5-2013

Error Covariance Matrix Estimation in High Dimensional Approximate Factor Models Using Adaptive Thresholding: A Simulation Study

Paul Chimenti
Clemson University, pchimen@g.clemson.edu

Error Covariance Matrix Estimation in High Dimensional Approximate Factor Models Using Adaptive Thresholding: A Simulation Study

---

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Mathematical Sciences

---

by
Paul J. Chimenti
May 2013

Accepted by:
Dr. Jun Luo, Committee Chair
Dr. Patrick Gerard
Dr. Colin Gallagher

Abstract

Approximate factor models are popular in finance and economics. A key to effectively utilizing such a model is to accurately estimate the error covariance matrix. Errors related to certain predictors are expected to be correlated and this must be modeled effectively. Adaptive thresholding is a method for estimating the error covariance matrix of such a model. This method is described in detail and a simulation study sheds light on the behavior of this method under different sample sizes and parameterizations.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The approximate factor model used in economics and finance is a case where it is possible to encounter data for which the number of predictor variables is close to or larger than the sample size. In practice, this can occur when evaluating stocks. There may be thousands of stocks in a portfolio which is to be analyzed, but the sample size may not be that large. For most estimation methods, this situation is problematic. However, adaptive thresholding is a technique which can be used to address this problem. In the proceeding, the factor model and adaptive thresholding technique will be introduced. Following that, some estimation methods for similar models will be reviewed. This will allow for a better understanding of the current method and will also supplement the material being studied. After that point, theoretical properties of the adaptive thresholding technique will be demonstrated and these properties will be verified via simulation. Finally, conclusions will be drawn and implications for future work will be discussed.

The factor model is defined as

$$y_{it} = \mathbf{b_i'f_t} + u_{it} \tag{1.1}$$

where y_{it} is the observation for the ith asset at time t; $\mathbf{b_i}$ is a $K \times 1$ vector of factor loadings: $\mathbf{f_t}$ is a $K \times 1$ vector of common factors and u_{it} is the error associated with y_{it} with $i = 1, ..., p$ and $t = 1, ..., T$. The factor loadings are correlation coefficients between the factors and the variables being studied.

Chapter 2

Related Work

Linear regression continues to be a popular method of data analysis due to the relative ease of calculation and interpretability of results. Many methods are available for use in research and several will be discussed here. The ordinary least squares estimator and ridge estimator will be summarized under the assumption that the sample size is larger than the dimension of the data ($n > p$). The objective functions which are minimized will be presented in a loss + penalty format. The derivations of the expected value and mean squared error will also be presented. A description of the mechanics of semi-parametric regression including the effect of a differencing matrix will also be explained.

2.1 Linear Model

A linear model is defined by

$$Y = X\beta + \epsilon \tag{2.1}$$

where $Y = (y_1, y_2, ..., y_n)'$ is the vector of response variables, X is the $n \times p$ design matrix whose first column is $(1, 1, ..., 1)'$ to account for the intercept β_0 and whose remaining columns are the data gathered for the predictor variables, $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)'$ is the vector of error terms resulting from $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})'$ which is the $p \times 1$ vector of regression coefficients which measure the effect of each predictor X_i on Y.

2.2 Ordinary Least Squares Estimator(OLS)

The ordinary least squares technique is subject to many limiting assumptions. First, the relationship between the response vector of response variables Y and the matrix of predictor variables X must be approximately linear. Otherwise a different model may be appropriate. It is necessary that the error terms are uncorrelated and commonly assumed that the error terms are normally distributed with mean 0 and variance σ^2. That is $E[\epsilon_i, \epsilon_j] = 0$ for all $i \neq j$ and $\epsilon|X \sim N(0, \sigma^2 I)$. The normality assumption is not necessary for this method but leads to desirable results. A corresponding assumption that will also be used is that $Y \sim N(X\beta, \sigma^2 I)$. It must also be assumed that their is no linear dependence between the predictor variables (the columns of X). This assumption can be overcome by other models which include a penalty term on the coefficients but is needed for OLS.

Since β is not a known value, it must be estimated. We do so by finding the value which minimizes the sum of the square of the error terms $\left(\sum_{i=1}^{n} \epsilon_i^2\right)$. Since $\epsilon_i = y_i - \beta_0 - x_{1,1}\beta_1 - \cdots - x_{1,p-1}\beta_{p-1}$, this minimum has been formulated as

$$\begin{aligned}\hat{\beta}_{LS} &= \arg\min_{\beta} \|Y - X\beta\|^2 \\ &= (X'X)^{-1}X'Y\end{aligned}$$

where $\|\cdot\|$ is the L^2 norm for $\mathbb{R}^n$. That is for $V = (v_1, v_2, ..., v_n)'$, $\|V\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$. Once $\hat{\beta_{LS}}$ has been found, $\widehat{\epsilon}$ can easily be calculated. Here $\|Y - X\beta\|^2$ is the L_2 loss function. Note that for OLS there is no penalty function.

For the OLS estimator, one desirable property is that it is unbiased. That is $E(\hat{\beta}_{LS}) = \beta$.

Observe

$$\begin{aligned} E(\hat{\beta}_{LS}) &= E[(X'X)^{-1}X'Y] \\ &= E[(X'X)^{-1}X'(X\beta + \epsilon)] \\ &= \beta + (X'X)^{-1}X'E(\epsilon) \\ &= \beta \end{aligned}$$

as $E(\epsilon) = 0$. It is also important to measure the variability of a given estimator. A lower amount of variability is a desirable quality. In this case, the true covariance of $\hat{\beta}_{LS}$ can be derived as

$$\begin{aligned} \text{Cov}(\hat{\beta}_{LS}) &= \text{Cov}((X'X)^{-1}X'Y) \\ &= (X'X)^{-1}X'\text{Cov}(Y)((X'X)^{-1}X')' \\ &= (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

However, σ^2 is not known and therefore must be estimated. Therefore $\widehat{\text{Cov}}(\hat{\beta}_0) = \hat{\sigma}^2(X'X)^{-1}$ is obtained where

$$\hat{\sigma}^2 = \frac{\|Y - \hat{Y}\|^2}{n-p} = \frac{Y'Y - \hat{\beta}'_{LS}X'Y}{n-p} = \frac{Y'(I-H)Y}{n-p}$$

where $H = X(X'X)^{-1}X'$ and is commonly referred to as the projection (or hat) matrix ([Kutner et al., 2005]). Again $\|\cdot\|$ is the L^2 norm on $\mathbb{R}^n$ previously described. If the assumption of uncorrelated errors fails, then a different method may be optimal. Assume that rather than $\epsilon|X \sim N(0, \sigma^2 I)$ we have that $\epsilon|X \sim N(0, \Sigma)$ where $\Sigma \neq \sigma^2 I$. Here we multiply on the left by the matrix $\Sigma^{-\frac{1}{2}}$ where we know $\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}} = I$. Thus we have

$$\Sigma^{-\frac{1}{2}}Y = \Sigma^{-\frac{1}{2}}X\beta + \Sigma^{-\frac{1}{2}}\epsilon$$

which can be estimated using OLS. Now our desired assumption is satisfied as

$$\begin{aligned}\text{Cov}(\Sigma^{-\frac{1}{2}}\epsilon) &= \Sigma^{\frac{1}{2}}\text{Cov}(\epsilon)\Sigma^{-\frac{1}{2}} \\ &= \Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}} \\ &= I.\end{aligned}$$

Remark: It is important to note that the above calculations involving $(X'X)^{-1}$ can only be carried out when $X'X$ is invertible. In this case, as long as $n > p$ (and as previously assumed there is no linear dependence between predictors), this matrix is guaranteed to be invertible. The $n < p$ case will be discussed later.

2.3 Ridge Estimator

One way of overcoming multicollinearity or an ill-conditioned $X'X$ matrix, is by performing ridge regression. The model

$$Y = X\beta + \epsilon$$

is still used. However, we estimate our vector of parameters as

$$\begin{aligned}\hat{\beta}(k) &= \arg\min_{\beta}\{\|Y - X\beta\|^2 + k\|\beta\|^2\} \\ &= (X'X + kI_p)^{-1}X'Y \\ &= WX'Y\end{aligned}$$

with $k \geq 0$. k is referred to as the ridge tuning parameter. Observe that this is simply the OLS objective with the added L_2 norm penalty term $k\|\beta\|^2$. This can also be expressed in terms of the OLS extimator. That is

$$\begin{aligned}\hat{\beta}(k) &= [I_p + k(X'X)^{-1}]^{-1}\hat{\beta}_0 \\ &= Z\hat{\beta}_{LS} \qquad (2.2)\end{aligned}$$

$$(2.3)$$

This form can be used to explore properties of $\hat{\beta}(k)$ [Hoerl and Kennard, 1990]. First, $\hat{\beta}(k)$ is a biased estimator. This is clear as

$$\begin{aligned}E[\hat{\beta}(k)] &= E[Z\hat{\beta}_{LS}] \\ &= ZE[\hat{\beta}_{LS}] \\ &= Z\beta\end{aligned}$$

where Z has been previously defined in (1). Another property is

$$\begin{aligned}\mathrm{Cov}[\hat{\beta}(k)] &= \mathrm{Cov}(Z\hat{\beta}_0) \\ &= \mathrm{Cov}(Z(X'X)^{-1}X'Y) \\ &= Z(X'X)^{-1}X'\mathrm{Cov}(Y)X(X'X)^{-1}Z' \\ &= \sigma^2 Z(X'X)^{-1}Z'\end{aligned}$$

where again Z has been previously defined in (2).

2.4 Semi-parametric Model and Difference Based Technique

For the semi-parametric model, Y, X and ϵ can be as previously described in (1). Also let $t = (t_1, t_2, ..., t_n)'$ be a vector of explanatory variables which have bounded support, say the unit interval, and have been reordered so that $\leq t_1 \leq t_2 \leq \cdots \leq t_n \leq 1$. Also assume that t has a smooth regression relationship with X. Thus X is a function of t with bounded first derivative. Let f be a function with bounded first derivative. Under these assumptions, we can define the semi-parametric model as

$$y_i = x_i\beta + f(t_i) + \epsilon_i \tag{2.4}$$

for $i = 1, 2, ..., n$. By using a difference-based approach, the effect of f can be removed and the $\beta_i's$ can be estimated[Yatchew, 2003]. For example, a first order difference is

$$\{y_i - y_{i-1)}\} = \beta\{x_i - x_{i-1}\} + \{f[t_i] - f[t_{i-1}]\} + \{\epsilon_i - \epsilon_{i-1}\}.$$

But since the first derivative of f is bounded,

$$\{f[t_i] - f[t_{i-1}]\} \to 0 \text{ as } i \to \infty$$

and thus

$$y_i - y_{i-1} \cong \beta(x_i - x_{i-1}) + (\epsilon_i - \epsilon_{i-1})$$

can be used to estimate β using the aforementioned ordinary least squares method. To see more clearly why a bounded first derivative is sufficient, assume that the t_is are equally spaced on the unit interval and that $f' \leq L$. Thus by the mean value theorem, for some $t_i^* \in [t_{i-1}, t_i]$ we have

$$f(t_i) - f(t_{i-1}) = f'(t_i^*)(t_i - t_{i-1}) \leq \frac{L}{n} \implies y_i - y_{i-1} = \epsilon_i - \epsilon_{i-1} + O(\tfrac{1}{n}).$$

In the case that the t_is have a density function which is bounded away from 0, then $t_i - t_{i-1} \cong O_P\left(\frac{1}{n}\right)$ and $y_i - y_{i-1} \cong \epsilon_i - \epsilon_{i-1} + O_P\left(\frac{1}{n}\right)$. Thus the reordering and bounded first derivative of f are sufficient to remove the nonparametric effect[Yatchew, 2003].

For higher order differences, a differencing matrix is used and is defined as follows. For an mth order differencing sequence, the matrix D has dimension $(n-m) \times n$ and is of the form

$$D = \begin{pmatrix} d_0 & d_1 & d_2 & \cdots & d_m & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & d_0 & d_1 & d_2 & \cdots & d_m & 0 & \cdots & \cdots & 0 \\ \vdots & & \ddots & & & & & & & \vdots \\ \vdots & & & \ddots & & & & & & \vdots \\ 0 & \cdots & \cdots & 0 & d_0 & d_1 & d_2 & \cdots & d_m & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & d_0 & d_1 & d_2 & \cdots & d_m \end{pmatrix}$$

where $d_0, d_1, ..., d_m$ satisfy

$$\sum_{j=0}^{m} d_j = 0 \text{ and } \sum_{j=0}^{m} d_j^2 = 1.$$

Observe that the first constraint guarantees removal of the non-parametric effect in large samples and the second constraint ensures the same variance in the new model as the original model [Yatchew, 2000]. Thus the mth order difference is

$$DY = DX\beta + Df(t) + D\epsilon$$
$$\cong DX\beta + D\epsilon$$

as the effect of f has been removed [Lokshin, 2006]. This new model can be used to find estimators of β with desirable properties. Some more insight in to this technique can be gained by examining optimal differencing weights. The following table contains optimal differencing sequences for $m = 1, 2, ..., 10$ in the sense that they achieve minimum asymptotic variance [Hall et al., 1990]. These weights do not have analytic expressions but can be approximated (in this case to four decimal places).

Table 2.1: Optimal Differencing Weights

m	d_0	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
1	0.7071	-0.7071									
2	0.8090	0.5000	-0.3090								
3	0.1942	0.2809	0.3832	-0.8582							
4	0.2708	-0.0142	0.6909	-0.4858	-0.4617						
5	0.9064	-0.2600	-0.2167	-0.1774	-0.1420	-0.1103					
6	0.9200	-0.2238	-0.1925	-0.1635	-0.1369	-0.1126	-0.0906				
7	0.9302	-0.1965	-0.1728	-0.1506	-0.1299	-0.1107	-0.0930	-0.0768			
8	0.9381	-0.1751	-0.1565	-0.1389	-0.1224	-0.1069	-0.0925	-0.0791	-0.0666		
9	0.9443	-0.1578	-0.1429	-0.1287	-0.1152	-0.1025	-0.0905	-0.0792	-0.0687	-0.0588	
10	0.9494	-0.1437	-0.1314	-0.1197	-0.1085	-0.0978	-0.0877	-0.0782	-0.0691	-0.0606	-0.0527

For this semiparametric model, we can define the difference based OLS estimator

$$\hat{\beta}_D = (U'U)^{-1}U'W$$

where $U = DX$ and $W = DY$. Note that after differencing, the errors are no longer uncorrelated. However ordinary least squares can still be used in lieu of generalized least squares in order to obtain the following reusult. It has been shown that as $m \to \infty$ and $\frac{m}{n} \to \infty$ where

the columns of X and t are independent and several other assumptions are satisfied, that

$$\sqrt{n}(\hat{\beta}_1 - \beta) \xrightarrow{L} N(0, \sigma^2 \Sigma_X^{-1}) \tag{2.5}$$

where Σ_X^{-1} is the non-singular covariance matrix of X([Wang et al., 2011]). This result does not require the first derivative of f to be bounded but actually loosens the assumption. For this result we can assume that $f \in \Lambda^\alpha(M)$ where

$$\Lambda^\alpha(M) = \{g : \forall x, y \in [0,1], k = 0, ..., \lfloor\alpha\rfloor - 1, |g^{(k)}(x)| \leq M, \text{ and } |g^{(\lfloor\alpha\rfloor)}(x) - g^{(\lfloor\alpha\rfloor)}(y)| \leq M|x-y|^{\alpha'}\}$$

where $\lfloor\alpha\rfloor$ is the greatest integer less than α and $\alpha' = \alpha - \lfloor\alpha\rfloor$ for $\alpha > 0$. We must also assume that for $k = 1, 2, ..., m$ with $c_k = \sum_{i=0}^{m-k} d_i d_{i+k}$ that

$$\sum_{k=0}^{m} c_k^2 = O(m^{-1}) \text{ as } m \to \infty.$$

2.5 An $n < p$ (or High Dimension) Application

In most cases, the sample size n will be greater than the number of predictors p. However there are cases, such as microarray data, which exhibit the property that $n < p$. For this application, there may only be a limited number of participants in a study, say 70, but possibly 10000 genes which are analyzed. Clearly we have a situation where n is much less than p. In cases of high dimension, ordinary least squares estimation is out of the question, as the $X'X$ matrix will be singular. Therefore, we must appeal to other methods.

Dimension reduction is one possibility. Methods such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) exist for determining the amount of information lost when comparing models [Akaike, 1974], [Akaike, 1977], [Schwarz, 1978]. If the dimension of the data can be reduced so that $n > p$, then the problem is solved and OLS can be used. If sufficient data reduction is not possible, then other methods should be used. Ridge regression can be implemented as the penalty term will ensure the existence $(X'X + kI_p)^{-1}$.

2.5.1 Variable Selection

Variable selection is an important issue when building statistical models. This is especially true when $n < p$ since reducing the number of variables makes the analysis much easier to carry out. For microarray data, where there the number of genes being studied to predict an outcome such as cancer, variable selection is critical. A thresholding technique can be used to carry out this variable selection.

Any of the above techniques will yield values of $\hat{\beta}_i$ that are greater than 0 for all i. However, that does not mean that they should all be included in the model. In fact, very few of these should actually be included. To determine which predictors are included, a threshold is determined. This is a value for which any coefficient which is less than the thresholding value is determined to be zero. That is, the true value of β is 0 even though $\hat{\beta}_i \neq 0$. More explicitly, there must exist a constant c such that for all $\hat{\beta}_i < c$ it can be determined that $\beta_i = 0$ and can therefore be excluded from the model. Likewise if $\hat{\beta}_i \geq c$ it will be included in the model. Thus the thresholding technique can be used to define a new vector of predictors $\hat{\beta}^*$ where

$$\hat{\beta}_i^* = \begin{cases} \hat{\beta}_i & \text{if } \hat{\beta}_i \geq c \\ 0 & \text{if } \hat{\beta}_i < c \end{cases}$$

for all i.

Chapter 3

Adaptive Thresholding in Approximate Factor Models

Let

$$\mathbf{B} = (\mathbf{b_1}, ..., \mathbf{b_p})', \mathbf{y_t} = (y_{1t}, ..., y_{pt})', \mathbf{u_t} = (u_{1t}, ..., u_{pt})'$$

then the factor model previously described can be expressed in the more compact form

$$\mathbf{y_t} = \mathbf{B}\mathbf{f_t} + \mathbf{u_t}$$

where $E(\mathbf{u_t}|\mathbf{f_t}) = 0$. Unlike other models which assume independent errors, this model expects some correlation between error components. The remainder of this paper will focus on estimation of these errors. For notational purposes, let $\| \mathbf{A}\|_F$, $\|\mathbf{A}\|_\infty$, and $\|\mathbf{A}\|$ represent the Frobenius norm, elementwise norm, and operator norm of a matrix $\mathbf{A}$ respectively. Also let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ be the minimum and maximum eigenvalues of a matrix $\mathbf{A}$ respectively. Lastly note that $\| \mathbf{A}\|_F = \mathrm{tr}^{1/2}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\|_\infty = \max_{i,j} |A_{ij}|$, and $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$.

3.1 The Adaptive Thresholding Method

For this model, it is important to estimate the error covariance matrix $\mathbf{\Sigma_u} = \text{Cov}(\mathbf{u_t}) = (\sigma_{ij})_{p\times p}$ for each t. This is because many factors may be correlated and it is important to understand how they relate to each other. For instance, it is important to understand how housing prices shift in relation to one another. Also in economic and financial applications, it is possible for $p > T$. This leads to a covariance matrix which will not be invertible. Thus a method is needed that leads to an invertible matrix. Adaptive thresholding provides such a method.

For values of p close to or larger than T, it is important for $\mathbf{\Sigma_u}$ to be sparse. This allows for effective estimation. In practice this assumption is reasonable. Although some factors are expected to be correlated, many will also be completely uncorrelated. Thus it is fine to assume that many off-diagonal elements of $\mathbf{\Sigma_u}$ are 0. We impose this assumption in the following way. Define

$$m_T = \max_{i\leq p} \sum_{j\leq p} I(\sigma_{ij} \neq 0).$$

Thus m_T can be bounded by assuming

$$m_T^2 = o\left(\frac{T}{K^2 \log p}\right).$$

We construct the residual covariance matrix by first estimating

$$\hat{\mathbf{\Sigma}}_\mathbf{u} = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_t \hat{u}_t = (\hat{\sigma}_{ij})_{p\times p}$$

where $\hat{u}_{it} = y_{it} - \hat{\mathbf{b}}_\mathbf{i}'\mathbf{f_t}$ with $\hat{\mathbf{b}}_\mathbf{i}$ being the OLS estimator of $\mathbf{b_i}$. Once $\hat{\mathbf{\Sigma}}_\mathbf{u}$ is calculated, the thresholding estimator of [Fan et al., 2011] $\hat{\mathbf{\Sigma}}_\mathbf{u}^\tau$ can be defined in the following way. Let

$$\hat{\mathbf{\Sigma}}_\mathbf{u}^\tau = \hat{\sigma}_{ij}^\tau = \hat{\sigma}_{ij} I(|\hat{\sigma}_{ij}| \geq \sqrt{\hat{\theta}_{ij}}\omega_T) \text{ where } \hat{\theta}_{ij} = \frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{it}\hat{u}_{jt} - \hat{\sigma}_{ij})^2. \tag{3.1}$$

where only ω_T remains to be defined.

3.2 Assumption 3.1

(i) $(\mathbf{u_1}, ..., \mathbf{u_T})$ are independent and identically distributed with mean zero vector and covariance matrix Σ_u.

(ii) There exist constants $c_1, c_2 > 0$ such that $c_1 < \lambda_{\min}(\Sigma_u) < \lambda_{\max}(\Sigma_u) < c_2$, and $c_1 < \mathrm{Var}(u_{it}u_{jt}) < c_2$ for all $i \leq p, j \leq p$.

(iii) There exist $r > 0$ and $b > 0$, such that for any $s > 0$ and $i \leq p$,

$$P(|u_{it}| > s) \leq \exp(-(s/b)^r).$$

Also, assume there exists a positive sequence a_T such that

$$\max_{i \leq p} \frac{1}{T} \sum_{t=1}^{T} |u_{it} - \hat{u_{it}}|^2 = O_p(a_T^2).$$

3.3 Theorem 3.1

Let $\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau}$ be as defined in (3.1) and

$$\omega_T = C \max\left\{\frac{\log p}{T}, a_T\right\}$$

for some $C > 0$. Also, let $\max_{i,t} |u_{it} - \hat{u}_{it}| = o_p(1)$, $a_T = o_p(1)$, and $(\log p)^{4/r-1} = o(T)$. Then under our assumptions,

(i)

$$\|\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau} - \mathbf{\Sigma}_u\| = O_p(m_T \omega_T)$$

(ii) $\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau}$ is positive definite and

$$\|(\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau})^{-1} - \mathbf{\Sigma}_u^{-1}\| = O_p(m_T \omega_T).$$

(ii) is especially significant because without thresholding, the covariance matrix of u_{ij} is singular when using the usual methods when $p > T$.

Chapter 4

Simulation Study Design

The above adaptive thresholding technique for estimating the error covariance matrix in approximate factor models developed by ([Fan et al., 2011]) was studied. To do so, a covariance matrix needed to be simulated and then a sample covariance matrix estimated. The difference between the two matrices was quantified to give a measurement of the effectiveness of the method. The details of the simulation are as follows.

First the size of the covariance matrix needed to be determined. In one set of simulations, T was set equal to $2p$ as this method allows both T and p to grow in tandem. $p = 20, 50, 100$ were tested. For the next set, T was set equal to $0.5p$ and the same values of p were used. Assigning these values of p and T allows the affect of increasing T to be analyzed. The expectation was that this method would improve as T was increased. That is, the difference between the known covariance matrix and the estimated covariance matrix should decrease as p and T increased. After explaining the steps of the simulation and estimation steps, the meaning of the term difference will be made clear. Before that, it is important to understand how the matrices were constructed.

For each repetition a covariance matrix was simulated by generating a sparse, symmetric, positive definite matrix Σ_u. This is the error covariance matrix which the adaptive thresholding method would be used to estimate. A residual vector was generated using the multivariate normal distribution with mean zero vector and covariance matrix Σ_u. f was generated from a multivariate normal distribution with mean zero vector and covariance matrix I. b was generated

by constructing each row by sampling from a p-dimensional vector containing 4 samples from a standard normal and the rest zeros. Once these values were generated, y was calculated using (5).

The next step was to calculate $\hat{\Sigma}_u^\tau$. First, $\hat{b}$ was estimated using ordinary least squares. This was done by using f and y to solve the normal equations. Once $\hat{b}$ was estimated, $y_{it} - \hat{b}_i f_t$ could be solved to yield $\hat{u}$. Then $\hat{\Sigma}_u^\tau$ was estimated using the adaptive thresholding technique previously described. θ was calculated as described above and $\omega_T = C\sqrt{\log p/T}$ was used where $C = 1, 2, ..., 5$ were all used. This C value is not restricted to be an integer, but this range of numbers was chosen to provide an overview of the behavior of this technique as C was increased.

The last step was determining a measure of accuracy of $\hat{\Sigma}_u^\tau$. To judge the validity of the model, $\|\hat{\Sigma}_u^\tau - \Sigma_u\|$ was calculated. For demonstration purposes $\|\hat{\Sigma}_u^\tau - \Sigma_u\|_F$ and $\|\hat{\Sigma}_u^\tau - \Sigma_u\|_\infty$ were also calculated. This process was repeated 100 times for each value of p and C. The mean of the norms of the 100 differences were calculated and recorded in the following tables.

Chapter 5

Simulation Results, Conclusions and Discussion

5.1 Results

The reults for both cases ($T = 2p$ and $T = 0.5p$) are tabulated below. The difference between $\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau}$ and $\mathbf{\Sigma}_u$ have been calculated under the three norms previously defined in section 3. They are organized with increasing dimension (and sample size) ascending from left to right. This should make it easier to follow the pattern of performance as p increases. Plots of these results are provided in appendix b if a more visual representation is desired.

Table 5.1: Difference between $\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau}$ and $\mathbf{\Sigma}_{\mathbf{u}}$ when $T = 2p$

Mean of 100 simulations				
		p=20	p=50	p=100
$\omega_T = \log p/T$	$\|\cdot\|$	2.312202	1.752818	1.339495
	$\|\cdot\|_\infty$	0.9309726	0.6511156	0.4852132
	$\|\cdot\|_F$	4.06545	4.625193	5.226393
$\omega_T = 2\log p/T$	$\|\cdot\|$	3.085599	2.073209	1.475796
	$\|\cdot\|_\infty$	1.230345	0.69867	0.5222568
	$\|\cdot\|_F$	3.990561	2.994523	2.446076
$\omega_T = 3\log p/T$	$\|\cdot\|$	4.513952	3.099808	2.272725
	$\|\cdot\|_\infty$	1.932484	1.575513	0.7242664
	$\|\cdot\|_F$	6.93451	4.322454	3.157791
$\omega_T = 4\log p/T$	$\|\cdot\|$	5.00246	4.543812	3.255887
	$\|\cdot\|_\infty$	1.951687	1.976027	1.684817
	$\|\cdot\|_F$	7.870693	10.26875	4.750544

Table 5.2: Difference between $\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau}$ and $\mathbf{\Sigma}_{\mathbf{u}}$ when $T = 0.5p$

Mean of 100 simulations				
		p=20	p=50	p=100
$\omega_T = \log p/T$	$\|\cdot\|$	5.041122	3.591173	2.827728
	$\|\cdot\|_\infty$	1.953074	1.368366	1.002036
	$\|\cdot\|_F$	7.97274	9.658299	11.06939
$\omega_T = 2\log p/T$	$\|\cdot\|$	4.858307	4.471217	3.408291
	$\|\cdot\|_\infty$	1.957913	1.95218	1.896769
	$\|\cdot\|_F$	7.839591	9.387605	6.87009
$\omega_T = 3\log p/T$	$\|\cdot\|$	4.815195	4.990087	4.782022
	$\|\cdot\|_\infty$	1.954222	1.980297	1.988614
	$\|\cdot\|_F$	7.769241	11.51958	15.211
$\omega_T = 4\log p/T$	$\|\cdot\|$	4.832899	4.813586	4.993023
	$\|\cdot\|_\infty$	1.95907	1.98033	1.989294
	$\|\cdot\|_F$	7.826615	11.44008	15.81804

5.2 Conclusions

5.2.1 $T = 2p$

The results of this study are very interesting. The first aspect to evaluate is the performance as p increased. When using the operator and infinity norms, $\|\hat{\Sigma}_u^\tau - \Sigma_u\|$ decreased with increasing p for every value of C. However, if measured using the Frobenius norm, this was not always the case. For $C = 1$, the performance actually declined with increasing p. For $C = 2, 3$, performance improved with increasing p when evaluated using the Frobenius norm. For $C = 4$, the behavior of the estimator was erratic. Performance sharply declined from $p = 20$ to $p = 50$. However from $p = 50$ to 100, performance improved again. Thus if $p = 50$ had not been tested, it would have appeared that performance would simply improve with increasing p. Mistakes in data entry or a statistical anomaly can be ruled out as this simulation was run several times for quality assurance.

In general these results are reassuring as to the efficacy of this method. The samples sizes used were relatively small compared to the sample sizes seen in practice. The shown improvement in both the operator in infinity norms show that this method will improve with increasing sample size. Some exploratory simulations with larger sample sizes also indicated that as samples became large, even the Frobenius norm would show measurable improvement.

5.2.2 $T = 0.5p$

For this set of simulations it is important to note that the sample size is much smaller than in the previous section. This means that if the estimator truly does improve with increasing sample size, it will be more difficult to observfe such differences. With that being said, it is still possible to observe a trend for this case.

$C = 4$ was not a poor parameterization regardless of which norm is used for evaluation. Performance declined as p increased. $C = 1, 2, 3$ showed the desired results for the infinity and operator norms. However, only $C = 2$ showed any promise for the Frobenius norm. For this parameterization, the measured difference followed the same pattern as when $C = 4$ for the previous case. Performance declined and then improved as p increased. It can be noted

that exploratory simulations with low repetitions indicated that the difference would continue to shrink under the Frobenius norm as p continued to increase for $C = 2$.

The results of this case are very promising. Taking into consideration that only 4 parameterizations were tested and the desired performance observed for most cases, this validates the use of the adaptive thresholding method. Although better performance under the Frobenius norm would be even more convincing, it is highly likely that a paremeterization which optimizes this technique can be found for that norm.

5.3 Further Discussion

As previously stated, C is not restricted to be an integer. Thus there may be some non-integer value which outperforms $C = 1$. Also, this paramterization may only be optimal for the given data set. Data sets with different characteristics might lead to a different optimal choice for C. That being said, in this case it is safe to conclude that if $C = 1$ is not the best value, it is probably close. Due to the behavior of the estimator for smaller and larger values of C, there appears to be a trend.

The greatest cause of concern with respect to these results is the erratic behavior of the estimator when the Frobenius-norm is used for evaluation. Both cases had parameterizations where the results were not clear. The idea that the estimator improved with growing p was based on the fact that performance was measured for 3 values of p. It is possible, that for each value of C, sample sizes could be found which provide the same bizarre behavior as this example. However this is probably unlikely. More experiments with sample sizes in the range used in this experiment would make for a convincing argument if the same trends were observed.

Another interesting aspect of this method is its behavior when estimation techniques other than ordinary least squares are applied. A ridge estimator may be preferable in certain settings. Certainly for the $p > T$ case ordinary least squares is probably not the best method. A ridge estimator would be better suited for this case as it yields a unique solution when ordinary least squares doew not. For that reason, it is important to understand how this adaptive thresholding technique will perform when a ridge estimator has been used to estimate $\hat{\mathbf{b}}$. It is

likely that this method will be improved by introducing a ridge estimator much like the difference based technique from section 2.4 was improved in that way. This method may perform better or worse in general than ordinary least squares. It is also possible that optimal values of C may vary depending on the method of estimation of $\hat{\mathbf{b}}$.

Appendices

Appendix A Proofs

Lemmas A.1, A.2, A.3, and A.4 are used to prove Theorem 3.1. Thus they are stated and proven first. They are followed by the proof of Theorem 3.1. In the following we consider the operator norm $\|\mathbf{A}\|^2 = \lambda_{\max}(\mathbf{A}'\mathbf{A})$. All results proven by [Fan et al., 2011].

A.1 Lemma A.1

Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric, positive semi-definite matrices, and $\lambda_{\min}(\mathbf{B}) > c_T$ for a sequence $c_T > 0$. If $\|\mathbf{A} - \mathbf{B}\| = o_p(c_T)$, then $\lambda_{\min}(\mathbf{A}) > c_T/2$ and

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| = O_p(c_T^{-2})\|\mathbf{A} - \mathbf{B}\|.$$

Proof. Suppose both $\mathbf{A}$ and $\mathbf{B}$ are $m \times m$. For any $\mathbf{v} \in \mathbb{R}^m$ such that $\|v\| = 1$, $|\mathbf{v}'(\mathbf{A} - \mathbf{B})\mathbf{V}| \leq |\mathbf{v}\|^2$. Since $\|\mathbf{A} - \mathbf{B}\| = o_p(c_T)$, for all large T, $\mathbf{v}'\mathbf{A}\mathbf{v} \geq \mathbf{v}'\mathbf{B}\mathbf{v} - 0.5c_T \geq \lambda_{\min}(\mathbf{B}) - 0.5c_T > 0.5c_T$. Hence, $\lambda_{\min}(\mathbf{A}) \geq 0.5c_T$. Also,

$$\begin{aligned}
\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| &= \|\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}\| \\
&\geq \lambda_{\min}(\mathbf{A})^{-1}\|\mathbf{A} - \mathbf{B}\|\lambda_{\min}(\mathbf{B})^{-1} \\
&= O_p(c_T^{-2})\|\mathbf{A} - \mathbf{B}\|.
\end{aligned}$$

Note that when c_T is a constant $O_p(c_T^{-2})\|\mathbf{A} - \mathbf{B}\| = o(a_T)$.

□

A.2 Lemma A.2

Let random variables Z_1 and Z_2 satisfy the exponential-type tail condition (assumption 3.1 (iii)). That is, there exist $r_1, r_2 \in (0, 1)$ and $b_1, b_2 > 0$ such that for all $s > 0$ and for $i = 1, 2$

$$P(|Z_i| > s) \leq \exp(1 - (s/b_i)^{r_i}).$$

Then for some r_3 and b_3 both greater than 0 and any $s > 0$,

$$P(|Z_1 Z_2| > s) \leq \exp(1 - (s/b_3)^{r_3}). \tag{A.1}$$

Proof. For any $s > 0$, let $M = (s b_2^{r_1/r_2}/b_1)^{r_1/(r_1+r_2)}, b = b_1 b_2$ and $r = r_1 r_2/(r_1 + r_2)$. Thus,

$$\begin{aligned} P(|Z_1 Z_2| > s) &\leq P(M|Z_1| > s) + P(|Z_2| > M) \\ &\leq \exp(1 - (s/b_1 M)^{r_1}) + \exp(1 - (M/b_2)^{r_2}) \\ &= 2\exp(1 - (s/b)^r) \end{aligned}$$

Now let $r_3 \in (0, r)$ and $b_3 > \max\{(r_3/r)^{1/r}, (\log 2)^{1/r} b\}$, then $F(s) = (s/b)^r - (s/b_3)^{r_3}$ is increasing when $s > b_3$. Therefore, $F(s) > F(b_3) > \log 2$ when $s > b_3$. Hence when $s > b_3$,

$$P(|Z_1 Z_2| > s) \leq 2exp(1 - (s/b)^r) \leq exp(1 - (s/b_3^{r_3}))$$

and when $s \leq b_3$,

$$P(|Z_1 Z_2| > s) \leq 1 \leq \exp(1 - (s/b_3)^r).$$

□

A.3 Lemma A.3

Under assumption 3.1, $a_T = o(1)$ and $\log p = o(T)$,

(i) $\max_{i,j \leq p} \left| \frac{1}{T} \sum_{t=1}^{T} u_{it} u_{jt} - \sigma_{ij} \right| = O_p\left(\frac{\log p}{T}\right)$

(ii) $\max_{i,j \leq p} |\hat{\sigma}_{ij} - \sigma_{ij}| = O_p\left(\max\left\{\frac{\log p}{T}, a_T\right\}\right)$

Proof. (i) By assumption 2.1 (iii) and Lemma A.2, $u_{it} u_{jt}$ satisfies the exponential type tail condition. ([Merlevede et al., 2009]) showed that by Bernstein's inequality, there exist constants

$C_1, ..., C_5 > 0$ that depend only on b and r such that for any $i, j \leq p$, and $\gamma = r/4$,

$$P\left(\left|\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - \sigma_{ij}\right| \geq s\right) \leq T\exp\left(-\frac{(Ts)^{\gamma}}{C_1}\right) + \exp\left(-\frac{T^2s^2}{C_2(1+TC_3)}\right) + \exp\left(-\frac{(Ts)^2}{C_4T}\exp\left(\frac{(Ts)^{\gamma(1-\gamma)}}{C_5(\log Ts)^{\gamma}}\right)\right)$$

Bonferroni's method can be applied to yield

$$P\left(\max_{i,j\leq p}\left|\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - \sigma_{ij}\right| > s\right) \leq p^2 \max_{i,j\leq p} P\left(\left|\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - \sigma_{ij}\right| > s\right).$$

Since $(\log p)^{4/r-1} = o(T)$, as long as $s > \sqrt{(\log p)/T}$, for all large T,

$$p^2T\exp\left(-\frac{(Ts)^{\gamma}}{C_1}\right) + p^2\exp\left(-\frac{(Ts)^2}{C_4T}\exp\left(\frac{(Ts)^{r(1-r)}}{C_5(\log Ts)^{r}}\right)\right) = o(1).$$

Also, as long as $s^2T > 6C_2C_3\log p$, for all large T,

$$p^2\exp\left(-\frac{T^2s^2}{C_2(1+TC_3)}\right) = O(1)$$

which implies the desired result.

(ii) By part (i) and the triangle inequality

$$\max_{i,j\leq p}|\hat{\sigma}_{ij} - \sigma| \leq O_p\left(\frac{\log p}{T}\right) + \max_{i,j\leq p}|\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{it}\hat{u}_{jt} - u_{it}u_{jt})|.$$

It will now be shown that $\mathbf{A} \equiv \max_{i,j\leq p}|\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{it}\hat{u}_{jt} - u_{it}u_{jt})| = O_p(a_T)$. By the triangle and Caucy-Schwarz inequalities we have

$$
\begin{aligned}
A &\leq \max_{i,j\leq p} \frac{1}{T}\left|\sum_{t=1}^{T}(\hat{u}_{it}-u_{it})(\hat{u}_{jt}-u_{jt})\right| + 2\max_{i,j\leq p}\frac{1}{T}\left|\sum_{t=1}^{T}u_{it}(\hat{u}_{jt}-u_{jt})^2\right| \\
&\leq \max_{i\leq p}\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{it}-u_{it})^2 + 2\sqrt{\max_{i\leq p}\frac{1}{T}\sum_{t=1}^{T}u_{it}^2}\sqrt{\max_{i\leq p}\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{it}-u_{it})^2} \\
&\leq O_p(a_T^2) + 2\sqrt{o_p(1)\max_{i\leq p}\sigma_{ii}}\sqrt{a_T^2}.
\end{aligned}
$$

Hence the desired result follows.

□

A.4 Lemma A.4

There exist $C_1, C_2 > 0$ such that with probability approaching one,

$$
C_1 \leq \min_{i,j}\hat{\theta}_{ij} \leq \max_{i,j}\hat{\theta}_{ij} \leq C_2.
$$

Proof. (i) For any i, j we can add and subtract terms to yield

$$
\begin{aligned}
\hat{\theta}_{i,j} &= \frac{1}{T}\sum_{t}(\hat{u}_{it}\hat{u}_{jt} - \frac{1}{T}\sum_{l}(\hat{u}_{il}\hat{u}_{jl})^2 \\
&\leq \frac{2}{T}\sum_{t}(\hat{u}_{it}\hat{u}_{jt}-\hat{\sigma}_{ij})^2 + 2\max_{i,j}(\sigma_{ij} - \frac{1}{T}\sum_{l}(\hat{u}_{il}\hat{u}_{jl})^2)^2 \\
&\leq \frac{2}{T}\sum_{t}(\hat{u}_{il}\hat{u}_{jl}-\sigma_{ij})^2 + o_p(1),
\end{aligned}
$$

where $o_p(1)$ does not depend on i or j by Lemma A.3. By adding and subtracting terms we have

$$\begin{aligned}\sum_t(\hat{u}_{it}\hat{u}_{jt}-\sigma_{ij})^2 &\le 4\sum_t(\hat{u}_{it}-u_{it})^2\hat{u}_{jt}^2+4\sum_t(\hat{u}_{jt}-u_{jt})^2\hat{u}_{it}^2+2\sum_t(u_{it}u_{jt}-\sigma_{ij})^2\\ &\le 4\max_{it}|\hat{u}_{it}-u_{it}^2(2\max_j\sum_t(\hat{u}_{jt}-u_{jt})^2+3\max_j\sum_t u_{jt}^2)+2\sum_t(u_{it}u_{jt}-\sigma_{ij})^2\\ &= o_p(1)(o_p(Ta_T^2)+\max_j\sum_t u_{jt}^2)+2\sum_t(u_{it}u_{jt}-\sigma_{ij})^2.\end{aligned}$$

Since $(\mathbf{u_t})_{t\ge1}$ are independent and identically distributed random vectors with exponential tails on each component, the same arguments as those in the proof of Lemma 2 in [Cai and Liu, 2011] imply that

$$\max_{i,j}|\frac{1}{T}(u_{it}u_{jt}-\sigma_{ij})^2-\mathrm{Var}(u_{it}u_{jt})|=o_p(1),$$

and $\mathrm{Var}(u_{it}u_{jt})$ is bounded away from both zero and infinity. Therefore, $\frac{1}{T}\sum_t((u_{it}u_{jt}-\sigma_{ij})^2$ is bounded away from zero and infinity with probability approaching one. In addition, by Lemma A.3(i), with probability approaching one,

$$\max_j\frac{1}{T}\sum u_{jt}^2\le o_p(1)+\max_j\sigma_{jj}\le\max_j\sigma_{jj}.$$

To summarize, $\max_{ij}\hat{\theta}_{ij}$ is bounded away from infinity with probability approaching one.

(ii) By adding and subtracting terms, we obtain

$$\begin{aligned}\sum_t(u_{it}u_{jt}-\sigma_{ij})^2 &\le 4\sum_t(u_{it}u_{jt}-\hat{u}_{it}\hat{u}_{jt})^2+4\sum_t(\hat{u}_{it}\hat{u}_{jt}-\frac{1}{T}\sum_t\hat{u}_{il}\hat{u}_{jl})^2+4(\sigma_{ij}-\frac{1}{T}\sum_t\hat{u}_{il}\hat{u}_{jl})^2\\ &\le 8\sum_t u_{it}^2(u_{jt}-\hat{u}_{jt})^2+8\sum_t\hat{u}_{jt}^2(u_{it}-\hat{u}_{it})^2+4T\hat{\theta}_{ij}+o_p(T)\\ &\le 16\max_{it}|\hat{u}_{it}-u_{it}|^2(\max_j\sum_t(\hat{u}_{jt}u_{jt})^2+\max_j\sum_t u_{jt}^2)+4T\hat{\theta}_{ij}+o_p(T),\end{aligned}$$

where $o_p(T)$ does not depend on i, j due to Lemma A.3. As is demonstrated in part (i),

$$16 \max_{it} |\hat{u}_{it} - u_{it}|^2 (\max_j \sum_t (\hat{u}_{jt} - u_{jt})^2 + \max_j \sum_t u_{jt}^2) = o_p(T)$$

and

$$\frac{1}{T} \sum_t (u_{it} u_{jt} - \sigma_{ij})^2 \geq C$$

uniformly in i, j for some $C > 0$ with probability approaching one. This establishes the result.

□

A.5 Proof of Theorem 3.1

Proof. (i) For the operator norm, the triangle inequality still holds:

$$\|\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau} - \mathbf{\Sigma}_{\mathbf{u}}\| \leq \|\mathbf{\Sigma}_{\mathbf{u}}^{\tau} - \mathbf{\Sigma}_{\mathbf{u}}\| + \|\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau} - \mathbf{\Sigma}_{\mathbf{u}}^{\tau}\|,$$

where

$$\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau} = (\sigma_{ij}^{\tau}), \sigma_{ij}^{\tau}) = \sigma_{ij} I(|\sigma_{ij}| > \sqrt{\hat{\theta}_{ij}} \omega_T)$$

and $\omega_T = C \max(\sqrt{\log p / T}, a_T)$ for some $C > 0$. We bound $\|\mathbf{\Sigma}_{\mathbf{u}}^{\tau} - \mathbf{\Sigma}_{\mathbf{u}}\|$ and $\|\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau} - \mathbf{\Sigma}_{\mathbf{u}}^{\tau}\|$ separately.

First of all, for symmetric matrix $\mathbf{A} = (a_{ij}), \|\mathbf{A}\| \leq \max_i \sum_{j=1}^p |a_{ij}|$. Therefore we have

$$\begin{aligned} \|\mathbf{\Sigma}_{\mathbf{u}}^{\tau} - \mathbf{\Sigma}_{\mathbf{u}}\| &\leq \max_{i \leq p} \sum_{j=1}^p |\sigma_{ij}| I(|\sigma_{ij}| \leq \omega_T \hat{\theta}_{ij}^{1/2}) \\ &\leq \max_i \sum_{j : \sigma_{ij} \neq 0} \omega_T \hat{\theta}_{ij}^{1/2} = O_p(\omega_T m_T), \end{aligned}$$

where the last inequality is due to $\hat{\theta}_{ij}$ being bounded above uniformly in i, j with probability approaching one, according to Lemma A.4.

On the other hand,

$$\begin{aligned}\|\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau} - \mathbf{\Sigma}_{\mathbf{u}}^{\tau}\| \leq & \max_{i\leq p} \sum_{j=1}^{p} |\sigma_{ij}| I(|\hat{\sigma}_{ij}| \leq \omega_T \hat{\theta}_{ij}^{1/2}, |\sigma_{ij}| > \omega_T \hat{\theta}_{ij}^{1/2}) \\ & + \max_{i\leq p} \sum_{j=1}^{p} |\sigma_{ij} - \hat{\sigma}_{ij}| I(|\hat{\sigma}_{ij}| > \omega_T \hat{\theta}_{ij}^{1/2}, |\sigma_{ij}| > \omega_T \hat{\theta}_{ij}^{1/2}) \\ & + \max_{i\leq p} \sum_{j=1}^{p} |\hat{\sigma}_{ij}| I(|\sigma_{ij}| \leq \omega_T \hat{\theta}_{ij}^{1/2}, |\hat{\sigma}_{ij}| > \omega_T \hat{\theta}_{ij}^{1/2}).\end{aligned}$$

Since by Lemma A.4 $\hat{\theta}_{ij}^{1/2}$ is bounded away from both zero and infinity uniformly in i, j, all three terms on the right hand side can be bounded in a similar way as in the broof of Theorem in [Bickel and Levina, 2008], correspoding to the case $q = 0$. Therefore the details are omitted, which are available from the authors. Here we only show a key different step in the proof, which is,

$$\max_{i\leq p} \sum_{j=1}^{p} I(|\hat{\sigma}_{ij} - \sigma_{ij}| \geq (1-r)\omega_T \hat{\theta}_{ij}) = O_p(1) \tag{1}$$

for any $r \in (0, 1)$. This implies that

$$\begin{aligned}& \max_{i\leq p} \sum_{j=1}^{p} |\hat{\sigma}_{ij} - \sigma_{ij}| I(|\hat{\sigma}_{ij}| \geq \omega_T \hat{\theta}_{ij}, |\sigma_{ij}| \leq r\omega_T \hat{\theta}_{ij}) \\ & \leq O_p(\omega_T) \max_{i\leq p} \sum_{j=1}^{p} I(|\hat{\sigma}_{ij}| \geq \omega_T \hat{\theta}_{ij}, |\sigma_{ij}| \leq r\omega_T \hat{\theta}_{ij}) \\ & = O_p(\omega_T).\end{aligned}$$

To show (1), let $C_1 > 0$ be such that $P(\min_{ij} \hat{\theta}_{ij} \leq C_1) = o(1)$, whose existence is guaranteed by Lemma A.4. Since $\max_{ij} |\hat{\sigma}_{ij} - \sigma_{ij}| = O_p(\omega_T)$, for any $\epsilon, M > 0$, and sufficiently large $C > 0$,

$$
\begin{aligned}
&P(\max_{i \leq p} \sum_{j=1}^{p} I(|\hat{\sigma}_{ij} - \sigma_{ij}| \geq (1-r)\omega_t \hat{\theta}_{ij}) > M) \\
&\leq P(\max_{ij} |\hat{\sigma}_{ij} - \sigma_{ij}| \geq (1-r)\omega_t \hat{\theta}_{ij}) \\
&\leq P\left(\frac{\max_{ij} |\hat{\sigma}_{ij} - \sigma_{ij}|}{\max\{\sqrt{(\log p)/T}, a_T\}} \geq (1-r) C C_1 \right) + o(1) < \epsilon,
\end{aligned}
$$

which yields the result.

(ii) Since both $\hat{\mathbf{\Sigma}}_{\mathbf{u}}^{\tau}$ and $\mathbf{\Sigma}_{\mathbf{u}}$ are symmetric and $\lambda_{\min}(\mathbf{\Sigma}_{\mathbf{u}}) > C$ for some $C > 0$, the result follows immediately from Lemma A.1.

□

Appendix B Plots of Simulation Results

The following plots demonstrate the behavior of the adaptive thresholding technique for estimating the error covariance matrix $\mathbf{\Sigma}_u$. The first set of plots are for the case where $T = 2p$ and the second set of plots are for the case where $T = 0.5p$. Each case has two types of plots. The first type show the behavior for each value of C as p increases where $\omega_T = C\frac{\log p}{T}$ is the thresholding value. Once it is determined that a parameterization improves as p increases, that particular parameterization can be compared to other well performing parameterizations to determine the best value of C for a particular sample size.

B.1 $T = 2p$

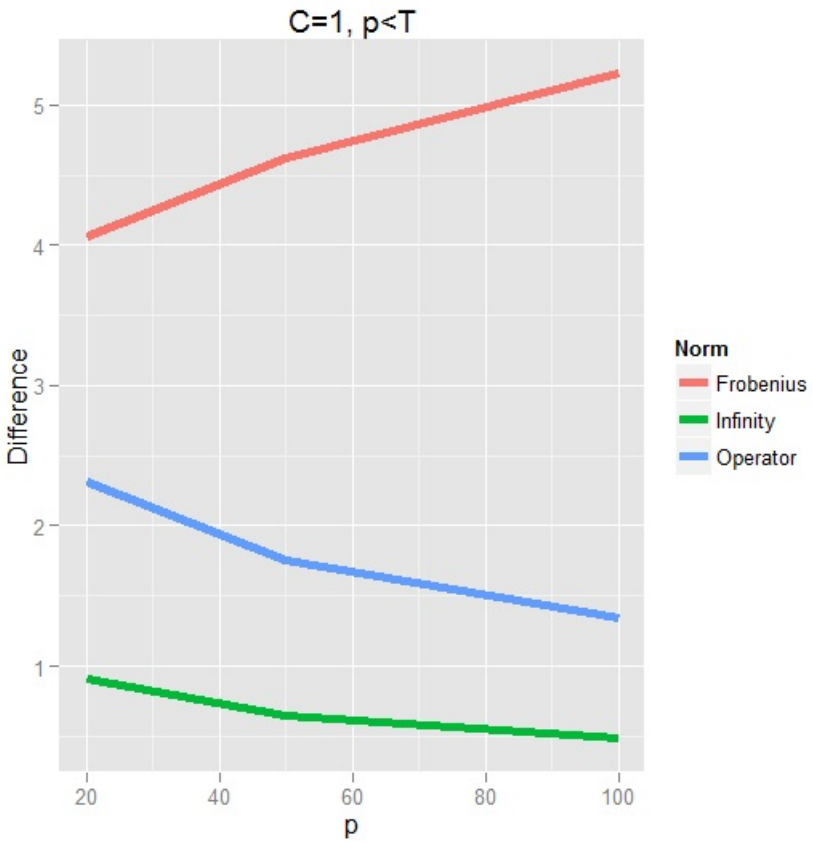

Figure 1: $\omega_T = \sqrt{\log p/T}$

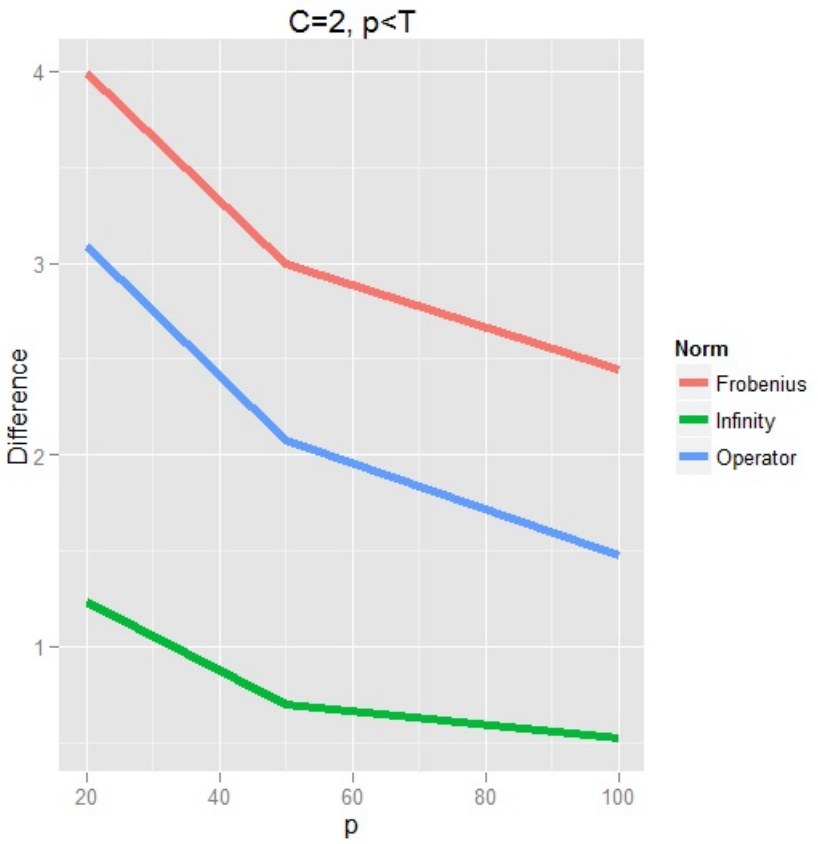

Figure 2: $\omega_T = 2\sqrt{\log p/T}$

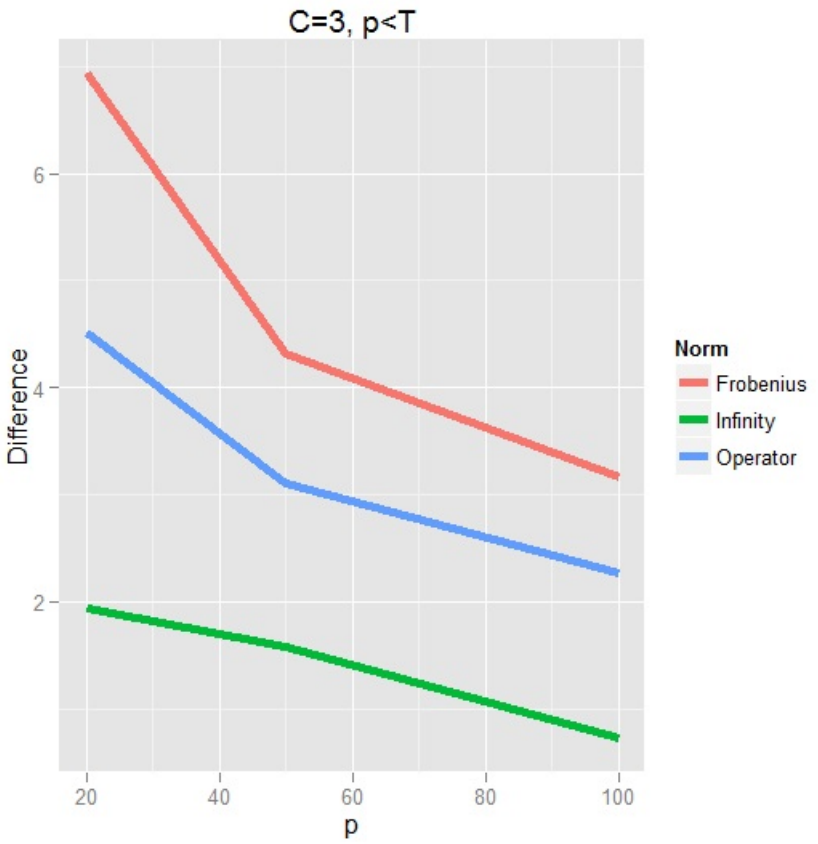

Figure 3: $\omega_T = 3\sqrt{\log p/T}$

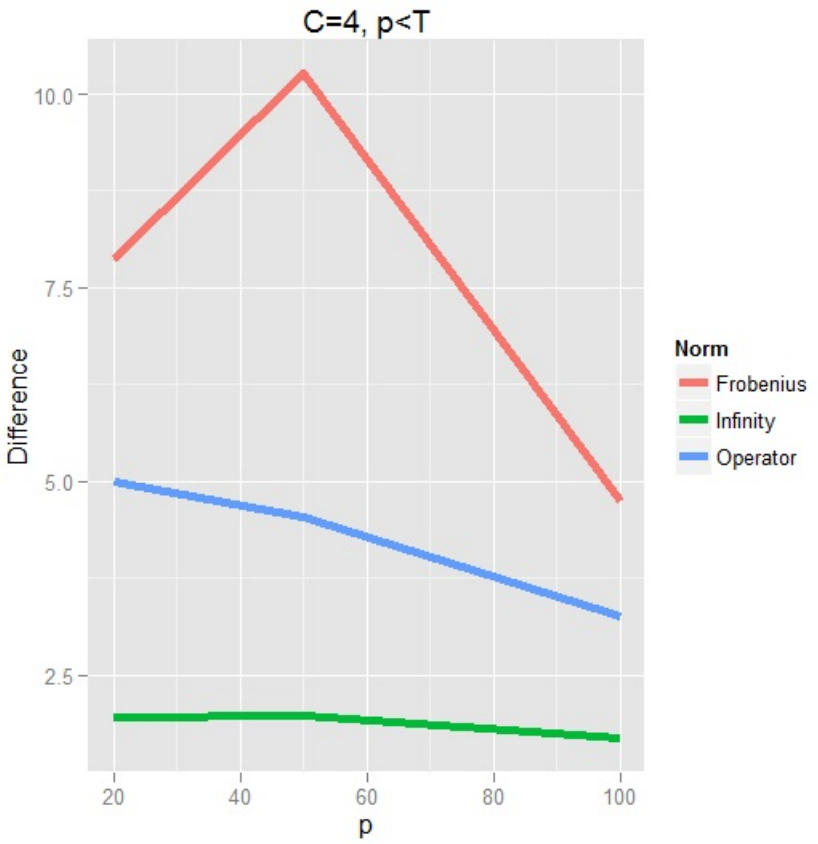

Figure 4: $\omega_T = 4\sqrt{\log p/T}$

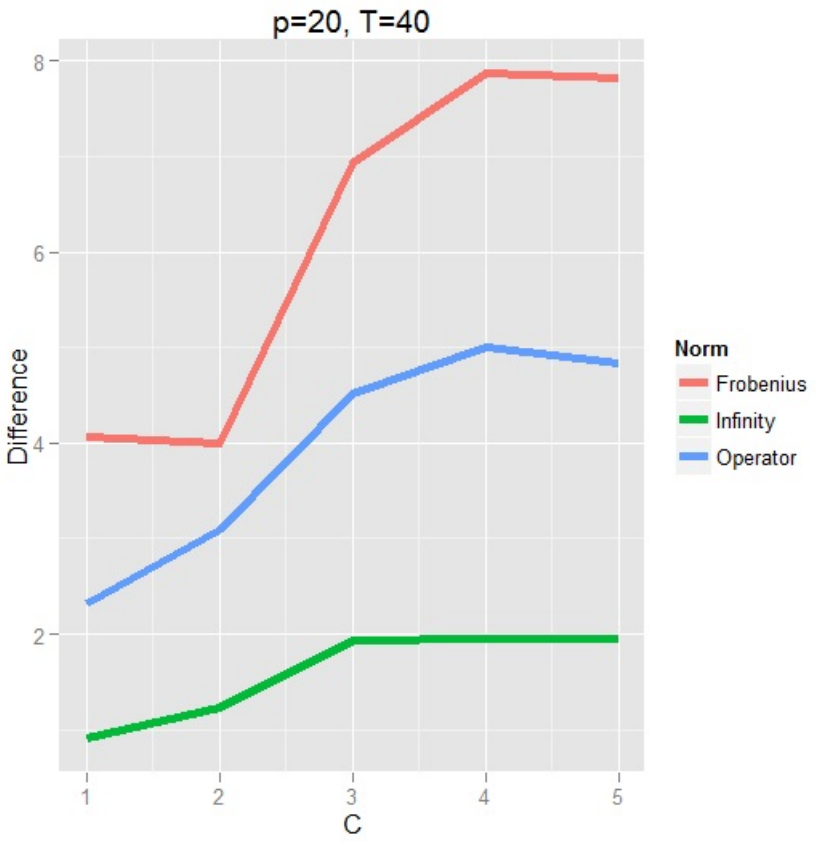

Figure 5: Performance of various C values when $p = 20$

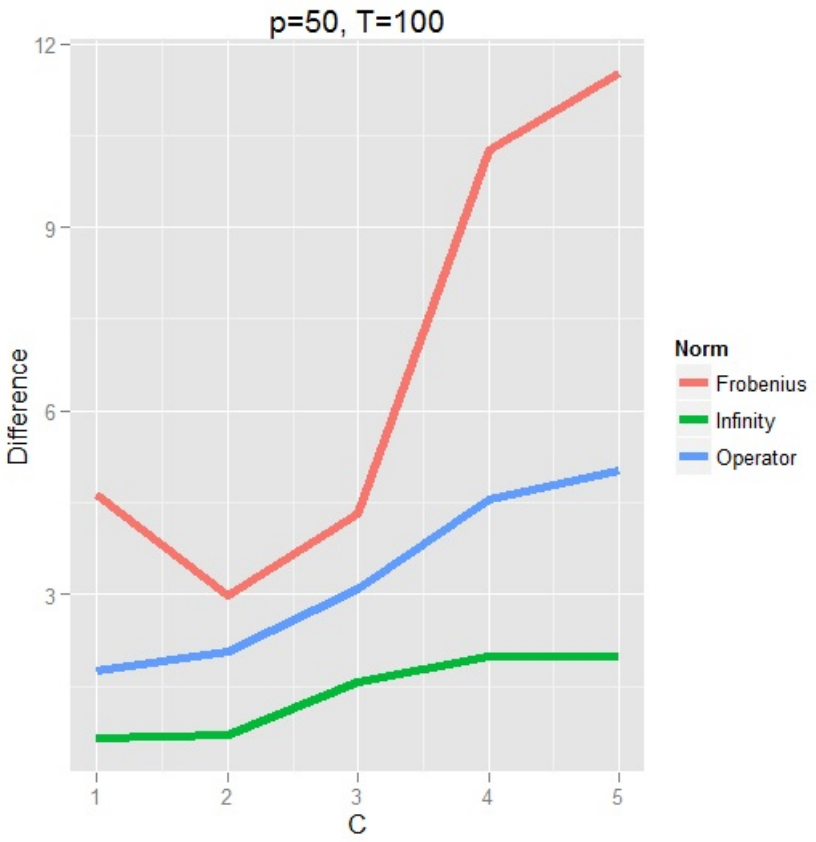

Figure 6: Performance of various C values when $p = 50$

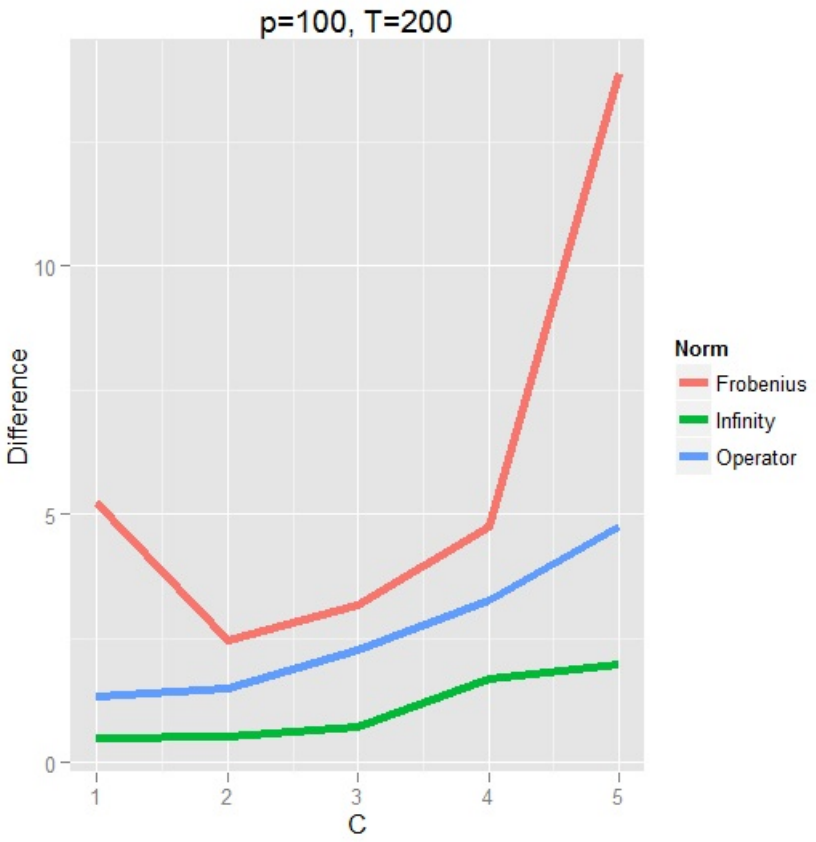

Figure 7: Performance of various C values when $p = 100$

B.2 $T = 0.5p$

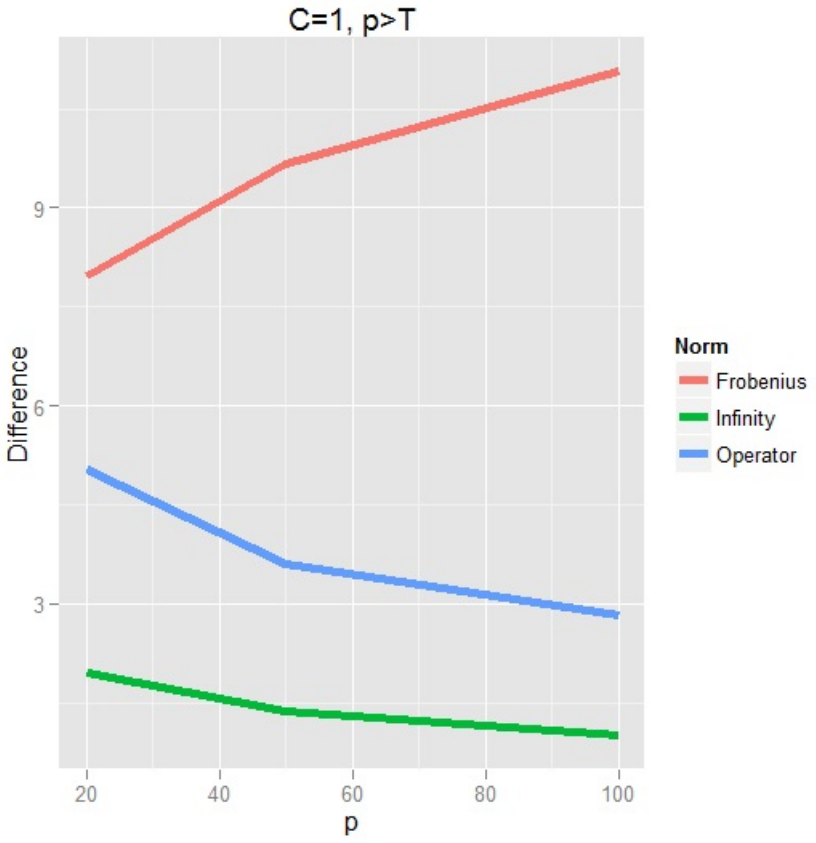

Figure 8: $\omega_T = \sqrt{\log p/T}$

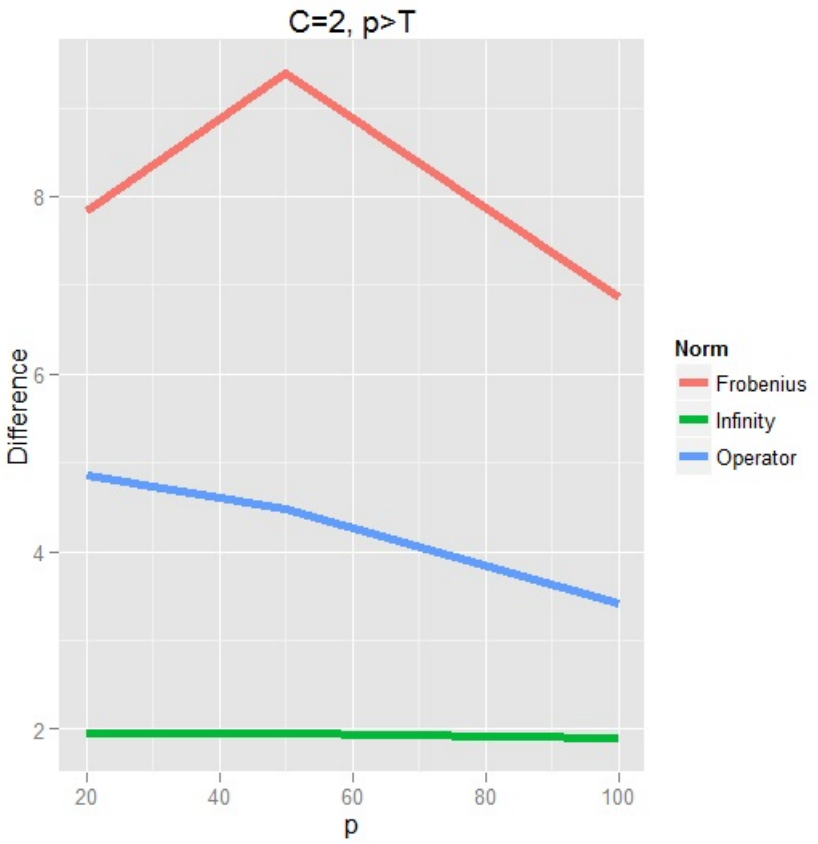

Figure 9: $\omega_T = 2\sqrt{\log p/T}$

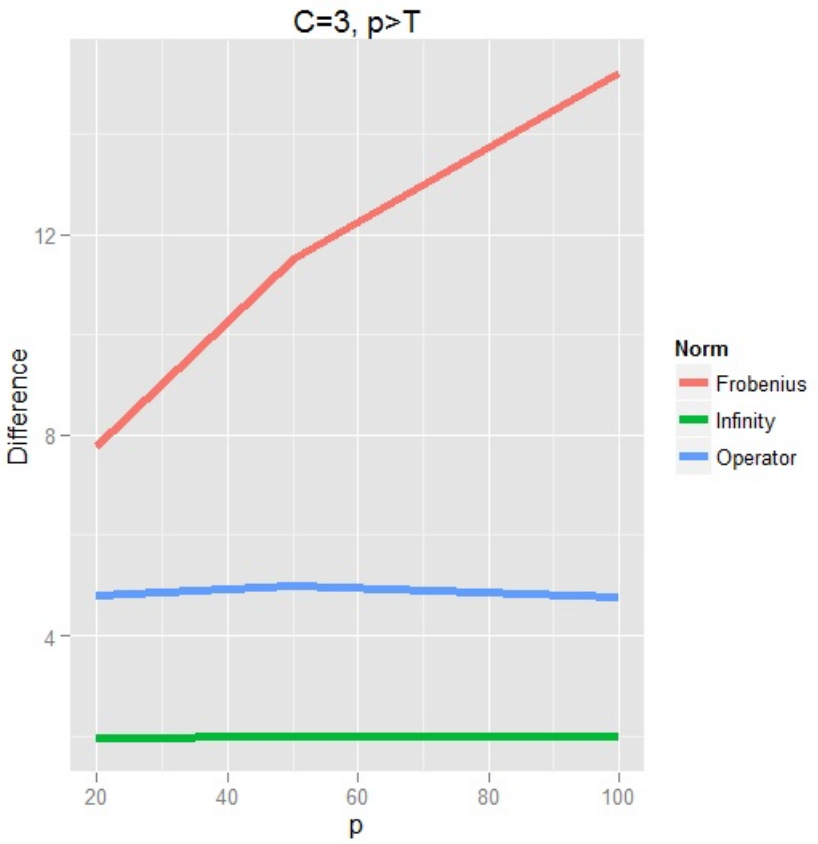

Figure 10: $\omega_T = 3\sqrt{\log p/T}$

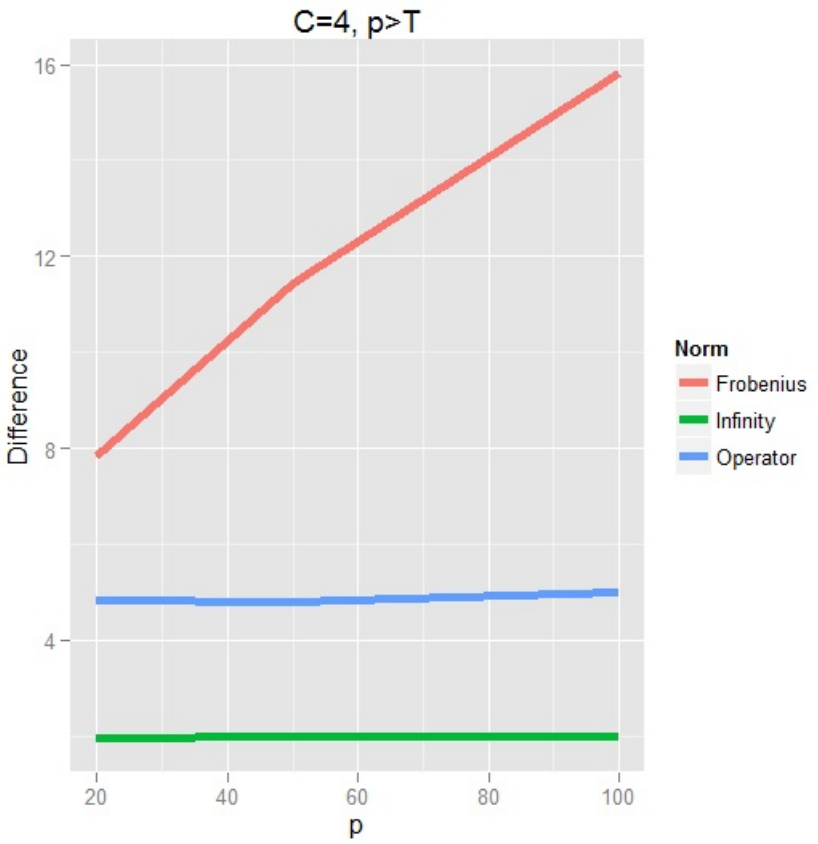

Figure 11: $\omega_T = 4\sqrt{\log p/T}$

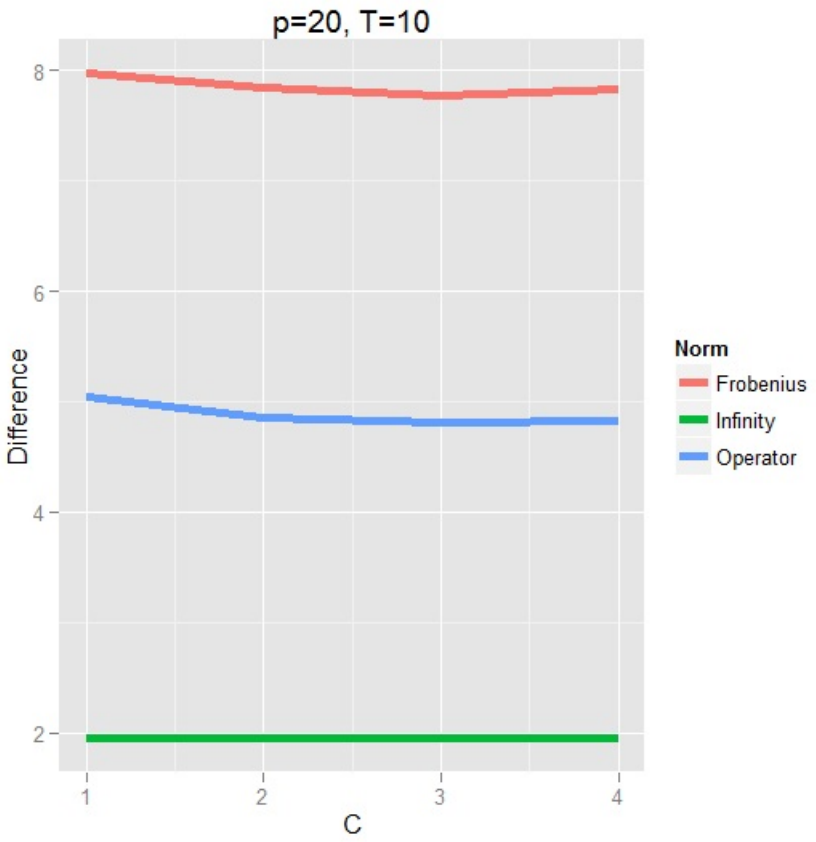

Figure 12: Performance of various C values when $p = 20$

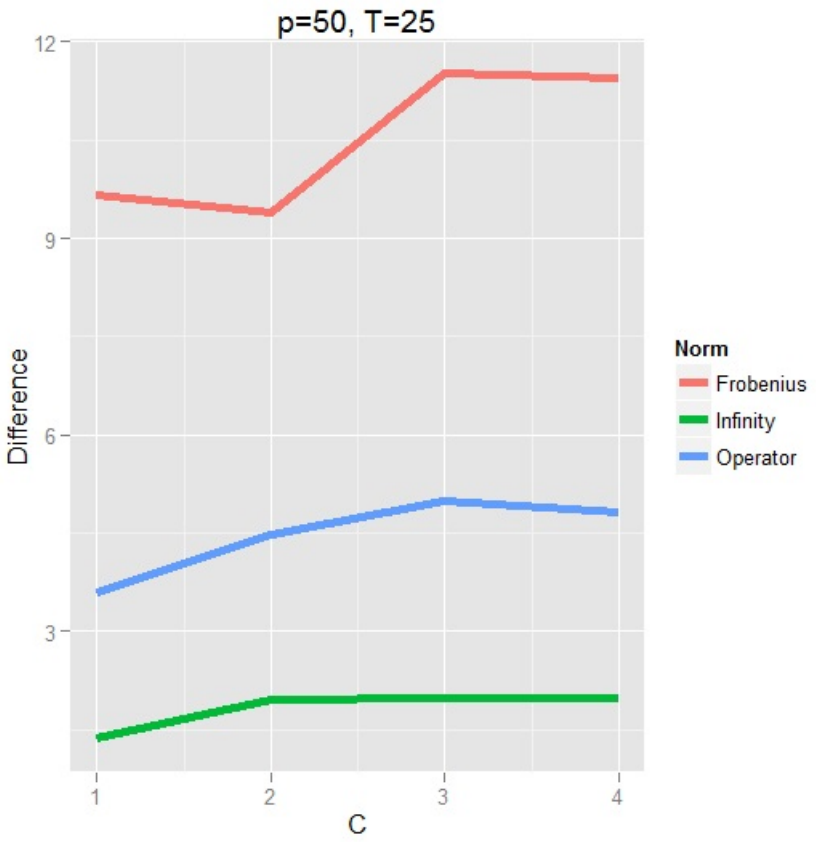

Figure 13: Performance of various C values when $p = 50$

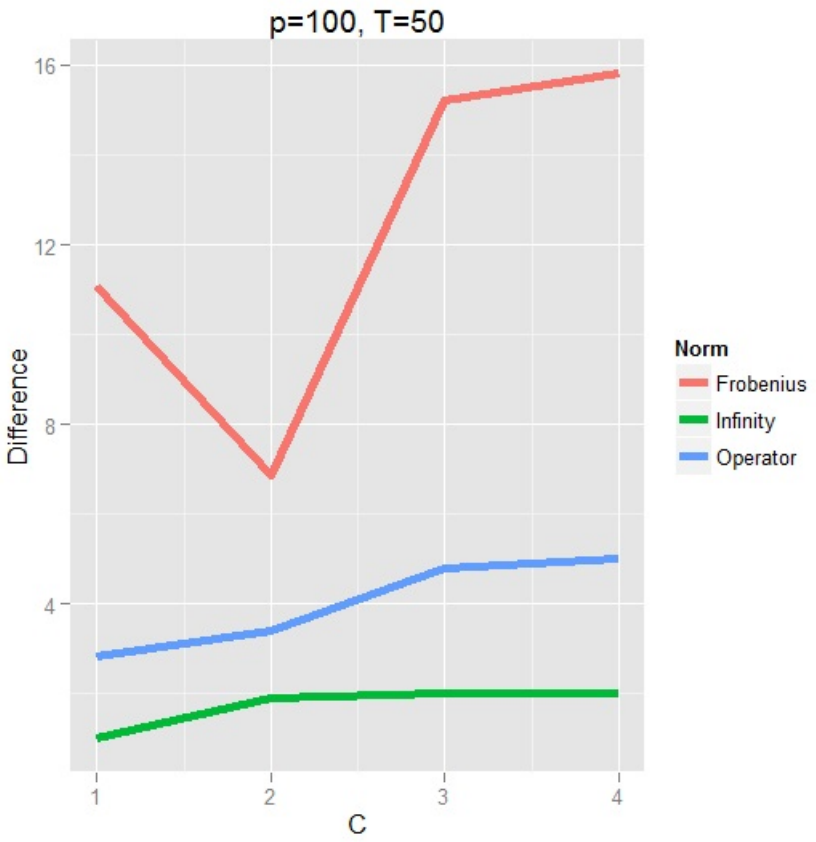

Figure 14: Performance of various C values when $p = 100$

Bibliography

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

[Akaike, 1977] Akaike, H. (1977). *On entropy maximization principle*. North-Holland.

[Bickel and Levina, 2008] Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.*, 36:25772604.

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

[Fan et al., 2011] Fan, J., Liao, Y., and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models.

[Hall et al., 1990] Hall, P., Kay, J., and Titterington, D. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77:521–528.

[Hoerl and Kennard, 1990] Hoerl, A. and Kennard, R. (1990). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):56–67.

[Kutner et al., 2005] Kutner, M., Nachtsheim, N., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill.

[Lokshin, 2006] Lokshin, A. (2006). Difference-based semiparametric estimation of partial linear regression models. *The Stata Journal*, 6(3):377–383.

[Merlevede et al., 2009] Merlevede, F., Peligrad, M., and Rio, E. (2009). A bernstein type inequality and moderate deviations for weakly dependent sequences.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

[Wang et al., 2011] Wang, L., Brown, L., and Cai, T. (2011). A difference based approach to the semiparametric partial linear model. *Electronic Journal of Statistics*, 5:619–641.

[Yatchew, 2000] Yatchew, A. (2000). Scale economies in electricity distribution - a semiparametric analysis. *Journal of Applied Econometrics*, 15(2):187–210.

[Yatchew, 2003] Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press.