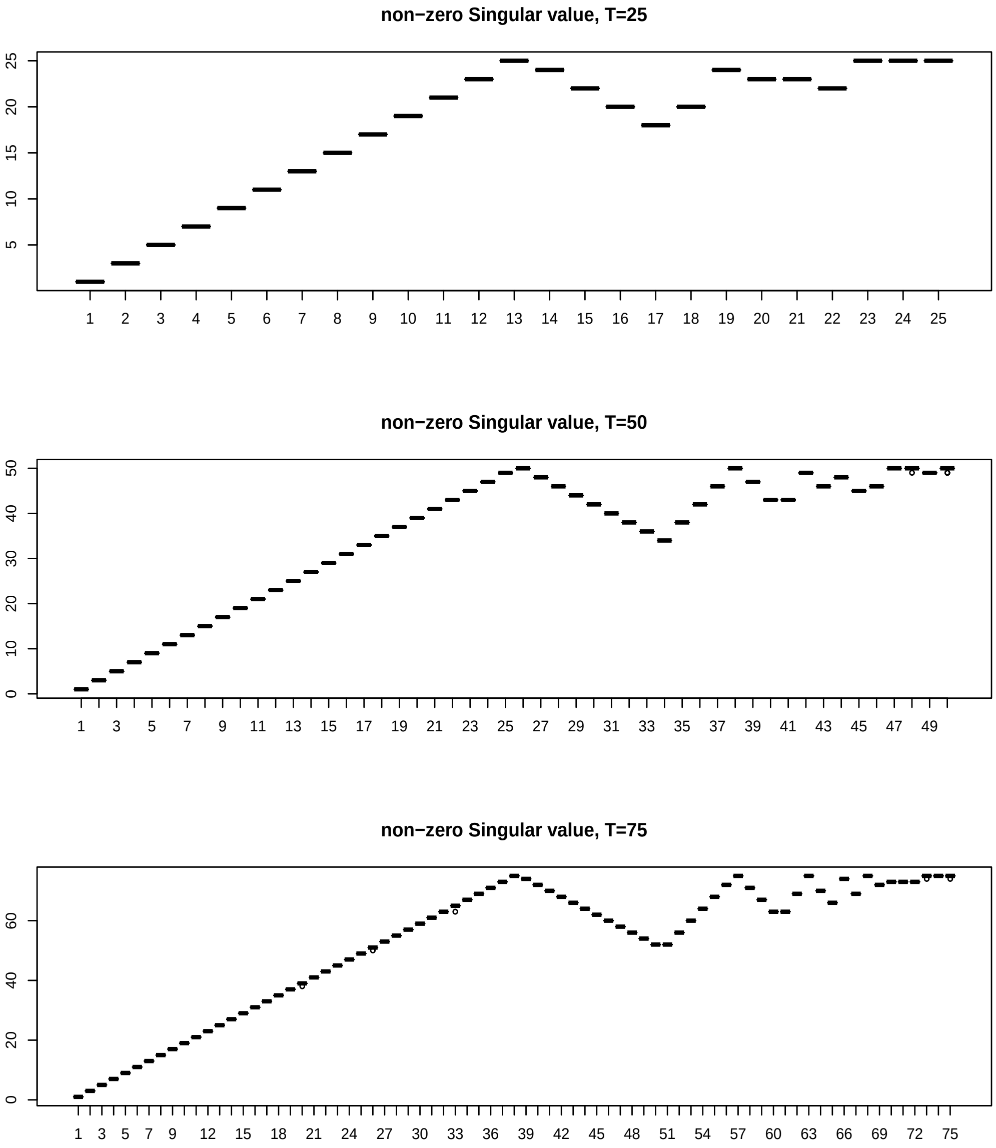


Figure 9: Number of non-zero Singular Values:Truncation

inverse of the covariance matrix estimate is needed in GLS estimation. The next candidate is banding. Figure 10 shows the cases with three different dimensionality, i.e. $n = 25, 50, 75$. The plot shows very interesting behavior of singular values.

The number of non-zero singular values increases gradually and hits the full rank matrix when the banding is half the dimensionality and then oscillates thereafter. The same behavior is shown under different dimensionality. More interesting observation is that almost all the random repetitions give the same result so that given the same banding criteria the same number of non-zero singular values are returned.


non−zero Singular value, T=25

non−zero Singular value, T=50

non−zero Singular value, T=75

Figure 10: Number of non-zero Singular Values:Banding

3.3.1 Banding Strategy

Given this simulation result, we can try FGLS with different banding schemes. In this simulation, we assume three regressors, i.e. $k = 3$. The procedure is as follows.

(1) Randomly generate X, β from $N(0.0022, 0.00006944)$ and $U(1, 2)$, respectively.

(2) Generate Cholesky lower triangular matrix Q by drawing random non-zero elements from $U(-0.5, 0.5)$ to make $\Sigma = QQ'$

(3) Randomly generate ϵ from multivariate normal distribution $N(0, \Sigma)$.

(4) Following the data generating process (22).

(5) Run OLS and get residuals of ee'.

(6) FGLS using banded ee'.

Summation of squared error is computed in each case as $[\beta - \hat{\beta}_{FGLS}]'[\beta - \hat{\beta}_{FGLS}]$. The box plot comparing four banding strategies and OLS is provided in figure 11. As shown in the plot, OLS estimator is better than any other estimators. Among other banded strategies, *Band25* is the best and almost as good as OLS estimator.

From this simple experiment suggests that FGLS has no advantage over OLS. The reason is the following. We have only one observation for each residual. Even if the errors are heavily correlated to each other, it is impossible to tell the difference between zero and non-zero correlation with one observation. Therefore the only information we can get from ee' is variance. In previous simulation setting, the diagonals of true covariance matrix Σ are similar to each other by construction. In order to verify the claim, diagonals are set to be spread out intentionally. In order to see the behavior GLS with connection to how far the diagonals are spread out, we set three cases as below.

- $\epsilon_i \sim log(1 + i) \times N(0, 1)$
- $\epsilon_i \sim i \times N(0, 1)$
- $\epsilon_i \sim exp(0.1 \times i) \times N(0, 1)$

The simulation results are plotted from figure 12 to figure 14.

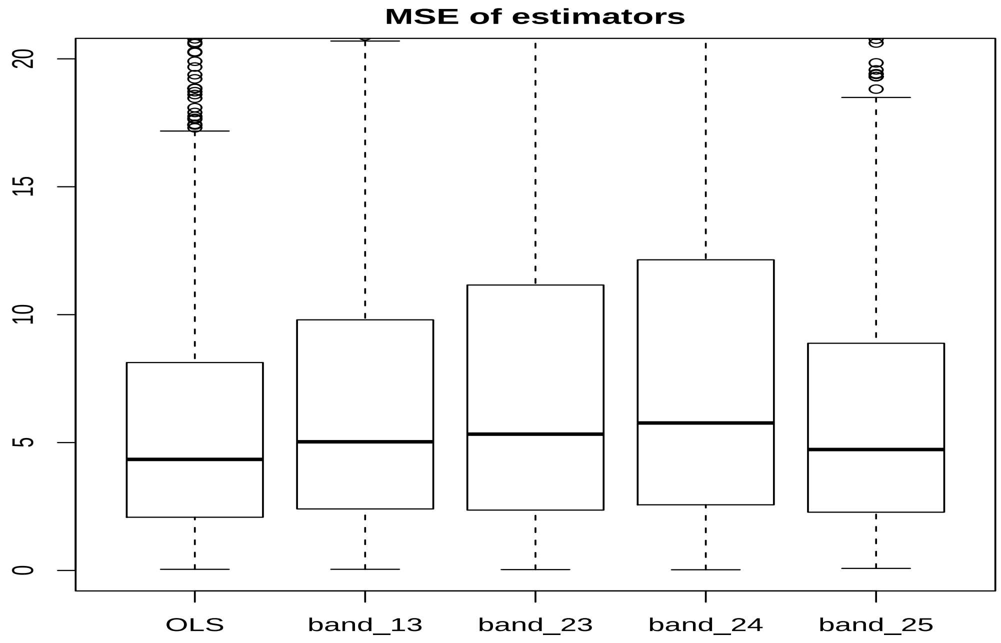


Figure 11: Box plot:$[\beta - \hat{\beta}_{FGLS}]'[\beta - \hat{\beta}_{FGLS}]$ Comparing with OLS

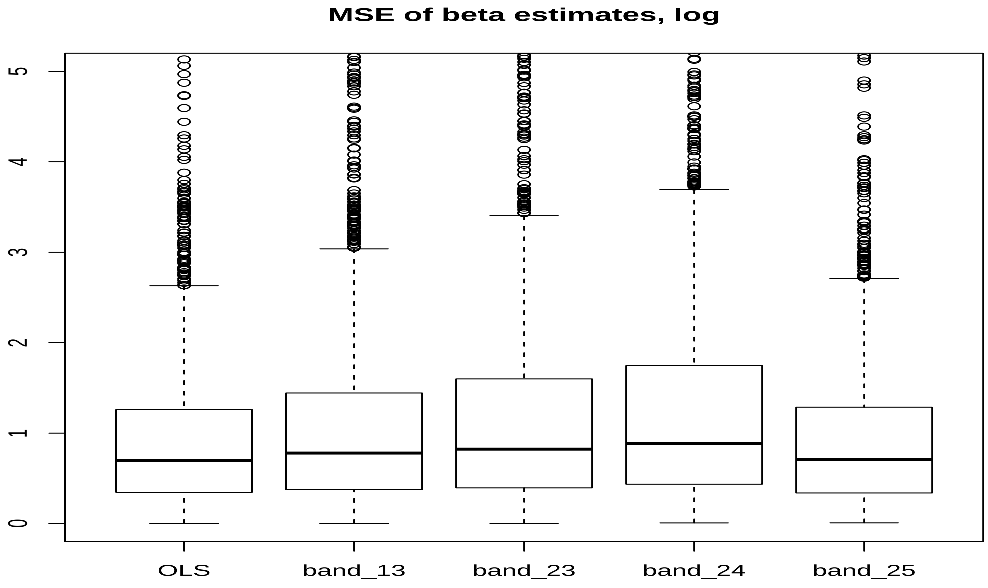


Figure 12: MSE: logarithm diagonals

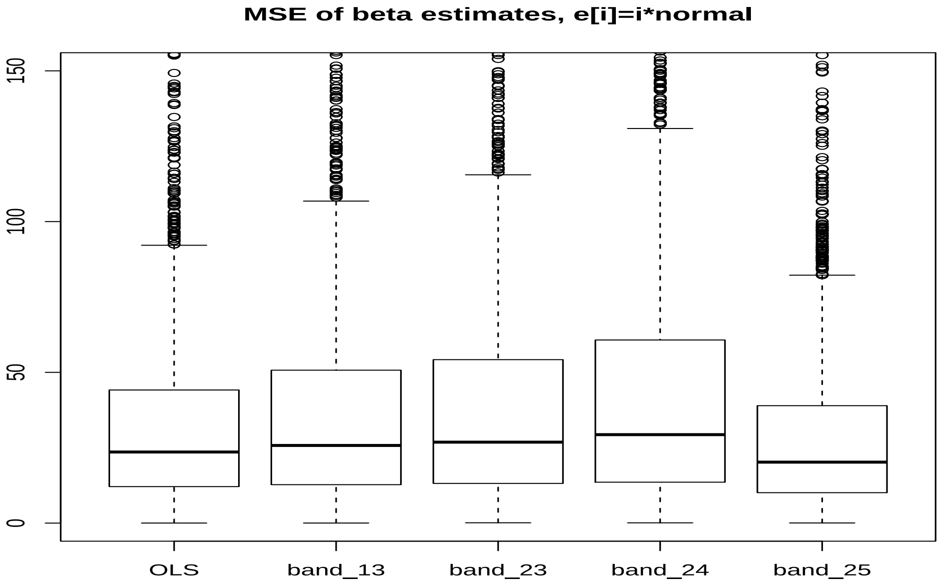


Figure 13: MSE: linear diagonals

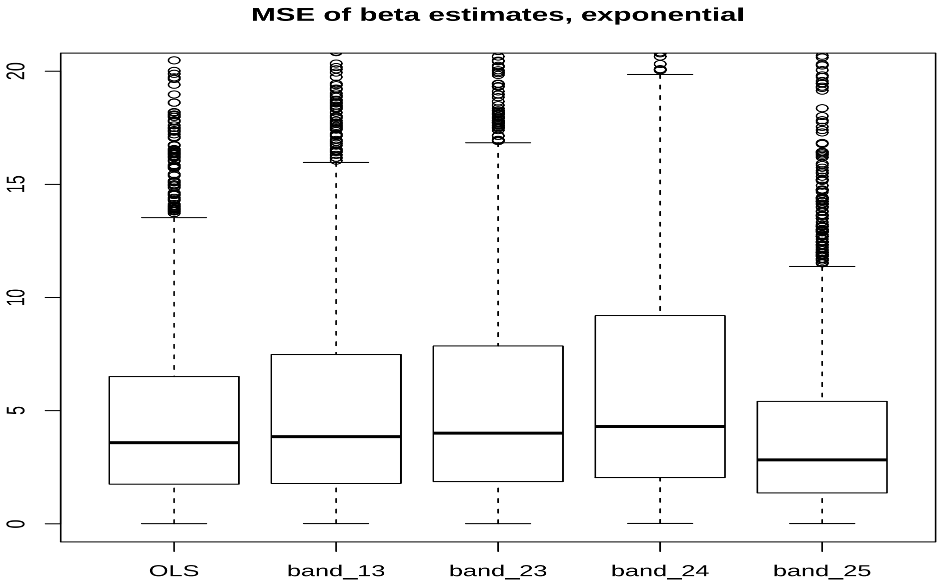


Figure 14: MSE: exponential diagonals

Clearly, the relative performance of FGLS to OLS improves and *Band25* outperforms OLS. Moreover we can see that as the degree of spread-out of diagonal elements changes from logarithm to linear and then to exponential, the relative performance gets better. The rationale of the spread-out diagonal will be explored in the next section. Putting this issue aside for a while, we will move onto sparsity.

3.3.2 Sparsity of Σ

Estimating covariance matrix is specifically challenging for it contains too many parameters. High dimensional problem gets even more challenging because the number of parameter is increasing quadratically. In reality, correlations of random variables of our interest may not be all significantly large in magnitude. Some of them could be no different from zero. Portfolio construction with many asset makes a suitable example. A portfolio consisting of different asset classes would have sparse covariance matrix in its return. Several examples of different asset classes such as forward and interest rate products are uncorrelated by construction. Even within the same asset classes, say stocks, it is usually easy to find two stocks whose correlation is very small. From numerical point of view, sparse matrix is beneficial because many of the off-diagonal elements representing covariances are zeros and we have much less parameters to estimate. Combining these two aspects of sparsity, studying sparse covariance matrix may present useful insights. We first explore the sparsity and its effects on GLS estimates to see the performance of banding strategy by simulations. And then we will move on to the analytical considerations to view the conditions on covariance matrix which make GLS estimates more efficient. Also simulation studies will follow to support the analytical conjectures.

3.3.2.1 Simulations on Sparsity

In this simulation, we'd like to examine the effects of sparsity of true covariance matrix on efficiency of linear regression estimators. We set three simulation schemes. The first one is the common case, where we only have one observation for each x_{ij}, y_i, ϵ_i . The second case is hypothetical situation where only ϵ_i's are observed multiple times. In other words, we have (22) where there are multiple observations on y_i which comes from multiple observations on ϵ_i. This hypothetical setting is explored in order to see the potential behavior as the number of observations increasing. The third simulation setting is the two-pass procedure. This is a popular method in empirical asset pricing and usually the second procedure adopts GLS with sample covariance matrix for the transformation matrix. In this simulation studies, we compare the banding strategy with OLS for the first two cases and GLS with sample covariance matrix in the third case since OLS and GLS with sample covariance matrix are the possible alternative in each case.

The measure for estimating error is the same as before. After obtaining the estimates for $\beta_1, \beta_2, \beta_3$ using different methods, we compute the mean squared error as $(\hat{\beta}_1 - \beta_1)^2 + (\hat{\beta}_2 - \beta_2)^2 + (\hat{\beta}_3 - \beta_3)^2)$. The sparsity is defined as the percentage of zeros in off-diagonal elements of true covariance matrix. The simulations are conducted under dimensionality of 25. The experiment is repeated 1,000 times for each sparsity (the first and second setting) or for each number of observations (the third setting).

The simulations include sparsity 30% through 90%. The box plot and pair plot of sparsity 50% is given in figure 15 and 16. Since all the other sparsity settings demonstrate similar results, plots for those are not reported here. First of all, only banding 25, i.e. diagonal case, is comparable to OLS. All the other banding strategies are worse off than OLS. The next point is that the performance of FGLS is getting worse as more banding is conducted but it becomes better when it comes to the extreme banding, or exactly diagonals. It is easy to see that with one observation on error term given, covariance estimate is meaningless in comparison with the estimate for variance. However, it is quite puzzling why more banding leads to worse FGLS estimate.

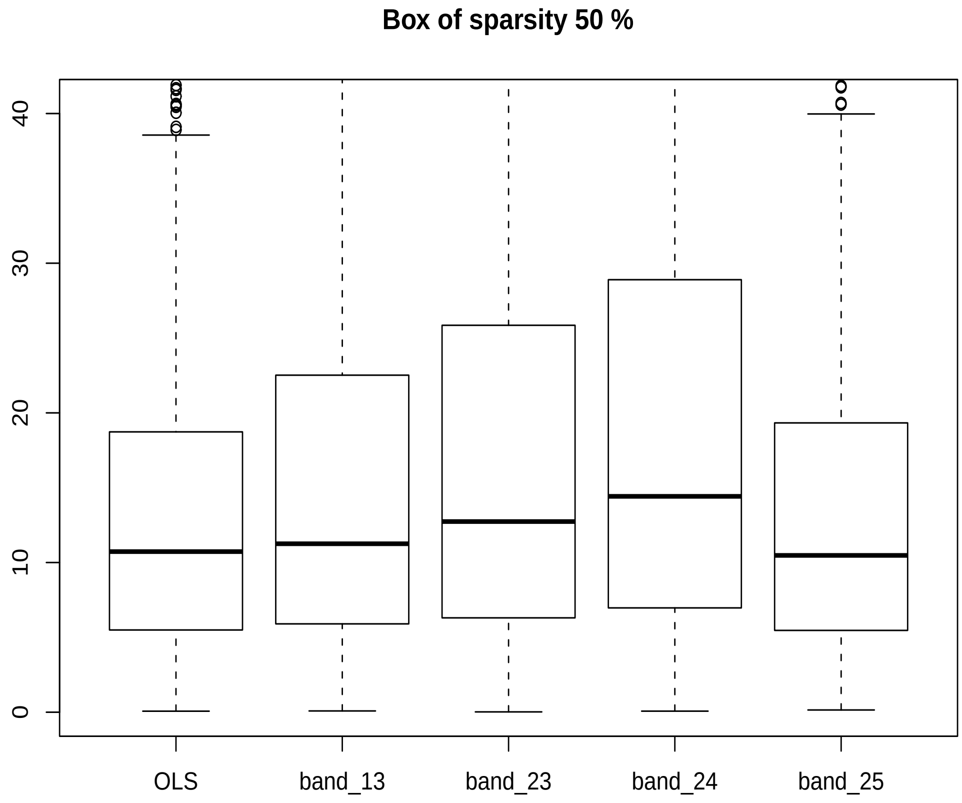


Figure 15: Boxplot of MSE, sparsity=50%

In order to compare FGLS with banding strategy to that of OLS performance, we consult table 15. The FGLS in the table indicates the feasible GLS that employs the fully banding strategy which suppresses all the off-diagonal elements to zeros. Mean, median and standard deviation ratios of FGLS MSE to OLS MSE are provided.

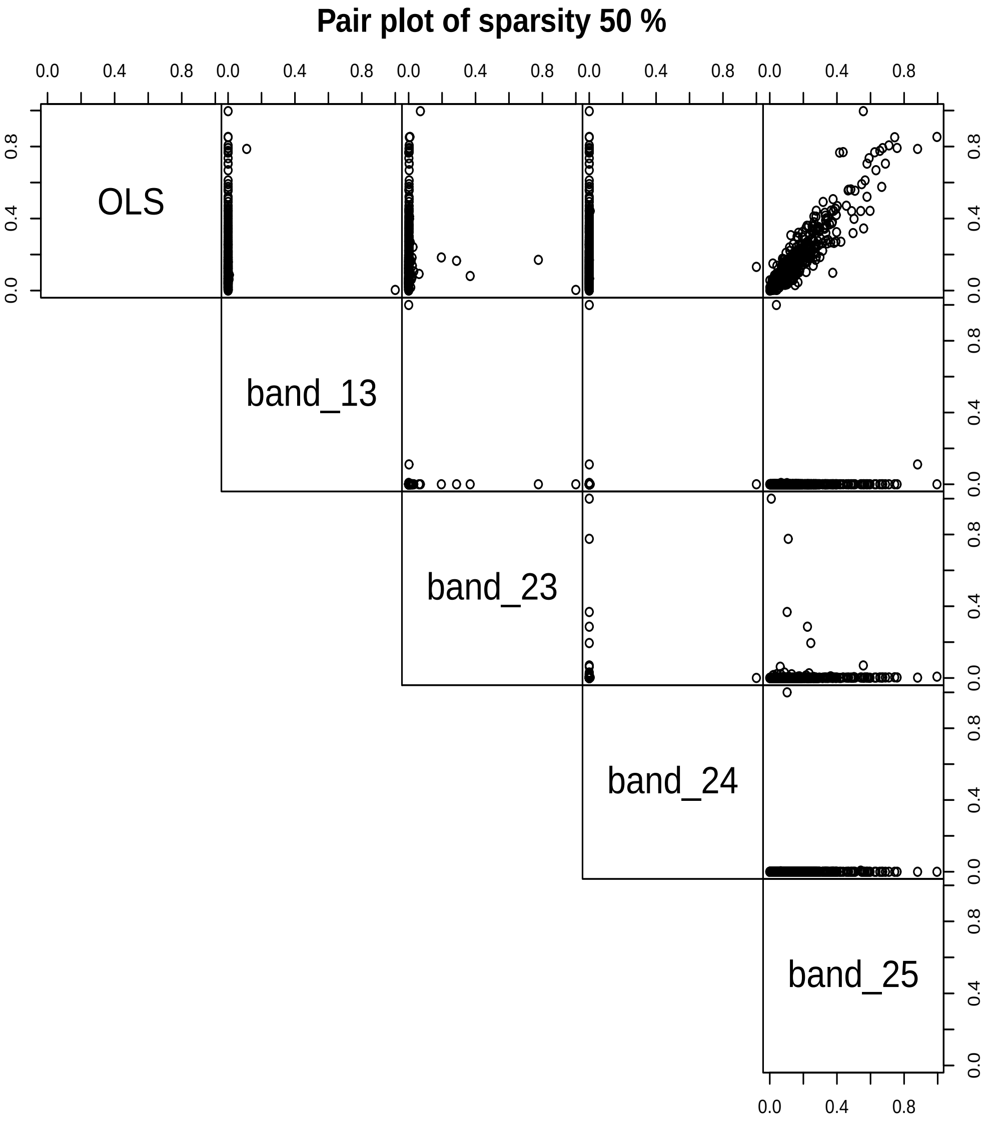


Figure 16: Pairplot of MSE, sparsity=50%

Although FGLS performances get better as sparsity increases, it appears not to be significant in any measures. Since the banding strategy is built based on one observation, it is worth taking a glance at the hypothetical situation where we have more than one observations on errors. We observe three sample sizes, $n = 20, 50, 100$ and report relative MSE ratio of fully banded FGLS to OLS in table 16.

Table 15: MSE ratio of FGLS to OLS with different sparsity

Sparsity	30%	50%	70%	90%
Mean	1.01	1.00	0.99	0.98
Median	1.02	0.98	1.00	0.99
Standard Deviation	1.02	1.01	1.00	0.97

Table 16: MSE ratio of FGLS to OLS: multiple observation case

		30%	50%	70%	90%
n=20	Mean	0.96	0.95	0.92	0.86
	Median	0.95	0.89	0.87	0.79
	Standard Deviation	0.96	0.99	0.98	0.88
n=50	Mean	0.96	0.95	0.92	0.86
	Median	0.95	0.88	0.86	0.83
	Standard Deviation	0.94	0.91	0.90	0.83
n=100	Mean	0.92	0.87	0.79	0.75
	Median	0.94	0.86	0.82	0.75
	Standard Deviation	0.89	0.87	0.81	0.86

Banded FGLS here is obtained by applying banding strategy to sample covariance matrix. As expected, as the number of observation increases from 20 to 100, the relative performance of fully banded FGLS to OLS is getting better. Furthermore, the relative performance of FGLS behaves more nicely as the sparsity increases. In case of sample size 100, the mean squared error is reduced to 75% of OLS, which verifies that sparsity of covariance matrix matters in banding strategy.

The last simulation setting is two-pass procedure with panel data. Note that panel data has both time-series and cross-sectional data set where regression coefficients are not constant over cross-sectional direction. This is a good example of multivariate problem in finance application. If there exist common factors as driving force to each individual stock return, we can model this problem as panel structure. In this simulation study, two parameters are of our interest: sparsity and sample size. The sample size is referred to the number of sample in time-series data. Again, three factor model is assumed. The simulation model is as follows.

$$y = X\beta + \epsilon \tag{28}$$

where y, ϵ is $T \times n$ matrices, X is $T \times k$ regressor matrix, and β is $k \times n$ coefficient matrix. Each row of ϵ is assumed to be from multivariate normal distribution. In simulation, we selected $MN(0, \Sigma)$ for each row of ϵ. Coefficients β, regressor X are randomly and independently drawn from $U(0, 2)$ and $N(0.0022, 0.00006944^2)$, respectively. Three factor model and intercept implies $k = 4$, n is set to 25. We define sparsity as the percentage of zeros in off-diagonals. By setting different time series sample size T and sparsity of Σ, the behavior of MSE in two pass procedure is evaluated. FGLS in the second-pass are conducted using sample covariance matrix and several banding strategies. The full banding strategy works the best and a part of simulation comparison with sample covariance matrix can be found in figure 17 through figure 20. As are enough to show the idea, sparsity 30%, 90% and sample size 160, 700 are provided in figures. Full comparison with all the other alternatives are demonstrated in the box plot of 18.

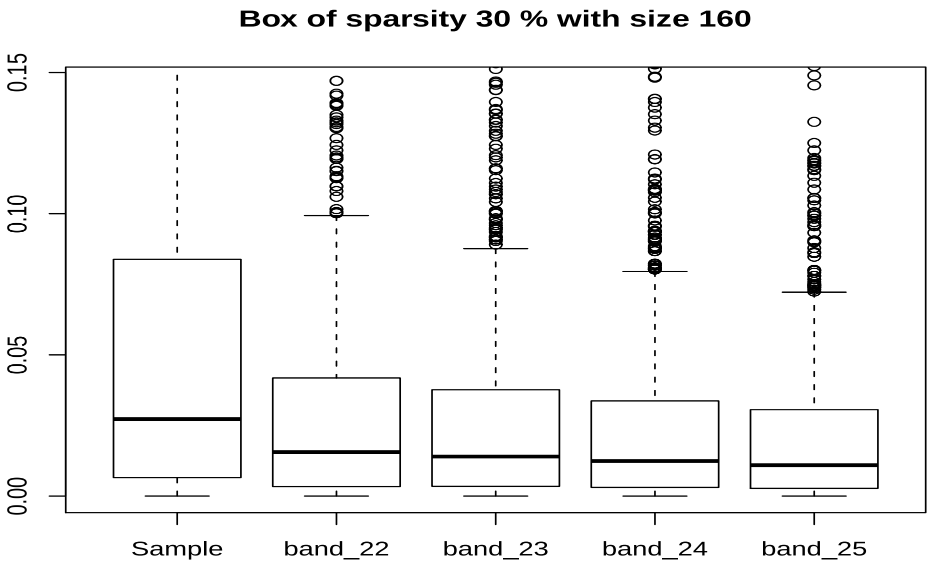


Figure 17: Boxplot of MSE in two-pass GLS, sparsity=30%, n=160

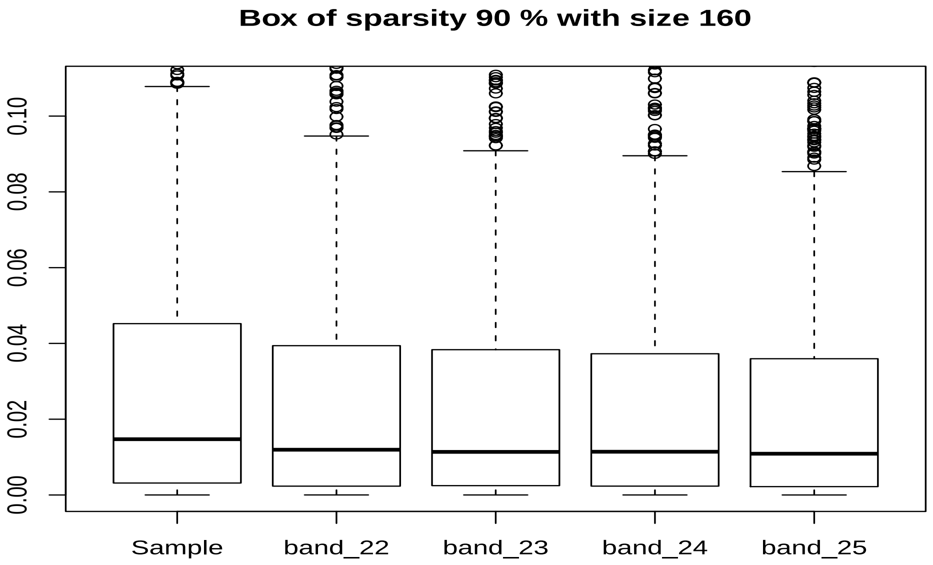


Figure 18: Boxplot of MSE in two-pass GLS, sparsity=90%, n=160

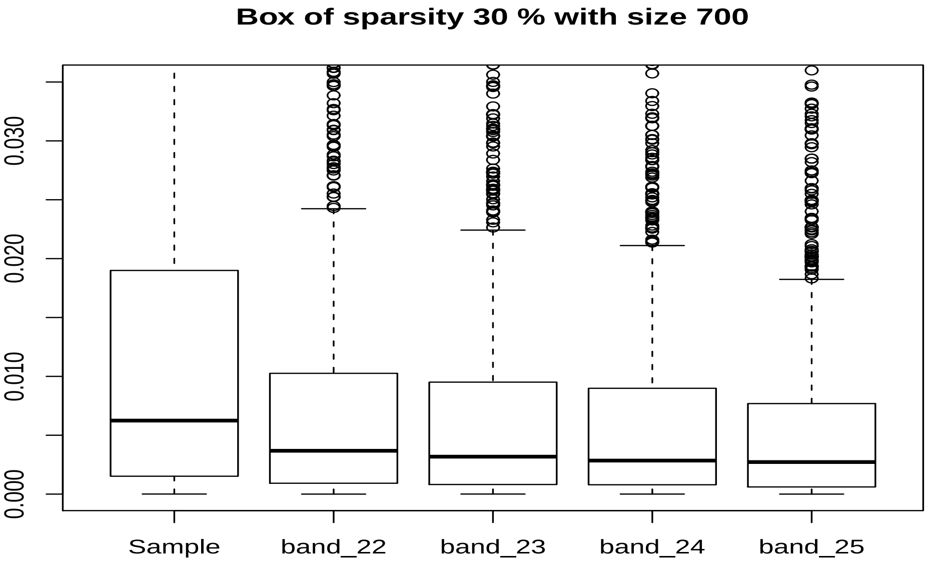


Figure 19: Boxplot of MSE in two-pass GLS, sparsity=30%, n=700

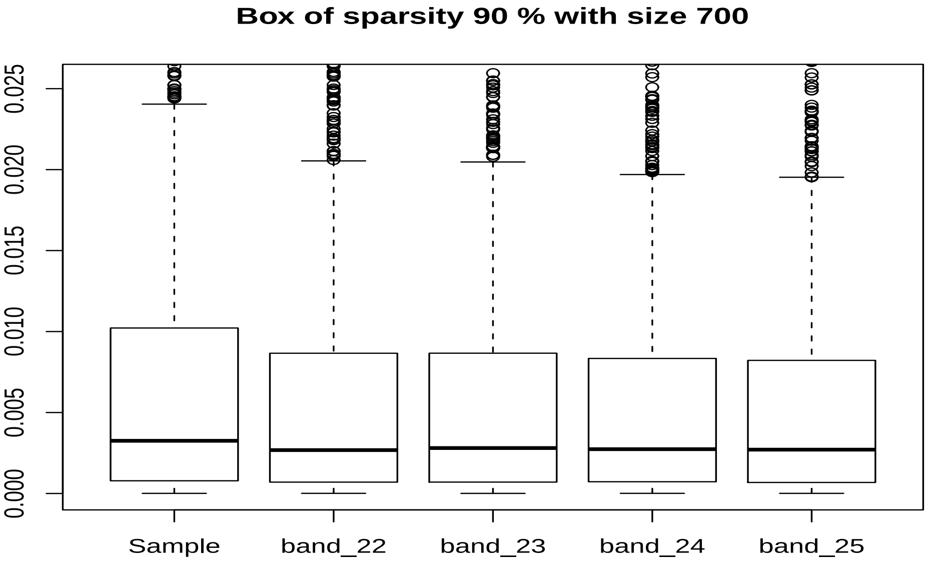


Figure 20: Boxplot of MSE in two-pass GLS, sparsity=90%, n=700

Table 17: MSE ratio of FGLS in two-pass procedure: MSE with full banding strategy to sample covariance matrix

		30%	50%	70%	90%
n=160	Mean	0.35	0.44	0.56	0.75
	Median	0.41	0.48	0.62	0.74
	Standard Deviation	0.32	0.42	0.53	0.76
n=330	Mean	0.38	0.45	0.58	0.79
	Median	0.41	0.49	0.61	0.76
	Standard Deviation	0.37	0.43	0.54	0.82
n=700	Mean	0.41	0.48	0.58	0.81
	Median	0.44	0.50	0.60	0.83
	Standard Deviation	0.41	0.44	0.54	0.79

First of all, all the banding strategies outperform sample covariance matrix. We also can see that more banding results in better estimation, agreeing with previous two simulation setting. The obvious finding is that the large sample size makes relative outperformance of full banding to be small. This is consistent with general statistical convergence idea of sample covariance matrix. Sparsity truly matters in estimation of panel data. Both from table 17 and figures 17 through 20, as we have more sparse covariance matrix, the MSE ratio of banding strategy to sample covariance matrix increases. The possible explanation is the following. If covariance matrix has many non-zero off-diagonals, the estimates of those parameters by sample covariance matrix are misleading the whole GLS estimate, which offsets the effects of suppressing zeros for off-diagonal estimates. In panel data using two-pass procedure, FGLS with fully banding strategy is suggested when time-series sample size is small and the covariance matrix is believed to be less sparse.

In three simulations, we learned that sparsity of true covariance matrix really matters. Combining the first and the second simulation setting, we have seen that GLS estimation improves when sparsity increases comparing with OLS performance.

The potential reason might be that banding strategy is getting closer to the true covariance matrix by construction because banding is to suppress off-diagonals to zeros which shares the main feature of sparse covariance matrices. In panel data, sample covariance matrix and banding strategies are compared. As noted previously, the sample size and sparsity affect the relative performance of banding strategy over sample covariance matrix. Meanwhile, in all cases, the FGLS with full banding strategy shows the best efficiency.

3.4 When does the relative efficiency of GLS to OLS improves?

In previous sections, we have seen the simulation results showing that FGLS using fully banding strategy provides improved estimation over OLS or FGLS with sample covariance matrix. In this section, the focus is shed on the conditions under which outperformance of GLS becomes even more. Unlike the preceding sections, population version of GLS and OLS are considered. Analytical derivation along with guided simulation verifies our claims. We first analyze diagonal elements and then move onto factor decomposition of true covariance matrix Σ and their effects on GLS efficiency.

3.4.1 First glance at GLS efficiency

In order to find and verify the conditions that improve GLS efficiency, we start with simple cases and develop the argument into more general cases. Since OLS is a special case of GLS, if Σ is the identity matrix I, then estimates of OLS and GLS are the same. For the simplicity of calculation, we take a constant regressor into account for a while.

$$y = \mu + \epsilon \tag{29}$$

where y, μ and ϵ are all $n \times 1$ vectors. ϵ is a random vector distributed with mean zero and covariance matrix Σ. In addition to identity matrix I, we have special covariance matrices that produce the same GLS and OLS estimates.

Proposition 3.4.1. *Let Σ is a form of $\begin{pmatrix} 1 & a & \cdots & a \\ a & 1 & \cdots & a \\ \vdots & & & \vdots \\ a & \cdots & & 1 \end{pmatrix}$ with a constant a. Then GLS estimate is the same as OLS.*

Proof. Σ of stated form has a explicit expression of its inverse. See [44]. Since Σ^{-1} is $c \begin{pmatrix} 1 & d & \cdots & d \\ d & 1 & \cdots & d \\ \vdots & & & \vdots \\ d & \cdots & & 1 \end{pmatrix}$ for some scalar c and d, it is easily derived that

$$\hat{\mu}^{GLS} = (\mathbf{1}^T \Sigma^{-1} \mathbf{1})^{-1} \mathbf{1}^T \Sigma^{-1} y = \bar{y} = \hat{\mu}^{OLS}$$

□

Therefore, the GLS estimate with this special covariance matrix form is not dependent on the covariances a. It is interesting that correlations do not have any influence on GLS estimates when the variances are the same.

The next case is when correlations are all zeros and the variances are not the same: heteroscedasticity with zero-correlation case. The following proposition shows that GLS is more efficient than OLS and that the more variances are spread out, the better GLS's relative efficient it gets.

Proposition 3.4.2. *Let $\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & & \sigma_n \end{pmatrix}$ then GLS estimator is more efficient than OLS. Moreover, as $\frac{\sigma_i}{\sigma_{i+1}}$ increases, the relative efficiency of GLS to OLS* *becomes better. In other words, the more the variances are spread out, the better the relative efficiency of GLS to OLS becomes.*

Proof. The GLS estimate is computed as,

$$\begin{aligned}
\hat{\mu}^{GLS} &= (\mathbf{1}^T\Sigma^{-1}\mathbf{1})^{-1}\mathbf{1}^T\Sigma^{-1}y \\
&= \frac{1}{\sigma_1^{-1}+\cdots+\sigma_n^{-1}}[\sigma_1^{-1}\cdots\sigma_n^{-1}]\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\
&= \sum_{i=1}^{n}\frac{\sigma_i^{-1}y_i}{\sigma_1^{-1}+\cdots+\sigma_n^{-1}} \\
&= \sum_{i=1}^{n} w_i y_i
\end{aligned}$$

where, $w_i = \frac{\sigma_i^{-1}}{\sigma_1^{-1}+\cdots+\sigma_n^{-1}}$. Since we have $var(\hat{\mu}^{OLS}) = var(\bar{y}) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i$, efficiency of GLS and OLS is compared as follows.

$$\begin{aligned}
var(\hat{\mu}^{GLS}) &= var(\sum_{i=1}^{n} w_i y_i) \\
&= \sum_{i=1}^{n}(\frac{\sigma_i^{-1}}{\sigma_1^{-1}+\cdots+\sigma_n^{-1}})^2\sigma_i \\
&= \frac{1}{\sigma_1^{-1}+\cdots+\sigma_n^{-1}} \\
&\le \frac{\sigma_1+\cdots+\sigma_n}{n^2} = var(\hat{\mu}^{OLS})
\end{aligned}$$

To verify the last line, we assume without loss of generality that $\sigma_1 \ge \cdots \ge \sigma_n$. If n^2 is less than $\frac{1}{\sigma_1^{-1}+\cdots+\sigma_n^{-1}}(\sigma_1+\cdots+\sigma_n)$ then we are done.

$$
\begin{aligned}
& n^2 - \frac{1}{\sigma_1^{-1} + \cdots + \sigma_n^{-1}} (\sigma_1 + \cdots + \sigma_n) \\
= \; & n^2 - (1 + \frac{\sigma_1}{\sigma_2} + \cdots + \frac{\sigma_1}{\sigma_n} && (30) \\
& \frac{\sigma_2}{\sigma_1} + 1 + \cdots + \frac{\sigma_2}{\sigma_n} && (31) \\
& \cdots \\
& \frac{\sigma_n}{\sigma_1} + \frac{\sigma_n}{\sigma_2} + \cdots + 1) && (32) \\
& && (33)
\end{aligned}
$$

Since for $i \neq j$,

$$\frac{\sigma_i}{\sigma_j} + \frac{\sigma_j}{\sigma_i} = \frac{\sigma_i^2 + \sigma_j^2}{\sigma_i \sigma_j} \geq 2,$$

we verify that (33) is negative, thus GLS is more efficient than OLS.

Moreover, we can see that if σ_i's are more spread out, the more efficient GLS is. To see this, let $\sigma_2 < \sigma_2'$

$$\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1 \sigma_2} - \frac{\sigma_1^2 + \sigma_2'^2}{\sigma_1 \sigma_2'} = \frac{(\sigma_2' - \sigma_2)(\sigma_1^2 - \sigma_2' \sigma_2)}{\sigma_1 \sigma_2 \sigma_2'} < 0.$$

Considering the equation (33), it is straightforward that the more variances are spread out, the better relative efficiency of GLS over OLS we have. □

Two special cases of covariance matrix form in constant regressor model have been seen. Next argument is still based on constant regressor model as (29), but the covariance structure has no restriction, i.e. any matrices that are symmetric and positive semi-definite are allowed.

Let $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \vdots & \ddots & & \\ & & & \\ \sigma_{n1} & \cdots & & \sigma_{n1} \end{pmatrix}$. First of all, we have

$$var(\hat{\mu}^{OLS}) = var(\bar{y}) = \frac{1}{n^2} \sum_i^n \sum_j^n \sigma_{ij}. \quad (34)$$

Now the GLS estimator is

$$\begin{aligned}\hat{\mu}^{GLS} &= (\mathbf{1}^T\Sigma^{-1}\mathbf{1})^{-1}\mathbf{1}^T\Sigma^{-1}y \\ &= \frac{1}{\sum_i^n S_i}[S_1 \cdots S_n]\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},\end{aligned}$$

where S_i is the i-th column sum of matrix Σ^{-1}. This leads us to variance of GLS as

$$var(\hat{\mu}^{GLS}) = \frac{1}{(\sum_i^n S_i)^2}\sum_{i,j=1}^{n} S_i S_j \sigma_{ij}. \tag{35}$$

Comparing equation (34) with equation (35) we can see that the first is the arithmetic average while the latter is the weighted average of σ_{ij}. The variances of OLS and GLS will be similar if the weights are similar, or $S_i \approx S_j$. We have already seen that covariance matrix with the same off-diagonals with diagonal of 1 gives the same GLS and OLS estimator. This is a special case of $S_i = S_j$, and it makes a good example of the argument.

Looking inside the summand of equation (35), we have

$$var(\hat{\mu}^{GLS}) \propto S_1^2\sigma_{11} + \cdots + S_n^2\sigma_{nn} + S_1 S_2 \sigma_{12} + \cdots . \tag{36}$$

We can find two observations here.

- Σ with non-positive σ_{ij} leads to more efficient GLS estimator than Σ with all-positive.
- The more σ_{ii}'s are spread out, the more efficient GLS estimator is.

The first one is obvious while the second one takes more consideration. The logic is as follows.

Let $\Sigma^{-1} = \{\xi_{ij}\}_{i,j}^{n}$, then $\xi_{11} \leq \cdots \leq \xi_{nn}$, since $\sigma_{ii} \geq \cdots \geq \sigma_{nn}$. Others being equal, S_i is increasing function of ξ_{ii} and if the order of ξ_{ii}'s has inverse relation with σ_{ii} then equation (35) shows that smaller weight is for greater σ_{ii} and greater weight is for smaller σ_{ii}. Therefore, with fixed $\sum_{i=1}^{n} \sigma_{ii}$, the more σ_{ii}'s are spread out the less the GLS variance is. We will look into detail about this later on.

Finally general case will be studied. The regressors of linear model are no longer assumed to be constant and covariance matrix is set to be any legitimate forms. The only restriction we made on regression model (22) is $k = 1$, one-regressor model.

$$y = X\beta + \epsilon,$$

with ϵ has covariance matrix of Σ, and $X = [x_1, x_2, \cdots, x_n]'$.
Similarly as before, the GLS estimator is

$$\begin{aligned}
\hat{\beta}^{GLS} &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y \\
&= \frac{1}{\sum_{i,j=1}^{n} x_i x_j \xi_{ij}} \sum_{i,j=1}^{n} x_i \xi_{ji} y_j \\
&= \frac{1}{\sum_{i,j=1}^{n} x_i x_j \xi_{ij}} \sum_{j}^{n} W_j y_j,
\end{aligned}$$

where, W_j is weighted j-th column sum of Σ^{-1} with x_i. The variance of GLS estimate is computed as

$$\begin{aligned}
var(\hat{\beta}^{GLS}) &= var(\frac{1}{\sum_{i,j=1}^{n} x_i x_j \xi_{ij}} \sum_{j}^{n} W_j y_j) &(37)\\
&= \frac{1}{(\sum_{i=1}^{n} W_i)^2} \sum_{i,j=1}^{n} W_i W_j \sigma_{ij}. &(38)
\end{aligned}$$

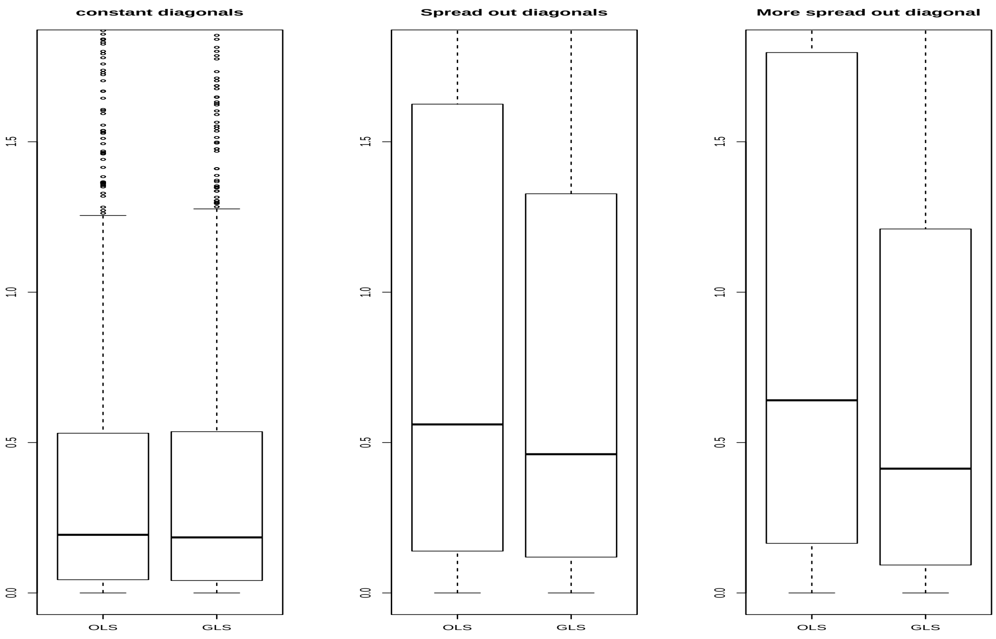


Figure 21: Boxplot of MSE for OLS and GLS estimates with different diagonal structure: constant regressor

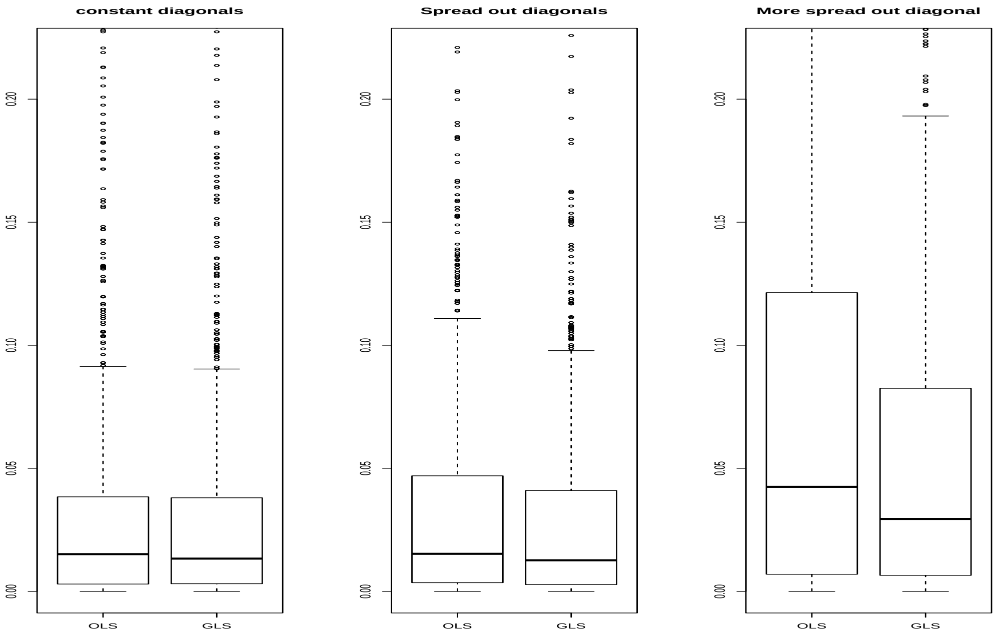


Figure 22: Boxplot of MSE for OLS and GLS estimates with different diagonal structure: non-constant three regressor

With fixed X, we can reach the same conjecture as in constant regressor case: the more σ_{ij}'s are spread out, the more efficient GLS estimate is. The simulation results support our conjecture both in constant and nonconstant regressor models. Figure 21 and 22 are the box plots of the MSE of GLS, MSE with different diagonal structure in covariance matrices. The simulations is set as before except that true covariance matrix is used for transformation matrix in GLS because we like to see the behavior of GLS in population version, not estimation counterpart. Another note is that figure 22 is the result of three regressor model which, in fact, generalizes our claim to multivariate regression case. In these simulations, the diagonals of true covariance matrices are assumed in three ways: constant, spread-out, and more spread-out. Spread-out setting is to force the diagonals of σ_{ii} to be $0.5i$ and more spread-out case to be $2i$. The figures show that more spread-out diagonals in Σ provides better relative performance of GLS estimates in terms of MSE. Even the three-regressor case agrees with our conjecture.

Although the analytical calculations and simulation experiments regarding variances of GLS estimators in various settings shed an intuition about covariance matrix Σ, it is not conclusive. Direct analytical proof seems difficult here but we can explore more into the structure of Σ and Σ^{-1} to verify our conjecture more clearly. The next subsection provides an argument that supports our claim.

3.4.2 On diagonals of Σ and Σ^{-1}

Let us change the point of view upon covariance matrix into numerical side. Since Σ is symmetric and positive-semi definite, it can be decomposed as UDU', where U is a matrix of eigenvectors with $UU' = U'U = I$ and diagonal matrix D of eigenvalues such that $\lambda_1 < \cdots < \lambda_n$. From different aspect, if V is a random vector uncorrelated to each other with covariance matrix of D, then U can be seen as a linear transformation matrix for V into Z: $Z = UV$, $cov(Z) = \mathbb{E}(UVV'U') = UDU' = \Sigma$. Let's look into the expression of diagonal elements of Σ, the variances of Z:

$$\begin{aligned} \sigma_{11} &= \lambda_1 u_{11}^2 + \cdots + \lambda_n u_{1n}^2 \\ \sigma_{22} &= \lambda_1 u_{21}^2 + \cdots + \lambda_n u_{2n}^2 \\ \cdots . \end{aligned} \tag{39}$$

Since U is an orthogonal, $u_{j1}^2 + \cdots + u_{jn}^2 = 1$, σ_{ii} can be seen as a weighted average with weights u_{j1}^2. If the transformation matrix U is such a matrix that makes variances of Z spread out each other, say, $\sigma_1 < \cdots < \sigma_n$, then we can guess that big weights on smaller λ_i's for smaller σ_{jj} and big weights on bigger λ_k for σ_{ll}'s and so on. Therefore, fixing the eigenvalues, the degree of spread-out of variances of Z is decided by transformation matrix U. Therefore, if U is a matrix that preserves the order of $\lambda_i's$ into the order of $\sigma_i's$, then the order of ξ_{ii} of Σ^{-1} will be reversed because $\Sigma^{-1} = UD^{-1}U'$.

The reasoning so far is somewhat abstract and the numerical examples help understand the behavior of U and $\Sigma = UDU'$. Both of the matrices in the example are 4×4 and have increasing diagonals. For the purpose of the comparison, the first one has $\sigma_{ii} = 0.5i$ while the second has $\sigma_{ii} = 0.7i$ with $i = 1, 2, 3, 4$. All the other elements of Σ are randomly generated and spectral decomposition is conducted in such a way that eigenvalues are organized in ascending order. Therefore the more weights on the small eigenvalues is given, the smaller the resulting σ_{ii}. Let's look at the following example.

$$U_{0.5} = \begin{pmatrix} 0.858 & 0.434 & 0.134 & -0.236 \\ -0.375 & 0.695 & 0.565 & 0.236 \\ 0.0702 & -0.529 & 0.804 & -0.258 \\ 0.341 & -0.219 & 0.117 & 0.906 \end{pmatrix},$$

$$U_{0.7} = \begin{pmatrix} 0.969 & -0.117 & 0.112 & -0.184 \\ -0.120 & -0.826 & 0.509 & 0.205 \\ -0.084 & 0.474 & 0.844 & -0.232 \\ 0.197 & 0.277 & 0.120 & 0.932 \end{pmatrix}.$$

Square of elements (1,1) of the two matrices are the weights for the smallest eigenvalue to generate σ_{11}. Square of (2,2) are for the second smallest eigenvalues and so on. Looking up for equation (39), the rows of $U_{0.5}$ and $U_{0.7}$ are the weights in weighted average of eigenvalues λ's. The first observation is that the diagonals of the matrices are the biggest in magnitude so that the corresponding eigenvalues are given the most weights in calculating the corresponding σ_{ii}. The next observation is that as the σ_{ii} are more spread out, the more the diagonals of U are spread-out. In other words, as the variances are spread out, the weights on the corresponding λ increases, thus it is likely to reverse the order of ξ_{ii} because $\Sigma^{-1} = UD^{-1}U'$ and diagonals of D^{-1} is in reversed order of diagonals of D. In fact, here is the corresponding Σ and Σ^{-1} of $U_{0.7}$.

$$\Sigma = \begin{pmatrix} 0.70000 & -0.04316 & -0.2596 & 0.4137 \\ -0.04316 & 1.40000 & -0.5317 & -0.5094 \\ -0.25967 & -0.53173 & 2.1000 & 0.3075 \\ 0.41375 & -0.50944 & 0.3075 & 2.8000 \end{pmatrix}$$

$$\Sigma^{-1} = \begin{pmatrix} 1.68751 & 0.05374 & 0.26157 & -0.26831 \\ 0.05374 & 0.83660 & 0.20057 & 0.12224 \\ 0.26157 & 0.20057 & 0.56878 & -0.06463 \\ -0.26831 & 0.12224 & -0.06463 & 0.42613 \end{pmatrix}$$

The example shows the reasoning about the diagonals of Σ via spectral decomposition. It is not analytically proved but worth looking into it for it helps us understand the numerical standpoint of covariance matrices and their inverses. The bottom line of this analysis is as follows. If there exists an orthogonal matrix U that transforms random vector V with diagonal covariance matrix D to a new random vector $Z = UV$. Then the covariance matrix of Z becomes UDU', and if matrix U has a property that preserves the order of magnitude of diagonals of D as in equation (39), then the order diagonals of inverse matrix $UD^{-1}U'$ is reversed. Our claim or conjecture is that more spread-out diagonals of UDU' likely come from the prescribed characteristic of U, which is shown by numerical examples.

Now we move onto the next argument about column summation of Σ^{-1}. Σ^{-1}, the inverse of covariance matrix, is also symmetric and positive semi-definite matrix. This can be easily seen by spectral decomposition. Therefore, Σ^{-1} also can serve as a covariance matrix. For this reason, we are given $\Sigma^{-1} = \{\xi_{ij}\}_{i,j=1}^{n}$ with $\xi_{ij} = \pi_{ij}\sqrt{\xi_{ii}\xi_{jj}}$ where π_{ij} is a correlation coefficient. Simply substituting this expression to the Σ^{-1}, the column summation S_i's are written as below.

$$\begin{aligned} S_1 &= \sqrt{\xi_{11}}(\sqrt{\xi_{11}} + \pi_{12}\sqrt{\xi_{22}} + \cdots + \pi_{1n}\sqrt{\xi_{nn}}) \\ S_2 &= \sqrt{\xi_{22}}(\pi_{21}\sqrt{\xi_{11}} + \sqrt{\xi_{22}} + \cdots + \pi_{2n}\sqrt{\xi_{nn}}) \\ &\cdots \end{aligned} \tag{40}$$

The diagonal elements of ξ_{ii} are the multipliers of corresponding column sum. The remaining part in bracket is the weighted summation of all diagonal elements with correlation coefficients π_{ij}. From this point of view, we can see that the magnitude of S_i is decided mainly by ξ_{ii}, providing the parts in the bracket in equations (40) are not different to each other. Putting in more mathematical form, let's assume that correlation coefficients π_{ij}'s are independently distributed with mean M and variance V. Then the followings are obtained.

$$\begin{aligned}
\mathbb{E}(S_1) &= \xi_{11} + M(\sqrt{\xi_{22}} + \cdots + \sqrt{\xi_{nn}}) \\
\mathbb{E}(S_1^2) &= \xi_{11} + M(\sqrt{\xi_{22}} + \cdots + \sqrt{\xi_{nn}}) + V\xi_{11}(\xi_{22} + \cdots + \xi_{nn}) \\
\mathbb{E}(S_2) &= \xi_{22} + M(\sqrt{\xi_{11}} + \cdots + \sqrt{\xi_{nn}}) \\
\mathbb{E}(S_2^2) &= \xi_{22} + M(\sqrt{\xi_{11}} + \cdots + \sqrt{\xi_{nn}}) + V\xi_{22}(\xi_{11} + \cdots + \xi_{nn})
\end{aligned} \tag{41}$$

Therefore, if $\xi_{11} \geq \xi_{22} \geq \cdots \geq \xi_{nn}$, then the column summation of Σ^{-1} has a relation as below.

$$\mathbb{E}(S_i) \leq \mathbb{E}(S_j) \tag{42}$$

Taking the previous analysis on the relationship between orders of σ_{ii}'s and ξ_{ii}'s into account, we have come to an useful conclusion. Since σ_{ii}'s and ξ_{ii}'s have reversed order in magnitude, $\sigma_{ii} \geq \sigma_{jj}$ leads to $\mathbb{E}(S_i) \leq \mathbb{E}(S_j)$. Now recall that variances of OLS and GLS estimates.

$$\begin{aligned}
var(\hat{\mu}^{OLS}) &= var(\bar{y}) = \frac{1}{n^2}\sum_i^n \sum_j^n \sigma_{ij} \\
var(\hat{\mu}^{GLS}) &= \frac{1}{(\sum_i^n S_i)^2} \sum_{i,j=1}^n S_i S_j \sigma_{ij}
\end{aligned}$$

The variance of GLS is the weighted average of σ_{ij} with weights , $\frac{S_i Sj}{(\sum_i^n S_i)^2}$. In this subsection we first argue that σ_{ii}, the diagonals of Σ are likely to have the reversed order of ξ_{ii}'s, the diagonals of Σ^{-1}. In addition, we show that $\mathbb{E}(S_i) \leq \mathbb{E}(S_j)$ if $\xi_{ii} \leq \xi_{jj}$. By the expression for variance of GLS estimate above, we can see that big σ_{ii} is likely to result in small ξ_{ii} and then leads to small S_{ii}, and finally small weights of S_{ii} is put on big σ_{ii} in variance of GLS. For the smaller σ_{jj}, the opposite logic can be applied. Therefore, if we ignore the cross-product in $var(\hat{\mu}^{GLS}) = \frac{1}{(\sum_i^n S_i)^2} \sum_{i,j=1}^n S_i S_j \sigma_{ij}$, as in sparse covariance matrix case, the more spread-out the diagonals of Σ are, the smaller the variance of GLS estimates are obtained. Since OLS estimates are invariant with respect to covariance matrices, the relative efficiency of GLS becomes better as the degree of spread-out of diagonals of covariance matrix Σ increases. Finally, we have come to the conjecture as follows.

Conjecture 1. *Let the linear model* (22) *with covariance matrix* $\Sigma_0 = \{\sigma_{ij}\}_{i,j=1}^n$ *be given. We can write* $\Sigma_0^{-1} = \{\xi_{ij}\}_{i,j=1}^n = \{\pi_{ij}\sqrt{\xi_{ii}\xi_{jj}}\}_{i,j=1}^n$ *with* $\pi_{ij} = 1$ *for all* $i = j$*. Let's assume that* π_{ij}*'s are independently and identically distributed. Suppose another covariance matrix* Σ_1 *is given to the same model, and diagonals of* Σ_1 *are more spread-out than* Σ_0*. Then we claim that*

$$var(\hat{\beta}^{GLS}|\Sigma_1) \leq var(\hat{\beta}^{GLS}|\Sigma_0). \tag{43}$$

As notified earlier, the conjecture is not analytically proved in this thesis, but we can partly verify the claim via guided simulation. The following are the steps of the simulation we are about to show. The first step is about generating covariance matrix.

1. Generate uniform random numbers from $U(-0.8, 0.8)$ and use them as the entries of lower triangular matrix L.

2. Make positive-semi definite matrix by $A = LL'$.

3. Let the diagonals of A be $\{d_i\}_{i=1}^n$ and diagonal matrix $B = diag(\frac{1}{\sqrt{d_1}}, \cdots, \frac{1}{\sqrt{d_n}})$.

4. Then we get correlation matrix $\Omega = BAB$, or covariance matrix with constant diagonals.

5. In order to get spread-out diagonal covariance matrix with correlations and sum of diagonals fixed, generate a sequence s_i and get new diagonals $d_i = \frac{s_i}{\sum_{i=1}^n s_i}$ by $\Sigma = F\Omega F$ where $F = diag(\sqrt{d_1}, \cdots, \sqrt{d_n})$

- We use $s_i = i$ and $s_i = i^2$, $i = 1, 2, \cdots, n$.

The next step is about generating the data in linear model using pre-generated covariance matrix Σ from the previous steps.

1. Generate data by the model, $y = X\beta + \epsilon$ where ϵ is randomly drawn from Σ, β from $\frac{1}{\sqrt{k}} \times U(0, 1)$, X from $N(0, 1)$ independently.

2. We experiment with three cases, $k = 1, 2, 3$.

3. By doing this, we maintain signal-noise ratio across the number of factors: 1-1.

Once covariance matrix and data X and y are generated we computed GLS and OLS estimates and compare with true β's. The estimating errors are computed by $\gamma_O = \hat{\beta}_{OLS} - \beta$, $\gamma_G = \hat{\beta}_{GLS} - \beta$. We will look at MSE's, i.e. $\gamma_O'\gamma_O$ and $\gamma_G'\gamma_G$. The simulation is run with 1,000 repetitions for each diagonals structure s_i, and the statistics of $\gamma_O'\gamma_O/\gamma_G'\gamma_G$ are reported in tables 18 and 19.

For both of averages and standard deviations, the ratios are increasing significantly as the degree of spread-out of diagonals of Σ increases. This simulation results

Table 18: Ratio of squared errors of OLS to GLS: AVERAGE

Diagonal Sequence	$s_i = 1$	$s_i = i$	$s_i = i^2$
One Factor	20.56	55.03	273.70
Two Factors	14.22	28.58	117.11
Three Factors	9.91	19.50	64.48

Table 19: Ratio of squared errors of OLS to GLS: Standard Deviation

Diagonal Sequence	$s_i = 1$	$s_i = i$	$s_i = i^2$
One Factor	15.93	30.88	144.82
Two Factors	9.93	20.76	62.66
Three Factors	7.04	14.41	38.55

support our claim in conjecture 1. Now we conducted slightly different simulation to see if our claim is valid when real data set is used. The same simulation procedure is taken as described above.

The simulation results agree with the intuitions from our analytical derivations. It is difficult to set up an empirical studies concerning various diagonals of covariance matrix because we do not have ex ante knowledge about covariance matrix. Therefore, a hypothetical GLS case is set up. This computer experiment is basically the same simulation as before except that the covariance matrix is constructed from real data. We used CRSP ex-dividend daily stock return data in 2008. After cleaning the data set we have 6,299 stocks available for the experiment. We randomly pick 25 stocks from these and calculate sample covariance matrix and then use it as the true covariance matrix for the simulation. The procedure is repeated for 1,000 times with three stock picking criteria.

- Pick randomly 25 stocks and construct covariance matrix.

- Pick randomly 25 stocks each from distinct industry group using SIC 2-digit code.

- Pick Randomly 25 stocks each from distinct size group using ”share outstanding × share price”.

Table 20: MSE ratio of OLS to GLS: AVERAGE, real stock return data

Criteria	Random	By industry	By size
One Factor	5.12	3.21	5.66
Two Factors	4.99	3.39	5.36
Three Factors	4.61	3.24	4.59

Table 21: MSE ratio of OLS to GLS: Standard deviation, real stock return data

Criteria	Random	By industry	By size
One Factor	26.59	10.20	33.11
Two Factors	27.24	12.50	40.00
Three Factors	26.40	11.09	22.83

The ratios of OLS to GLS regarding MSE of each estimate are reported in tables 20 and 21. The results are interesting in that it is consistent with our anticipation. First of all, GLS estimates are more efficient than OLS in all cases. Among them the GLS estimates using ”by size” criteria produce the most efficient GLS relative to OLS estimates, which is consistent with our belief that the variances of stock returns have something to do with size. ”By industry” is the least efficient GLS relative to OLS, even less efficient than ”by random” case, which makes sense because stocks from different industry may have very small correlations and this will make GLS similar to OLS. In real data, though it is hypothetical setting, the result supports our claim. The next section, we take a look at the factor structure of covariance matrices.

3.5 Factor analysis of Σ

In this section, we will explore the factor covariance matrix. Recently factor analysis is studied in the context of covariance matrix as in [20] showing that high dimensional covariance matrix estimation is better estimated by factor modeling. As is argued in the paper, covariance matrix estimation via factor modeling is very promising because we only need to care for a handful of factors instead of $p(p+1)/2$ parameters. Factor model is widely exercised in finance. Fama-French three factor model in [17], momentum factor model in [9] where linear factors are modeled by using observable factors. On the other hand, [14] and [37] model unobservable statistical factors using stock return data via factor analysis and principal component analysis. Similar techniques are introduced to fixed income market in [40] and [28].

Previous research on factor analysis shows that three principal components explain over 95% of stock and forward market movements. Here, we explore covariance matrix of ϵ in equation (22) via factor modeling. Earlier research only focuses on covariance matrix of return itself, but this paper looks into covariance matrix of errors. As noted above, the higher dimensional covariance matrix may make the most of factor structure by which we can detour too-many-parameter problem.

3.5.1 Factor analysis and linear model

Let's assume the following model,

$$y = X\beta + \epsilon,$$

where error ϵ is a $n \times 1$ random vector with covariance matrix Σ. Now, we model the error with linear common factor F and factor loading L. See [3] or [33] for more detail on factor analysis.

$$\epsilon = LF + e, \tag{44}$$

where $e = [e_1, \cdots, e_n]'$ and each component is independent to each other. ϵ, L, F are $n \times 1$, $n \times r$, $r \times 1$, respectively with $r \leq n$. We have this assumption because the number of factor is usually smaller than the dimensionality. As noted earlier, in many finance literature it is found that only three principal components account for most of the randomness. In our model, even smaller number of factors will capture most of the movements because our covariance matrix is for error from linear model, where regressor X is believed to explain significant portion of the movement of y, and error ϵ is the rest. F is the common factor and orthogonal, i.e. $FF' = I$. L is the factor loading and will be written as $[l_1, l_2, \cdots, l_n]'$. With standard assumptions in orthogonal factor analysis, we have that $\mathbb{E}(\epsilon\epsilon') = \Sigma = LL' + D$ with diagonal matrix D.

We start our argument on factor analysis and GLS estimates with a simple setting of $r = 1$ and $D = \lambda I$ with constant λ. Number of factor r being set to be one is not too much simplification as error is the remaining effect after conditioning y with X in the linear model. The covariance matrix is expressed as

$$Cov(\epsilon) = \Sigma = LL' + \lambda I. \tag{45}$$

Since $r = 1 < n$, LL' is not invertible. By so called, Sherman-Morrison formula, the inverse matrix of Σ is explicitly known. Applying the formula to equation (45),

$$\Sigma^{-1} = \frac{1}{\lambda}[I - \frac{1}{\lambda + \sum_{i=1}^{n} \frac{l_i^2}{\lambda^2}} LL'], \tag{46}$$

where l_i is the i-th component of factor loading vector L. In standard factor analysis, we call l_i^2 communality and call λ specific variance. This is easily seen from direct calculations of equation (46) by substituting $L = [l_1, l_2, \cdots, l_n]'$. Providing the sequence of l_i^2 does not converge to zero, $n \to \infty$ means $\Sigma^{-1} \to \frac{1}{\lambda}I = D^{-1}$. Of course, this is not mathematically rigorous in that we have not defined matrix norm yet. But still, we have an intuitive ground that $n \to \infty$ leads to the fact that inverse of covariance matrix Σ^{-1} may converge to inverse of specific variance matrix D^{-1}. If the argument is right, the inverse of high dimensional covariance matrix is approximated by inverse of specific variance matrix. This is consistent with our previous simulation result of fully banded matrix, which suppresses off-diagonals to zeros.

Now we move onto the case where specific matrix takes a general form, i.e. $D \neq \lambda I$. D is a diagonal matrix but no longer contains constant diagonal elements. By Sherman-Morrison formula, we have,

$$\Sigma^{-1} = D^{-1} - \frac{D^{-1}LL'D^{-1}}{1 + L'D^{-1}L}.$$

Plugging in each elements to the formula and obtain,

$$\Sigma^{-1} = \begin{pmatrix} \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2} & 0\cdots & \\ \vdots & \ddots & & \\ 0 & \cdots & & \frac{1}{\lambda_n} \end{pmatrix} - \frac{1}{w_n} \begin{pmatrix} \frac{1}{\lambda_1^2} l_1^2 & \frac{1}{\lambda_1\lambda_2} l_1 l_2 & \cdots & \frac{1}{\lambda_1\lambda_n} l_1 l_n \\ \frac{1}{\lambda_2\lambda_1} l_2 l_1 & \frac{1}{\lambda_2^2} l_2^2 & 0\cdots & \\ \vdots & \ddots & & \\ \frac{1}{\lambda_n\lambda_1} l_n l_1 & \cdots & & \frac{1}{\lambda_n^2} l_n^2 \end{pmatrix}, \tag{47}$$

where $w_n = 1 + \sum_{i=1}^{n} \frac{l_i^2}{\lambda_i}$.

Again, providing that $\sum_{i=1}^{n} \frac{l_i^2}{\lambda_i}$ does not converge to zero, it seems that $\Sigma^{-1} \rightarrow D^{-1}$, which gives the same idea as $D = \lambda I$. Since it is not analytically conclusive, we examine the behavior of $H = D^{-1} - \Sigma^{-1}$ via simulation. Before going into the detail of simulation, we first need to overview matrix norms because matrix convergence only makes sense under pre-specified norms.

3.5.2 Matrix norms and behavior of H

In order to argue matrix convergence or distance between two matrices, matrix norms have to be defined. Short overview of matrix norms followed by simulation result of norms of H are discussed here.

$\|A\|$ denotes the norm of matrix A if the following conditions are satisfied. The discussion of more intensive matrix norms and its computations can be found in [22] and [27].

- $\|A\| \geq 0$ and $\|A\| = 0$ if and only if $A = 0$.
- For any constant α, $\|\alpha A\| = |\alpha| \|A\|$.
- $\|A + B\| \leq \|A\| + \|B\|$ for any conformable matrices A and B.

Given the conditions of matrix norms above, there are many norms available and we focus on three of them useful in our applications here.

The first example is Frobenius norm which is not only intuitive but also straightforward to compute. The definition is

$$\begin{aligned}\|A\|_F &= \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}|a_{ij}|^2} \\ &= \sqrt{trace(A'A)} \\ &= \sqrt{\sum_{i=1}^{min(n,m)} \phi_i^2},\end{aligned} \tag{48}$$

where a_{ij} and ϕ_i are (i,j)-th entry and i-th singular value of matrix A.

The second example is the operator norm, which is a special case of induced norm. Induced norm is defined as $\|A\|_p = \max_{x\neq 0}\frac{\|Ax\|_p}{\|x\|_p}$, and operator norm is the induced norm where $p=\infty$. Operator norm is known to be computed conveniently as below.

$$\|A\|_{op} = \max_i\{\phi_i : \phi_i's \textit{ are the eigenvalues of } A\}$$

The last example is the maximum norm, which is simply the maximum absolute value of the entries of the matrix.

$$\|A\|_{max} = \max_i\{|a_{ij}|\}$$

Among the three norms introduced above, Frobenius norm is the most intuitive and analytically computable. The proposition below shows that Frobenius norm of matrix H is bounded.

Proposition 3.5.1. *Let $H = D^{-1} - \Sigma^{-1}$ and Σ is constructed by factor, i.e. $\Sigma = LL' + D$ as in equation (47). Suppose $\frac{1}{\lambda_i}$ and l_i^2 are bounded from both sides, then $\|H\|_F$ is bounded regardless of dimensionality n.*

Proof. By Sherman-Morrsion formula applying to Σ, we have

$$\begin{aligned}\|H\|_F^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}^2 \\ &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{l_i^2 l_j^2}{\lambda_i^2\lambda_j^2}}{(\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{l_i^2 l_j^2}{\lambda_i\lambda_j}) + (\sum_{i=1}^{n}\frac{l_i^4}{\lambda_i^2}) + 1}.\end{aligned}$$

Suppose inverse of specific variance $\frac{1}{\lambda_j}$ is bounded by M and m, l_i^2 is bounded by C and c, then we have

$$\frac{m^2 c}{MC} \leq \lim_{n\to\infty} \|H\|_F^2 \leq \frac{M^2 C}{mc}. \tag{49}$$

Therefore $\|H\|_F \geq 0$ is bounded as well. □

As the dimensionality is increasing, the norm of H is bounded. This means that the relative distance between D^{-1} and Σ^{-1} is getting close to each other as the dimension increases. Proposition 3.5.1 does not tell us about the convergence, and the convergence under the other norms are not shown, thus the simulation is conducted to support the idea. The simulation procedure is taken by the steps below. In this simulation, we take number of factor r is set from 1 to 3.

1. Generate $\Sigma = LL' + D$ by random draw factor loading L and diagonal matrix D, each component drawn from uniform distribution.

2. Compute $H = D^{-1} - \Sigma^{-1}$ and measure different norms: Operator Norm, Frobenius Norm and Max norm.

3. Repeat the procedure with dimension n increased.

4. We will take a look at the Norms of H as dimension increases.

5. The number of factor r is changed from 1 to 3.

The simulation results are provided in figure 23 through 25. Under all three norms, H looks not only bounded but also converging.

In addition to the convergence of norms of H, we find that as the number of factor increases, the distance between Σ^{-1} and D^{-1} increases, three lines are parallel to one another. Therefore, D^{-1} is a good approximate especially when the number of factor is small and the dimensionality is large. This is a theoretical support that the inverse of high dimensional covariance matrix can be replaced by diagonal matrix, such as GLS transformation matrix. We have already seen that fully banding strategy outperforms sample covariance matrix in various GLS circumstances, and the factor analysis in this chapter serves an explanation for the behaviors.

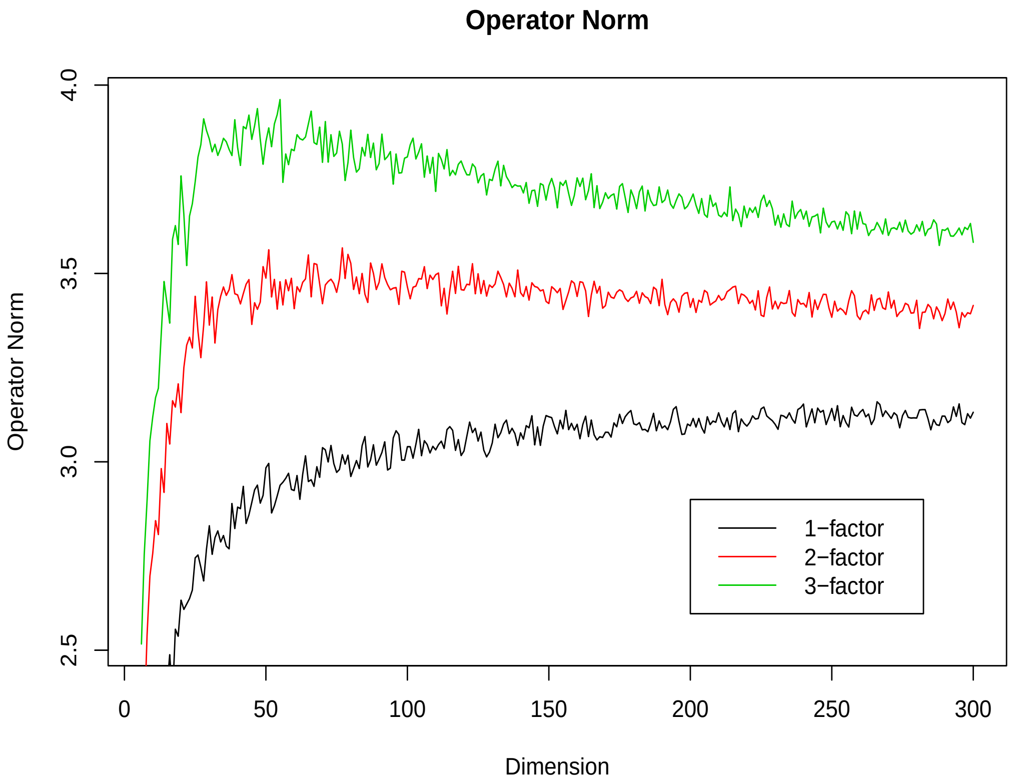


Figure 23: Plot with Operator norm of H as n increases

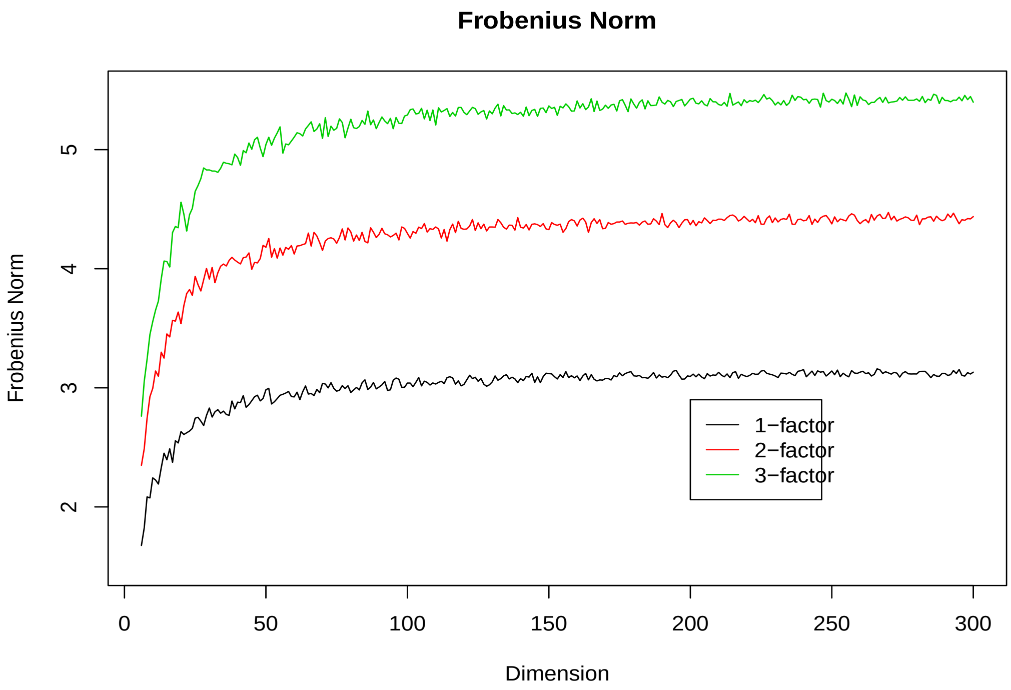


Figure 24: Plot with Frobenius norm of H as n increases

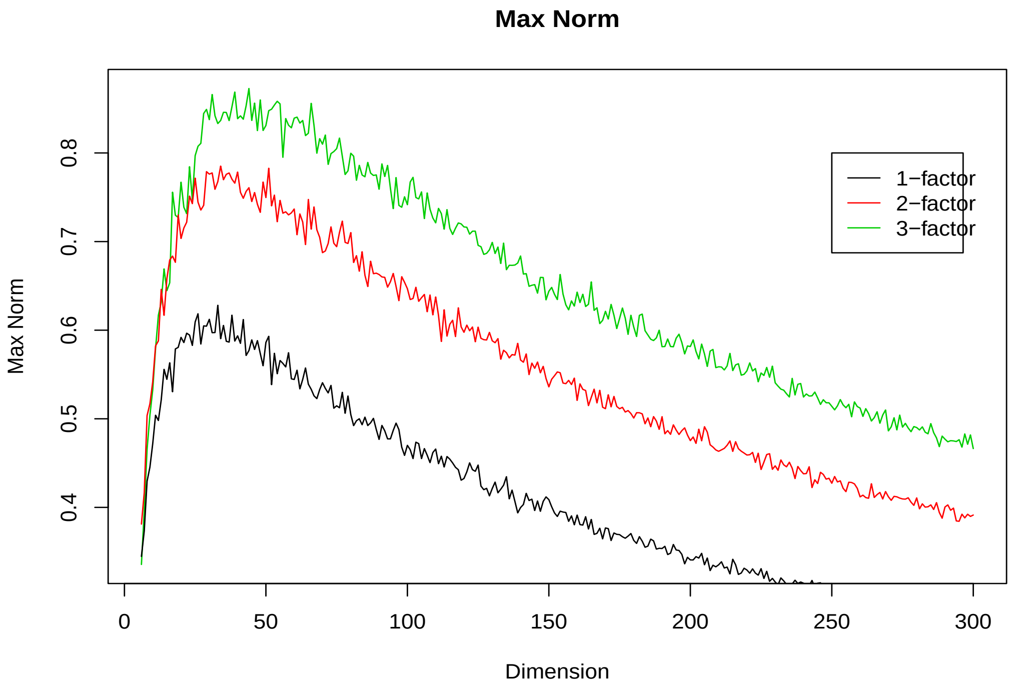


Figure 25: Plot with max norm of H as n increases

3.5.3 Communality and specific variance

As noted previously, if a covariance matrix is modeled with factor analysis, the diagonal elements of covariance matrix are decomposed into communality and specific variances. Communality is the diagonals of LL' and specific variances are those of D. It is an interesting question what relationship between these two plays a role in GLS estimation. In this subsection, the ratio of $\frac{l_i}{\lambda_i}$ is explored in the context of GLS estimate efficiency.

Let's assume one-regressor with one-factor Σ model again for simplicity. The variance of GLS estimation can be written as,

$$\begin{aligned} var(\hat{\beta}_{GLS}) &= (X'\Sigma^{-1}X)^{-1} \\ &= (X'(D^{-1} - \frac{D^{-1}LL'D^{-1}}{1+L'D^{-1}L})X)^{-1} \\ &= [\sum_{i=1}^{n} \frac{x_i^2}{\lambda_i} - \frac{(\sum_{i=1}^{n} \frac{l_i}{\lambda_i} x_i)^2}{1+\sum_{i=1}^{n} \frac{l_i^2}{\lambda_i}}]^{-1}. \end{aligned}$$

Let $\alpha = \frac{l_i^2}{\lambda_i}$ interpreted as variance ratio of communality to specific. Let us further assume that α is constant. Then we have

$$\begin{aligned} var(\hat{\beta}_{GLS}) &= (X'\Sigma^{-1}X)^{-1} \\ &= \Phi^{-1}, \end{aligned}$$

where $\Phi = \frac{(nz-g)\alpha^2+\{(n+1)z-g\}\alpha+z}{1+\alpha n}$, $z = \sum_{i=1}^{n} x_i^2$ and $g = (\sum_{i=1}^{n} x_i)^2$. In order to see the efficiency of GLS estimation with respect to α, differentiation is taken.

$$\frac{\partial \Phi}{\partial \alpha} = \frac{n(nz-g)\alpha^2 + 2(nz-g)\alpha + z - g}{(1+\alpha n)^2} \tag{50}$$

First note that $nz-g = \sum_{i=1}^{n-1}(x_i - x_{i+1})^2 \geq 0$. If $z-g$ is bounded, we can see that Φ is increasing function of α as long as n is large. Therefore, $var(\hat{\beta}_{GLS})$ is decreasing function of α. This is the proof of the following proposition.

Proposition 3.5.2. *Regression model* (22) *with one regressor is given. Suppose $\Sigma = LL' + D$ with $n \times 1$ vector L. In addition, let's assume that communality-specific ratio $\alpha = \frac{l_i}{\lambda_i}$ is constant and $\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2$ is bounded. Then there exists $N \in Z^+$ such that variance of GLS estimator for the regression model* (22) *is a decreasing function of α for all $n \geq N$.*

The point of the proposition is clear. If the dimensionality is high enough, then large portion of factor loading l_i in diagonals of covariance matrix leads to more efficient GLS estimation. In other words, the more portion of the covariance matrix of error in regression model captured by common factor, the better the GLS performance becomes. In order to expand proposition 3.5.2 to more general situation, we set up a simulation as below.

1. Set α level: from 0.05 to 1
2. After compute corresponding $\mathbb{E}(l)$ and $\mathbb{E}(\lambda)$, randomly generate $l_i \sim U(\mathbb{E}(l) - 0.25, \mathbb{E}(l) + 0.25)$, $\lambda_i \sim U(\mathbb{E}(\lambda) - 0.25, \mathbb{E}(\lambda) + 0.25)$.
3. Simulate $\Sigma = LL' + D$.
4. After random data generating process $y = X\beta + \epsilon$ with three regressors and estimate β with OLS and GLS.
5. For each α repeat the process for 1,000 times and observe relative performance of GLS to OLS via average MSE or variance of MSE.

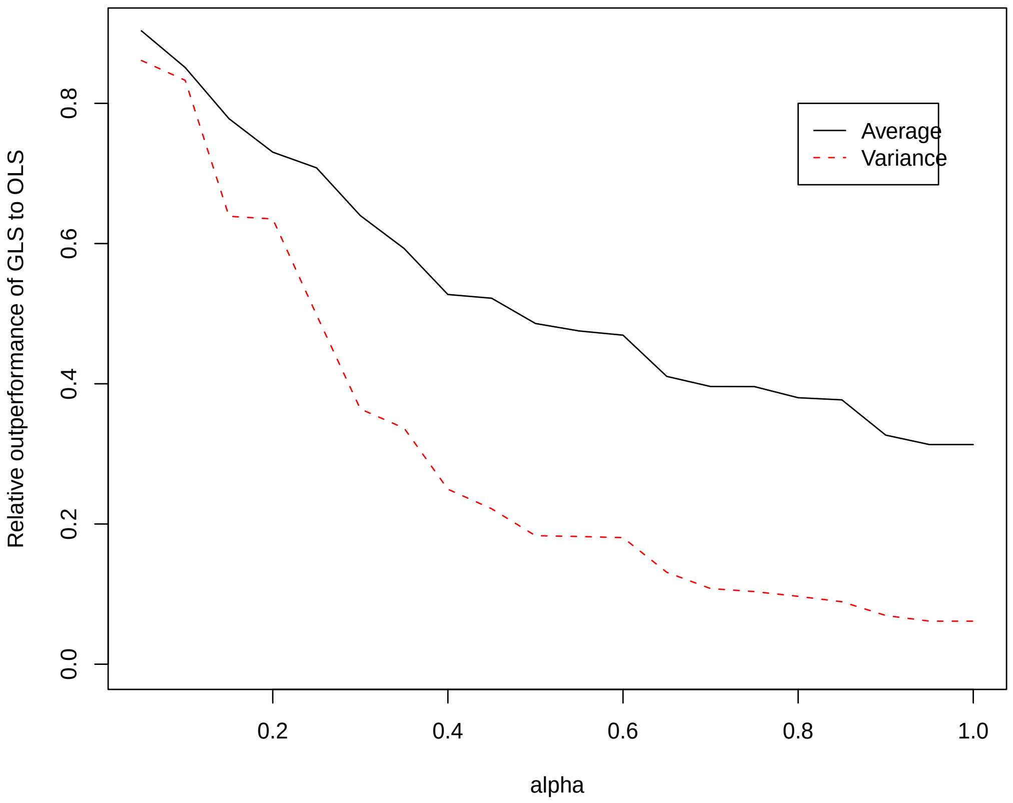


Figure 26: Plot: average and variances of GLS estimation as α increases

Figure 26 is the plot of average and variance of relative performance of GLS to OLS. Averages and variances are both decreasing as α increase, which suggests that proposition 3.5.2 can be extended to multivariate regression model with non-constant α ratio. In summary, the communality to specific variance ratio plays a role in GLS estimation and communality portion of the diagonals of Σ is a decreasing function of efficiency of GLS estimates.

3.6 Summary

The main purpose of this chapter is to explore covariance matrix structure in the context of generalized least squares estimation. We have started the discussion with feasible generalized least squares, which is practically difficult because knowledge about covariance structure is required. Therefore, we proposed banding strategy as the estimator of covariance matrix, among which fully banded strategy, i.e. diagonals of matrix ee' with e being residual of OLS, shows the best performance. In various simulations, we have shown that diagonal matrix serves the most efficient covariance matrix estimator.

We also examined the conditions under which relative performance of GLS estimate to OLS becomes better in terms of efficiency. Inspired by analytical derivations, guided simulations suggest that the more diagonals of true covariance matrix spread out, the better the GLS estimates over OLS. By taking a close look at the behaviors of diagonals of Σ and column summation of Σ^{-1}, our conjecture is supported.

Changing the view to the problem, the factor covariance matrix is introduced. Unlike preceding research, our focus lies in factor analysis to covariance matrix of error in regression model (22), in the context of GLS estimation. From both analytical calculation and simulation, Σ^{-1} is very well approximated by inverse of specific variance D^{-1} especially in higher dimensional case. Furthermore, communality to specific variance ratio and its effect on GLS estimation is studied.

We first raise the question about FGLS, and we proposed fully banded strategy for covariance matrix estimation. From several perspectives, such as sparsity, diagonal spread-out and factor modeling, the diagonal matrix is shown to be a good approximation, which supports our proposal of fully banding strategy.

CHAPTER IV

CONCLUSION

4.1 Summary and Conclusion

High dimensional covariance estimation has not been properly studied for its growing importance in practice because it is very challenging to estimate one. Especially in panel data, if sample size is relatively small to the dimensionality, the estimating error is unbearably large. In this thesis, by focusing on simulation methods, high covariance matrix estimation and its effects on asset pricing and generalized least squares are explored.

In chapter 2, we have shown that modified Stein (MST) method in estimating precision matrix works better than all the other techniques. Both in parameter estimation and model specification test, MST has the least squared error. Two-pass procedure and Hansen-Jagannathan distance are mainly considered in our computational experiments. In addition to the existing techniques, we propose a heuristic estimator of diagonal variance matrix for covariance matrix. In our simulation study of model specification testing, the new method works better than all the other ones. Throughout the chapter it is shown that small sample size relative to dimensionality is very crucial in sample covariance matrix. Although it is most frequently used in financial applications, it gives too large errors to bear with. In model specification test, it gives 99% rejections with true model given when the sample size is 160 and dimensionality is 100, for example. Even if the sample size increases to 330, the rejection rate is too high to be used in practice. Our method of diagonal variance matrix, on the other hand, provides almost the same rejection rate as the theoretical suggestions.

In chapter 3, we slightly change the point of view into the generalized least squares (GLS). GLS takes inverse of covariance matrix as a transformation matrix, and thus it is very crucial to estimate covariance matrices. The chapter examined the efficiency of banding strategy as a new method for covariance matrix estimator. The fully banded matrix, in fact, shows impressive performances in estimation errors, which is consistent with the result from chapter 2. Diagonal variance matrix is the same as fully banded strategy.

Throughout the chapter, we study effects of covariance matrix estimation on efficiency of GLS estimates with analytic efforts along with simulation experiments. Due to the nature of problem being multivariate, simplified version of the problems are analytically taken into account. General case was shown by simulations, on the other hand. We found that spread-out diagonals of covariance matrix are essential in improving GLS efficiency. Furthermore, factor covariance matrix gives us even more in-depth intuition about the diagonals in that communality-specific variance ratio is another key point in GLS efficiency. These evidences partly explain the answer to the simulation result in chapter 2 that diagonal variance matrix works very well.

In this thesis, linear asset pricing model and generalized least squares are mainly taken for our simulation settings. Factor analysis reveals that diagonal matrix is a reasonable estimator for covariance matrix. It is worth noting that our findings is about the performance of diagonal variance matrix in our specific settings. Further study on the performance of diagonal variance matrix in different situations would expand our understanding on high dimensional covariance matrices.

4.2 Future Research

Although covariance and precision matrix was intensively explored in chapter 2 and chapter 3, the study does not contain much real data work. In chapter 3, we already used real stock return data to see if our claim about degree of spread-out of diagonals in covariance matrix, and it turned out to perform well. Therefore, our research is expected to extend to a financial application with real data. Finding optimal weights on different asset classes in asset allocation problem can be one example. Hedging problem with multiple assets involved are another important extension we can think of. Since correlations are the crucial factor in hedging, high dimension with short period of hedging horizon may be a good place where our covariance matrix estimators can be utilized.

Theoretical side regarding convergence rate is also an important extension. Since this thesis is mainly showing the result using simulations, many of the essential research questions are remained as conjectures. Some of the results are only shown in univariate special cases, so there exists theoretical room to fill in for future research.

APPENDIX A

ADDITIONAL PLOTS FOR CHAPTER 2

In Chapter 2, computer experiment with simulations are conducted to check the performances of various precision matrix estimation. Especially section 2.3 is about two-pass procedures. The following figures are the pair plots of MSE in two-pass procedure with different covariance structures, which is not provided in the main body of the thesis due to the space limit.

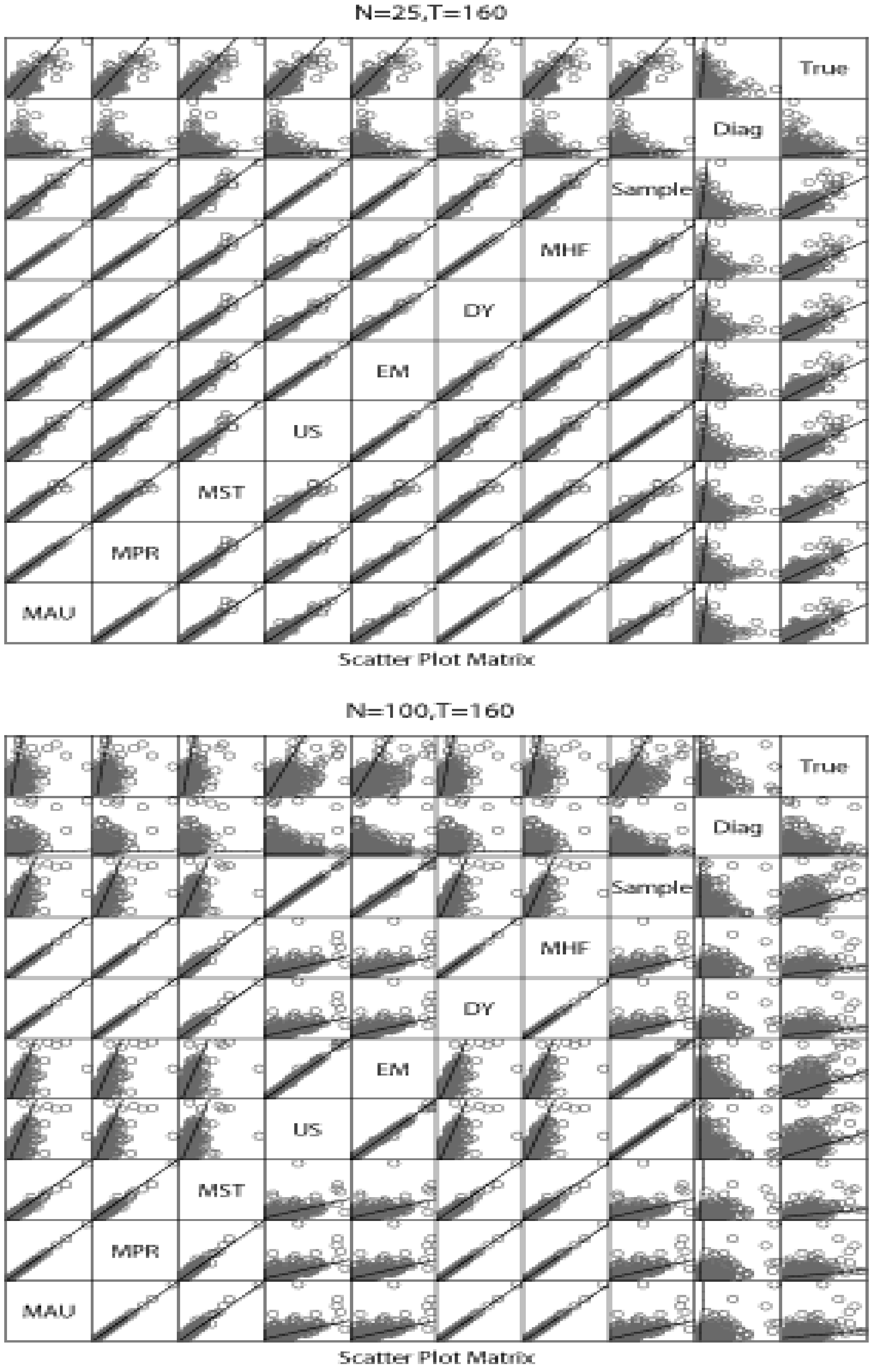
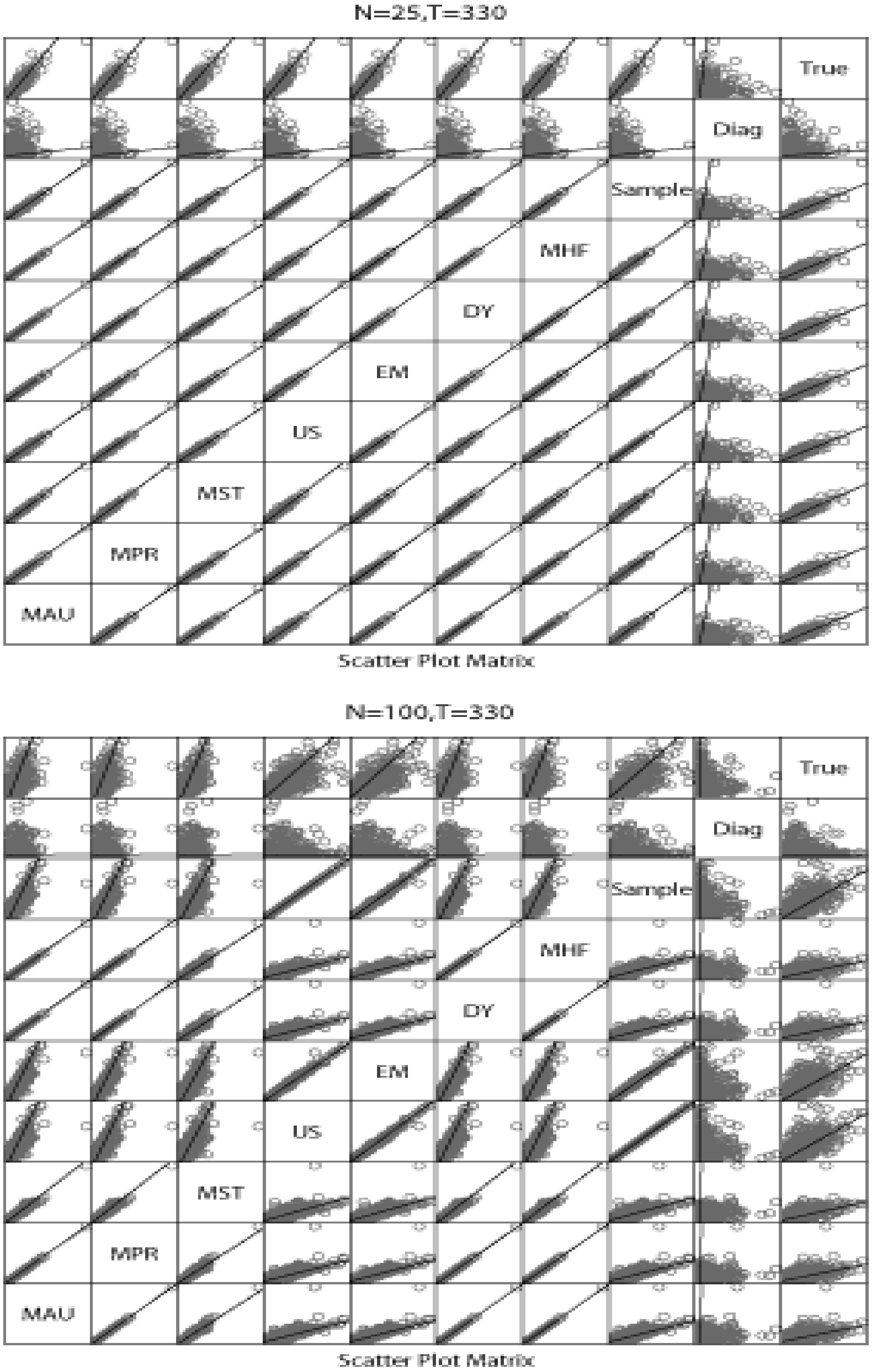


Figure 27: AR(1), T=160


N=25,T=330

N=100,T=330

Figure 28: AR(1), T=330

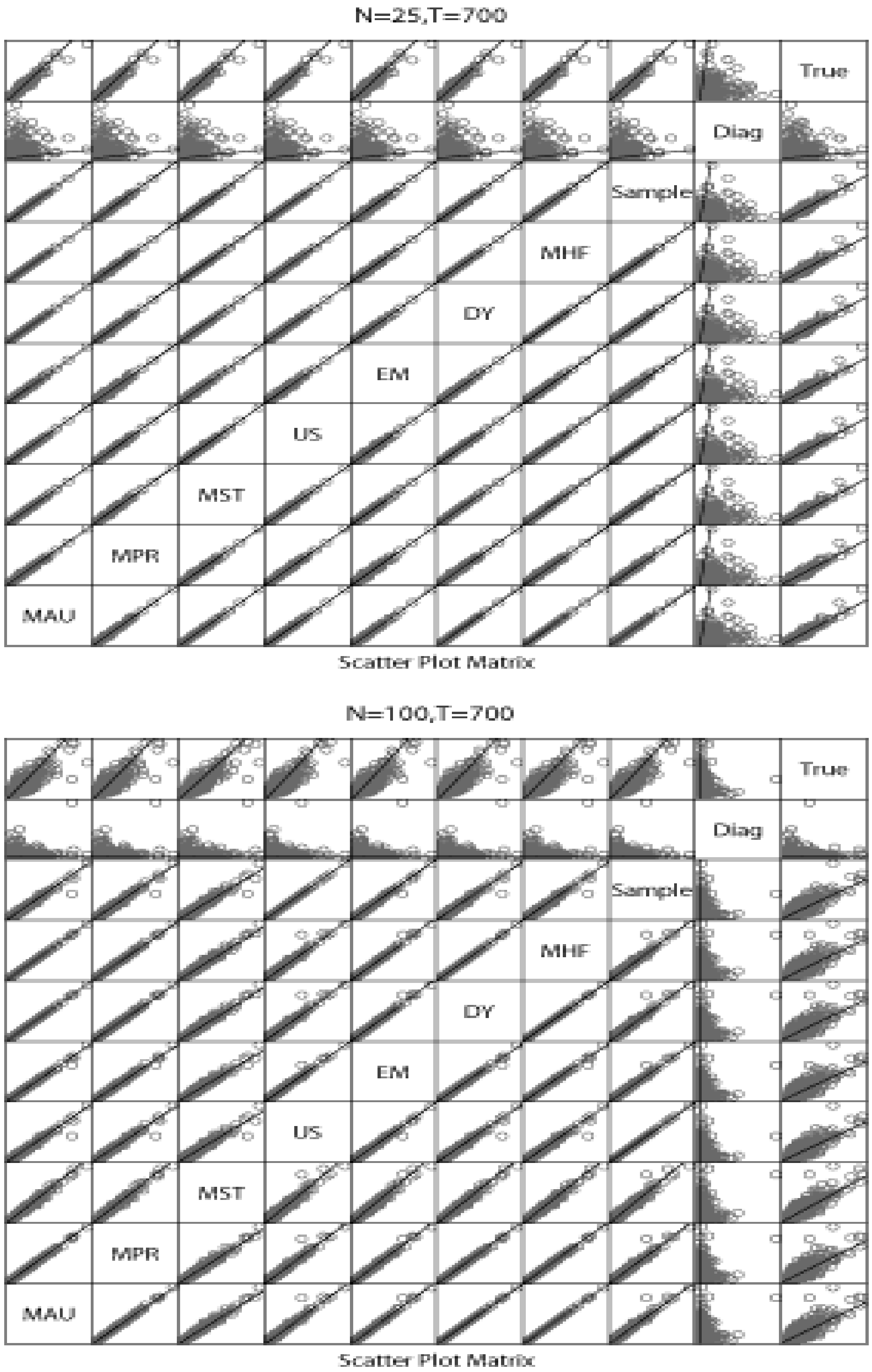


Figure 29: AR(1), T=700

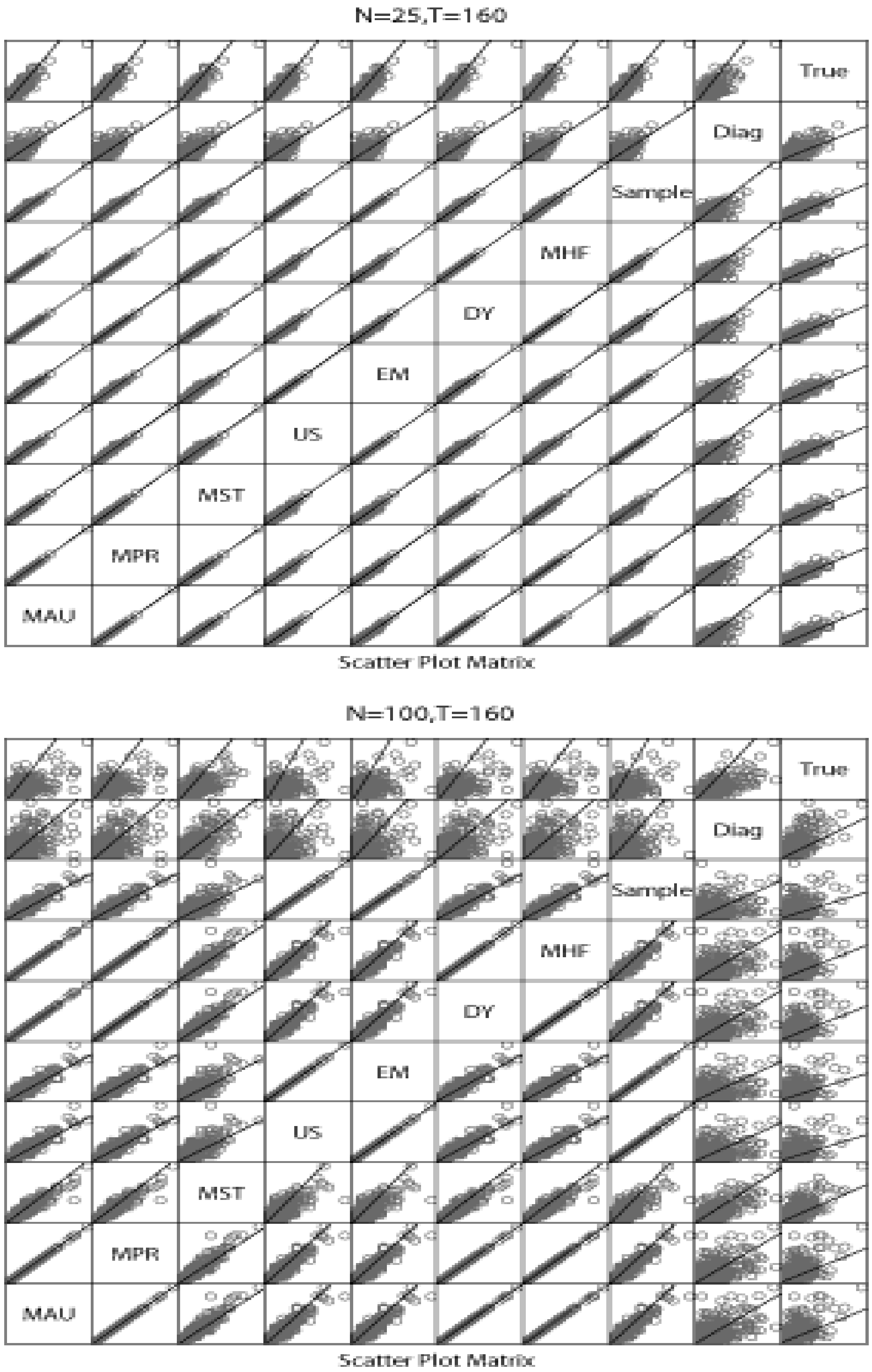
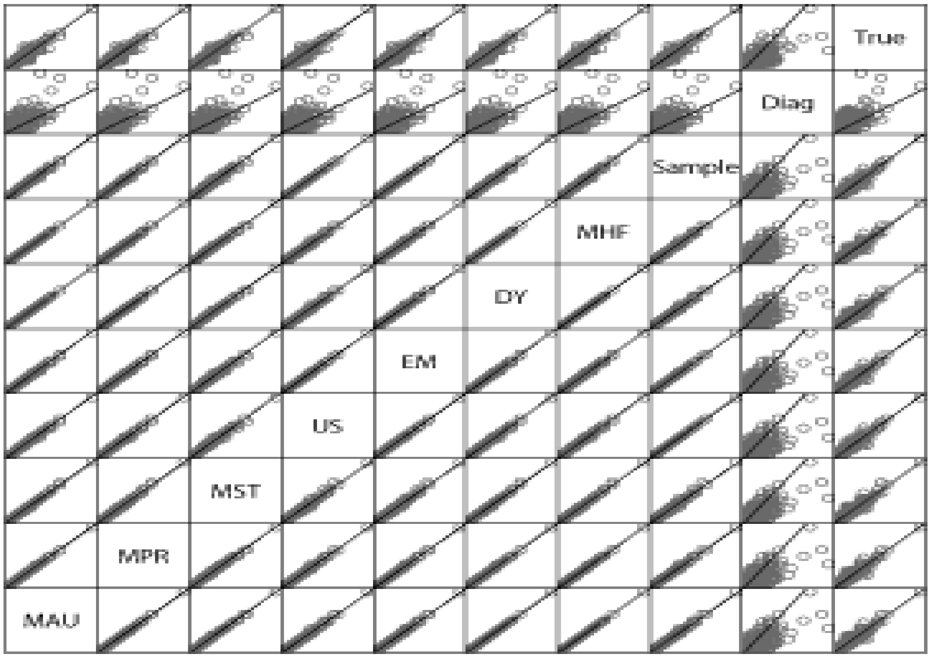


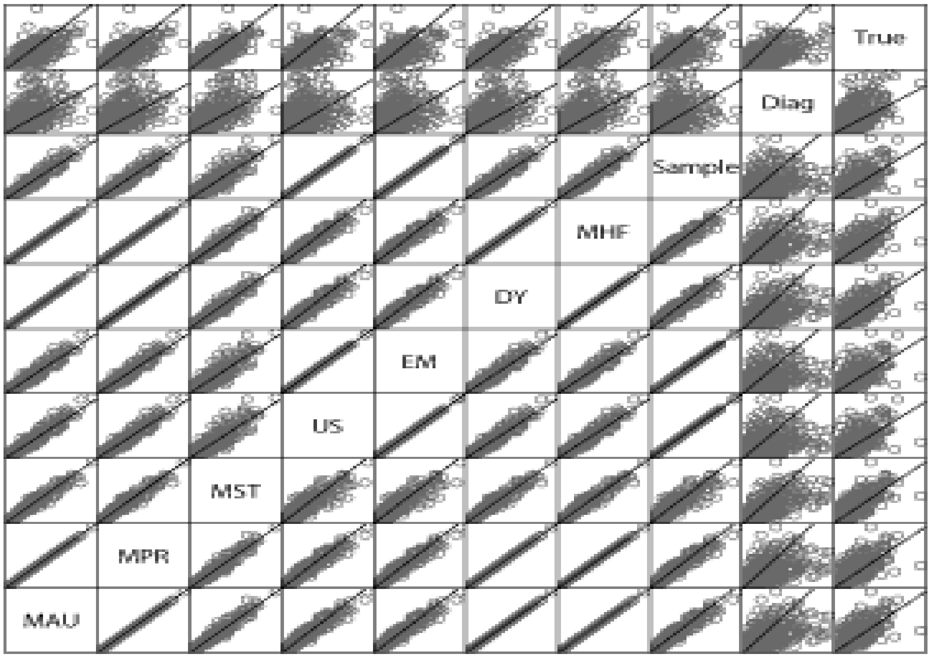
Figure 30: AR(2), T=160

N=25,T=330



Scatter Plot Matrix

N=100,T=330



Scatter Plot Matrix

Figure 31: AR(2), T=300

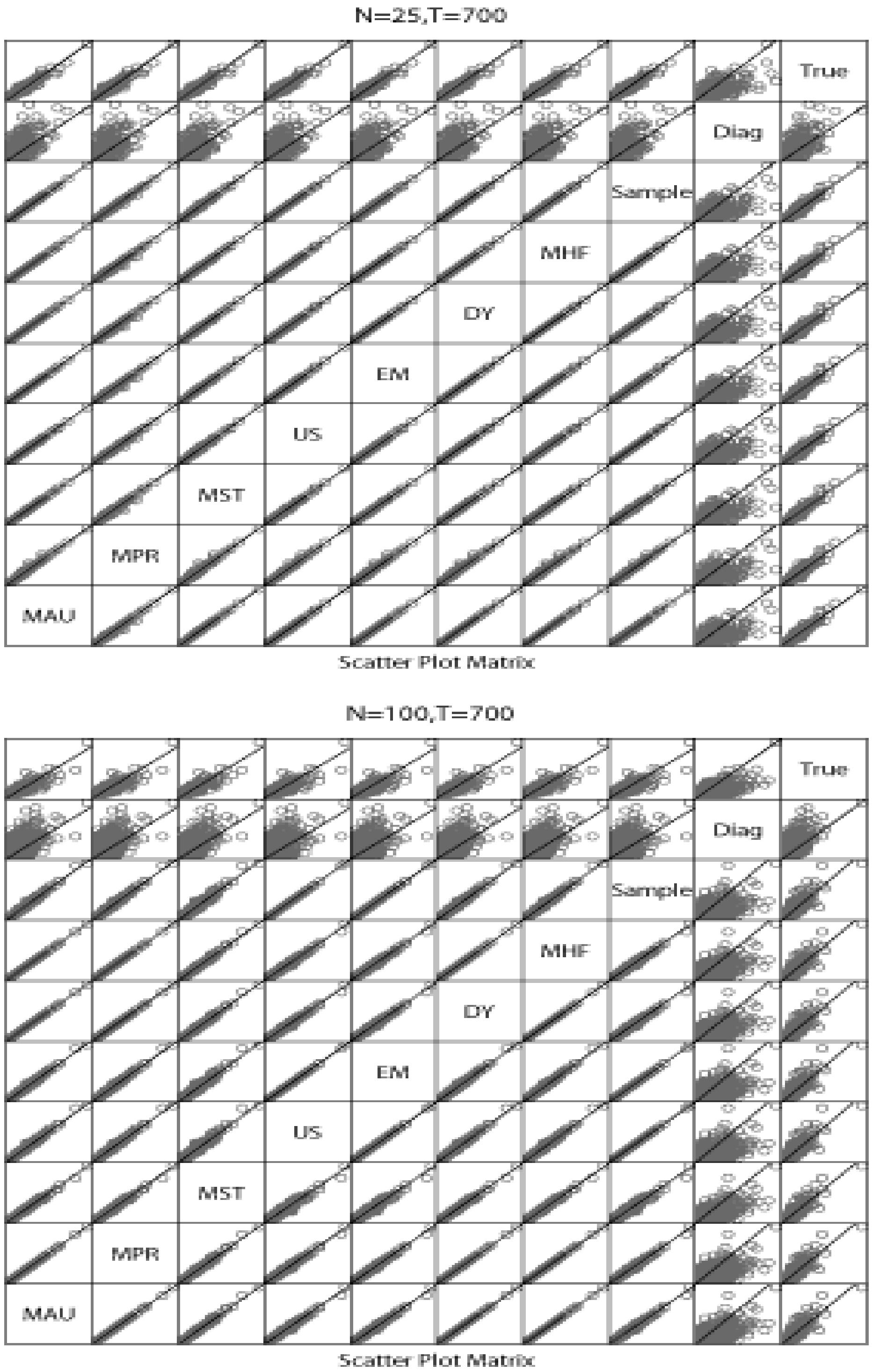


Figure 32: AR(2), T=700

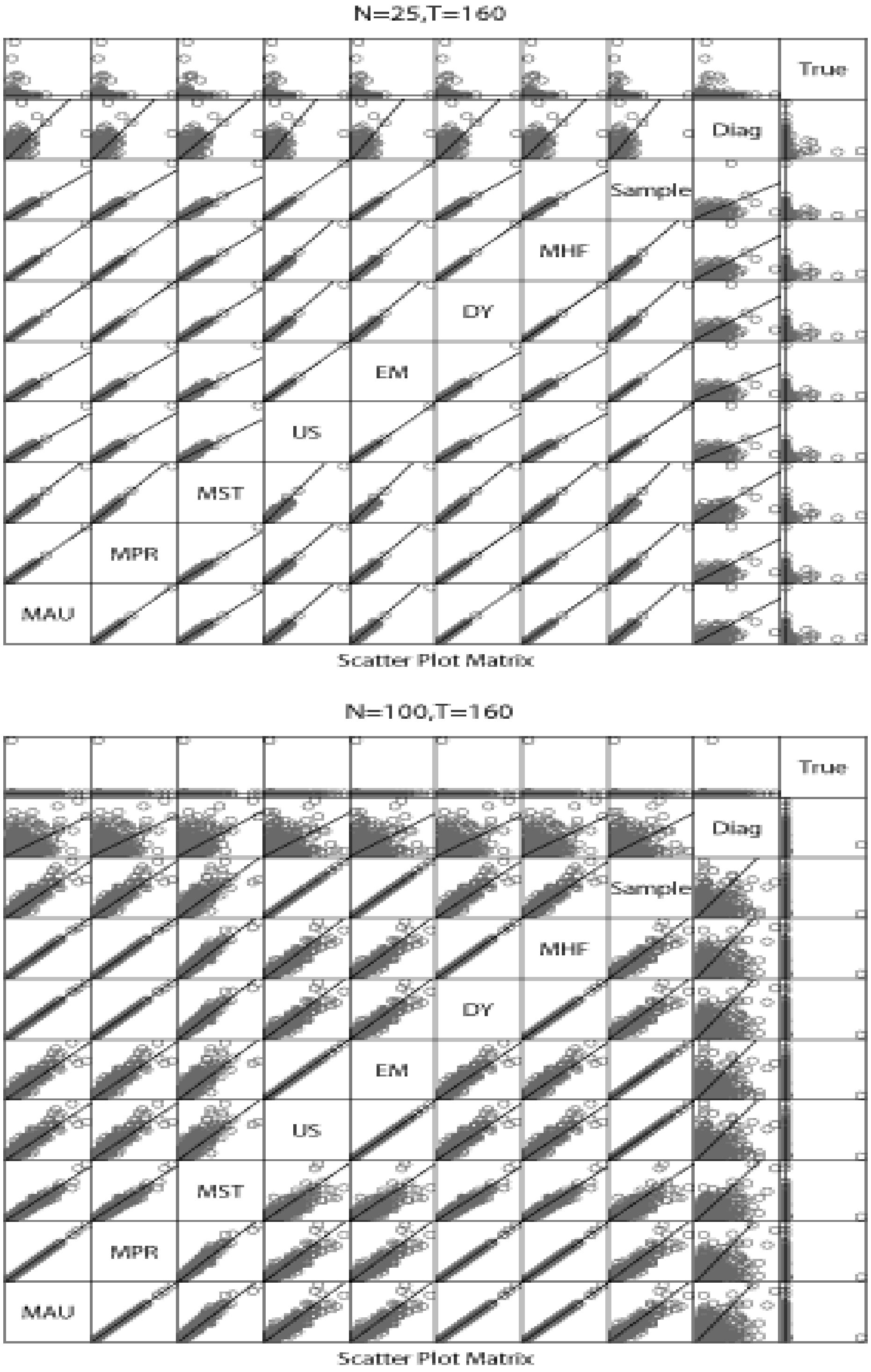


Figure 33: Heteroscedastic random correlations, T=160

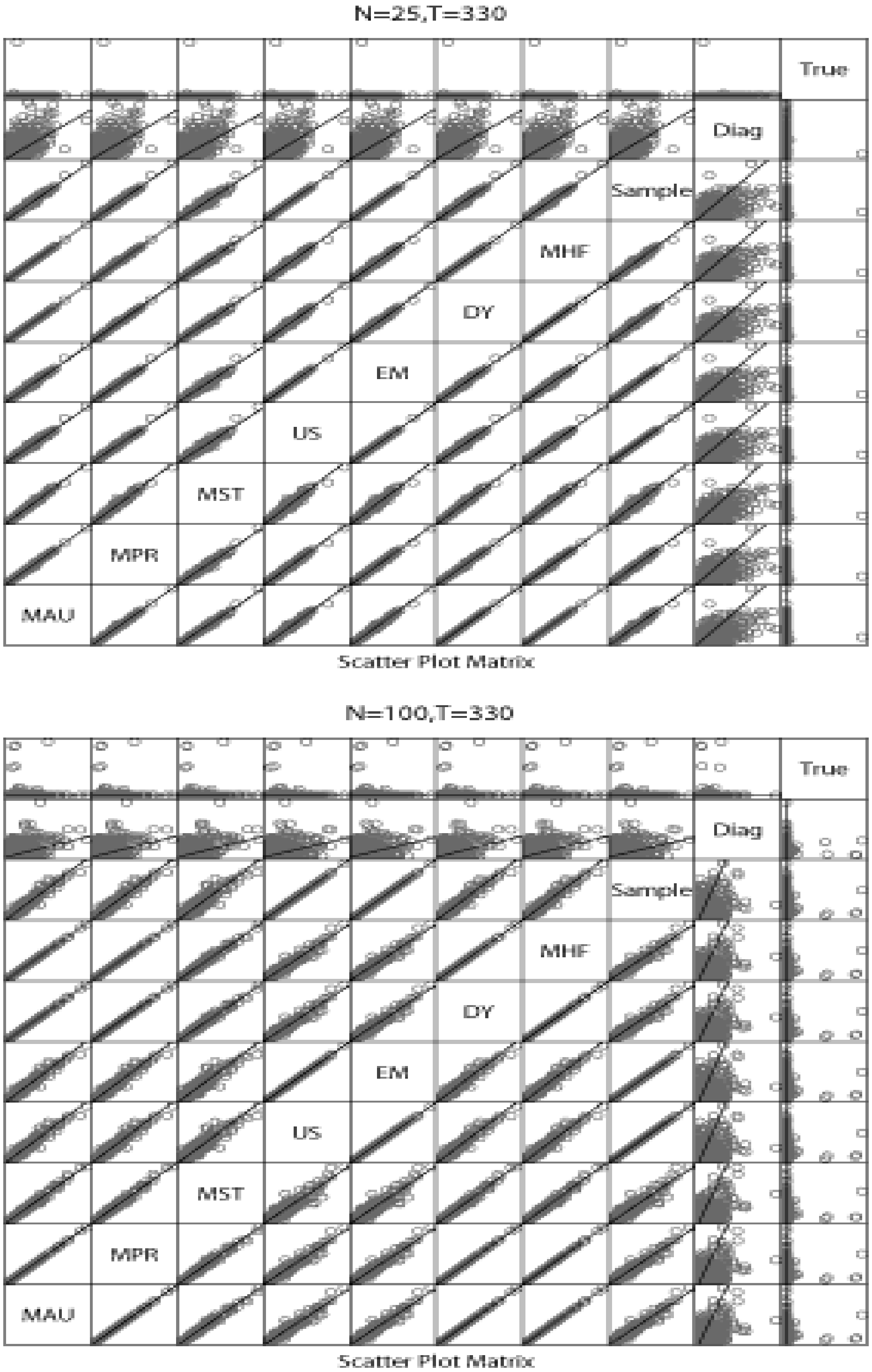


Figure 34: Heteroscedastic random correlations, T=330

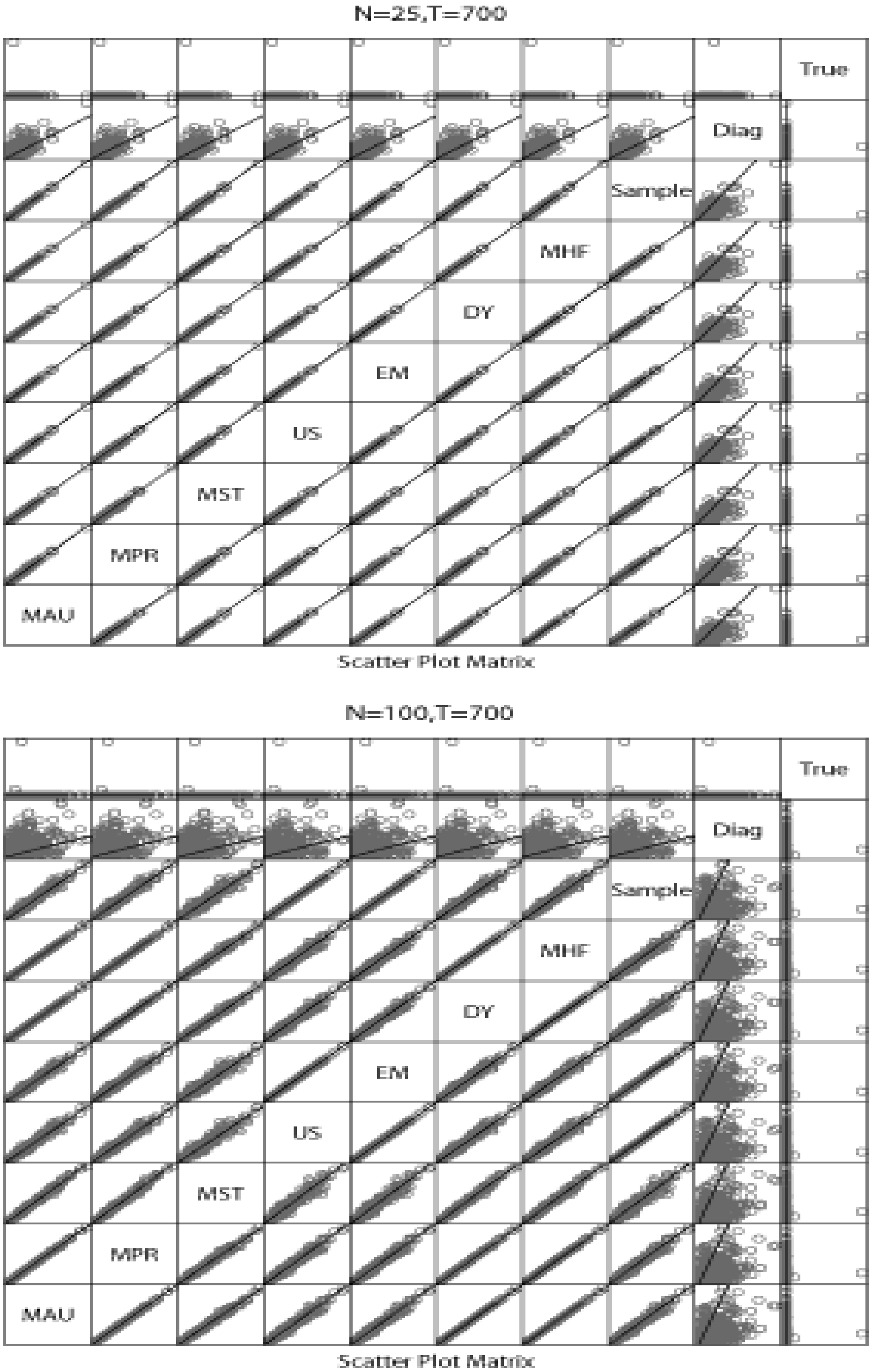


Figure 35: Heteroscedastic random correlations, T=700

APPENDIX B

ADDITIONAL PLOTS FOR CHAPTER 3

Followings are the miscellaneous plots which supports the ideas represented in chapter 3.

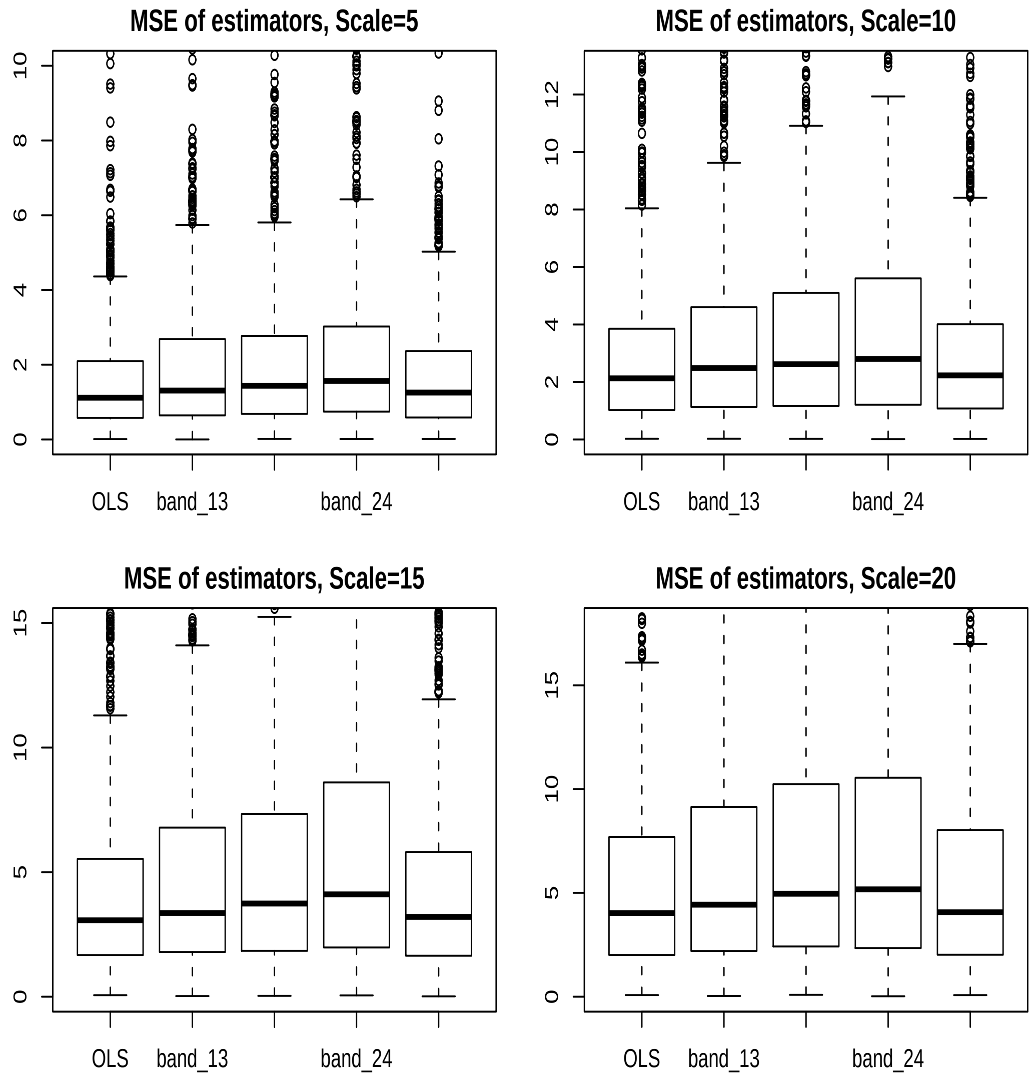
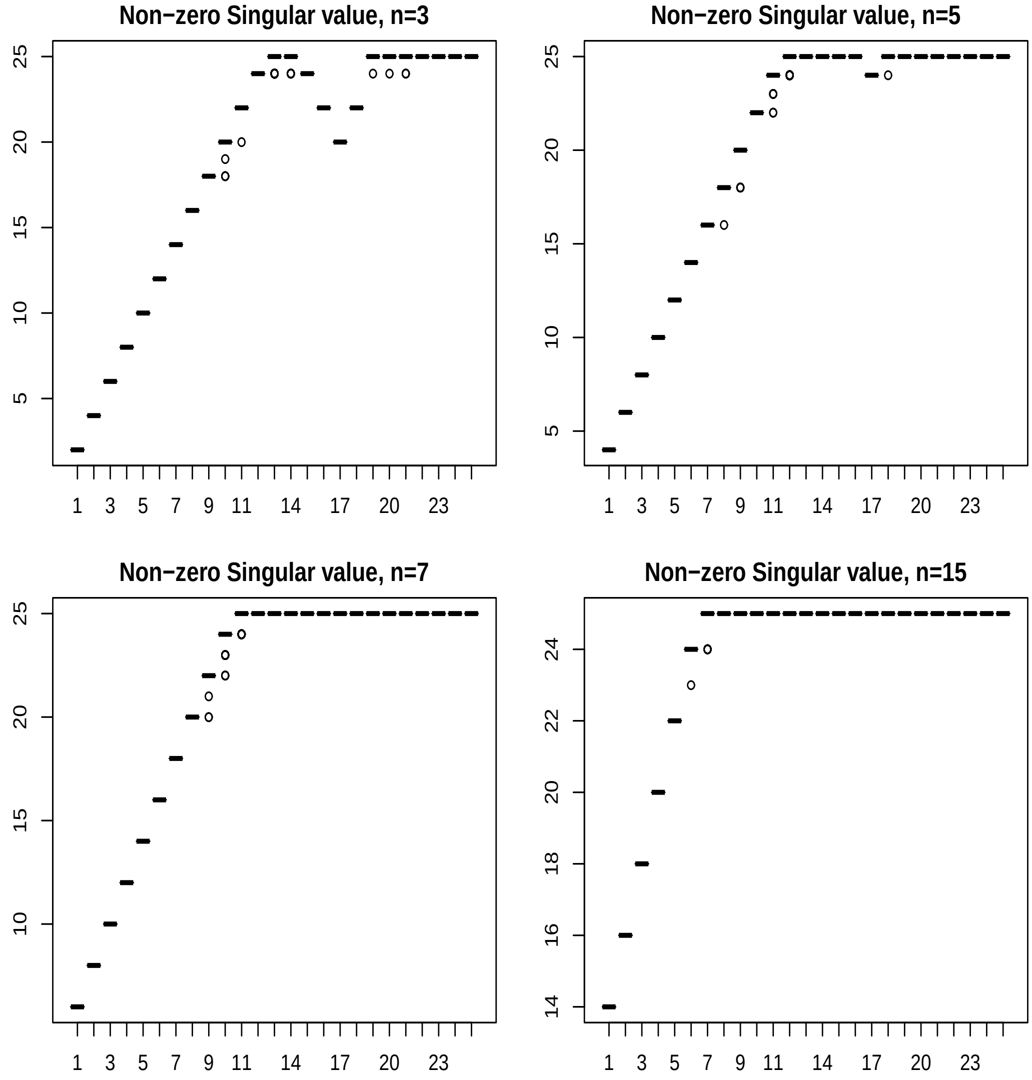


Figure 36: Estimating error with OLS and various banding strategies

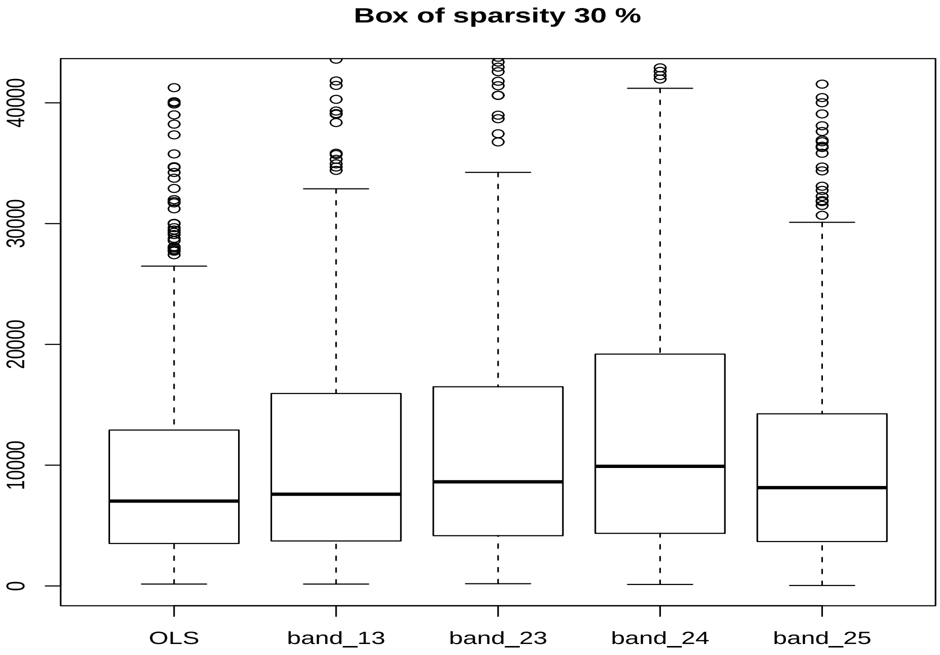

Non−zero Singular value, n=3

Non−zero Singular value, n=5

Non−zero Singular value, n=7

Non−zero Singular value, n=15

Figure 37: Number of Nonzero singular values of banding strategies in multiple observations case



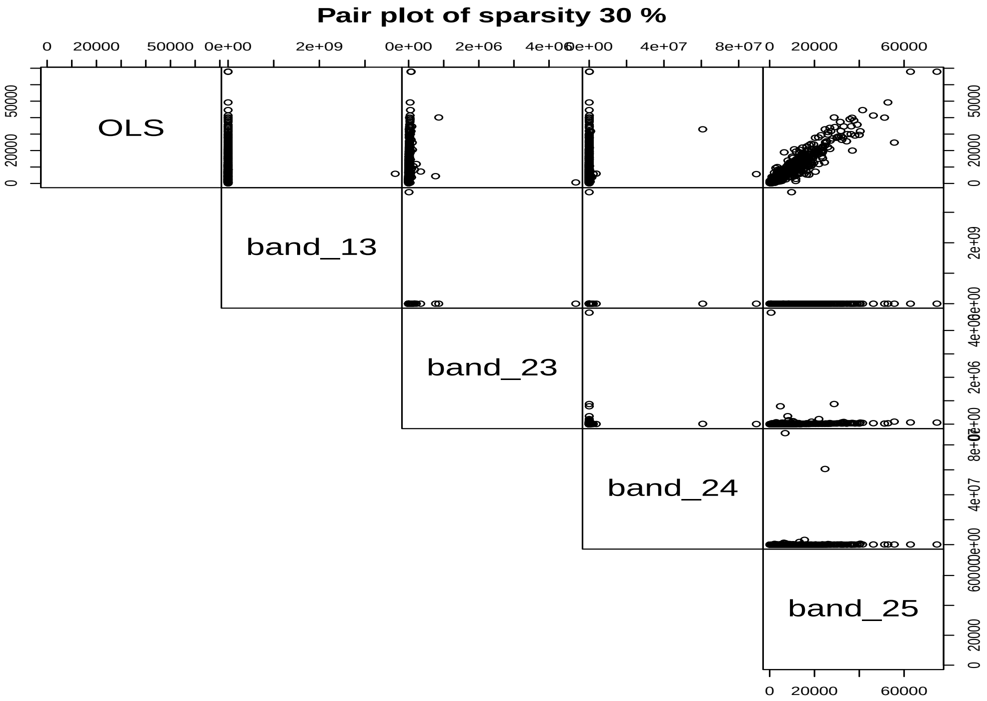


Figure 38: box and pair plots when sparsity of covariance matrix is 30%

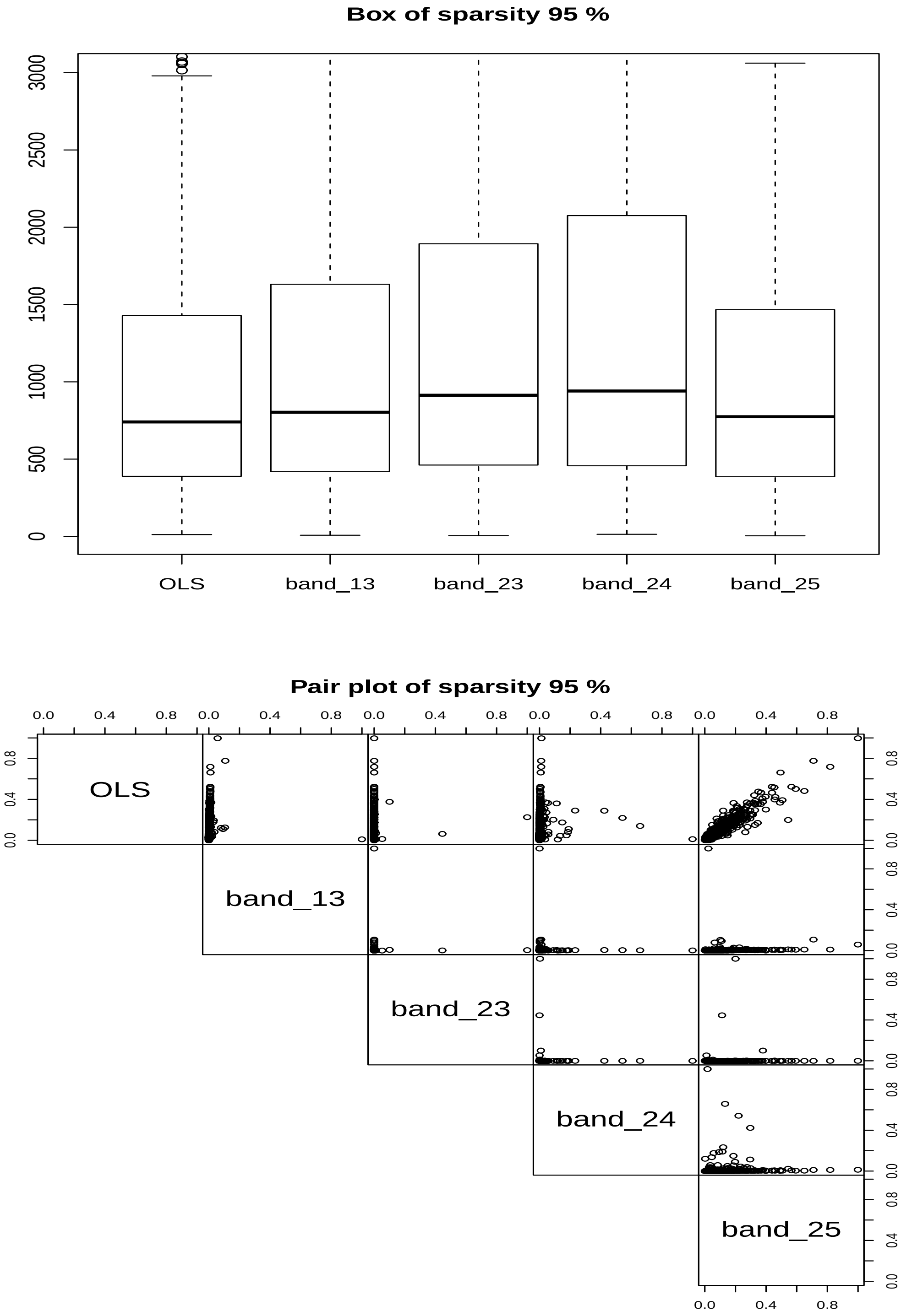


Figure 39: box and pair plots when sparsity of covariance matrix is 95%

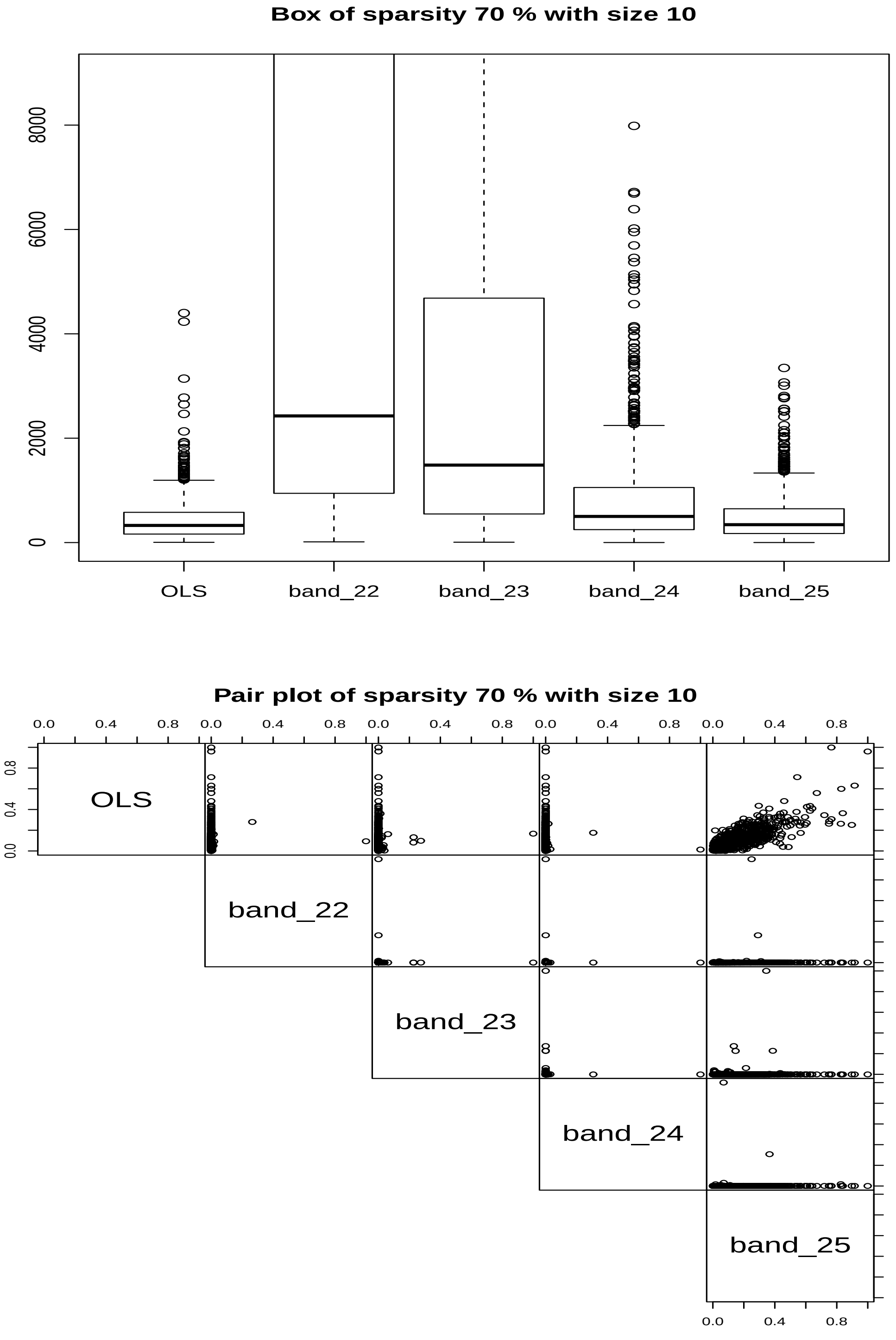


Figure 40: box and pair plots of panel data case when sparsity of covariance matrix is 70% and sample size is 10.

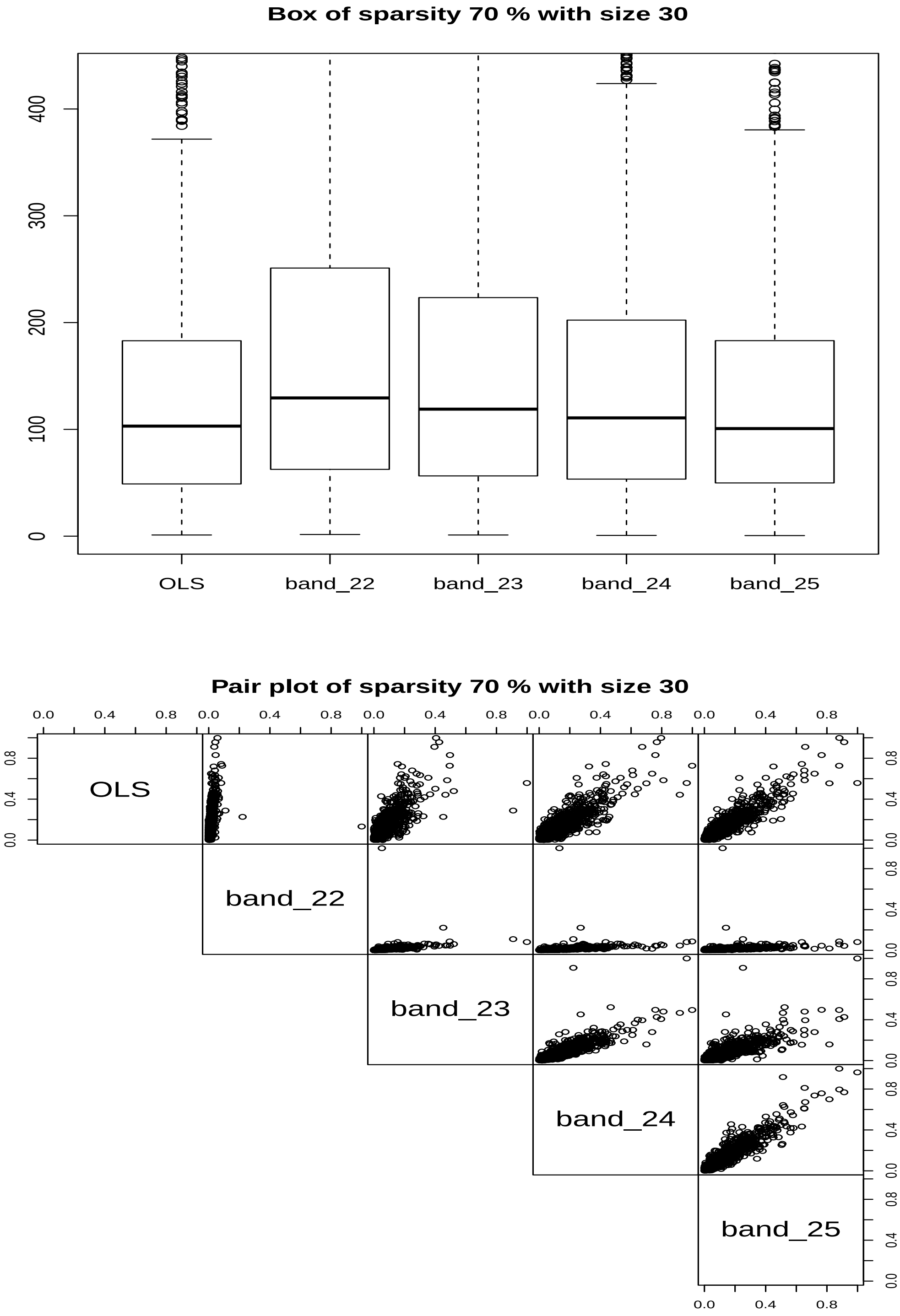


Figure 41: box and pair plots of panel data case when sparsity of covariance matrix is 70% and sample size is 30.

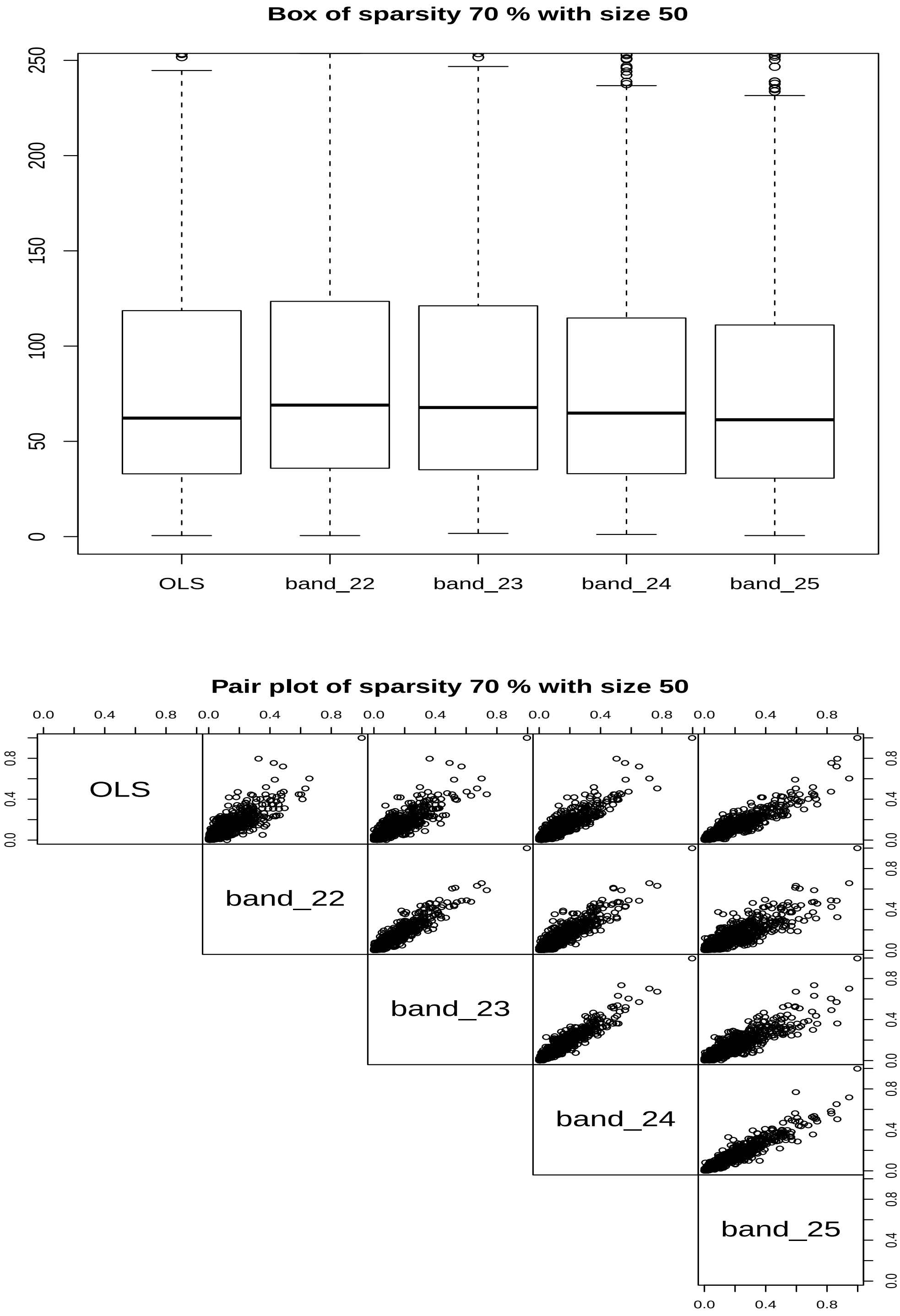


Figure 42: box and pair plots of panel data case when sparsity of covariance matrix is 70% and sample size is 50.

REFERENCES

[1] Seung C. Ahn and Christopher Gadarowski. Small sample properties of the GMM specification test based on the Hansen-Jagannathan distance. *Journal of Empirical Finance*, 11:109–132, 2004.

[2] A. C. Aitken. On least squares and linear combinations of observations. *Proceedings of Royal Statistical Association*, 55:42–48, 1935.

[3] T. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 2003.

[4] M. Blume and I. Friend. A new look at the capital asset pricing model. *Journal of Finance*, 28:19–34, 1973.

[5] T. Bollersleve. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[6] M. Burnside, C.and Eichenbaum. Small-sample properties of gmm-based wald test. *Journal of business and economic statistics*, 14:294–308, 1996.

[7] J.Y. Campbell and J.H. Cochrane. Explaining the poor performance of consumption-based asset pricing models. *Journal of Finance*, 55:2863–2878, 2000.

[8] J.Y. Campbell, A.W. Lo, and A.C. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1997.

[9] M. Carhart. On persistence in mutual fund performance. *Journal of Finance*, 52:57–82, 1997.

[10] D. Chchrane and G.H Orcutt. Application of least squares regression to relationships containing auto- correlated error terms. *Journal of the American Statistical Association*, 44:32–61, 1949.

[11] N. Chen, R. Roll, and S. Ross. Economic forces and the stock market. *Journal of Business*, 59:383–403, 1986.

[12] J.H. Cochrane. *Asset Pricing*. Princeton University Press, 2005.

[13] G. Connor and R. Korajczyk. Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics*, 15:373–394, 1986.

[14] R. Connor and Korajczyk R. Risk and return in an equilibrium APT: Application of a new test methodology. *Journal of Financial Economics*, 21:255–290, 1988.

[15] G. W. Douglas. Risk in the equity market: An empirical appraisal of market efficiency. *Yale Economic Essays*, 9:5–45, 1969.

[16] R.F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of uk inflation. *Econometrica*, 50:987–1007, 1982.

[17] E. Fama and K.R. French. The cross-section of expected stock returns. *Journal of Finance*, 47:427–466, 1992.

[18] E. Fama and K.R. French. Multifactor explanations of asset pricing anomalies. *Journal of Finance*, 51:55–84, 1996.

[19] Eugen Fama and J. MacBeth. Risk, returns and equilibrium: Empirical tests. *Journal of Political Economy*, 81:607–636, 1973.

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

[21] Foerster S.R. Ferson, W.E. Finite sample properties of the generalized method of moments in tests of conditional asset pricing models. *Journal of Financial Economics*, 36:29–55, 1994.

[22] G. Golub and C. Van Loan. *Matrix computations*. Johns Hopkins University Press, 1996.

[23] William H. Greene. *Econometric Analysis*. Prentice Hall, 2002.

[24] Lars Peter Hansen and Ravi Jagannathan. Assessing specification errors in stochastic discount factor models. *Journal of Finance*, 52:557–590, 1997.

[25] L.P. Hansen. Large sample properties of generalized method of moment estimators. *Econometrica*, 50:1029–1054, 1982.

[26] R.J. Hodrick and X.Y. Zhang. Evaluating the specification errors of asset pricing models. *Journal of Financial Economics*, 62:327–376, 2001.

[27] R. Horn and Johnson C. *Matrix Analysis*. Cambridge University Press, 1985.

[28] Barber J. and Copper M. Immunization using principal component analysis. *Journal of Portfolio Management*, 23:99–105, 1996.

[29] K. Jacobs and K.Q. Wang. Idiosyncratic consumption risk and the cross section of asset returns. *Journal of Finance*, 59:2211–2252, 2004.

[30] R. Jagannathan and Z. Wang. An asymptotic theory for estimating beta-pricing models using cross-sectional regression. *Journal of Finance*, 53:1258–1309, 1998.

[31] R. Jagannathan and Z. Wang. Empirical evaluation of asset-pricing models: A comparison of the sdf and beta methods. *Journal of Finance*, 57:2337–2367, 2002.

[32] Ravi Jagannathan and Zhenyu Wang. The conditional CAPM and the cross-section of expected returns. *Journal of Finance*, 51:3–53, 1996.

[33] R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, 2007.

[34] R. Kan and C. Zhang. Gmm tests of stochastic discount factor models with useless factors. *Journal of Financial Economics*, 54:103–127, 1999.

[35] T. Kariya and H. Kurata. *Generalized Least Squares*. Wiley, 2004.

[36] Oliver Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10:603–621, 2003.

[37] B. Lehmann and D. Modest. The empirical foundations of the arbitrage pricing theory. *Journal of Financial Economics*, 21:213–254, 1988.

[38] M. Lettau and S. Ludvigson. Resurrecting the capm: a cross-sectional test when risk premia are time-varying. *Journal of Political Economy*, 109:1238–1287, 2001.

[39] J. Litner. The valuation of risk assets and the selection of risky investments in stock portfolios ans capital budgets. *Riview of Economics and Statistics*, 47:13–37, 1965.

[40] R. Litterman and J. Scheinkman. Common factors affecting bond returns. *Journal of Fixed Income*, 1:54–61, 2008.

[41] G.S. Maddala. *Introduction to Econometrics*. Prentice Hall, 1998.

[42] Yu Ren and Katsumi Shimotsu. Specification test based on the Hansen-Jagannathan distance with good small sample properties. *Working Paper*, 2006.

[43] S. Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13:341–360, 1976.

[44] R. Shayle Searle. *Matrix algebra useful for statistics*. Wiley, 1982.

[45] Jay Shanken. On the estimation of beta-pricing models. *The Review of Financial Studies*, 5:1–33, 1992.

[46] W. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19:425–442, 1964.

[47] Hisayki Tsukuma and Yoshihiko Konno. On improved estimation of normal precision matrix and discriminant coefficients. *Journal of Multivariate Analysis*, 97:1477–1500, 2006.

[48] M. Vassalou. Nes related to future gdp growth as a risk factor in equity returns. *Journal of Financial Economics*, 68:47–73, 2003.

VITA

Soo-Hyun Kim graduated with B.A. in economics in 2000 from Seoul National University, Seoul, Republic of Korea. After college, he served Republic of Korea Airforce (ROKAF) as a supply planning officer for three years, and then was discharged with 1st Lieutenant in June 2003.

Soo-Hyun came to Unitied States to continue his study in economics in July 2003. He started his graduate study at economic department at Yale University. He earned M.A. in international and development economics in May 2004, and then obtained another master's degree in statistics next year at Yale University. He started Ph.D. program at Georgia Institute of Technology in August 2005.

During his Ph.D. study, he worked at the office of economic analysis at US Securities and Exchange Commission (SEC), from May 2008 to March 2009. He conducted statistical analysis in financial markets. He also worked as teaching assistant and research assistant for faculty members of ISyE. As his research interests lies in financial statistics, he course worked in Quantitative and Computational Finance program throughout his Ph.D. program to have earned M.S. in QCF in May 2009. He is expected to graduate with doctorate in applied statistics in August of 2010 and to start work for Samsung Asset Management as quantitative investment strategist.