Contributions to high-dimensional data analysis: some applications of the regularized covariance matrices

A thesis submitted in partial fulfilment of the requirements for the

degree of

Doctor of Philosophy

in

Statistics

at Massey University, Albany

New Zealand.

Insha ULLAH

March 2015

Abstract

High-dimensional data sets, particularly those where the number of variables exceeds the number of observations, are now common in many subject areas including genetics, ecology, and statistical pattern recognition to name but a few. The sample covariance matrix becomes rank deficient and is not invertible when the number of variables are more than the number of observations. This poses a serious problem for many classical multivariate techniques that rely on an inverse of a covariance matrix. Recently, regularized alternatives to the sample covariance have been proposed, which are not only guaranteed to be positive definite but also provide reliable estimates. In this Thesis, we bring together some of the important recent regularized estimators of the covariance matrix and explore their performance in high-dimensional scenarios via numerical simulations. We make use of these regularized estimators and attempt to improve the performance of the three classical multivariate techniques in high-dimensional settings.

In a multivariate random effects models, estimating the between-group covariance is a well known problem. Its classical estimator involves the difference of two mean square matrices and often results in negative elements on the main diagonal. We use a lasso-regularized estimate of the between-group mean square and propose a new approach to estimate the between-group covariance based on the EM-algorithm. Using simulation, the procedure is shown to be quite effective and the estimate obtained is always positive definite.

Multivariate analysis of variance (MANOVA) face serious challenges due to the undesirable properties of the sample covariance in high-dimensional problems. First, it suffer from low power and does not maintain accurate type-I error when the dimension is large as compared to the sample size. Second, MANOVA relies on the inverse of a covariance matrix and fails to work when the number of variables exceeds the number of observation. We use an approach based on the lasso regularization and present a comparative study of the existing approaches including our proposal. The lasso approach is shown to be an improvement in some cases, in terms of power of the test, over the existing high-dimensional methods.

Another problem that is addressed in the Thesis is how to detect unusual future observations when the dimension is large. The Hotelling T^2 control chart has traditionally been used for this purpose. The charting statistic in the control chart rely on the inverse of a covariance matrix and is not reliable in high-dimensional problems. To get a reliable estimate of the covariance matrix we use a distribution free shrinkage estimator. We make use of the available baseline set of data and propose a procedure to estimate the control limits for monitoring the individual future observations. The procedure do not assume multivariate normality and seems robust to the violation of multivariate normality. The simulation study shows that the new method performs better than the traditional Hotelling T^2 control charts.

Acknowledgements

My Ph.D project was supported by Higher Education Commission (HEC) of Pakistan. I gratefully acknowledge the financial support of the HEC.

I am grateful to my supervisor, Dr. Beatrix Jones for being my supervisor. She gave me the opportunity to learn from her and work together with her. Indeed, it is because of her, I managed to accomplish the task. Thanks Beatrix for your constant support, patience, suggestions and constructive criticisms. You made it a fruitful journey for me. I would like to thank my co-supervisor Prof. James Curran for his valuable feedbacks.

I have greatly benefited from the group of statistician working at the Institute of Natural and Mathematical Sciences, Massey University. I would like to thank all members of the group.

I also acknowledge the support of my family and friends. Thanks to Hoor and Hamood for being patient with me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

High-dimensional data, particularly those where the number of observed variable, p, is greater than the sample size, n, is becoming increasingly prevalent in many subject areas. For example, because of the high-throughput technology, a greater number of features can be observed in a microarray data set whereas the sample size cannot be increased. In this kind of high-dimensional data set, the sample covariance is either not invertible (if $n < p$) because of its rank deficiency; or, the inverse of a covariance matrix is unstable (n is comparable to p) and is not reliable.

The situation is challenging to any multivariate statistical procedure, especially those that rely on the inverse of the covariance matrix, and has attracted the attention of many researchers in the recent years. Regularized procedures for estimating the covariance matrix have been proposed. These include ridge regularization; the shrinkage method that shrinks the sample covariance towards a target; and the lasso regularization among others. These regularized estimates are not only guaranteed to be positive definite (even when $n < p$) but have also been proven to be more reliable than the sample covariance when n is comparable to p. However, it has been only recently that the regularized covariance matrix is used as an ingredient in other statistical techniques. In this Thesis, we bring together some of the important high-dimensional covariance estimation methodologies and extend their utility to three different multivariate statistical techniques. The contents of each chapter are outlined here.

1. In Chapter 2, we cover some of the important procedures to estimate the high-dimensional covariance matrices (or the inverse covariance matrices). These procedures include Moore-Penrose generalized inverse, shrinkage estimation of the covariance matrices, and the procedures based on the penalized likelihood approach (ridge, lasso, adaptive lasso, and SCAD regularization). The performance of different regularization procedures, to estimate the true covariance matrix, is assessed via simulation studies. To quantify the accuracy of each estimator, we use three different loss functions. Some other factors that can potentially affect the behavior of different regularization procedures are taken into account. The lasso estimator, although computationally expensive and assume multivariate normality, maintains the highest accuracy in most of the cases. The shrinkage estimator, on the other hand, is computationally inexpensive and does not make distributional assumption about the underlying set of data.

2. In Chapter 3, we address the problem associated with a multivariate random effect model that is used when a same set of characteristics is measured in several different groups. To fit the model, one need to estimate the within-group covariance and the between-group covariance. The estimation of the between-group covariance involves the difference of two mean square matrices: the between-group mean square and the within-group mean square. For a sufficiently large sample size, both mean squares are individually non-negative definite, however, their difference is not and often results negative elements on the diagonal. The probability of negative elements on the main diagonal increases as we increase p. This makes the analysis impracticable and the model cannot be fitted unless we have a very large sample size. In this part of the Thesis, we propose a strategy that overcome the problem. The difference of the two mean square matrices to obtain the between-group covariance is avoided, rather an EM-algorithm is used. The positive definiteness of the covariance matrices is ensured by using the lasso-regularized estimates. The performance of the method is illustrated by a number of simulated examples and a real glass chemical composition data set.

3. Chapter 4, is allocated to Multivariate Analysis of Variance (MANOVA). The traditionally available MANOVA tests such as Wilks lambda and Pillai-Bartlett trace start to suffer from low power and do not preserve accurate Type-I error rates, as the number of variables approaches the sample size.

Moreover, as the number of variables exceeds the number of available observations, these statistics are not available for use. Using regularized estimates of covariance matrices not only allow the use of MANOVA test in high-dimensional situations but has also been shown to exhibit high power. In this part of the Thesis, we bring together the previously used approaches for high-dimensional MANOVA and present an approach based on the lasso regularization. The comparative performance of the different approaches has been explored via an extensive simulation study. The MANOVA test based on the lasso regularization performs better in terms of power of the test in some cases. The methods are also applied to real data set of soil compaction profiles at various elevation ranges.

4. In Chapter 5, we present an overview of Hotelling T^2 control charts and highlight their inapplicability in high-dimensional settings. These charts have been used to monitor a stochastic process for out-of-control signals. The Phase-I analysis involves the clean up process of historical data, calculating baseline parameters and establishing control limits for Phase-II analysis. Once the control limits are established, the next Phase is to monitor the process for special cause. For each individual observation (i.e. the sub-group size is 1), Hotelling T^2 statistic is calculated and an out-of-control signal is issued if it goes beyond the control limits. A problem arises when the number of variables, p, approaches the number of baseline observations n: the Hotelling T^2 control chart becomes unreliable and even impractical when $n < p$. In this part of Thesis, we devise a procedure to improve the process monitoring in the high-dimensional setting. We use a shrinkage estimate of the covariance matrix as an estimate of the baseline parameter. A leave-one-out re-sampling procedure is used to obtain independent T^2 values. The upper control limit for monitoring the future observations is then calculated from kernel smoothed empirical distribution of the independent T^2 values. The performance of the proposed approach is tested, and compared to the Hotelling T^2 and the hypothetically "best possible" results, via an extensive simulation study. The procedure outperforms the standard Hotelling T^2 method and gives comparable results to the one based on true parameters. The procedure is also applied to a real gene expression data set and a chemical process data that has been analyzed in literature to demonstrate the principal component approach for process monitoring.

5. Finally, we conclude in Chapter 6 by providing a discussion about the main findings of this Thesis, and highlight areas of future work.

Chapter 2

Regularized estimation of high-dimensional covariance matrices

2.1 Introduction

The covariance matrix is the key input for most of the classical multivariate statistical techniques. Some of these techniques are Principal Component Analysis, Multivariate Analysis of Variance, Linear Discriminant Analysis, and Gaussian Graphical Models. Consider an $n \times p$ matrix $\mathbf{Y}$ of observations. The n rows of $\mathbf{Y}$ have a p-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{1 \leq i,j \leq p}$ i.e. $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Without loss of generality we assume that the observations are centered, so that, $\boldsymbol{\mu} = \mathbf{0}$. The log-likelihood function for estimating the covariance matrix is given by

$$l(\boldsymbol{\Sigma}; \mathbf{Y}) = Const - \frac{n}{2} \log \, det\,(\boldsymbol{\Sigma}) - \frac{1}{2} \sum_{i=1}^{n} \mathbf{Y}^t \boldsymbol{\Sigma}^{-1} \mathbf{Y}, \tag{2.1}$$

where $det\,(\mathbf{A})$ is the determinant of a matrix $\mathbf{A}$ and $\mathbf{A}^t$ denotes the transpose of a matrix $\mathbf{A}$. The global maximizer of $l(\boldsymbol{\Sigma}; \mathbf{Y})$ is the sample estimate of the covariance matrix given by $\widehat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{ij})_{1 \leq i,j \leq p} = \frac{1}{n} \mathbf{Y}^t \mathbf{Y}$. The maximum likelihood estimate $\widehat{\boldsymbol{\Sigma}}$ or its related unbiased estimate $\mathbf{S} = \frac{n}{n-1} \widehat{\boldsymbol{\Sigma}} = (s_{ij})_{1 \leq i,j \leq p}$ is a widely-employed estimator of the covariance matrix.

High-dimensional data sets, where the sample size, n, is smaller relative to the dimension, p, are now common in many fields. Examples of high-dimensional data include gene expression arrays, high resolution images, and high-frequency financial data. The classical multivariate statistical techniques are designed to deal with the applications where n is large relative to p, and face significant challenges when n is comparable to p. One obvious reason is because these techniques rely on accurately estimated covariance matrices or on the inverse of it. The two undesirable properties of the sample covariance matrix in high-dimensional applications are well-known. First, for a fixed p, as we decrease n, the spread of the eigenvalues of the sample covariance matrix increases; therefore, $\widehat{\mathbf{\Sigma}}$ becomes unstable (Ledoit & Wolf, 2004). Consequently, the traditional multivariate techniques can be misleading. Second, the sample covariance matrix is singular, if $n < p$. As a result, those multivariate techniques that rely on the inversion of a sample covariance are not applicable at all. The behavior of sample covariance matrix relative to the true and some regularized alternatives is demonstrated in Figure 2.3 for a fixed $p = 40$ and $n \in \{20, 40, 1000\}$. For more discussion about this Figure, the readers are referred to Section 2.5.

To overcome these issues, different methods of regularizing the sample covariance matrix have been proposed in the literature. Some of these methods are restricted to the cases where sample covariance matrices are invertible ($n \geq p$). For example, an estimator that is inspired by the empirical Bayes approach is introduced by Haff (1980). Dey & Srinivasan (1985) derive an estimator based on the Stein's entropy loss function. These regularization techniques break down when $n < p$. Recently, new regularization procedure have been proposed with the emergence of high-dimensional data sets. These regularization procedures not only overcome the singularity issue of the sample covariance in $n < p$ setting but are also more stable. Pourahmadi (2013) reviews these recent regularization based estimation methods of the covariance matrix including banding, tapering, and thresholding estimation of the covariance matrix. Some of these regularization procedures are presented in the following sections.

2.2 Moore-Penrose generalized inverse

When the number of variables, p, is more than the number of observations, n, then some of the eigenvalues of the sample covariance matrix are zero and it is not invertible. In this situations, Moore-Penrose generalized inverse is often used (Penrose, 1955). It is an approximation to the true inverse covariance matrix and is based on the singular value decomposition. The singular value decomposition of the sample covariance, $\widehat{\mathbf{\Sigma}}$, is given by

$$\widehat{\mathbf{\Sigma}} = \mathbf{UDV}^t \tag{2.2}$$

where the columns of $\mathbf{U}$ and $\mathbf{V}$ are, respectively, the orthonormal eigenvectors of $\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}^t$ and $\widehat{\mathbf{\Sigma}}^t\widehat{\mathbf{\Sigma}}$, and $\mathbf{D}$ is diagonal with the square root of the eigenvalues from $\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}^t$ (or $\widehat{\mathbf{\Sigma}}^t\widehat{\mathbf{\Sigma}}$) on the main diagonal. Note that the diagonal elements of $\mathbf{D}$ are in descending order and the columns of $\mathbf{U}$ and $\mathbf{V}$ are ordered according to their respective eigenvalues as well. The Moore-Penrose generalized inverse is obtained by restricting $\mathbf{D}$ to non-zero eigenvalues. That is, it reduces the dimension of $\mathbf{D}$ from p to the rank of $\widehat{\mathbf{\Sigma}}$. Those columns in $\mathbf{U}$ and $\mathbf{V}$ that correspond to zero eigenvalues are also eliminated. The generalized inverse is then calculated using

$$\widehat{\mathbf{\Sigma}}^{-1} = \mathbf{VD}^{-1}\mathbf{U}^t \tag{2.3}$$

It can be shown that $\widehat{\mathbf{\Sigma}}^{-1}$ is the shortest length least-squares solution of $\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}^{-1} = \mathbf{I}$ and whenever $rank(\widehat{\mathbf{\Sigma}}) \geq p$, the Moore-Penrose generalized inverse reduces to the standard matrix inverse (Golub & Kahan, 1965). In this Thesis we use the built-in R function *ginv()* in “MASS” package to calculate Moore-Penrose generalized inverse (Venables & Ripley, 2002).

2.3 Shrinkage Estimate

The idea of shrinkage estimation dates back to 1960s when Stein demonstrated that the performance of an estimator may be improved via shrinking it towards a structured target (Stein, 1956; James & Stein, 1961). Ledoit & Wolf (2004) consider the idea of shrinkage estimation and propose a shrinkage estimator of a covariance matrix that is the convex linear combination of a sample covariance and a target matrix. They provided a procedure for finding the optimal shrinkage intensity, which asymptotically minimizes the expected quadratic loss function, $E\|\widehat{\mathbf{\Sigma}}_\rho - \mathbf{\Sigma}\|^2$, where $\|\mathbf{A}\|^2$ is the squared matrix norm of $\mathbf{A}$. The expected quadratic loss function measures the mean-squared error summed over elements; an estimator with minimal mean-squared error is desired. The Ledoit-Wolf estimator is shown to be well-conditioned in high dimensional problems. It is important to note that the estimator does not make any distributional assumptions about the underlying distribution of the data and its performance advantages are, therefore, not restricted to Gaussian assumptions.

Consider the maximum likelihood estimate of a high-dimensional covariance matrix $\widehat{\mathbf{\Sigma}}$ and let $\mathbf{T} = (t_{ij})_{1 \le i,j \le p}$ be a target estimate towards which we want to shrink our estimate. The target estimator, $\mathbf{T}$, is required to be positive definite and its specification needs some assumptions about the structure of the true covariance matrix, $\mathbf{\Sigma}$. For example, Ledoit & Wolf (2004) uses a diagonal matrix as a structured target estimate (presuming that all the variances are equal and all the covariances are zero) which is also positive definite. A Steinian-class of shrinkage estimators is obtained by the convex linear combination of $\widehat{\mathbf{\Sigma}}$ and $\mathbf{T}$, given by

$$\widehat{\mathbf{\Sigma}}_\rho = \rho\mathbf{T} + (1-\rho)\widehat{\mathbf{\Sigma}} \tag{2.4}$$

where $\rho \in [0,1]$ is the shrinkage intensity. Note that, for $\rho = 0$ we get $\widehat{\mathbf{\Sigma}}_\rho = \widehat{\mathbf{\Sigma}}$ while for $\rho = 1$, we have $\widehat{\mathbf{\Sigma}}_\rho = \mathbf{T}$. The regularized estimate, $\widehat{\mathbf{\Sigma}}_\rho$, obtained in this way is more accurate and statistically efficient than the estimators $\widehat{\mathbf{\Sigma}}$ and $\mathbf{T}$ in the problems with n comparable to p (see Ledoit & Wolf, 2004).

2.3.1 Computation of the shrinkage intensity ρ

The value of ρ is critical to choose because it in turn determines the properties of the shrinkage estimate, $\widehat{\mathbf{\Sigma}}_\rho$. Schäfer & Strimmer (2005b) take the formulation of Ledoit & Wolf (2003) and derive analytic expressions to compute the optimal shrinkage intensity for six commonly used targets. To be more specific, they follow Ledoit & Wolf (2003) and minimize the risk function

$$R(\rho) = E\|\widehat{\mathbf{\Sigma}}_\rho - \mathbf{\Sigma}\|^2 \tag{2.5}$$

to compute the value of ρ. Minimizing (2.4) with respect to ρ, the following expression has been obtained for the optimal value of ρ

$$\hat{\rho}^* = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} Var(\hat{\sigma}_{ij}) - Cov(t_{ij}, \hat{\sigma}_{ij}) - Bias(\hat{\sigma}_{ij})E(t_{ij} - \hat{\sigma}_{ij})}{\sum_{i=1}^{p}\sum_{j=1}^{p} E\left[(t_{ij} - \hat{\sigma}_{ij})^2\right]}. \tag{2.6}$$

It is possible from the above expression to obtain a value of $\hat{\rho}^*$ that is either greater than 1 (over shrinkage) or even negative. This is avoided by using $\hat{\rho} = max(0, min(1, \hat{\rho}^*))$. If $\widehat{\boldsymbol{\Sigma}}$ is replaced by the unbiased estimate, $\mathbf{S}$, in (2.3) then the expression for ρ in (2.5) reduces to

$$\hat{\rho}^* = \frac{\sum_{i=1}^{p} \sum_{j=1}^{p} Var(s_{ij}) - Cov(t_{ij}, s_{ij})}{\sum_{i=1}^{p} \sum_{j=1}^{p} E\left[(t_{ij} - s_{ij})^2\right]}. \tag{2.7}$$

It is worth noting at this point, that the shrinkage intensity varies as we change the target estimator. Schäfer & Strimmer (2005b) provide a detailed discussion about the six commonly used targets. A natural choice for $\mathbf{T}$ is $\mathbf{I}$, the identity matrix (used by (Ledoit & Wolf, 2003) and (Ledoit & Wolf, 2004)) or its scalar multiple. This choice not only assumes sparsity which is more intuitive in high-dimensional applications but also remarkably simple because it require no parameters or one parameter to be estimated. Using the identity matrix as a target estimate reduces the expression in (2.6) to

$$\hat{\rho}^* = \frac{\sum_{i=1}^{p} Var(s_{ij}, i \neq j) + \sum_{i=1}^{p} Var(s_{ii})}{\sum_{i=1}^{p} (s_{ij}^2, i \neq j)}. \tag{2.8}$$

This target shrinks both the off-diagonal elements (covariances) and the diagonal elements (variances) of the sample covariance matrix, therefore alters the complete eigenstructure of the sample covariance matrix. Another, more complex choice, which has been the focus of Schäfer & Strimmer (2005b) is $\mathbf{S}_d$, where $\mathbf{S}_d$ is a diagonal matrix with diagonal elements of $\mathbf{S}$ on the main diagonal and zero elsewhere (it is complex because it requires p parameters to be estimated). This target shrinks only the off-diagonal elements, therefore, shrink only the eigenvalues and leave the eigenvectors unchanged. The expression in (2.6) for $\mathbf{T} = \mathbf{S}_d$ simplifies to

$$\hat{\rho}^* = \frac{\sum_{i=1}^{p} Var(s_{ij}, i \neq j)}{\sum_{i=1}^{p} (s_{ij}^2, i \neq j)}. \tag{2.9}$$

An advantage of using $\mathbf{I}$ (or its scalar multiple) and $\mathbf{S}_d$ as target estimates is that they are positive definite. Since (2.3) becomes a convex linear combination of positive definite target matrix, $\mathbf{T}$, and positive semidefinite matrix $\mathbf{S}$, therefore the obtained shrinkage estimate $\widehat{\boldsymbol{\Sigma}}_\rho$ is guaranteed to be positive definite.

In this Thesis, we use the function *cov.shrink()* with the default options, available in contributed R package “corpcor” (Schaefer et al., 2010), to calculate the shrinkage estimate of the covariance matrix. The *cov.shrink()* function shrinks the sample correlation matrix, $\mathbf{R} = (r_{ij})_{1 \le i,j \le p}$, towards the identity target estimate, $\mathbf{I}$. Replacing $\widehat{\mathbf{\Sigma}}$ by $\mathbf{R}$ and $\mathbf{T}$ by $\mathbf{I}$ in (2.3), the expression for shrinkage intensity in (2.6) simplifies to

$$\hat{\rho}^* = \frac{\sum_{i=1}^{p} Var(r_{ij}, i \neq j)}{\sum_{i=1}^{p} (r_{ij}^2, i \neq j)}. \tag{2.10}$$

The shrunken covariance is then obtained using the equation

$$\widehat{\mathbf{\Sigma}}_\rho = \mathbf{S}_d^{1/2} \widehat{\mathbf{R}}_\rho \mathbf{S}_d^{1/2}, \tag{2.11}$$

where $\widehat{\mathbf{R}}_\rho$ is the regularized version of $\widehat{\mathbf{R}}$. This formulation is more appropriate when variables are measured on different scales. Note that, the *cov.shrink()* function also allows the diagonal elements to shrink (this is the default option) with separate shrinkage intensity calculated by

$$\hat{\rho}^* = \frac{\sum_{i=1}^{p} Var(s_{ii})}{\sum_{i=1}^{p} (s_{ii}^2 - median(s))^2}, \tag{2.12}$$

where $median(s)$ is the median of sample variances.

2.4 Penalized Normal Likelihood

A likelihood-based approach, using penalized multivariate normal likelihood, provides another class of regularized estimators of the covariance matrices. The following log-likelihood function, based on a random sample, $\mathbf{Y}$, of size n from a multivariate normal distribution, $\mathbf{Y} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, has been optimized subject to the positive-definiteness constraint of $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = (\omega_{ij})_{1 \le i,j \le p}$

$$l(\mathbf{\Omega}) = Const - \frac{n}{2} \log\, det\,(\mathbf{\Sigma}) - \frac{1}{2} \sum_{i=1}^{n} \mathbf{Y}^t \mathbf{\Omega} \mathbf{Y} - \sum_{j=1}^{p} \sum_{k=1}^{p} p(\omega_{jk}), \tag{2.13}$$

where $p(a) > 0$ is a penalty function. Some well-known penalty functions include the ridge penalty, lasso penalty, adaptive lasso penalty, and SCAD penalty. These penalty functions are discussed in the following subsections.

2.4.1 Ridge Regularization

Ridge regularization which uses the L_2 penalty, has been adopted by Warton (2008). The solution to equation (2.13) with penalty function $p_\kappa(\omega_{jk}) = \kappa(\omega_{jk})^2$ is

$$\widehat{\mathbf{\Sigma}}_\kappa = \widehat{\mathbf{\Sigma}} + \kappa\mathbf{I}, \tag{2.14}$$

where $\widehat{\mathbf{\Sigma}}_\kappa$ is the ridge regularized estimator of covariance matrix and $\kappa > 0$ is a ridge parameter. For $\kappa = 0$, we simply get the maximum likelihood estimator.

Alternatively consider the sample estimate of the correlation matrix, $\mathbf{R}$ that can be obtained as

$$\widehat{\mathbf{R}} = \widehat{\mathbf{\Sigma}}_d^{-1/2}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}_d^{-1/2} \tag{2.15}$$

where $\widehat{\mathbf{\Sigma}}_d$ is the diagonal matrix with corresponding diagonal elements of $\widehat{\mathbf{\Sigma}}$ on the diagonal. The regularized version $\widehat{\mathbf{R}}_\rho$ of $\widehat{\mathbf{R}}$ can be obtained using the following convex linear combination of $\widehat{\mathbf{R}}$ and $\mathbf{I}$

$$\widehat{\mathbf{R}}_\rho = \rho\widehat{\mathbf{R}} + (1-\rho)\mathbf{I} \tag{2.16}$$

and the corresponding regularized estimate $\widehat{\mathbf{\Sigma}}_\rho$ is as

$$\widehat{\mathbf{\Sigma}}_\rho = \widehat{\mathbf{\Sigma}}_d^{1/2}(\rho\widehat{\mathbf{R}} + (1-\rho)\mathbf{I})\widehat{\mathbf{\Sigma}}_d^{1/2}, \tag{2.17}$$

where $\rho = 1/(1+\kappa) \in (0, 1]$ is the ridge parameter. For any choice of $\rho \in (0, 1]$ the regularized estimator $\widehat{\mathbf{\Sigma}}_\rho$ of the covariance matrix is guaranteed to be positive definite. An additional interesting property of $\widehat{\mathbf{R}}_\rho$ is that it can be derived as the penalized likelihood estimate of $\mathbf{R}$ for multivariate normal data, with a penalty term proportional to the trace of $\mathbf{R}^{-1}$ (see Warton (2008) for detailed proof).

2.4.1.1 Selection of the ridge parameter ρ

Warton (2008) use the normal likelihood in (2.1) as an objective function and the cross-validation procedure to obtain the optimum value of ridge parameter ρ. The n rows of $\mathbf{Y}$ are partitioned into K disjoint sub-samples i.e. $\mathbf{Y}^t = [\mathbf{Y}_1^t, \mathbf{Y}_2^t, ..., \mathbf{Y}_K^t]$, where $\mathbf{Y}_k^t$ has n_k rows for $k = 1, 2, ..., K$. The size of n_k is roughly the same for all K sub-samples i.e. $n_k \approx n/k$. For example, if $n = 25$ and we use 6-fold cross-validation then there are total 6 sub-samples. The size of the 5 sub-samples is 4 and the fourth sub-sample will have 5 observations (the rest). The kth sub-sample is retained as a holdout set and the rest of the data, $\mathbf{Y}^{\backslash k}$, is used as a training data. The sample size affects the penalty parameter; it is generally kept roughly the same across different sub-samples and we want the size of the training data set not too different from the original sample size, n. Let us write the index of the observations in kth-fold as T_k, the estimated covariance matrix of $\mathbf{Y}^{\backslash k}$ as $\widehat{\boldsymbol{\Sigma}}_\rho^{\backslash k}$ and the estimated mean vector of $\mathbf{Y}^{\backslash k}$ as $\widehat{\boldsymbol{\mu}}^{\backslash k}$, then the cross-validation score is calculated using

$$CV(\rho) = \sum_{k=1}^{K} \left[n_k \log\, det(\widehat{\boldsymbol{\Sigma}}_\rho^{\backslash k}) + \sum_{i \in T_k} (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}^{\backslash k})^t (\widehat{\boldsymbol{\Sigma}}_\rho^{\backslash k})^{-1} (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}^{\backslash k}) \right]. \qquad (2.18)$$

The optimal value of ρ is one which gives the highest score i.e. $\hat{\rho} = \arg\max_{\rho} CV(\rho)$. A built-in R function, *ridgeParamEst()*, to choose the optimal value of ρ is available in contributed R package “mvabund” (Y. Wang et al., 2012). In this thesis, we use *ridgeParamEst()* to obtain the optimal value of ρ.

2.4.2 Lasso Regularization

In a multivariate normal distribution, the inverse of a covariance matrix determines the conditional independence structure among variables. A zero off-diagonal element in the inverse covariance matrix means that the corresponding variables are conditionally independent given the rest. Identifying zero off-diagonal elements in the inverse covariance matrix are therefore termed as model selection for Gaussian graphical models (Cox & Wermuth, 1996). Denote the inverse covariance matrix by $\boldsymbol{\Omega}$, then the natural estimator of $\boldsymbol{\Omega}$ is $\widehat{\boldsymbol{\Sigma}}^{-1}$ or $\mathbf{S}^{-1}$. These estimators are unlikely to produce an estimated inverse covariance matrix, $\widehat{\boldsymbol{\Omega}}$, with exactly zero off-diagonal entries. However, in high-dimensional problems, we believe that there are frequently superfluous parameter estimates (they are zero in population) which make the model unnecessarily more complex and unstable. The prediction accuracy and model interpretability can be substantially increased by setting some of the parameter estimates to zero, which is called covariance selection introduced by Dempster (1972). Moreover, in high-dimensional problems, fitting a sparser model provides more power to accurately estimate the important parameters.

A well-known problem due to high-dimensionality is the collinearity among predictors in multiple regression. To remedy this issue, Tibshirani (1996) proposed the lasso in the regression setting, a popular model selection and shrinkage estimation method, which has the ability to shrink some coefficients towards zero and sets the others as exactly zero. It is, therefore, simultaneously selecting important variables and estimating their effects. This idea was extended by Yuan & Lin (2007) to the likelihood-based estimation of the inverse covariance matrix $\mathbf{\Omega}$. The log-likelihood function in (2.13) with penalty function, $p_\rho(\omega_{jk}) = \rho\,|a|$, where ρ is a penalty parameter and $|a|$ denotes the absolute value of a, has been optimized.

Friedman et al. (2008) have proposed the fastest algorithm, known as graphical lasso algorithm, to solve the lasso problem. Here are the details of the algorithm:

Partition the estimate $\mathbf{W} = \mathbf{S} + \rho\mathbf{I}$ of $\mathbf{\Sigma}$ and $\mathbf{S}$ as

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & w_{12} \\ w_{12}^t & w_{22} \end{pmatrix}, \qquad \mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & s_{12} \\ s_{12}^t & s_{22} \end{pmatrix} \tag{2.19}$$

then the lasso problem is given by

$$min_\beta \Big\{ \frac{1}{2} \|\mathbf{W}_{11}^{1/2}\beta - b\|^2 + \rho\|\beta\|_1 \Big\}, \tag{2.20}$$

where $\|\mathbf{A}\|_1$ is L_1 norm and $b = \mathbf{W}_{11}^{-1/2} s_{12}/2$. The algorithm works as follows:

1. Start with $\mathbf{W}$ without changing the diagonal in the following steps.
2. For each $j = 1, 2, ..., p,\ 1, 2, ..., p,\ ...$, permute the rows and columns of $\mathbf{W}$ and $\mathbf{S}$ in such a way so that the target column is always the last and solve the problem in (2.20), which gives a $p-1$ vector solution $\hat{\beta}$.
3. Fill in the corresponding row and column of $\mathbf{W}$ using $w = 2\mathbf{W}_{11}\hat{\beta}$.
4. Continue until convergence.

The algorithm gives the regularized estimate of the covariance matrix $\widehat{\mathbf{\Sigma}}_\rho = \mathbf{W}$ at the convergence. The regularized estimate of the inverse covariance matrix $\widehat{\mathbf{\Omega}}_\rho = \mathbf{W}^{-1}$ is also recovered after convergence utilizing the relation $\mathbf{W\Omega} = \mathbf{I}$ partitioned as

$$\begin{pmatrix} \mathbf{W}_{11} & w_{12} \\ w_{12}^t & w_{22} \end{pmatrix} \begin{pmatrix} \mathbf{\Omega}_{11} & \omega_{12} \\ \omega_{12}^t & \omega_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 \\ 0^t & 1 \end{pmatrix} \tag{2.21}$$

from this the expression

$$\hat{\omega}_{22} = \frac{1}{w_{22} - w_{12}^t\hat{\beta}} \tag{2.22}$$

and

$$\hat{\omega}_{12} = -\hat{\beta}\hat{\omega}_{22} \tag{2.23}$$

is derived. The lasso problem in step (2) of the above algorithm is solved by the coordinate descent algorithm (Friedman et al., 2007). For $j = 1, 2, ..., p$, $1, 2, ..., p$, ..., and with $\mathbf{V} = \mathbf{W}_{11}$ the following update is cycled through the predictors until convergence:

$$\hat{\beta}_j \leftarrow S(s_{12j} - 2\sum_{k \neq j} \mathbf{V}_{kj}\hat{\beta}_k, \rho)/(2\mathbf{V}_{jj}) \tag{2.24}$$

where $S(x,t) = sign(x)(|x| - t)_+$ is the soft-threshold operator with $(a)_+ = max(0,a)$. The algorithm stops when the average absolute change in $\mathbf{W}$ is less than $t.ave\left|\mathbf{S}^{\setminus diag}\right|$, where $\mathbf{S}^{\setminus diag}$ are the elements of the sample covariance matrix $\mathbf{S}$, excluding the diagonal elements, and t is a small positive constant. For example, the *glasso()* function in the contributed R package "glasso" uses $t = 0.0001$ (Friedman et al., 2013).

2.4.3 Adaptive lasso and SCAD penalty functions

Clearly the lasso penalty discourages superfluous parameters estimates from appearing in the model. However, it has been criticized for its linear increase in penalty, which produces stringly biased estimates for large parameters (Fan & Li, 2001). This problem is alleviated by Fan et al. (2009), who use the extended versions of lasso penalty known as the adaptive lasso (Zou, 2006) and the Smoothly Clipped Absolute Deviation (SCAD) penalty Fan & Li (2001). The adaptive lasso penalty is given by:

$$p_\rho(\omega_{ij}) = \frac{\rho}{|\tilde{\omega}_{ij}|^\gamma}|\omega_{ij}| \tag{2.25}$$

for some initial estimator $\tilde{\mathbf{\Omega}} = (\tilde{\omega}_{ij})_{1 \leq i,j \leq p}$ and some $\gamma > 0$. For adaptive lasso the initial estimator $\tilde{\mathbf{\Omega}}$ is required to be a consistent estimator of $\mathbf{\Omega}$. The choice of a consistent initial estimator is an issue for the adaptive lasso penalty. As mentioned earlier, the sample variance-covariance matrix for high-dimensional problems is not invertible in $p > n$ settings and therefore cannot be used as an initial estimator. Following Fan et al. (2009) we use a lasso estimate as an initial estimator and keep $\gamma = 0.5$.

The first-order derivative of SCAD penalty is:

$$p'_\rho(\tilde{\omega}_{ij}) \quad = \quad \begin{cases} \rho & \text{when } |\tilde{\omega}_{ij}| \leq \rho \\ \frac{(a\rho - |\tilde{\omega}_{ij}|)_+}{(a-1)} & \text{when } |\tilde{\omega}_{ij}| > \rho \end{cases} \tag{2.26}$$

and the resulting solution is given by:

$$\hat{\omega}_{ij} = \begin{cases} \text{sign}(\hat{\omega}_{ij})(|\hat{\omega}_{ij}| - \rho)_+ & \text{when } |\tilde{\omega}_{ij}| \leq 2\rho \\ \frac{\{(a-1)\hat{\omega}_{ij} - \text{sign}(\hat{\omega}_{ij})a\rho\}}{(a-2)} & \text{when } 2\rho < |\tilde{\omega}_{ij}| \leq a\rho \\ \hat{\omega}_{ij} & \text{when } \ |\tilde{\omega}_{ij}| > a\rho \end{cases} \tag{2.27}$$

for some $a > 2$ and $(x)_+ = \max(0, x)$. Fan & Li (2001) recommend to use $a = 3.7$ and it is later used by Fan et al. (2009). Unlike the lasso penalty, these penalty functions leave large coefficients not excessively penalized. Ultimately, they not only produce sparse solutions but also produce the parameter estimates as efficient as if the true model were known, i.e. they enjoy the so called oracle properties (Fan & Li, 2001). Figure 2.1 shows the regularized estimates against the maximum likelihood estimates for the elements of the inverse covariance matrix. Note that, the ridge penalty does not produce zero estimates and also excessively penalizes the large coefficients. The lasso penalty forces small coefficients to be exactly zero, however, it still produce bias by excessively penalizing the large coefficients. Adaptive lasso and SCAD penalty resolve both these issues i.e. they produce sparse solutions and also eliminate the bias issue for large coefficients. Since our objective is to estimate the covariance matrix rather than model selection (identifying zero elements in the inverse covariance matrix), the estimator based on the lasso penalty perform better than the estimator based on the adaptive lasso and the SCAD penalties (see simulations in Section 2.5). Fitch et al. (2014) have also noticed that the lasso regularization and even the adaptive lasso do not do a great job of model selection compared to model selection procedures.

In our analysis, we use the *glasso()* function to calculate lasso regularized covariance matrix. It is is available in the contributed R package “glasso” (Friedman et al., 2013). The *glasso()* function allows us to use weighted L_1 penalty like adaptive lasso and SCAD penalty. Note that, the lasso regularization (like shrinkage regularization) also allows to change the complete eigenstructure (eigenvalues and eigenvectors) by penalizing both diagonal and off-diagonal elements. This is the default option in the *glasso()* function.

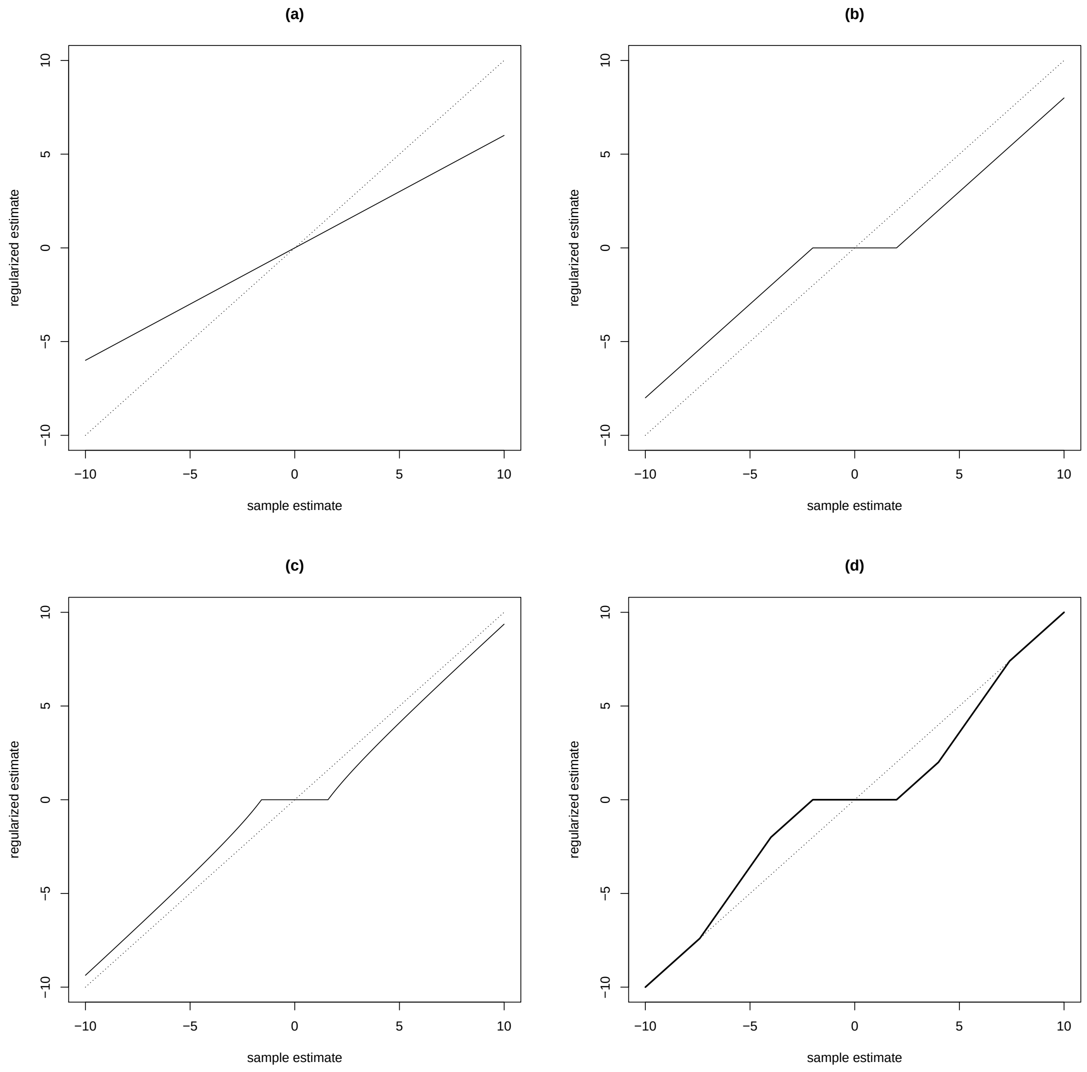


FIGURE 2.1: A schematic diagram showing the regularized estimates against the maximum likelihood estimates for the elements of the inverse covariance matrix (a) the ridge, (b) the lasso, (c) the adaptive lasso, (d) the SCAD regularized estimates.

2.4.4 Choosing the optimal value of the penalty parameter ρ

Like in ridge and shrinkage regularizations, the choice of ρ is critical for lasso regularization, as it controls the properties of the estimator. This is also essential for adaptive lasso and SCAD penalties to hold the aforementioned properties (Fan & Li, 2001; Fan et al., 2004; H. Wang et al., 2007). Larger values of ρ will produce sparser solutions. Smaller values of ρ, on the other hand, will encourage more non-zero off-diagonal elements to appear in the inverse covariance matrix. Traditionally, two automatic data-driven methods have been used to select the optimal value of tuning parameter:

1. Bayesian information criterion (BIC) proposed by H. Wang et al. (2007) and was further investigated by Gao et al. (2009) for Gaussian graphical models.
2. Cross-validation used by Friedman et al. (2008) and Fan et al. (2009).

BIC is computationally less expensive and its empirical performance is shown to be advantageous over cross-validation in Gaussian graphical models (Gao et al., 2009), where the objective is to identify zero off-diagonal elements in the inverse covariance matrix. In our experience, BIC generally chooses larger values of the regularization parameter than cross-validation and therefore forces more of the off-diagonal elements to be equal to zero. Since the lasso penalty heavily penalizes the large coefficients, it aggravates the problem if BIC is used to choose the penalty parameter. Cross-validation typically performs well with lasso penalty as compared to BIC, when the objective is to estimate covariances.

We follow Fan et al. (2009) and use K-fold cross-validation to select the optimal value of ρ via a grid search over a grid of values produced by $e^{v/10}$, where $v \in [-100, 10]$. To reduce the computational effort we stop the search at a value of ρ if the cross-validation score is decreasing over the next three consecutive values of ρ in the grid. We partition the n rows of $\mathbf{Y}$ into K disjoint sub-samples $\mathbf{Y}^t = [\mathbf{Y}_1^t, \mathbf{Y}_2^t, ..., \mathbf{Y}_K^t]$, where $\mathbf{Y}_k^t$ has n_k rows for $k = 1, 2, ..., K$. The size of n_k is roughly the same for all K sub-samples i.e assume that n is a multiple of K then $n_k = n/K$. The kth sub-sample is retained as a holdout set and the rest of the data, $\mathbf{Y}^{\setminus k}$, is used as a training data. Denote the covariance matrix of $\mathbf{Y}^{\setminus k}$ by $\widehat{\mathbf{\Sigma}}_\rho^{\setminus k}$ then the cross-validation score is calculated using

$$CV(\rho) = \sum_{k=1}^{K} n_k \left(\log det \left(\widehat{\mathbf{\Sigma}}_\rho^{\setminus k} \right) + tr(\mathbf{S}_k (\widehat{\mathbf{\Sigma}}_\rho^{\setminus k})^{-1}) \right), \tag{2.28}$$

where $\mathbf{S}_k$ is the sample covariance calculated from the kth sub-sample and $tr(\mathbf{A})$ is the trace of a matrix $\mathbf{A}$. The optimal value of ρ is one which gives the highest score i.e. $\hat{\rho} = \arg\max_\rho CV(\rho)$.

2.5 Numerical simulations

To evaluate the performance of the shrinkage, ridge, lasso, adaptive lasso, and SCAD regularizations, in how well they estimate the true covariance matrix, a large simulation study is conducted. We draw n observations from a p-variate normal distribution with mean vector, $\boldsymbol{\mu} = \mathbf{0}$, and covariance matrix, $\boldsymbol{\Sigma}$. Three different covariance structures are used: the exchangeable structure given by

$$\sigma_{ij} \quad = \quad \begin{cases} 1 & \text{when } i = j \\ b & \text{when } i \neq j \end{cases} \qquad \text{for} \quad 1 \leq i, j \leq p, \tag{2.29}$$

the AR(1) structure given by

$$\sigma_{ij} = b^{|i-j|} \qquad \text{for} \quad 1 \leq i, j \leq p, \tag{2.30}$$

and the method of Schäfer & Strimmer (2005a), which guarantees the generated matrix to be positive definite. We refer to the covariance matrices generated in this way as being from the random method. The algorithm to generate random covariances proceeds as follows:

1. Start with a null $p \times p$ matrix.
2. Choose randomly a suitable proportion of the off-diagonal positions, and fill them symmetrically with a value drawn from the uniform distribution between -1 and 1.
3. Set the rest of the off-diagonal elements as zero. To generate high-dimensional sparse inverse covariance matrices a smaller proportion of the off-diagonal positions would be required to fill in with non-zero elements in step 2.
4. Sum up the absolute values for each column plus a small constant and fill them in their respective diagonal positions. This is our inverse covariance matrix $\boldsymbol{\Omega}$.
5. The inverse of $\boldsymbol{\Omega}$ is the covariance matrix $\boldsymbol{\Sigma}$.

Figure 2.2 presents the off-diagonal elements of some typical correlation matrices generated using the random method. The Figure shows that for a fixed value of p as we increase the proportion of the non-zero elements in the off-diagonal positions, the size of the off-diagonal elements in the correlation matrix decreases.

The inverse covariance matrix of AR(1) structure is sparse while the inverse covariance for exchangeable structure does not have zeros on the off-diagonal positions. These choices of covariance structures allow us to test the methods at two extremes. The random method stands in the middle and allows us to control for the amount of sparsity.

Figure 2.3 gives the general picture of how these regularization procedures overcome the problem of over-dispersion of the eigenvalues of a sample covariance. It compares the eigenvalues of the three regularized estimates of the covariance matrix (shrinkage, ridge, and lasso) to the sample estimate and true covariance matrix, for $p = 40$ and $n \in \{20, 40, 1000\}$. The estimated eigenvalues of a true covariance matrix, generated using the random method with the proportion of the non-zero off-diagonal elements equals to 50%, are averaged over 1000 simulations. In general, for a very large sample size, the eigenvalues for the sample estimate of the covariance matrix and for the three regularized estimates are all close to the true eigenvalues values. The sample covariance over estimates the large eigenvalues and under estimates the small eigenvalues when p is large relative to n (see Ledoit & Wolf (2004) for theoretical demonstration). The regularization of a sample covariance matrix deflates the large eigenvalues and inflates the small eigenvalues; therefore, overcome the defect of the sample covariance. The lasso and the shrinkage estimate recover the true eigenvalues more accurately on the average, however, the ridge regularization does not do a great job in the simulation experiments conducted here.

We draw 1000 sample of size $n = 20$ from a multivariate normal distribution with different choices of $p \in [10, 20, 40, 80]$ and consider all the three covariance structures mentioned above. For both AR(1) and exchangeable covariance structures, we show the results for $b = 0.6$. The random covariance matrix is different for each of the 1000 samples and the proportion of non-zero positions is kept as 40%, 30%, 20%, and 10%, respectively, for $p = 10$, $p = 20$, $p = 40$, and $p = 80$. Three different loss functions are used to make the assessments. The first two loss functions, known as the entropy loss function and the quadratic loss function, are given by

$$loss_1 = tr\left(\mathbf{\Sigma}^{-1}\widehat{\mathbf{\Sigma}}\right) - log\,det\left(\mathbf{\Sigma}^{-1}\widehat{\mathbf{\Sigma}}\right) - p, \qquad (2.31)$$

where $tr(\mathbf{A})$ is the trace of $\mathbf{A}$ and $det(\mathbf{A})$ denotes the determinant of a matrix $\mathbf{A}$, and

$$loss_2 = tr\left(\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\Sigma}} - \mathbf{I}\right)^2, \tag{2.32}$$

respectively. These two loss functions are common to asses the performance of the covariance estimators (see for example, (James & Stein, 1961) and (Ledoit et al., 2012)). The third loss function, which measures how well the five competing procedures recover the eigenvalues of the true covariance matrix, is given by

$$loss_3 = \sum_{i=1}^{p} \left|\hat{\lambda}_i - \lambda_i\right|, \tag{2.33}$$

where λ is the true eigenvalue, $\hat{\lambda}$ is the respective estimated eigenvalue and $|a|$ denotes the absolute value of a. The value of each of the above three loss functions is 0 when $\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\Sigma}} = \mathbf{I}$ and is positive otherwise. An estimator, $\widehat{\boldsymbol{\Sigma}}$, with minimum average loss is considered the best.

The distributions of the three loss functions for the exchangeable, random, and AR(1) covariance structures are shown in Figures 2.4, 2.5, and 2.6, respectively. Since, the shrinkage, lasso, adaptive lasso, and SCAD regularization allow to penalize the diagonal elements, we arrange the results with diagonal elements unpenalized in column (a) of the figures while the results for the diagonal elements penalized are presented in column (b) of the figures. The ridge regularization does not allow to penalize the diagonal elements; therefore, the distributions in (b) are the replicates of the results in (a) for ridge regularization.

In general, the estimation error increases as we increase the number of variables. The lasso regularization maintains the highest accuracy in most of the cases, when the diagonal elements are not penalized (see column (a) of the figures). Its performance becomes weak when the diagonal elements are penalized that is more clear under the quadratic loss function. The diagonal elements (variances) in the covariance matrices are larger as compared to their respective off-diagonal elements and the lasso penalty increase linearly for large coefficients. The poor performance of the lasso penalty, when the diagonal elements are penalized, is due to its increased bias for large coefficients. The adaptive lasso and SCAD penalty do not heavily penalize the large elements, so the difference in their performance when the diagonal elements are not penalized to when the diagonal elements are penalized is not huge.

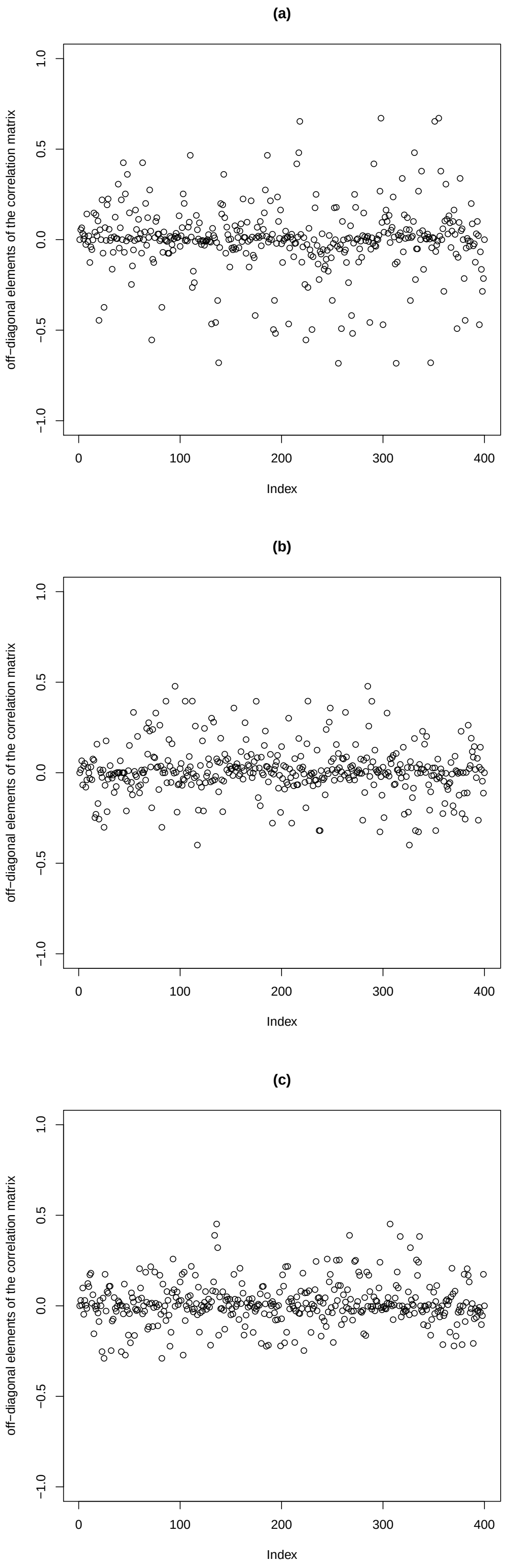


FIGURE 2.2: Off-diagonal elements of simulated correlation matrices for $p = 20$ using the algorithm of Schäfer & Strimmer (2005a). The proportion of non-zero off-diagonal elements in step 2 of the algorithm is (a) 20%, (b) 30%, and (c) 40%. For a fixed value of p as we increase the proportion of the non-zero elements in the off-diagonal positions, the size of the off-diagonal elements in the correlation matrix decreases.

Comparing the performance of shrinkage and ridge regularization, there is a little difference for exchangeable structure if the diagonal elements are unpenalized for shrinkage regularization. Penalizing the diagonal slightly improves the performance of shrinkage estimate over the ridge estimate with respect to the entropy loss function while the improvement is substantial with respect to the quadratic loss function (even better than lasso and its weighted versions). For random structure, however, ridge regularization performs well for small p and its performance becomes worse, under $loss_1$ and $loss_3$, for large p compare to the other counterparts. In our simulation study, the reason for the poor performance of ridge regularization is that it tends to underestimate the shrinkage intensity as compared to the shrinkage estimation.

2.6 Summary and conclusion

The Moore-Penrose generalized inverse is commonly used in applications where the sample covariance matrix is not invertible. It is obtained by restricting the dimensionality to the number of non-zero singular values and reduces to the standard matrix inverse whenever $rank(\widehat{\mathbf{\Sigma}}) \geq p$, which is known to be ill-conditioned when p is close to n. Shrinkage regularization is an improvement over Moore-Penrose generalized inverse in terms of mean square error (Schäfer & Strimmer, 2005). It shrinks the sample covariance matrix towards a target estimate and therefore converts the unstable but unbiased sample covariance into a biased but more stable estimate. The target matrix determines the properties of the shrinkage estimate and its specification requires some structural information about the true covariance matrix. If the specified target matrix is positive definite then the shrinkage estimate is guaranteed to be positive definite. The ridge regularization can be viewed as a special case of the shrinkage estimate because the target matrix is always identity matrix (it shrinks the sample correlation matrix towards an identity target matrix). It only shrinks the eigenvalues and leave the eigenvectors as unchanged. The shrinkage regularization, on the other hand, allows other target estimators and therefore together with shrinking the eigenvalues also allows to alter the eigenvectors. Another difference in the shrinkage and ridge regularization is that in ridge regularization the ridge parameter is chosen via cross-validation while shrinkage regularization minimizes the quadratic loss function to calculate shrinkage intensity.

A zero off-diagonal element in a true inverse covariance matrix can never be estimated as exactly equal to zero by the sample estimator (the inverse of the sample covariance) no matter how large the sample size. The shrinkage and ridge estimator do not hold this property either. The distinguishing property of the lasso (and its weighted versions: adaptive lasso and SCAD) regularization is that it sets some of the elements of the inverse covariance matrix as exactly zero. However, the lasso regularization has been criticized for its excessive penalty for large coefficients. This bias in the large coefficients becomes more clear when we penalize the diagonal elements as we saw in our simulation experiments. Adaptive lasso and SCAD penalty overcome this problem and they can produce sparse solutions without heavily penalizing the large coefficients. In our simulation, we found that lasso (if we do not penalize the diagonal elements) perform the best in most cases as long as the objective is to estimate the covariance matrix or its inverse (rather than model selection). Therefore, we do not consider adaptive lasso and SCAD penalty in the rest of the Thesis.

2.7 Contributions of the Chapter

A number of high-dimensional covariance estimation methodologies have been developed in recent decades. While each of these procedures has undergone some assessment in the literature, we bring together the most promising recent procedures (and the conventional Moore-Penrose generalized inverse), for a comprehensive comparison via simulation. Three different loss functions based on the eigen structure of the covariance matrix are used, and three different sparsity scenarios are considered. Our conclusions in this chapter guide us in our choice of regularization techniques in the rest of the thesis.

Note: The next three chapters: Chapter 3, Chapter 4, and Chapter 5 are formatted as papers. Some of the material presented in this chapter are repeated as needed.

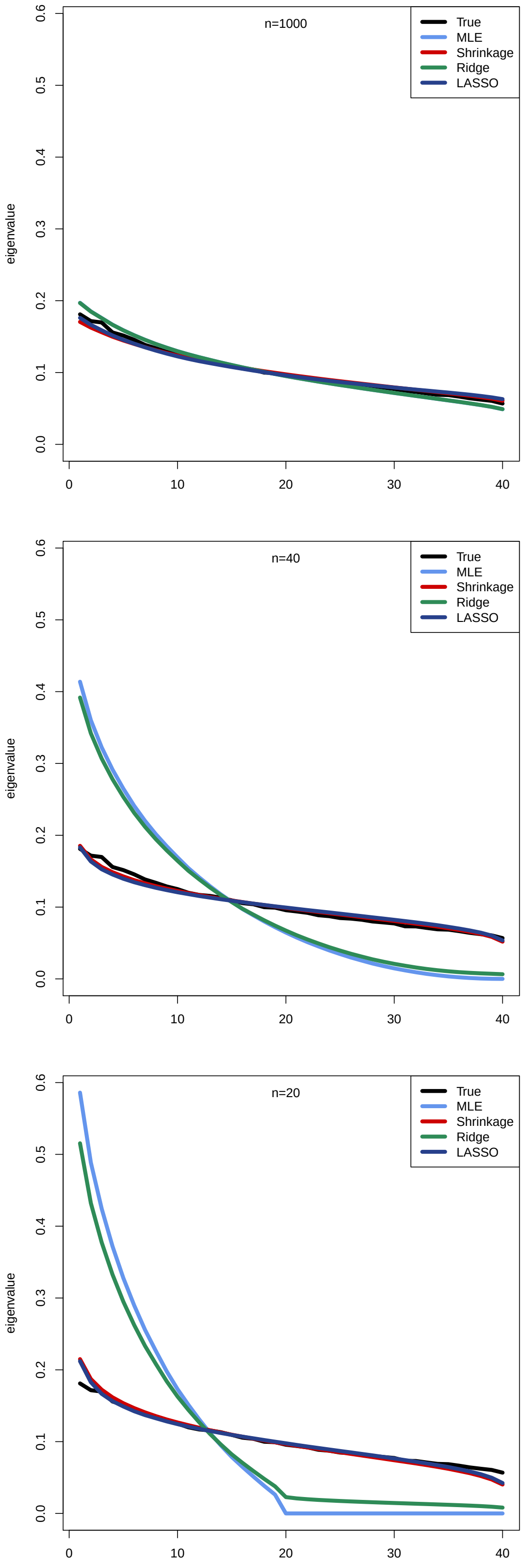


Figure 2.3: Ordered eigenvalues of a true and estimated covariance matrices. A true covariance matrix is generated using the random method with proportion of non-zero off-diagonal elements equals to 50% and $p = 40$. The covariance matrix is estimated 1000 times, using 1000 samples for each $n \in \{20, 40, 1000\}$ from a multivariate normal distribution and the average eigenvalues of the estimated covariance matrices are presented. The diagonal elements of the estimated covariance matrices are not penalized in any of the regularization procedures.

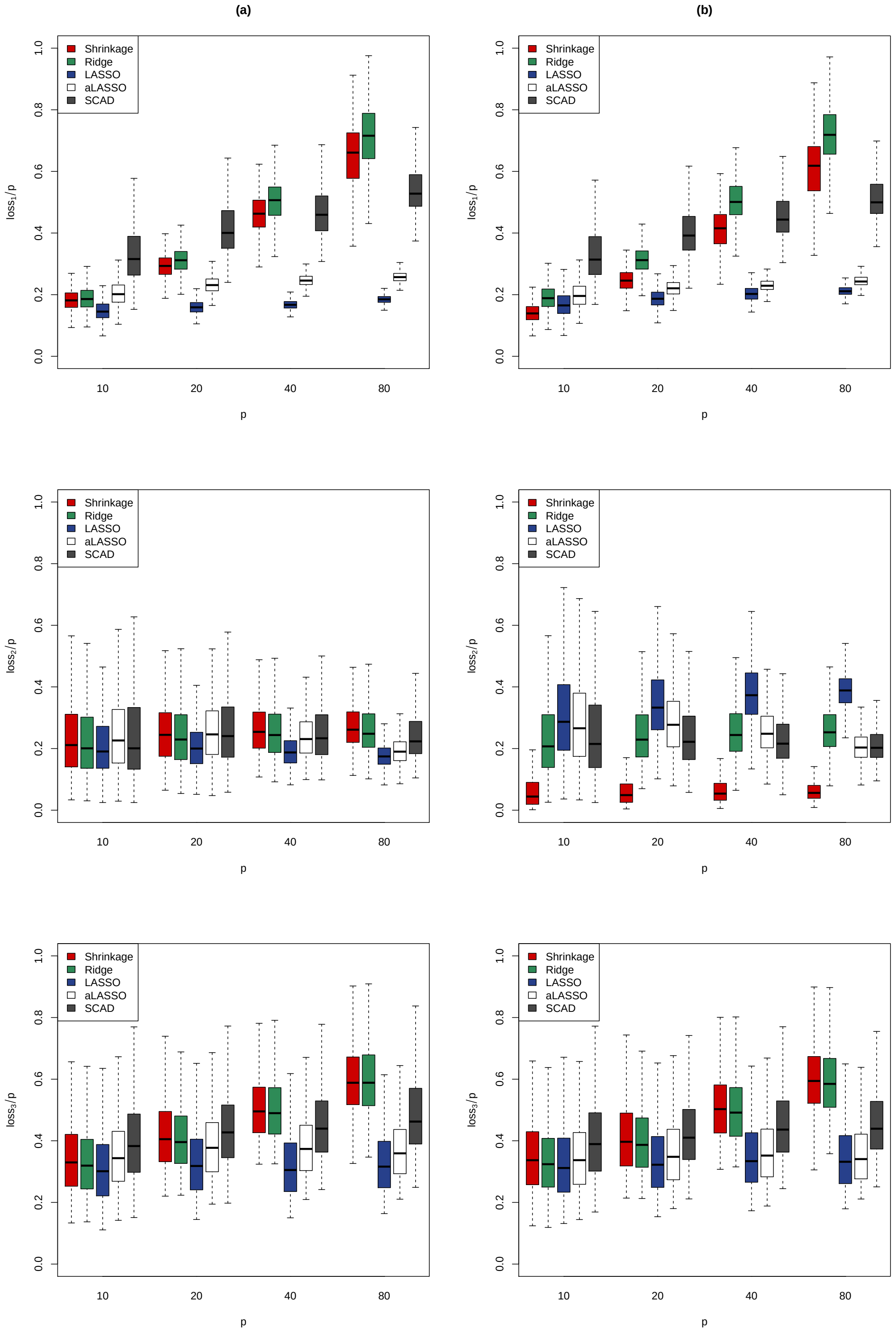


FIGURE 2.4: An exchangeable covariance structure is used with $b = 0.6$. The box-plots show the distributions of the three loss functions for the five competing procedures. The three loss functions are calculated for each of 1000 samples of size $n = 20$ from a multivariate normal distribution with $p \in \{10, 20, 40, 80\}$. Since the shrinkage, lasso, adaptive lasso, and SCAD regularization allow to penalize the diagonal, the diagonal elements in (a) are left unpenalized while in (b) they are penalized for all the four methods. As ridge regularization does not allow to penalize the diagonal elements; therefore, the distributions in (b) are the replicates of the results in (a) for ridge regularization.

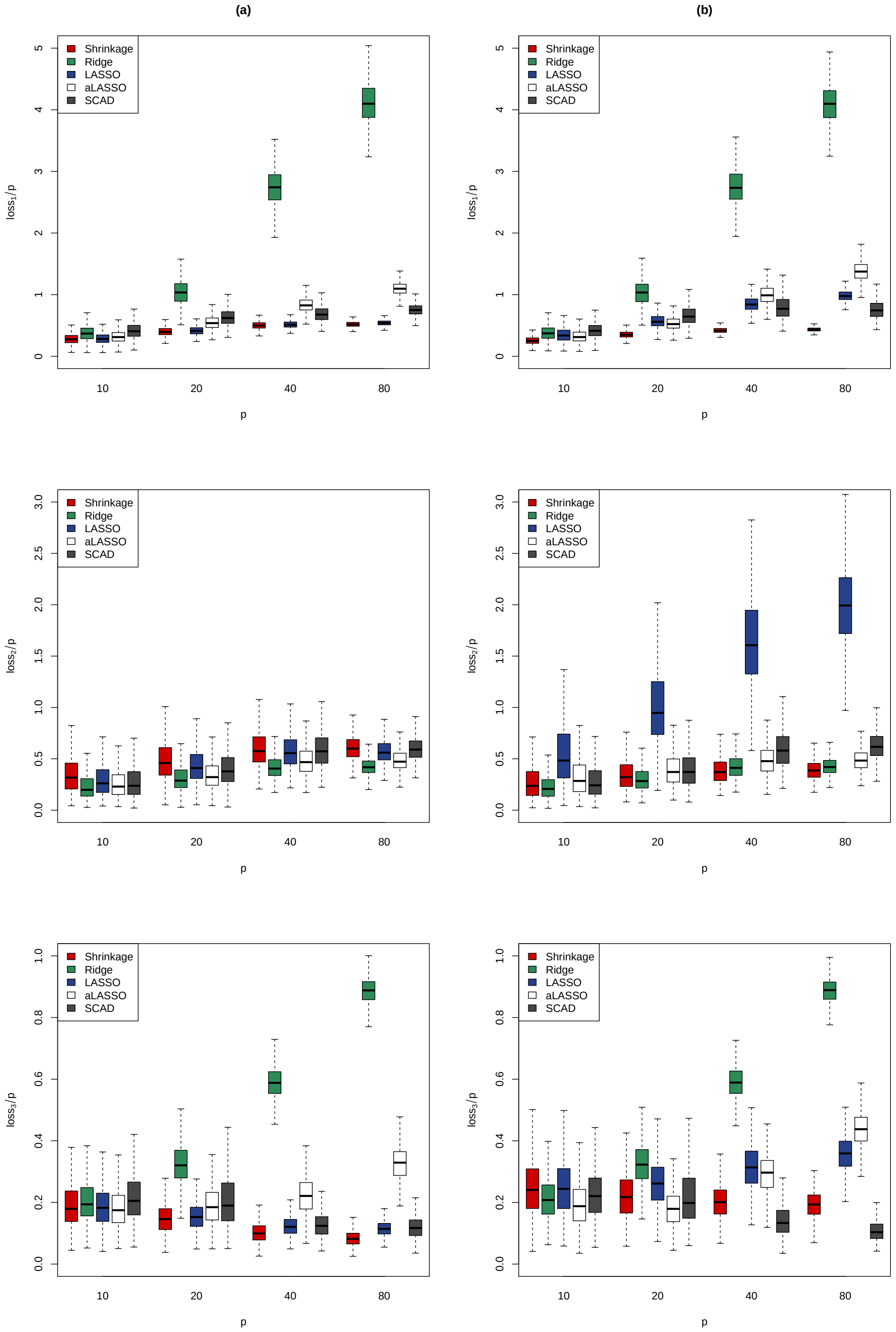


FIGURE 2.5: A random covariance structure is used with the proportion of non-zero edges equals to 50%, 30%, 20%, and 10%, respectively, for p equals to 10, 20, 40, and 80. The box-plots show the distributions of the three loss functions for the five competing procedures. The three loss functions are calculated for each of 1000 samples of size $n = 20$ from a multivariate normal distribution with $p \in \{10, 20, 40, 80\}$. Since the shrinkage, lasso, adaptive lasso, and SCAD regularization allow to penalize the diagonal, the diagonal elements in (a) are left unpenalized while in (b) they are penalized for all the four methods. As ridge regularization does not allow to penalize the diagonal elements; therefore, the distributions in (b) are the replicates of the results in (a) for ridge regularization.

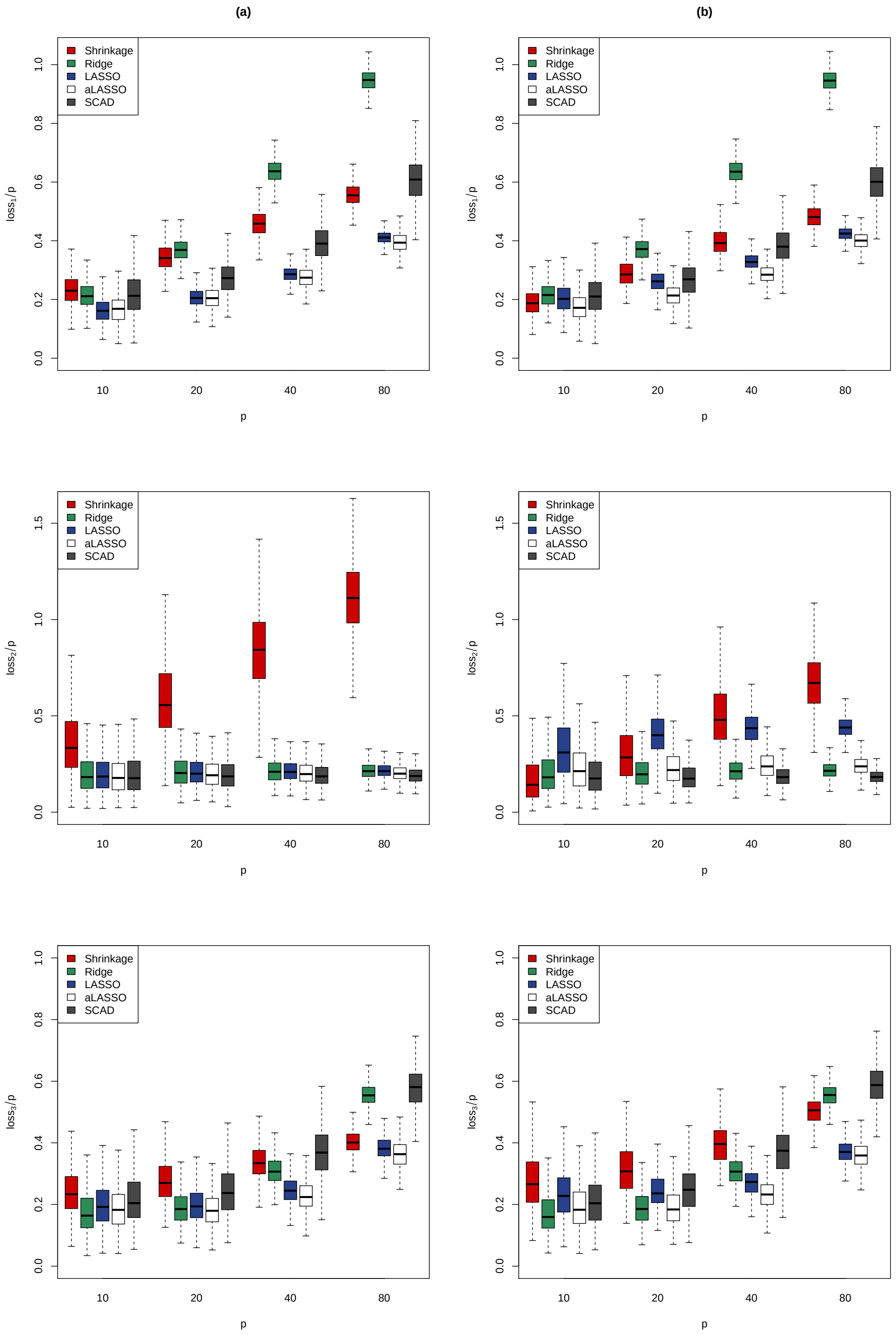


FIGURE 2.6: An AR(1) covariance structure is used with $b = 0.6$. The boxplots show the distributions of the three loss functions for the five competing procedures. The three loss functions are calculated for each of 1000 samples of size $n = 20$ from a multivariate normal distribution with $p \in \{10, 20, 40, 80\}$. Since the shrinkage, lasso, adaptive lasso, and SCAD regularization allow to penalize the diagonal, the diagonal elements in (a) are left unpenalized while in (b) they are penalized for all the four methods.As ridge regularization does not allow to penalize the diagonal elements; therefore, the distributions in (b) are the replicates of the results in (a) for ridge regularization.

Chapter 3

Hierarchical covariance estimation

3.1 Introduction

In many applied problems we come across multivariate data sets that come from multiple groups. One special case is when there are few observations in each group but many groups. An example of this kind can be found in Aitken & Lucy (2004) where multivariate replicate measurements are taken on the elemental composition of glass from different windows. A data set of similar nature has been collected by Bennett (2002): who made twenty replicate measurements of five elements on each of six different Heineken beer bottles. In both of these cases, since the within-group variation is because of measurement error our emphasis is on the between-group variation while controlling for the within-group variation. A multivariate random-effect model has been used by Aitken & Lucy (2004) to summarize the data.

We measure p variables from m different groups and there are r measurements from each group. Denote the $n = mr$ observations by $\mathbf{X}_{ij} = (\mathbf{X}_{ij1}, \mathbf{X}_{ij2}, \cdots, \mathbf{X}_{ijp})^t$, $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, r$. Let $\boldsymbol{\theta}_i$ be the mean vector of the ith group and $\mathbf{U}$ be the within-group covariance matrix. Then, given $\boldsymbol{\theta}_i$ and $\mathbf{U}$, the distribution of $\mathbf{X}_{ij}$ is assumed to be normal with $\mathbf{X}_{ij} \sim N_p(\boldsymbol{\theta}_i, \mathbf{U})$. Similarly, assume that $\boldsymbol{\mu}$ is the between-group mean vector and $\mathbf{C}$ is the between-group covariance matrix. Then $\boldsymbol{\theta}_i$ is assumed to be normally distributed with $\boldsymbol{\theta}_i \sim N_p(\boldsymbol{\mu}, \mathbf{C})$.

Denote the within-group sum-of-squares and cross-products matrix by $\mathbf{S}_w$, then $\mathbf{S}_w$ is given by

$$\mathbf{S}_w = \sum_{i=1}^{m} \sum_{j=1}^{r} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^t, \tag{3.1}$$

where $\bar{\mathbf{X}}_i$ is the mean of the i^{th} group, and the within-group covariance matrix is calculated by

$$\widehat{\mathbf{U}} = \frac{\mathbf{S}_w}{n - m}. \tag{3.2}$$

Similarly, denote the between-group sum-of-squares and cross-products matrix by $\mathbf{S}_b$, then $\mathbf{S}_b$ is obtained by

$$\mathbf{S}_b = \sum_{i=1}^{m} (\bar{\mathbf{X}}_i - \bar{\mathbf{X}})(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})^t, \tag{3.3}$$

where $\bar{\mathbf{X}}$ is the overall mean, and the between-group covariance matrix, $\mathbf{C}$, is estimated using

$$\widehat{\mathbf{C}} = \frac{\mathbf{S}_b}{m - 1} - \frac{\mathbf{S}_w}{r(n - m)}. \tag{3.4}$$

To analyze the data, one needs reliable estimates of both within-group and between-group covariance components. The problem of estimating the large covariance matrices has long been known in the literature of multivariate statistics. The estimates of covariance matrices become ill-conditioned (when n is slightly greater than p) or even singular (when n is less than p). The problem is exacerbated when we estimate the between-group covariance matrix. The standard estimation procedure in (3.4) involves the difference of two mean square matrices: the between-group mean square and the within-group mean square (we refer to this method of estimation as a standard method). For a sufficiently large sample size, both mean squares are individually guaranteed to be nonnegative definite, however, their difference is not and often results negative elements on the diagonal. This is pointed out by Hill & Thompson (1978), who shows that the probability of the negative variances increases with increasing the number of variables and also that how it affects the estimates of genetic covariance matrices. Others have addressed the same problem (see for example, Amemiya (1985), Shaw (1991), Phillips & Arnold (1999)).

To constrain the estimated covariance matrices to the parametric space, regularized alternatives have been proposed over time to deal with the problem for a single population, see for example Bickel & Levina (2007) and the references therein for detailed discussion. Dempster (1972) adopted parsimonious approach to estimate the covariance matrix. He called his approach covariance selection models, where the objective is to identify zero elements in the off-diagonal of the inverse covariance matrix. The resulting estimate is interpretable as well as regularized. Zero elements correspond to the pairs of variables that are conditionally independent given the others. We find this interpretability appealing and pursue an estimate of this type.

Tibshirani (1996) proposed the lasso in the regression setting, a popular model selection and shrinkage estimation method, which has the ability to force some parameters to be exactly zero. This idea was extended by Yuan & Lin (2007) to the likelihood-based estimation of the inverse covariance matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = (\omega_{ij})_{1\leq i,j \leq p}$. The following log-likelihood function based on a random sample of size n from a multivariate normal distribution $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$ has been optimized:

$$l(\mathbf{\Omega}) = Const - \frac{n}{2} \log |\mathbf{\Sigma}| - \frac{n}{2} tr\left(\widehat{\mathbf{\Sigma}}\mathbf{\Omega}\right) - \rho \sum_{i \neq j} |\omega_{ij}| , \tag{3.5}$$

where $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^t$ and ρ is the penalty parameter which shrinks some coefficients towards zero and sets the others as exactly zero. Larger value of ρ produces sparse solutions. Smaller value of ρ, on the other hand, encourages more non-zero off-diagonal elements to appear in the inverse covariance matrix. Traditionally, cross-validation, an automatic data-driven method, has been used to select the optimal value of the tuning parameter ρ (Friedman et al., 2008).

In this Chapter, our objective is to summarize a high-dimensional data set in which the same set of variables is measured in many different groups. The method of penalized likelihood is extended to the hierarchical covariance estimation via the EM algorithm in order to get reliable estimates for both the within-group and the between-group covariance structure. Our primary reason for using EM algorithm is to avoid the difference of two covariance components while calculating the between-group covariance. We use a positive definite estimate of the between-group covariance matrix as an initial estimate and update in such a way that the matrices remain positive definite. The second reason for using EM algorithm is that it allows us to use regularized estimate of the between-group covariance as we will see in the next section.

3.2 Hierarchical covariance estimation via EM algorithm

Since, $\mathbf{X}_{ij} \sim N_p(\boldsymbol{\theta}_i, \mathbf{U})$ and $\boldsymbol{\theta}_i \sim N_p(\boldsymbol{\mu}, \mathbf{C})$, the joint density can be written as:

$$f(\mathbf{X}_{i1}, \mathbf{X}_{i2}, ..., \mathbf{X}_{ir}, \boldsymbol{\theta}_i) = \prod_j f(\mathbf{X}_{ij}|\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i). \tag{3.6}$$

Using (3.6), the conditional distribution of $\boldsymbol{\theta}_i$ given $\mathbf{X}_{i1}, \mathbf{X}_{i2}, ..., \mathbf{X}_{ir}$ can be derived as follows:

$$\begin{aligned} f(\boldsymbol{\theta}_i|\mathbf{X}_{i1}, \mathbf{X}_{i2}, ..., \mathbf{X}_{ir}) \propto \exp\Bigg\{ -\frac{1}{2}\Big\{ \sum_j (\mathbf{X}_{ij} - \boldsymbol{\theta}_i)^t \mathbf{U}^{-1} (\mathbf{X}_{ij} - \boldsymbol{\theta}_i) + \\ (\boldsymbol{\theta}_i - \boldsymbol{\mu})^t \mathbf{C}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu}) \Big\} \Bigg\}. \end{aligned} \tag{3.7}$$

Expending the exponents and completing the quadratic form for $\boldsymbol{\theta}_i$ gives

$$\begin{aligned} f(\boldsymbol{\theta}_i|\mathbf{X}_{i1}, \mathbf{X}_{i2}, ..., \mathbf{X}_{ir}) &\propto \exp\left(-\frac{1}{2} (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)^t \left(\mathbf{C}^{-1} + r\mathbf{U}^{-1}\right)^{-1} (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i) \right) \\ &\sim N_p(\tilde{\boldsymbol{\theta}}_i, \mathbf{C}^{-1} + r\mathbf{U}^{-1}) \end{aligned} \tag{3.8}$$

where

$$\tilde{\boldsymbol{\theta}}_i = \left(\mathbf{C}^{-1} + r\mathbf{U}^{-1}\right)^{-1} \left(\mathbf{C}^{-1}\boldsymbol{\mu} + r\mathbf{U}^{-1}\bar{\mathbf{X}}_i\right). \tag{3.9}$$

An EM algorithm iterates between the Expectation-step and the Maximization-step. It is designed to compute the maximum likelihood estimate in the presence of incomplete data. In the Expectation-step, the missing data is estimated using the conditional distribution of the missing data given the observed data. In the Maximization-step, new estimates of the parameters are computed using the complete data from the Expectation-step. The process is repeated until a convergence is obtained that is when the change in the parameter estimates is smaller than a given specified threshold. For a detailed description of the EM algorithm see Dempster et al. (1977).

The EM algorithm has been used for the estimation of covariance components. An example of using EM algorithm to estimate the variance components in a univariate random effect model is discussed in Dempster et al. (1977). An iterative REML estimation procedure (Harville, 1977) is described by Calvin & Dykstra (1991) for balanced multivariate variance components models. Calvin (1993) extend the approach to estimate the covariance components in a general unbalaced multivariate mixed model, using an EM algorithm. He utilizes the algorithm of Calvin & Dykstra (1991) in the Maximization-step of the algorithm. In the Expectation-step, he consider the missing obsevations, the data that is needed to make the problem balanced and the conditional distribution of the missing data given the observed data is used to estimate the covariance components for the completed data set.

We present am EM algorithm that works in high-dimensional setting. Unlike Calvin (1993), who uses REML estimation procedure, we use lasso-regularized estimator to restrict the covariance components to the parametric space. Let k denotes the kth iteration for $k = 0, 1, 2, \ldots, \infty$, with $k = 0$ denotes the initialized values before the first iteration. The maximum likelihood estimate of the overall mean vector, $\boldsymbol{\mu}$, is given by

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} = \frac{1}{mr}\sum_{i=1}^{m}\sum_{j=1}^{r}\mathbf{X}_{ij}, \tag{3.10}$$

and the mean vector for the ith group, $\boldsymbol{\theta}_i$, is initialized as

$$\hat{\boldsymbol{\theta}}_i^{(0)} = \bar{\mathbf{X}}_i = \frac{1}{r}\sum_{j=1}^{r}\mathbf{X}_{ij}. \tag{3.11}$$

The estimate of the within-group covariance matrix, $\mathbf{U}$, at the kth iteration is obtained by

$$\widehat{\mathbf{U}}^{(k)} = \frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{r}(\mathbf{X}_{ij} - \hat{\boldsymbol{\theta}}_i^{(k)})(\mathbf{X}_{ij} - \hat{\boldsymbol{\theta}}_i^{(k)})^t. \tag{3.12}$$

The estimate of the between-group mean square is used as an estimate for the between-group covariance at the kth iteration, given by

$$\widehat{\mathbf{C}}^{(k)} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\boldsymbol{\theta}}_i^{(k)} - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\theta}}_i^{(k)} - \hat{\boldsymbol{\mu}})^t. \tag{3.13}$$

The estimated mean square, $\widehat{\mathbf{C}}$, obtained using (3.13) is likely to be rank deficient, or nearly so, if m and p are comparable. To avoid this we use a regularized estimate obtained by maximizing the penalized likelihood function

$$l(\mathbf{C}^{-1}) = Const - \frac{m}{2}\log|\mathbf{C}| - \frac{m}{2}tr\left(\widehat{\mathbf{C}}\mathbf{C}^{-1}\right) - \rho\sum_{i\neq j}\left|c_{ij}^{-1}\right|, \tag{3.14}$$

where c_{ij}^{-1} is the ijth element of $\mathbf{C}^{-1}$ and $\widehat{\mathbf{C}}$ as in (3.13). We use graphical lasso algorithm proposed by Friedman et al. (2008) to solve the problem presented in (3.14) and use cross-validation to choose ρ as described in Section 2.4.4. This provides a positive definite estimate of the between-group mean square. Note that, the optimum value of ρ is allowed to change over different iterations. The algorithm terminates at a local optima (see Figure 3.2).

In the **E-step**, given $\mathbf{X}$ and the estimates of $\boldsymbol{\mu}$, $\mathbf{U}$, and $\mathbf{C}$, we find the conditional expectation of the distribution of $\boldsymbol{\theta}_i$, as given in (3.9). At the **M-step**, we assume that $\boldsymbol{\theta}_i$ is known, which allows us to update the value of $\widehat{\mathbf{U}}$ using (3.12) and estimate $\mathbf{C}$ by maximizing $l(\mathbf{C}^{-1})$ given in (3.14). Here is the detailed description of the algorithm.

1. Initialize $\boldsymbol{\theta}_i$ as $\bar{\mathbf{X}}_i$ and estimate $\boldsymbol{\mu}$ using (3.10).
2. Obtain the estimate of $\mathbf{U}$ using (3.12).
3. Obtain the estimate of $\mathbf{C}$ from (3.14), using cross validation to select the value of ρ (see Figure 3.2).
4. Update the estimate of $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}^{(k)}$ using (3.9).
5. Repeat step 2-4 until convergence.

The algorithm terminates when $|l(\mathbf{C}^{(k+1)}) - l(\mathbf{C}^{(k)})|/|l(\mathbf{C}^{(k+1)})| < \tau$, where $l(\mathbf{C}^{(k)})$ is the penalized log-likelihood value at the kth iteration as in (3.14) and $\tau = 10^{-5}$.

3.3 Applications

In this section, we assess the performance of the method and report the results of our simulation study and a real data analysis.

3.3.1 Numerical experiments

In our simulation experiments, we first generate $\mathbf{U}$ and $\mathbf{C}$ using the random method of Schäfer & Strimmer (2005a), which guarantees the generated matrix to be positive definite. The method proceeds as follows:

1. Start with a null $p \times p$ matrix.
2. Choose randomly a suitable proportion of the off-diagonal positions (at least one position in each column), and fill them symmetrically with correlation values drawn from the uniform distribution between -1 and 1.
3. Set the rest of the off-diagonal elements as zero. To generate high-dimensional sparse inverse covariance matrices a smaller proportion of the off-diagonal positions would be required to fill in with non-zero elements in step 2.
4. Sum up the absolute values for each column plus a small constant and fill them in their respective diagonal positions. This is our inverse covariance matrix, $\mathbf{\Omega}$.
5. The inverse of $\mathbf{\Omega}$ is the true covariance matrix, $\mathbf{\Sigma}$.

The m p-dimensional group means, $\boldsymbol{\theta}_i$, are drawn from a multivariate normal distribution i.e. $\boldsymbol{\theta}_i \sim N_p(\boldsymbol{\mu}, \mathbf{C})$. The r observations, $\mathbf{X}_{ij}$, for each group are then generated from a multivariate normal distribution i.e. $\mathbf{X}_{ij} \sim N_p(\boldsymbol{\theta}_i, \mathbf{U})$. We consider three types of situations while taking into account the nature of the within-group and the between-group covariance matrices. At first, we examine the performance of the model in the situation when the between-group variation is dominant and the within-group covariance is relatively on smaller scale. This is the easiest case in the sense that the EM algorithm converges quickly and the estimated covariances are relatively closer to the true covariances. This is obtained by scaling down $\mathbf{U}$ in (3.15) and (3.17) by a factor of 1/30. In the second example, we allow the within-group covariance to spread out from a relatively smaller to a moderate size (we use $\mathbf{U}$ as it is in (3.15) and (3.17)). In the third case, we make the within-group variation dominant, which is comparatively hard case and the algorithm needs few more iterations to converge. This is achieved by rescaling $\mathbf{U}$ in (3.15) and (3.17) by 10. Note that, we keep $\mathbf{C}$ fixed as in (3.16) and (3.18), and rescale $\mathbf{U}$ to produce data of the three cases. The three cases: easiest, moderate, and hard, are illustrated in Figure 3.1, where we plot the data on the first two principal components with 30 groups represented by numbers 1-30 and with different colors. The number observations in each group are 20.

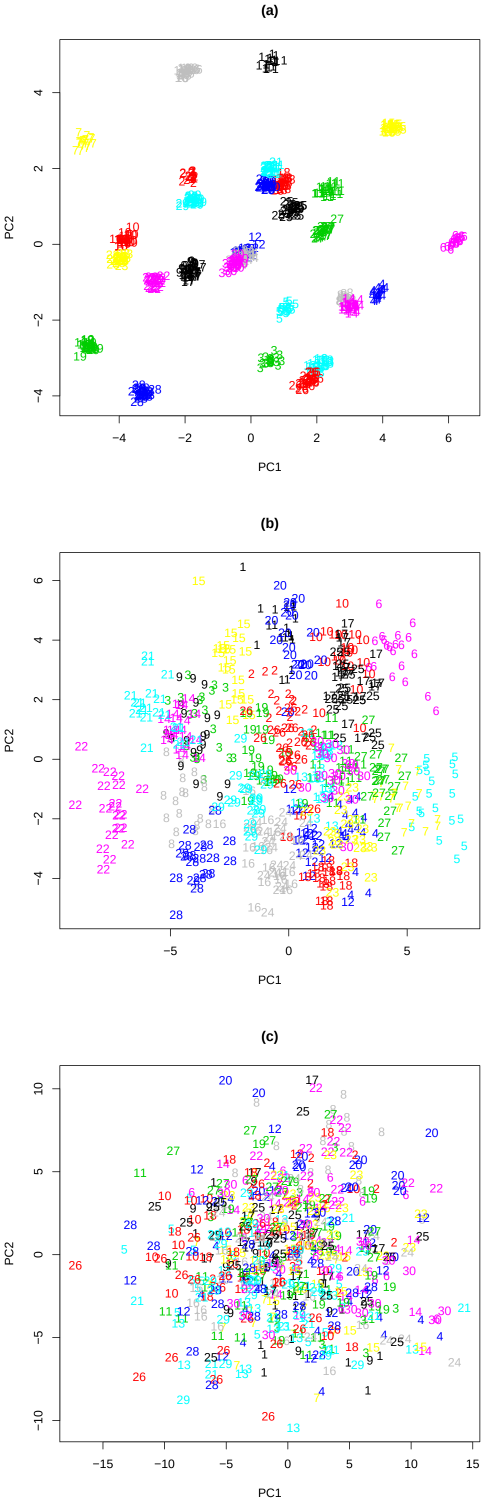


FIGURE 3.1: Simulated data on first two principle components for the three cases (a) easiest, (c) moderate, and (e) hard.

Two other factors that have the potential to affect the behavior of the model are the number of groups, m, and the number of observations in each group, r. We did the experiments for different combinations of m and r and report the results for $m \in \{10, 30\}$ and $r \in \{5, 20\}$. For each combination of m and r, we generate 1000 data sets and estimate the between-group covariance using the proposed method, therefore, obtain 1000 estimates for each of the 15 different elements (5 diagonal and 10 off-diagonal) of $\mathbf{C}$.

Figure 3.2 shows the typical cross-validation score obtained over the first five iterations of the proposed EM algorithm. The cross-validation scores becomes stable as the algorithm converges to a stationary value.

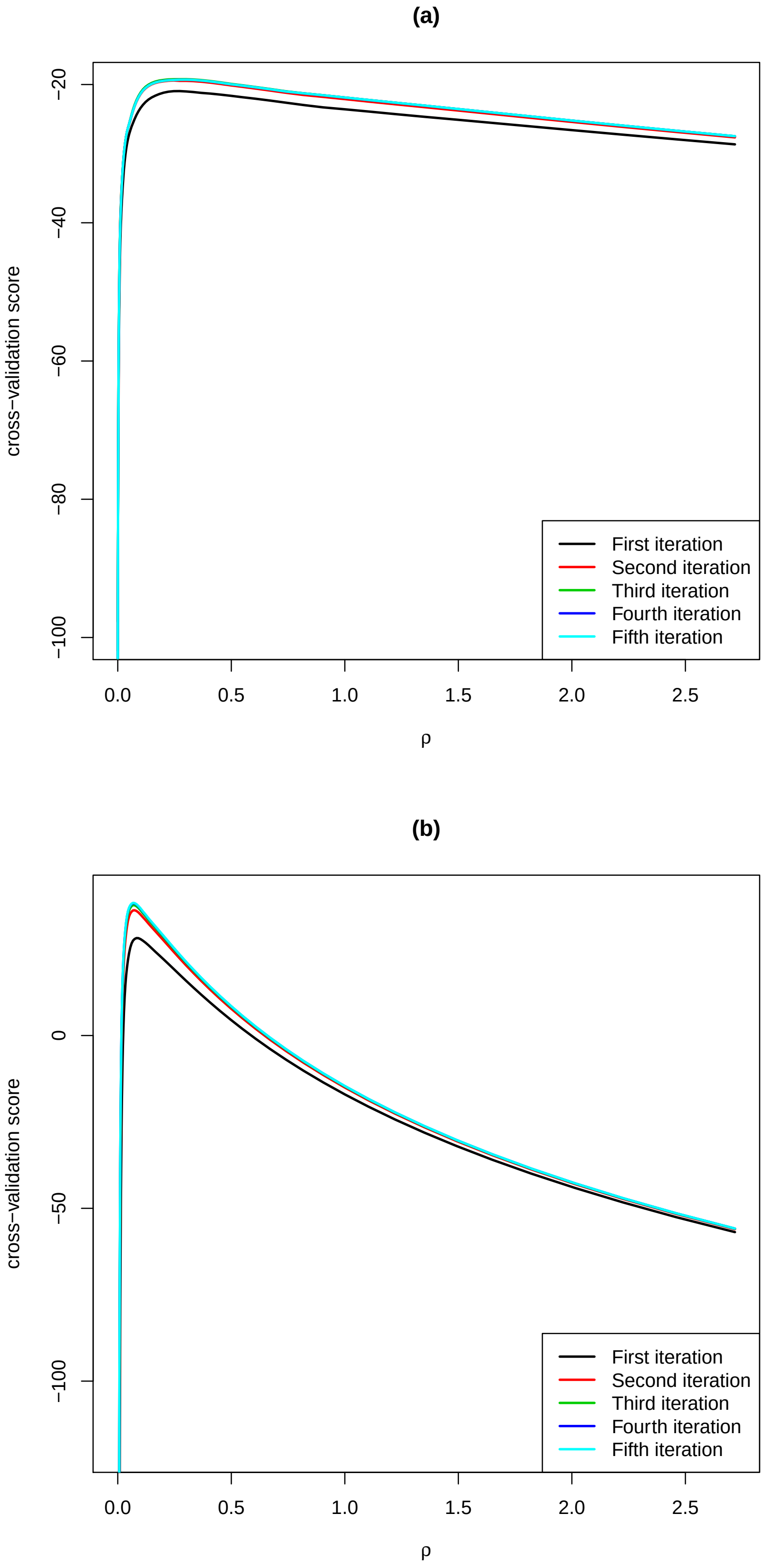


FIGURE 3.2: Typical cross-validation scores obtained over the first five iterations of the proposed EM algorithm. We use 100% non-zero off-diagonal elements in $\mathbf{U}^{-1}$ and $\mathbf{C}^{-1}$. The elements of $\mathbf{U}$ are of a moderate size. In (a) $p = 5$, $m = 10$, $r = 5$ and in (b) $p = 15$, $m = 10$, $r = 5$.

We judge the performance of the method in the following two ways:

1. First, we see how well the method estimates the elements of a between-group covariance matrix. Since the results are consistent across different choices of $\mathbf{U}$ and $\mathbf{C}$, generated using the above method, we give the results for a single draw of $\mathbf{U}$ and $\mathbf{C}$ with the dimension $p = 5$ and 100% non-zero elements in the $\mathbf{U}^{-1}$ and $\mathbf{C}^{-1}$, given by

$$\mathbf{U} = \begin{pmatrix} 0.74 & \text{-} & \text{-} & \text{-} & \text{-} \\ 0.41 & 0.90 & \text{-} & \text{-} & \text{-} \\ \text{-}0.49 & \text{-}0.24 & 1.12 & \text{-} & \text{-} \\ \text{-}0.09 & \text{-}0.38 & \text{-}0.05 & 0.71 & \text{-} \\ \text{-}0.46 & \text{-}0.50 & 0.35 & 0.31 & 0.85 \end{pmatrix} \tag{3.15}$$

and

$$\mathbf{C} = 5 \times \begin{pmatrix} 0.80 & \text{-} & \text{-} & \text{-} & \text{-} \\ \text{-}0.09 & 1.09 & \text{-} & \text{-} & \text{-} \\ 0.04 & \text{-}0.28 & 0.67 & \text{-} & \text{-} \\ \text{-}0.34 & 0.20 & 0.13 & 0.78 & \text{-} \\ 0.26 & 0.14 & \text{-}0.33 & \text{-}0.31 & 0.73 \end{pmatrix}. \tag{3.16}$$

Figures 3.3, 3.4, and 3.5 present the results for how well the proposed method estimate the elements of $\mathbf{C}$ in the three cases. The elements of $\mathbf{C}$, as given in (3.16), are estimated using 1000 different data sets for each combination of m and r. In all the three cases, as we increase m, the variability of the estimates decreases and the accuracy increases. Increasing the value of r (from 5 to 20) does not make a noticeable difference in the variability of the estimates, however, it does increase the accuracy of the estimated elements of the covariance matrix. As expected, the estimates becomes more inaccurate and less precise for the hard case as compared to the easy and moderate cases. The algorithm sometime does not reach the convergence if the diagonal elements are not penalized. This happens more frequently in the hard case. To avoid the convergence problem we penalize the diagonal elements in our simulations. The reason for underestimation of the diagonal elements in Figure 3.5 is that the lasso penalty heavily penalize the larger elements (that are usually the diagonal element in our simulation setup). This suggest to replace the lasso penalty by the adaptive lasso or SCAD penalty in the hard case, especially when m is very small.

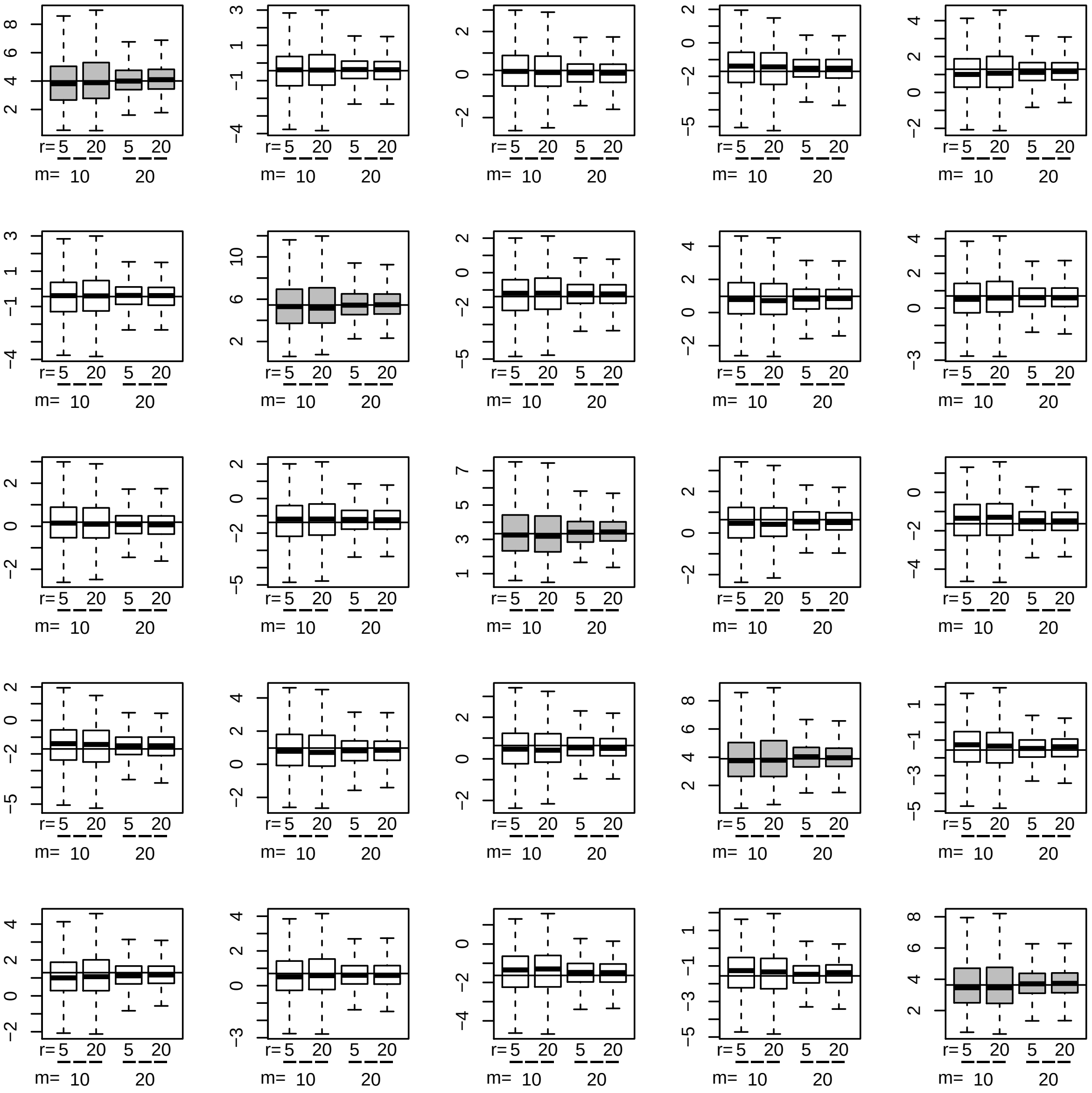


Figure 3.3: Estimates of different elements of the between-group covariance matrix in an easy case. Each box-plot is made up of 1000 estimates of the same element using 1000 different data sets. The gray boxes represent the estimate of five diagonal elements and the horizontal line in each panel represents the true value.

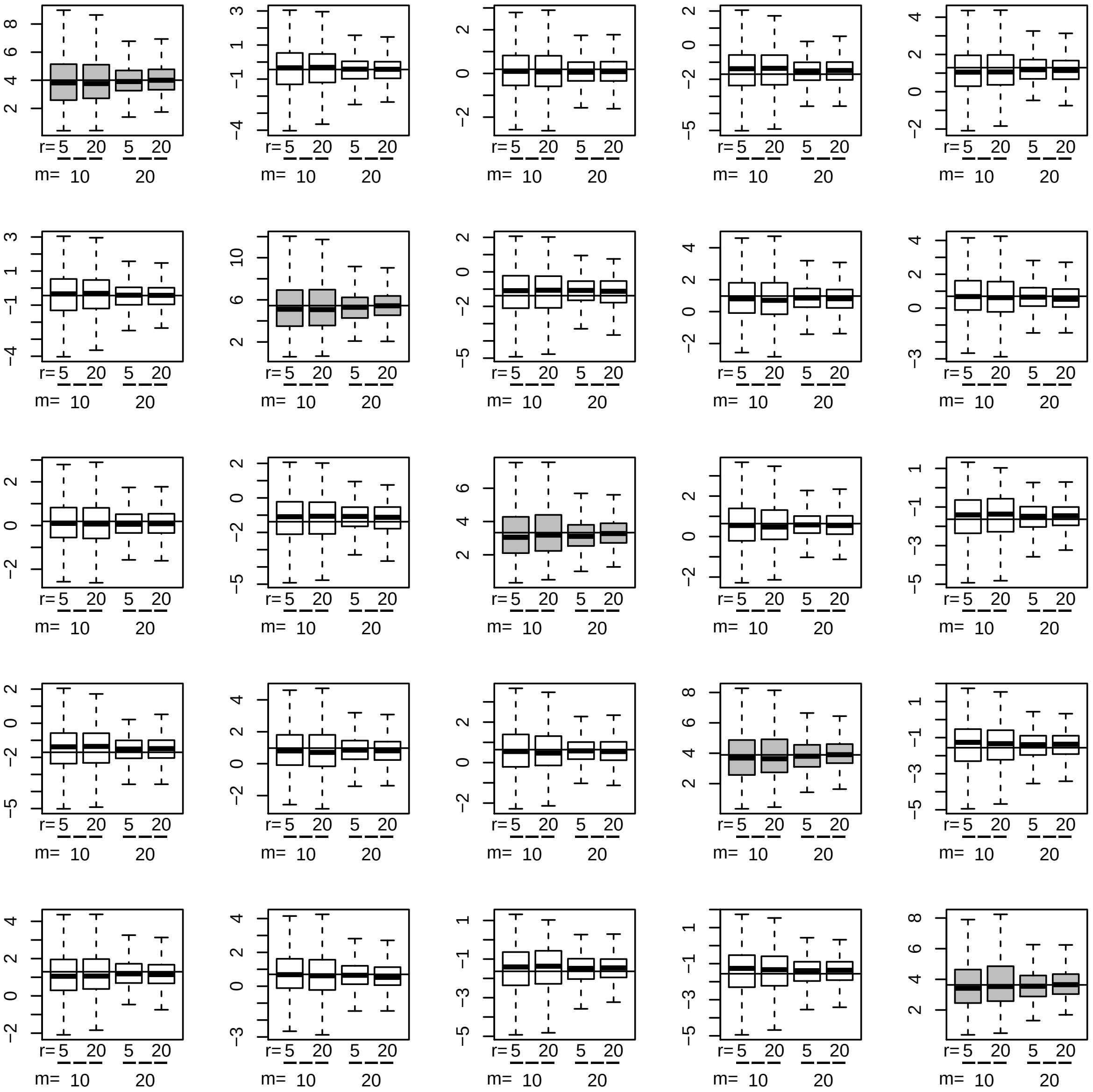


FIGURE 3.4: Estimates of different elements of the between-group covariance matrix in a moderate case. Each box-plot is made up of 1000 estimates of the same element using 1000 different data sets. The gray boxes represent the estimate of the five diagonal elements and the horizontal line in each panel represents the true value.

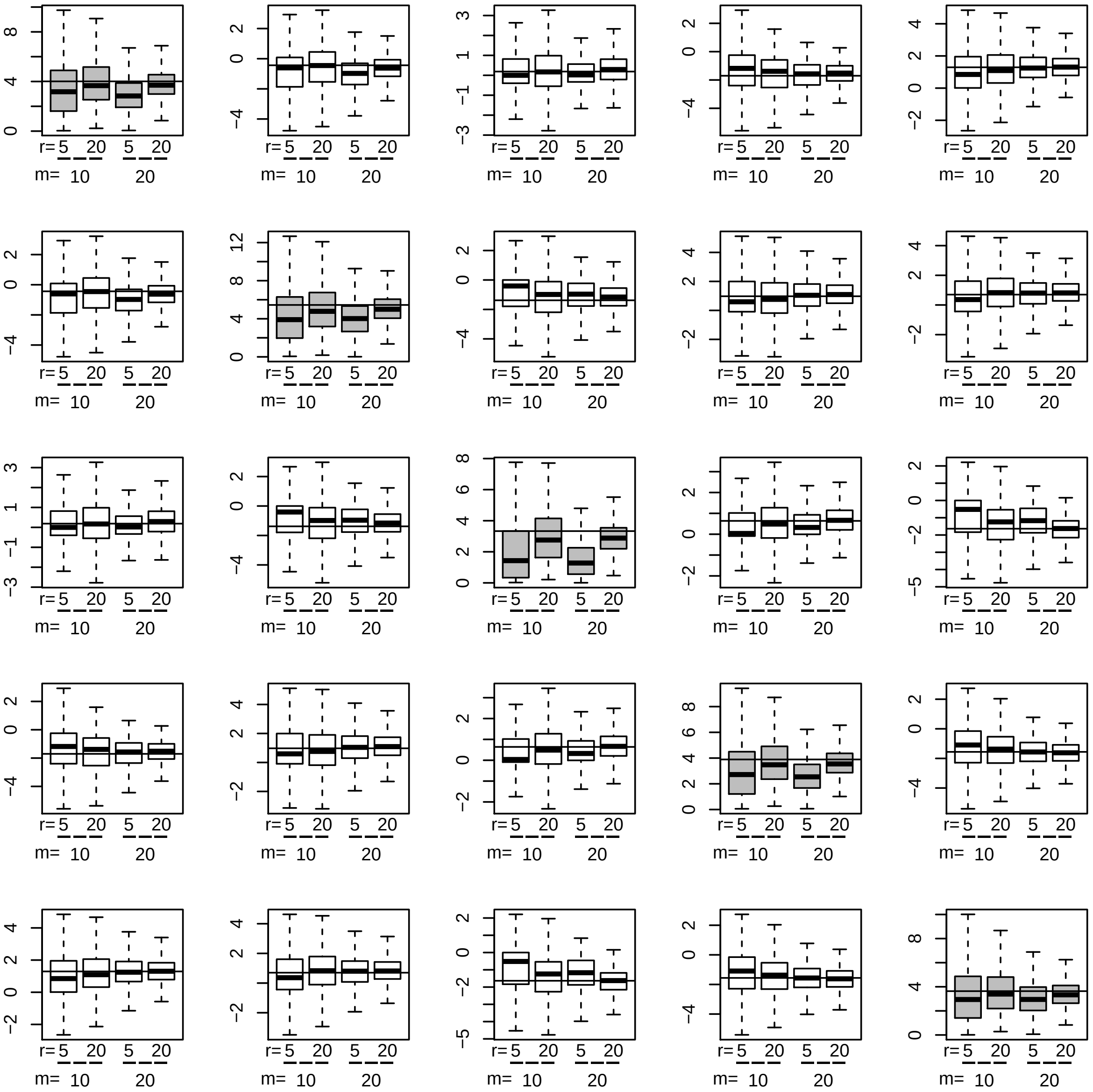


FIGURE 3.5: Estimates of different elements of the between-group covariance matrix in a hard case. Each box-plot is made up of 1000 estimates of the same element using 1000 different data sets. The gray boxes represent the estimate of five diagonal elements and the horizontal line in each panel represents the true value.

2. Here we compare the new method with the standard method in how well they estimate the true eigenvalues. We use another single draw of **U** and **C** with the dimension $p = 10$ and 50% non-zero elements in the U^{-1} and C^{-1},

given by

$$\mathbf{U} = \begin{pmatrix} 0.61 & - & - & - & - & - & - & - & - & - \\ -0.21 & 0.43 & - & - & - & - & - & - & - & - \\ 0.04 & -0.03 & 0.45 & - & - & - & - & - & - & - \\ -0.02 & 0.06 & -0.14 & 0.67 & - & - & - & - & - & - \\ -0.00 & -0.12 & 0.02 & -0.03 & 0.58 & - & - & - & - & - \\ -0.02 & 0.11 & -0.10 & 0.23 & -0.02 & 0.43 & - & - & - & - \\ -0.04 & -0.02 & -0.09 & 0.07 & 0.10 & 0.04 & 0.36 & - & - & - \\ -0.01 & -0.02 & 0.12 & -0.09 & 0.12 & 0.00 & 0.06 & 0.37 & - & - \\ -0.24 & 0.21 & -0.05 & 0.04 & -0.08 & 0.05 & 0.09 & -0.00 & 0.53 & - \\ 0.28 & -0.07 & 0.06 & 0.03 & -0.01 & 0.08 & -0.03 & -0.02 & -0.11 & 0.79 \end{pmatrix} \tag{3.17}$$

and

$$\mathbf{C} = 5 \begin{pmatrix} 1.60 & - & - & - & - & - & - & - & - & - \\ -0.50 & 1.26 & - & - & - & - & - & - & - & - \\ 0.56 & 0.19 & 1.26 & - & - & - & - & - & - & - \\ 0.99 & -0.75 & 0.31 & 1.27 & - & - & - & - & - & - \\ -1.18 & 0.22 & -0.84 & -0.72 & 1.68 & - & - & - & - & - \\ 0.56 & 0.19 & 1.25 & 0.30 & -0.84 & 4.06 & - & - & - & - \\ 0.10 & 0.65 & 0.78 & -0.15 & -0.38 & 0.78 & 2.20 & - & - & - \\ -0.56 & -0.19 & -1.25 & -0.30 & 0.84 & -4.05 & -0.78 & 5.58 & - & - \\ -0.78 & 0.29 & -0.77 & -0.79 & 0.78 & -0.77 & -0.31 & 0.77 & 1.81 & - \\ 0.14 & -0.89 & -0.55 & 0.39 & 0.14 & -0.55 & -1.59 & 0.55 & 0.07 & 2.12 \end{pmatrix}. \tag{3.18}$$

The results for how well the proposed method perform in comparison with the standard procedure are shown in Figures 3.6, 3.7, and 3.8. We estimate the eigenvalues of C, as given in (3.18), using 1000 different data sets for each combination of m and r. In each panel of the three figures, the box-plots for the smallest eigenvalues go below the zero line for the standard method. It happens for more eigenvalues as we go from easy to hard case and for less eigenvalues as we increase either the value of r or m (especially m). This makes it clear that the estimated between-group covariance obtained using the standard procedure often appears with negative eigenvalues. The new method not only provide positive definite estimate of the between-group covariance (non of the box-plots for the smallest eigenvalues goes below the zero line), but also gives improved estimates of the true eigenvalues that is the estimated eigenvalues are closer to true eigenvalues on the average (see Tables 3.1, 3.2, and 3.3).

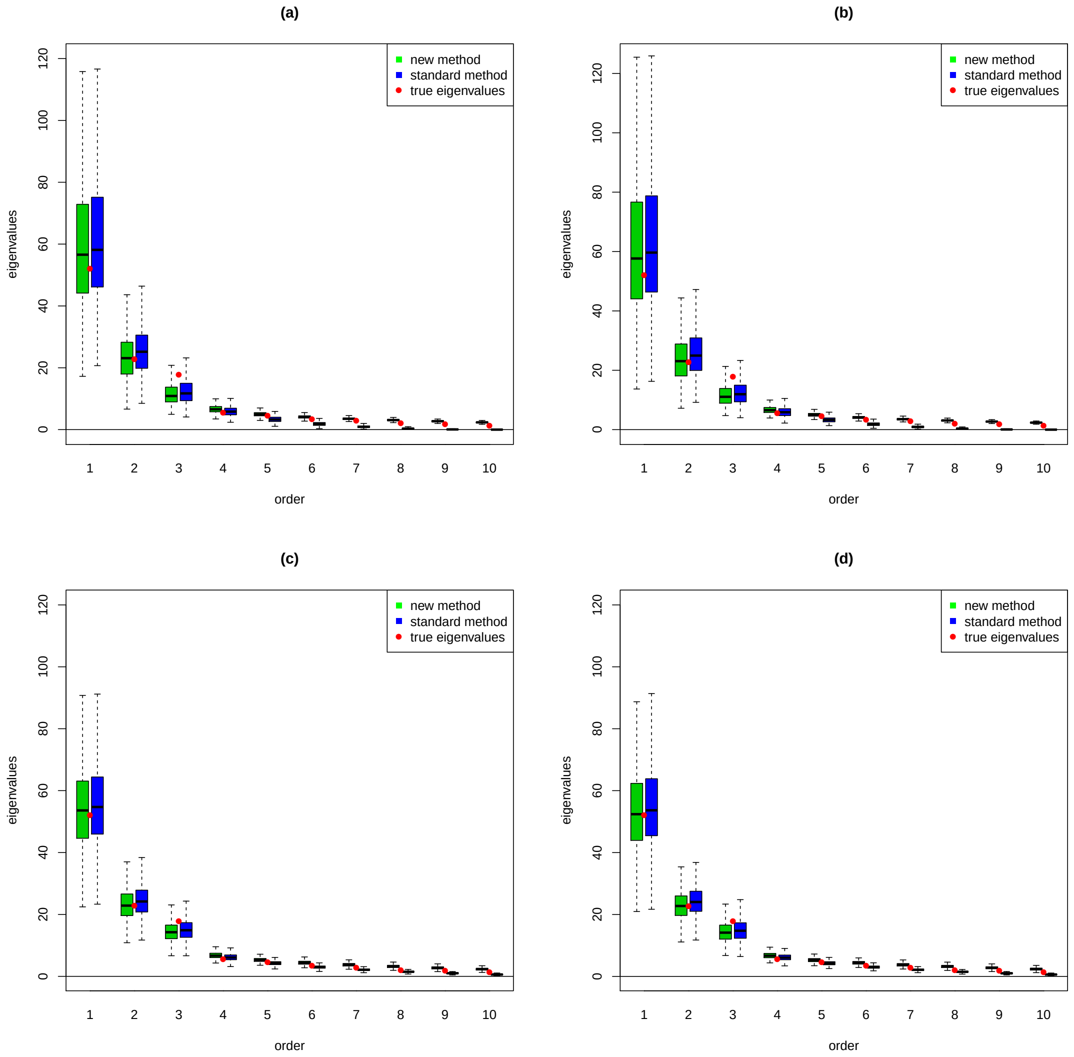


FIGURE 3.6: The distributions of the estimated eigenvalues of a between-group covariance matrix with 1000 different data sets. We use **C** as in (3.17) and **U** in (3.15) is scaled to the easy case: (a) $m = 10$ and $r = 5$, (b) $m = 10$ and $r = 20$, (c) $m = 30$ and $r = 5$, and (d) $m = 30$ and $r = 20$. See the mean squared errors of the estimated eigenvalues in Table 3.1.

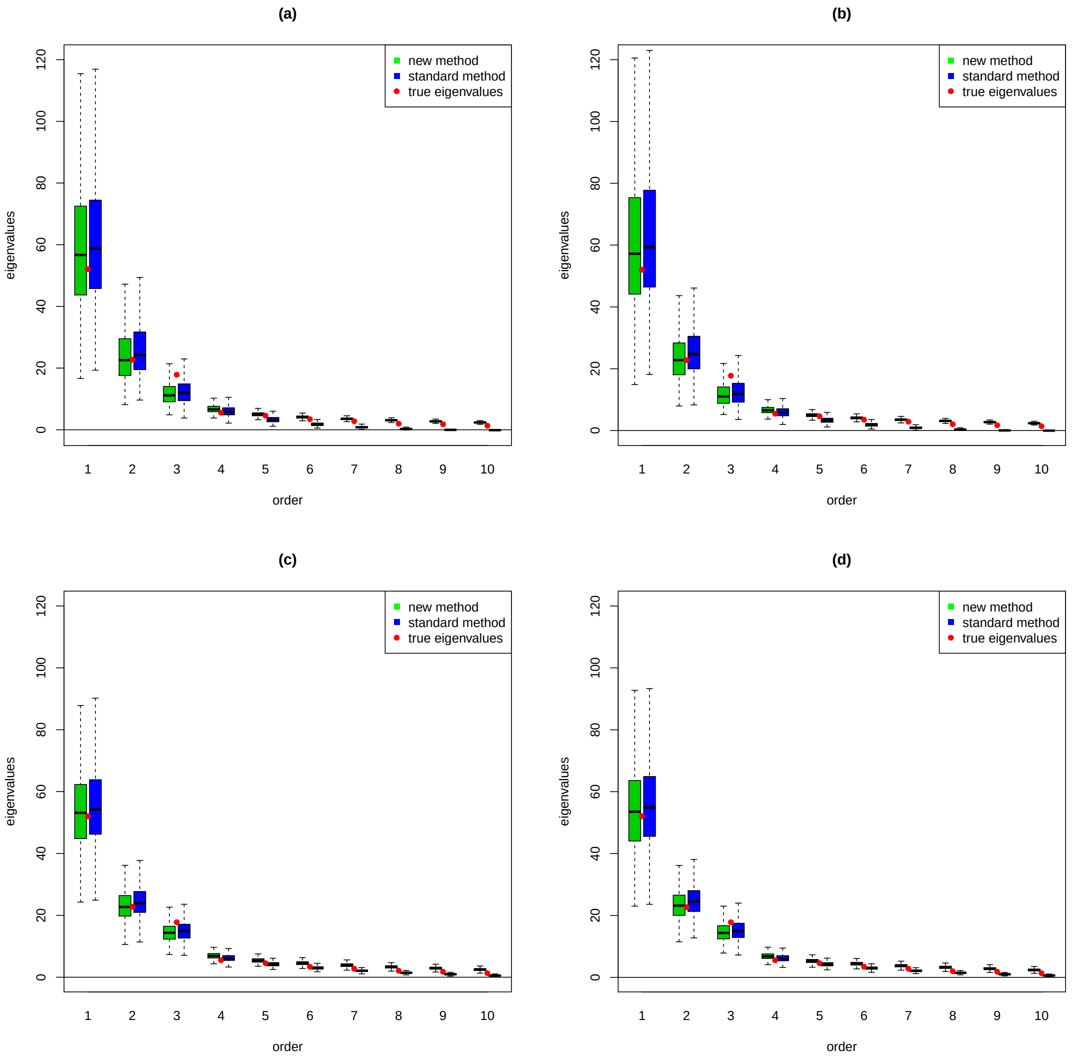


FIGURE 3.7: The distributions of the estimated eigenvalues of a between-group covariance matrix with 1000 different data sets. We use $\mathbf{C}$ as in (3.17) and $\mathbf{U}$ in (3.15) is scaled to the moderate case: (a) $m = 10$ and $r = 5$, (b) $m = 10$ and $r = 20$, (c) $m = 30$ and $r = 5$, and (d) $m = 30$ and $r = 20$. See the mean squared errors of the estimated eigenvalues in Table 3.2.

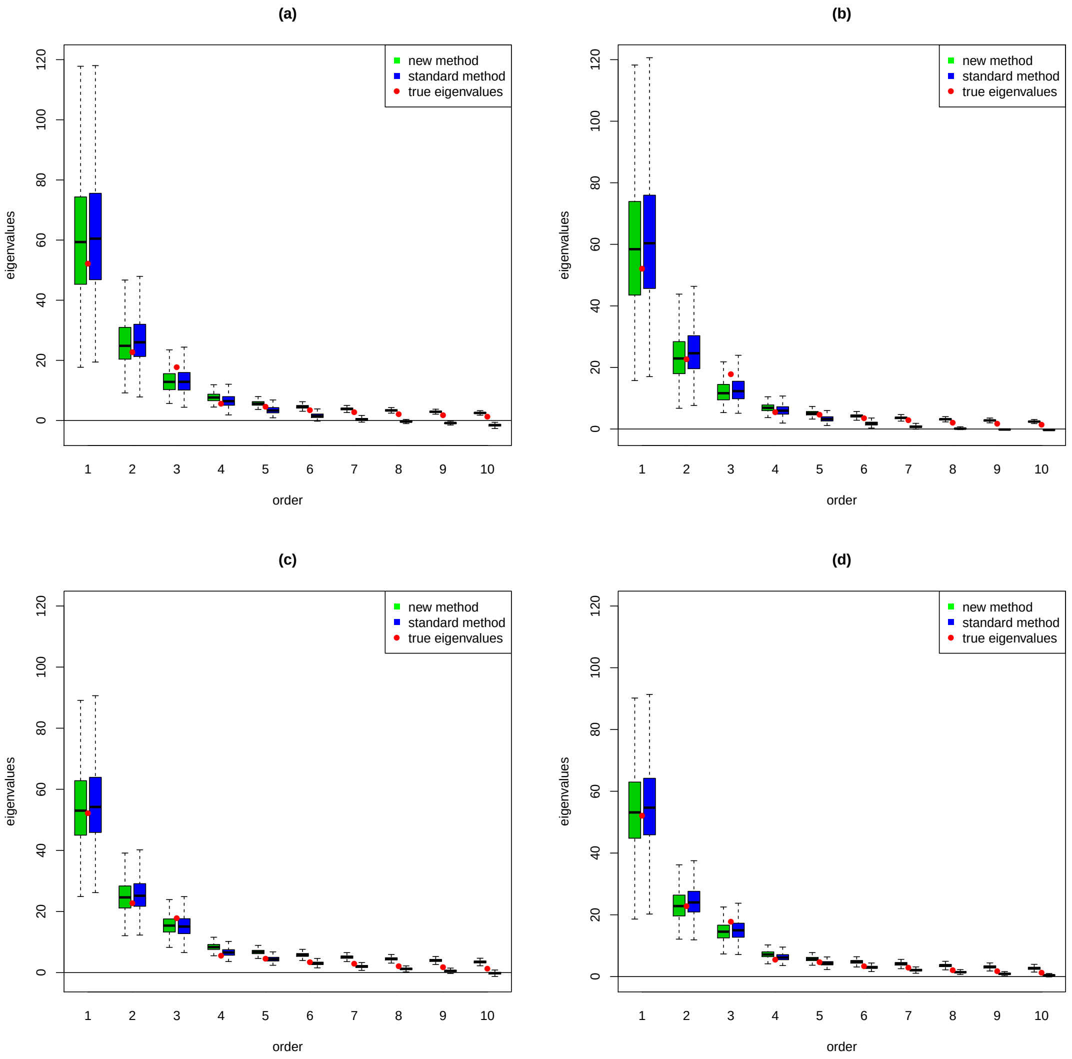


FIGURE 3.8: The distributions of the estimated eigenvalues of a between-group covariance matrix with 1000 different data sets. We use **C** as in (3.17) and **U** in (3.15) is scaled to the hard case: (a) $m = 10$ and $r = 5$, (b) $m = 10$ and $r = 20$, (c) $m = 30$ and $r = 5$, and (d) $m = 30$ and $r = 20$. See the mean squared errors of the estimated eigenvalues in Table 3.3.

3.3.2 Real data example

To further evaluate the proposed method, we use glass chemical composition data collected by Bennett (2002) and is also available in R packages “Hotelling” and “dafs”. The data are the measurements of elemental concentration of the five different elements namely Manganese, Barium, Strontium, Zirconium, and Titanium. Twenty replicate measurements are taken from six different Heineken beer bottles. Thus, there are 5 variables measured in 6 different groups with 20 replicates in

TABLE 3.1: Mean squared errors of the estimated eigenvalues presented in Figure 3.6.

m	r	method	order of the eigenvalues									
			1	2	3	4	5	6	7	8	9	10
10	5	standard	334.2	34.6	7.4	2.8	5.6	6.8	6.7	3.5	2.3	1.2
		proposed	275.6	26.4	5.4	2.3	0.9	0.3	0.2	0.7	1.2	1.5
	20	standard	259.0	38.2	8.9	2.5	5.5	7.0	6.8	3.6	2.3	1.2
		proposed	211.3	28.2	6.6	1.9	0.9	0.4	0.2	0.7	1.2	1.5
30	5	standard	74.3	12.3	3.5	1.7	1.3	1.4	1.4	0.6	0.4	0.3
		proposed	66.3	10.1	2.7	2.7	0.9	0.8	0.8	1.4	1.6	1.6
	20	standard	71.2	13.5	3.0	1.5	1.2	1.4	1.4	0.5	0.4	0.3
		proposed	64.2	11.0	2.3	2.7	1.0	0.9	0.9	1.4	1.7	1.6

TABLE 3.2: Mean squared errors of the estimated eigenvalues presented in Figure 3.7.

m	r	method	order of the eigenvalues									
			1	2	3	4	5	6	7	8	9	10
10	5	standard	284.3	38.6	8.1	3.0	5.7	7.0	6.9	3.9	2.6	1.4
		proposed	229.5	29.2	5.9	2.5	0.9	0.4	0.2	0.8	1.3	1.5
	20	standard	253.2	41.0	7.8	2.8	5.4	6.7	6.7	3.7	2.4	1.2
		proposed	204.3	31.5	5.8	2.3	0.8	0.4	0.2	0.8	1.2	1.6
30	5	standard	78.4	14.6	3.9	1.8	1.3	1.4	1.5	0.6	0.4	0.3
		proposed	67.2	12.0	3.0	3.4	1.2	1.1	1.1	1.7	1.9	1.8
	20	standard	75.0	13.5	3.5	1.5	1.3	1.4	1.5	0.6	0.4	0.3
		proposed	65.5	11.2	2.7	2.6	0.9	0.8	0.9	1.4	1.5	1.5

each group. The data is plotted on the first two principal components in Figure 3.9 where different groups are represented by numbers 1-6.

The between-group covariance obtained by the method followed by Aitken & Lucy (2004) is

$$\begin{pmatrix} 26.13 & - & - & - & - \\ 89.09 & 354.03 & - & - & - \\ 66.54 & 216.19 & 171.24 & - & - \\ 59.71 & 183.26 & 152.67 & 144.95 & - \\ 19.12 & 30.20 & 51.05 & 62.73 & 47.65 \end{pmatrix}$$

Table 3.3: Mean squared errors of the estimated eigenvalues presented in Figure 3.8.

m	r	method	order of the eigenvalues									
			1	2	3	4	5	6	7	8	9	10
10	5	standard	293.2	51.9	11.1	3.8	6.2	8.5	9.4	6.4	5.7	6.5
		proposed	247.5	45.3	10.2	4.9	1.0	0.5	0.4	1.3	1.7	1.9
	20	standard	280.2	42.2	9.0	3.2	5.9	7.5	7.5	4.3	3.0	1.8
		proposed	230.1	33.0	6.9	2.7	0.9	0.4	0.2	0.9	1.3	1.6
30	5	standard	89.6	15.5	3.9	2.2	1.4	1.7	2.0	1.1	1.2	1.5
		proposed	74.8	14.9	5.6	8.3	3.6	3.7	3.5	5.2	5.0	4.3
	20	standard	71.6	15.3	3.6	1.7	1.3	1.5	1.6	0.7	0.6	0.5
		proposed	63.1	12.8	3.0	3.8	1.3	1.4	1.4	2.2	2.3	2.2

which is not non-negative definite (two of the eigenvalues are negative) and the one obtained by the new method is

$$\begin{pmatrix} 26.58 & - & - & - & - \\ 90.05 & 358.61 & - & - & - \\ 67.26 & 220.23 & 175.00 & - & - \\ 60.02 & 186.20 & 154.56 & 148.18 & - \\ 21.09 & 37.47 & 57.49 & 66.50 & 60.41 \end{pmatrix}$$

which is positive definite. Because m is small, we use leave-one-out cross-validation to choose penalty parameter ρ.

Note that, the log-likelihood for some folds in the cross-validation does not exist for a very small value of ρ in the grid where the optimal value is expected. This is because the graphical lasso algorithm does not produce a positive definite between-group covariance matrix when ρ is very small. We use the smallest value of ρ that produce positive definite between-group covariance.

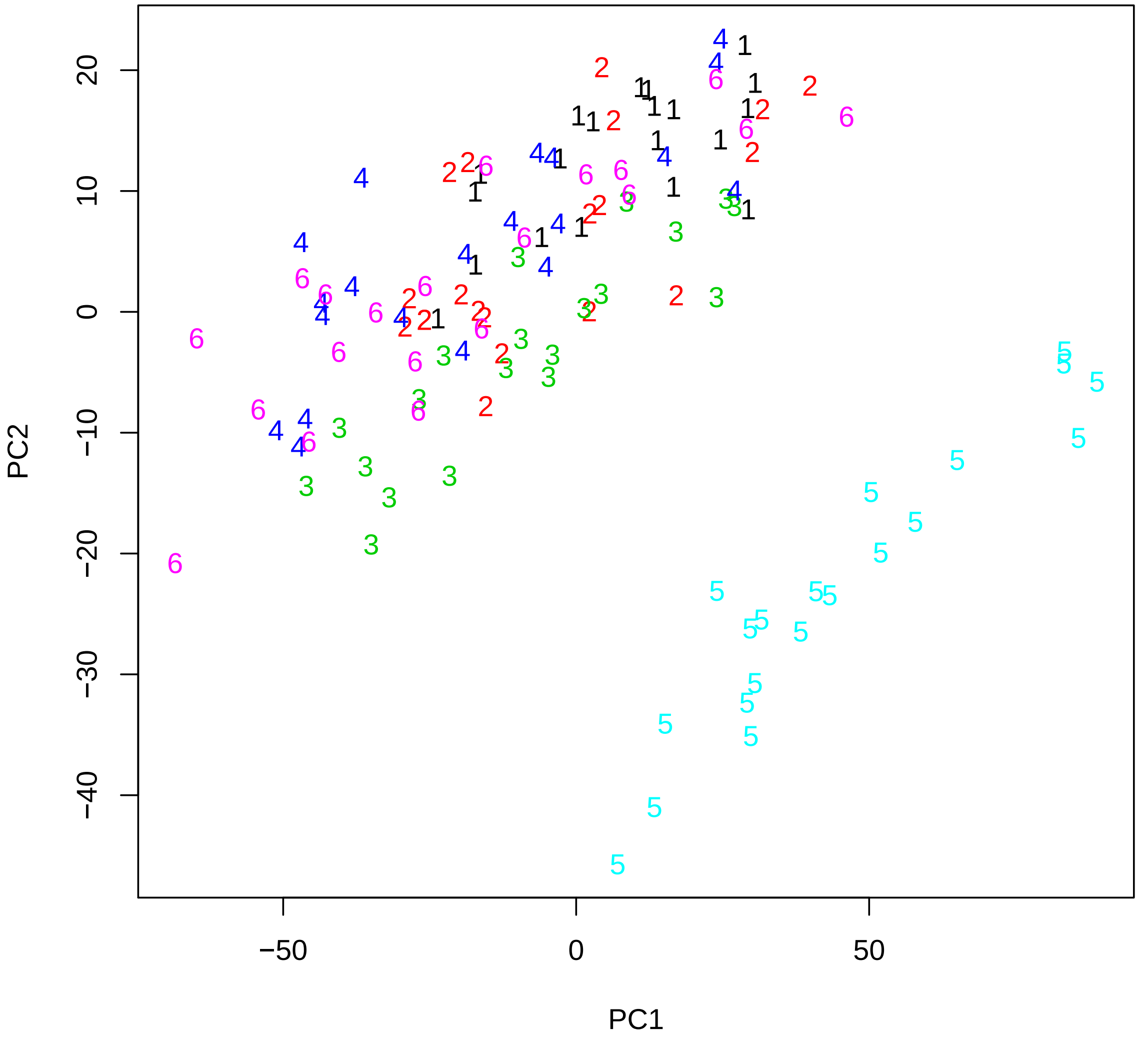


FIGURE 3.9: First two principle components of glass chemical composition data.

3.4 Conclusion

The standard method of estimation for the between-group covariance often results in negative variances. The main advantage of the proposed estimation procedure is that, in our examples, the estimated between-group covariance matrix obtained by using this method is positive definite. This is demonstrated by both the simulation experiments and the real world data example. Both the simulation study and analysis of the real world glass chemical composition data show that the developed method performs well even in the difficult situation where the within-group variation is substantial relative to the between-group variation. The estimates are more precise and less biased when there are more groups. Increasing the number of observations within groups also improve the accuracy of the estimates. It is also shown via simulation that the proposed method improve over the standard method in terms of accuracy in the estimated eigenvalues of the covariance matrix.

The algorithm needs less iterations to converge in easy case where the between-group covariance is dominant and is a good replacement in the situations where the traditional analysis of variance technique fails to work. In the hard case, where the within-group covariance is dominant, the algorithm some time fails to converge if we do not penalize the diagonal elements of the covariance matrix. Penalizing the diagonal elements using lasso penalty increase bias in the diagonal elements. The bias can be avoided if we replace lasso by adaptive lasso or SCAD penalty. We leave it to future work.

It should be noted that the results presented in this Chapter are based on a limited simulations and may not be true in general. However, our simulations suggest this is a promising technique worthy of further investigation.

3.5 Contributions of the Chapter

In this Chapter, we use the EM algorithm to optimize a lasso-penalized likelihood to create a procedure for estimation of the between-group covariance. To our knowledge, EM has not been used with a penalized likelihood before. We compare this estimator to one based on differencing the observed between and within group covariances. Simulation study demonstrates that the new procedure is an improvement in the sense that the estimated between-group covariance is positive definite.

Chapter 4

Regularized MANOVA for high-dimensional data

4.1 Introduction

The hypothesis of group effects on multiple response variables is simultaneously tested using MANOVA, the multivariate analogue of ANOVA. The classical MANOVA tests are large-sample approximations and perform well as long as we have a large number of observations, n, compared to the number of variables, p. As in other multivariate techniques, high-dimensionality poses serious problems for the classical MANOVA tests. First, it suffers from low power and does not maintain an accurate Type-I error rate, as p approaches n. Second, the classical test statistics involve the inversion of the sample covariance matrix, which is not possible if p exceeds n.

A common approach to improve the estimates of a high-dimensional covariance matrix is regularization. We present a procedure based on the lasso regularization of a covariance matrix. We compare the power of lasso-regularized MANOVA to four other competing procedures, including two ridge-type penalties that have not been assessed head-to-head before. Our consideration of lasso regularization is motivated by the fact that the estimated eigenvalues obtained from lasso regularization of a covariance are generally more accurate than those produced by the ridge regularization procedures previously used with MANOVA (see Figure 4.1). The lasso estimator of the inverse covariance matrix has also performed well in other contexts, such as model selection to produce a sparse inverse covariance (Hastie et al., 2009).

Denote an $n \times p$ matrix of observations by $\mathbf{Y}$, then the general linear multivariate model is given as

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times m}\mathbf{B}_{m\times p} + \boldsymbol{\epsilon}_{n\times p} \tag{4.1}$$

where $\mathbf{X}$ is an $n \times m$ design matrix, $\mathbf{B}$ is an $m \times p$ matrix of unknown parameters and $\boldsymbol{\epsilon}$ is an $n \times p$ matrix of errors. The n row vectors, $\boldsymbol{\varepsilon}_i$, of $\boldsymbol{\epsilon}$ are assumed to be independent observations drawn from a p dimensional multivariate normal distribution with mean vector zero and covariance matrix $\boldsymbol{\Sigma}_{p\times p}$ i.e. $\boldsymbol{\varepsilon}_i \sim N_p(0_{1\times p}, \boldsymbol{\Sigma}_{p\times p})$. Denote the inverse of a covariance matrix by $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = (\omega_{ij})_{1\leq i,j\leq p}$ then the log-likelihood for the multivariate normal distribution is well known to be

$$l(\mathbf{B}, \boldsymbol{\Sigma}; \mathbf{Y}) = Const - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y} - \mathbf{X}\mathbf{B})^t\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{X}\mathbf{B}) \tag{4.2}$$

The maximum likelihood estimates of $\mathbf{B}_{m\times p}$ and $\boldsymbol{\Sigma}_{p\times p}$ are

$$\widehat{\mathbf{B}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \tag{4.3}$$

and

$$\widehat{\boldsymbol{\Sigma}} = n^{-1}\mathbf{E} \tag{4.4}$$

respectively, where

$$\begin{aligned}\mathbf{E} &= (\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}})^t(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}) \\ &= \hat{\boldsymbol{\epsilon}}^t\hat{\boldsymbol{\epsilon}}\end{aligned} \tag{4.5}$$

is the $p \times p$ estimated matrix of the error sums of squares and cross-products (Seber, 2009). Consider the problem of testing the general linear hypothesis

$$H_0 : \mathbf{L}\mathbf{B} = \mathbf{0} \quad \text{versus} \quad H_1 : \mathbf{L}\mathbf{B} \neq \mathbf{0}, \tag{4.6}$$

where $\mathbf{L}$ is a $k \times m$ matrix of $rank(\mathbf{L}) = k \leq m$ specifying k linear combinations of the parameters. For instance, if the data comes from two groups and we want to test for the group effect, then the standard linear hypothesis would be

$$H_0 : \mathbf{B}_1 = \mathbf{B}_2 = \mathbf{0} \tag{4.7}$$

and in matrix notation it can also be written as

$$H_0 : \mathbf{LB} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{B}_0 \\ \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}. \tag{4.8}$$

Let $\mathbf{H}$ be the $p \times p$ matrix of the sums of squares and cross-products due the hypothesis defined by

$$\mathbf{H} = (\mathbf{L}\widehat{\mathbf{B}}_{m\times p})^t \left[\mathbf{L}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{L}^t\right]^{-1} \mathbf{L}\widehat{\mathbf{B}}_{m\times p}. \tag{4.9}$$

Under the assumption of normality, $\mathbf{E}$ is distributed according to a central Wishart distribution $W_p(\mathbf{\Sigma}_{p\times p}, q)$, where $q = n-k$ and $\mathbf{H}$ is distributed according to a non-central Wishart distribution $W_p(\mathbf{\Sigma}_{p\times p}, k, \mathbf{DD}^t)$, where $\mathbf{D}$ is a $p \times k$ matrix such that

$$\mathbf{D}^t\mathbf{D} = (\mathbf{LB})^t \left[\mathbf{L}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{L}^t\right]^{-1} \mathbf{LB}, \tag{4.10}$$

and $\mathbf{E}$ and $\mathbf{H}$ are independent. The likelihood ratio criterion Λ (also known as Wilks' Lambda) to test the hypothesis in equation (4.7) is given by

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{H}+\mathbf{E}|} = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}, \tag{4.11}$$

where $s = min(p, m)$ and $\lambda_1(.) \geq ... \geq \lambda_s(.)$ are the ordered non-zero eigenvalues of a $p \times p$ matrix $\mathbf{HE}^{-1}$ and $|\mathbf{A}|$ denotes the determinant of a matrix $\mathbf{A}$. Wilks' Lambda can be converted into an F-statistic using Rao's approximation; that is, the statistic

$$F = \frac{\nu_2}{\nu_1}(\Lambda^{-1/b} - 1), \tag{4.12}$$

where $\nu_1 = p(m-1)$, $\nu_2 = b\left[(n-1) - \frac{p+m}{2}\right] - \frac{p(m-1)-2}{2}$, and $b = \sqrt{\frac{p^2(m-1)^2-4}{p^2+(m-1)^2-5}}$, converges in distribution to the F-distribution with v_1 and v_2 degrees of freedom (Anderson, 2003). Three other well-known classical statistics are Bartlett-Nanda-Pillai's Trace, the Lawley-Hotelling Trace and Roy's Maximum Root Criterion. All these classical tests use large-sample approximations. We concentrate on two problems associated with the traditionally available MANOVA tests. First, these tests start to suffer from low power and do not maintain accurate Type-I error rates as the sample size n approaches the number of variables p. Second, the sample estimates of covariance matrices become singular when $p > n$. Since these tests rely on the inverse of a covariance matrix, they cannot be calculated when p exceeds n.

4.2 Previous works

Procedures have been proposed to regularize covariance matrices and a few of them have been adopted in the context of MANOVA. One way to account for singularity is to use a generalized inverse. In this situation, we do not have a closed form for the null distribution of the modified statistic; permutation methods are proposed as suitable alternatives to approximate the p-values. The performance of this test, however, is reported to be poor, see for example, Warton (2008). Another approach is to perform the classical test on the first q principal components, which account for most of the variation in the data (see Kong et al. (2006) and Tomfohr et al. (2005) for details). This approach performs well only if the group effect is along the first q eigenvectors.

Some more recent proposals are based on the idea of shrinking the elements of the covariance matrix towards some target estimator. These approaches include the Ledoit-Wolf shrinkage estimator (Ledoit & Wolf, 2003) and the ridge regularization as used by Warton (2008). Both approaches are very similar in that they both shrink the covariance or correlation matrix towards a pre-specified target estimator to get a corresponding regularized version. Let $\mathbf{S} = \frac{n}{n-1}\widehat{\mathbf{\Sigma}} = (s_{ij})_{1\leq i,j\leq p}$ be the unbiased estimate of the covariance matrix and $\mathbf{T} = (t_{ij})_{1\leq i,j\leq p}$ be the target estimate towards which we want to shrink $\mathbf{S}$. The Ledoit-Wolf shrinkage estimator is given by

$$\widehat{\mathbf{\Sigma}}_\rho = \rho\mathbf{S} + (1-\rho)\mathbf{T}, \tag{4.13}$$

where $\rho \in [0, 1]$ is the shrinkage intensity. Note that, for $\rho = 1$, we get $\widehat{\mathbf{\Sigma}}_\rho = \mathbf{S}$ while for $\rho = 0$, we have $\widehat{\mathbf{\Sigma}}_\rho = \mathbf{T}$. Ledoit & Wolf (2003) derive an analytical expression to calculate the value of ρ by minimizing the quadratic loss function

$$R(\rho) = E\|\widehat{\mathbf{\Sigma}}_\rho - \mathbf{\Sigma}\|^2, \tag{4.14}$$

where $\|\mathbf{A}\|^2$ is the squared matrix norm of $\mathbf{A}$. Schäfer & Strimmer (2005b) extend the work of Ledoit & Wolf (2003) and propose a method to calculate the value of ρ that only relies on the observed data. To be specific, they follow Ledoit & Wolf

(2003) and minimize the risk function in (4.14) with respect to ρ. The expression obtained for the optimal value of ρ is

$$\hat{\rho}^* = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} Var(s_{ij}) - Cov(t_{ij}, s_{ij})}{\sum_{i=1}^{p}\sum_{j=1}^{p} E\left[(t_{ij} - s_{ij})^2\right]}. \tag{4.15}$$

It is possible from the above expression to obtain a value of $\hat{\rho}^*$ that is either greater than 1 or even negative. This is avoided by using $\hat{\rho} = max(0, min(1, \hat{\rho}^*))$. Note that, the shrinkage intensity varies as we change the target estimator. Schäfer & Strimmer (2005b) provide detailed discussion about the various potential targets. A natural choice for $\mathbf{T}$ is $\mathbf{I}$, the identity matrix (used by (Ledoit & Wolf, 2003) and (Ledoit & Wolf, 2004)). This choice not only assumes sparsity, which is more intuitive in high-dimensional applications but also remarkably simple because it requires no parameter to be estimated. Another advantage of using $\mathbf{I}$ as target estimate is that it is positive definite. Since (4.13) becomes a convex linear combination of positive definite target matrix, $\mathbf{T}$, and positive semidefinite matrix $\mathbf{S}$, therefore the obtained shrinkage estimate, $\widehat{\mathbf{\Sigma}}_\rho$, is guaranteed to be positive definite. Replace $\mathbf{S}$ by the correlation matrix, $\mathbf{R} = (r_{ij})_{1\leq i,j\leq p}$, and $\mathbf{T}$ by $\mathbf{I}$ in (4.13), the expression for shrinkage intensity in (4.15) solves down to

$$\hat{\rho}^* = \frac{\sum_{i=1}^{p} Var(r_{ij}, i \neq j)}{\sum_{i=1}^{p}(r_{ij}^2, i \neq j)}. \tag{4.16}$$

The shrunken covariance is then obtained using the equation

$$\widehat{\mathbf{\Sigma}}_\rho = \mathbf{S}_d^{1/2}\widehat{\mathbf{R}}_\rho\mathbf{S}_d^{1/2}, \tag{4.17}$$

where $\widehat{\mathbf{R}}_\rho$ is the regularized version of $\widehat{\mathbf{R}}$. This formulation is more appropriate when variables are measured on different scales.

This shrinkage estimator has been studied by Tsai & Chen (2009) in the context of MANOVA and later adopted by Shen et al. (2011) to overcome the problem of singularity of the sample covariance and improve the power of the MANOVA test.

The penalized normal likelihood with L_2 penalty is given by

$$l(\mathbf{\Omega}) = Const + \frac{n}{2}\log|\mathbf{\Omega}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}-\mathbf{XB})^t\mathbf{\Omega}(\mathbf{Y}-\mathbf{XB}) - \kappa\sum_{j=1}^{p}\sum_{k=1}^{p}(\omega_{jk})^2 \tag{4.18}$$

where $\kappa > 0$ is the penalty parameter (Huang et al., 2006). Maximizing (4.18) with respect to $\mathbf{\Omega}$ reduces to a ridge-regularized covariance matrix given by

$$\widehat{\mathbf{\Sigma}}_\rho = \frac{1}{\rho}(\rho\widehat{\mathbf{\Sigma}} + (1-\rho)\mathbf{I}), \tag{4.19}$$

where $\mathbf{I}$ is a $p \times p$ identity matrix and $\rho = \frac{1}{1+\kappa} \in [0,1]$ is the regularization parameter. Alternatively consider the sample estimate of the correlation matrix $\mathbf{R}$ that can be obtained as

$$\widehat{\mathbf{R}} = \widehat{\mathbf{\Sigma}}_d^{-1/2}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}_d^{-1/2} \tag{4.20}$$

where $\widehat{\mathbf{\Sigma}}_d$ is a diagonal matrix with diagonal elements corresponding directly to the diagonal elements of $\widehat{\mathbf{\Sigma}}$ and zero elsewhere. The regularized version $\widehat{\mathbf{R}}_\rho$ of $\widehat{\mathbf{R}}$ can be obtained as

$$\widehat{\mathbf{R}}_\rho = \rho\widehat{\mathbf{R}} + (1-\rho)\mathbf{I} \tag{4.21}$$

and the corresponding regularized estimate $\widehat{\mathbf{\Sigma}}_\rho$ is obtained by

$$\widehat{\mathbf{\Sigma}}_\rho = \frac{1}{\rho}\widehat{\mathbf{\Sigma}}_d^{1/2}(\rho\widehat{\mathbf{R}} + (1-\rho)\mathbf{I})\widehat{\mathbf{\Sigma}}_d^{1/2}. \tag{4.22}$$

Warton (2008) has used (4.22) to estimate the covariance matrix. He used the normal likelihood in equation (4.2) as an objective function and cross-validation to obtain the optimum value of the ridge parameter, ρ.

In this part of the Thesis, we propose to use the lasso-regularized covariance matrix to replace $\widehat{\mathbf{\Sigma}}$. This approach to regularization was first introduced in Yuan & Lin (2007) for the purpose of model selection in Gaussian Graphical models. It forces some of the off-diagonal elements in the inverse covariance matrix to be exactly zero. Zero elements correspond to pairs of variables that are conditionally independent given the others (Dempster, 1972); the resulting estimate may therefore be interpretable. Although the goal of Yuan & Lin (2007) is to calculate a sparse estimate of the inverse covariance matrix, their method can also be used to regularize covariance matrices. We look at its performance both when the true inverse covariance is sparse and when it is dense to examine if the degree of sparsity has any effect on the quality of the estimate. The method is compared with the aforementioned four competing methods.

4.3 LASSO Regularization

Tibshirani (1996) proposed the lasso in a regression setting, and it has subsequently become a popular tool for model selection and estimation. It is based on penalizing the absolute size of the coefficients and will shrink some coefficients towards zero but set others as exactly zero; therefore, it simultaneously selects important variables and estimates their effects. This idea was extended by Yuan & Lin (2007) to estimation of the inverse covariance matrix, $\mathbf{\Omega}$. The following log-likelihood function based on a random sample from a multivariate normal distribution, $\mathbf{Y} \sim N_p(\mathbf{XB}, \mathbf{\Sigma})$, is optimized subject to positive-definiteness constraint for $\mathbf{\Omega}$:

$$l(\mathbf{\Omega}) = Const + \frac{n}{2}\log|\mathbf{\Omega}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}-\mathbf{XB})^t\mathbf{\Omega}(\mathbf{Y}-\mathbf{XB}) - \rho\sum_{j=1}^{p}\sum_{k=1}^{p}|\omega_{jk}|\,, \quad (4.23)$$

where ρ is a shrinkage parameter that controls the sparsity of $\widehat{\mathbf{\Sigma}}^{-1}$. An efficient solution to the problem in (4.23) is the graphical lasso algorithm proposed by Friedman et al. (2008).

4.3.1 Selection of ρ

The performance of regularized estimator depends on the choice of regularization parameter. We follow Fan et al. (2009) and use K-fold cross-validation to select the optimal value of ρ via a grid search over a grid of values produced by $e^{v/10}$, where $v \in [-100, 10]$. To reduce the computational effort, we stop the search at a value of ρ if the cross-validation score decreases over the next three consecutive values of ρ in the grid. We partition the n rows of $\hat{\boldsymbol{\epsilon}}$ into K disjoint sub-samples $\hat{\boldsymbol{\epsilon}}^t = \left[\hat{\boldsymbol{\epsilon}}_1^t, \hat{\boldsymbol{\epsilon}}_2^t, ..., \hat{\boldsymbol{\epsilon}}_K^t\right]$, where $\hat{\boldsymbol{\epsilon}}_k^t$ has n_k rows for $k = 1, 2, ..., K$. The size of n_k is roughly the same for all K sub-samples i.e assume that n is a multiple of K then $n_k = n/K$. The kth sub-sample is retained as a holdout set and the rest of the data, $\hat{\boldsymbol{\epsilon}}^{\setminus k}$, is used as training data. Denote the covariance matrix of $\hat{\boldsymbol{\epsilon}}^{\setminus k}$ by $\widehat{\mathbf{\Sigma}}_\rho^{\setminus k}$ then the cross-validation score is calculated using

$$CV(\rho) = \sum_{k=1}^{K} n_k\left(\log\, det\left(\widehat{\mathbf{\Sigma}}_\rho^{\setminus k}\right) + tr(\mathbf{S}_k(\widehat{\mathbf{\Sigma}}_\rho^{\setminus k})^{-1})\right), \quad (4.24)$$

where $\mathbf{S}_k$ is the sample covariance calculated from the kth sub-sample and $tr(\mathbf{A})$ is the trace of a matrix $\mathbf{A}$. The optimal value of ρ is one which gives the highest score i.e. $\hat{\rho} = \arg\max_{\rho} CV(\rho)$.

4.4 Simulation study

In this section, we demonstrate the performance of the lasso regularized MANOVA test in comparison to four competing procedures. We follow Warton (2008) and consider a number of different factors that can affect the performance of the tests. In each simulation, we draw a sample of size n observations from a p-variate normal distribution with a mean vector $\mathbf{B} = \mathbf{0}$ and a covariance matrix, $\mathbf{\Sigma}$. We make a shift of size δ for half of the sample therefore making two groups each of size $n/2$. The null hypothesis of no group effect $H_0 : \mathbf{LB} = \mathbf{0}$ is tested using $-2\log\Lambda$ as a test statistic. The estimate of $\mathbf{\Sigma}$, on which Λ is based, is computed using each of the following five competing regularization procedures:

Generalized Inverse: The test statistic is calculated using the Moore-Penrose generalized inverse as a plug-in estimator of the inverse covariance matrix. We use the built-in R function *ginv()* in "MASS" package to calculate Moore-Penrose generalized inverse (Venables & Ripley, 2002). Note that, the Moore-Penrose generalized inverse reduces to standard matrix inverse for a full rank sample covariance matrix.

Principal Component: We run the classical likelihood ratio test on the first q principal components of the standardized data. The value of q is chosen using eigenvalue-greater-than-one rule.

Shrinkage Estimator: The shrinkage estimator has been used to calculate the regularized estimate of inverse covariance matrix. We use the function *cov.shrink()* with default options, available in contributed R package "corpcor" (Schaefer et al., 2010), to calculate shrinkage estimate of the covariance matrix. The *cov.shrink()* function shrink the sample correlation matrix, $\mathbf{R}$, towards the identity target estimate, $\mathbf{I}$. Note that, the *cov.shrink()* function also allows to shrink the diagonal elements (this is the default option) with separate shrinkage intensity calculated by

$$\hat{\rho}^* = \frac{\sum_{i=1}^{p} Var(s_{ii})}{\sum_{i=1}^{p} (s_{ii}^2 - median(s))^2}, \tag{4.25}$$

where $median(s)$ is the median of sample variances.

Ridge Estimator: We use the ridge-regularized estimate of the inverse covariance matrix proposed by Warton (2008). A built-in R function, *ridgeParamEst()*, to choose the optimal value of ρ is available in contributed R package "mvabund" (Y. Wang et al., 2012). We use *ridgeParamEst()* to obtain the optimal value of ρ.

LASSO Regularization: We use lasso regularization to estimate the inverse covariance matrix as described in Section 2.2. We use the *glasso()* function to calculate lasso regularized covariance matrix. It is available in the contributed R package "glasso" (Friedman et al., 2013). Note that, we do not penalize the diagonal elements of the covariance matrix.

The classical procedure is viable for the principal component method; in other cases the reference distribution of the test statistic was approximated using 999 permutations of group labels. This is a Monte Carlo approximation to the exact test; so preserve an accurate Type-I error rates (Edgington & Onghena, 2007). Since the penalty parameters in shrinkage, ridge, and lasso regularization are controlled by the data in hand, they are allowed to change from one permutation to another.

We set $n = 20$ and $p \in [2, 30]$ in order to show the characteristics of all procedures for a range of p to n ratio. We did the experiments for two different covariance structures: the exchangeable structure given by

$$\sigma_{ij} \quad = \quad \begin{cases} 1 & \text{when } i = j \\ b & \text{when } i \neq j \end{cases} \qquad \text{for} \quad 1 \leq i, j \leq p. \tag{4.26}$$

and an AR(1) structure with

$$\sigma_{ij} = b^{|i-j|} \qquad \text{for} \quad 1 \leq i, j \leq p \tag{4.27}$$

An obvious reason for this selection is that for the AR(1) stricture, the inverse covariance matrix is sparse, while for the exchangeable structure it is dense, therefore comparing the procedure in two totally contrasting scenarios. For the AR(1), we do the experiments for $b \in \{0.4, 0.8\}$, while for the exchangeable structure we do the experiments for $b \in \{0.4, 0.7\}$. In the exchangeable covariance structure most of the error variation lies along one dimension. This becomes more extreme in high dimensions and with larger value of b. In the AR(1) structure the amount of error variation declines gradually from dominant to small eigenvectors if the dimension is low and the value of b is relatively small. We chose a suitable value of the shift parameter, δ, in order to examine cases where we expected moderate power. Since, most of the error variance lies along the dominant eigenvectors, a comparatively larger shift is required along those eigenvectors to identify the shift. The values of δ used in the simulation study are arranged in Table 4.1.

The orientation of the shift across different eigenvectors also affects the results. In our experiments, we placed the shift along all eigenvectors, the first, the second, and the pth eigenvectors in separate trials. In all of our experiments, we used the significance level $\alpha = 0.05$.

4.4.1 Simulation results

The simulation results are presented in Figures 4.2-4.5. Some simulation results for a slightly different value of δ (we increase the values of δ in Table 4.1 by 20%) are given in Appendix A. Note that, the point for each value of p on a power curve is the average of the results from 1000 simulated datasets. The power pattern does not seem to be heavily dependent on covariance structures. Overall, the power decreased as we increased p for all five competing procedures except when the shift is produced in all dimensions. This is because a shift cumulates across all p dimensions and a large value of p makes it more pronounced. The principal component method performed well when the shift is along first few eigenvectors. The lasso, ridge, and shrinkage approaches all perform well, with none universally superior across scenarios. Lasso regularization is always the best (by a narrow margin) when the shift is along all eigenvectors, consistent with its accurate recapture of all the eigenvalues in Figure 4.1. The power of the method based on the generalized inverse decreases sharply for $2 \leq p \leq 17$. In this range the generalized inverse reduces to the maximum likelihood estimate of the inverse covariance matrix, which is well known to be unstable as p approaches to n. Many scenarios show the generalized inverse power curve bouncing back at $p = 18$. This was also observed by Schäfer & Strimmer (2005a) who attribute it to a "dimensionality resonance effect" (see Schäfer & Strimmer (2005a) and references therein for details).

Figure 4.6 presents the computational time comparison. Clearly, lasso is computationally very expensive and its computational time increases dramatically with increasing number of variables for a fixed sample size. The computational time of ridge regularization is moderate and far less than lasso regularization. Principal components, generalized inverse and shrinkage are the least computationally expensive procedures. The computational time is also dependent on the structure of the covariance matrix for lasso regularization. The lasso is designed for sparse situations, and on average it takes longer to fit for the exchangeable structure than AR(1), with the difference increasing with the number of variables. The effect of changing the covariance structure on computational time is not noticeable for ridge and shrinkage regularization and therefore we show only the time for exchangeable structure. The computational times for principal components and generalized inverse are not shown on the graph but lie below the shrinkage estimate.

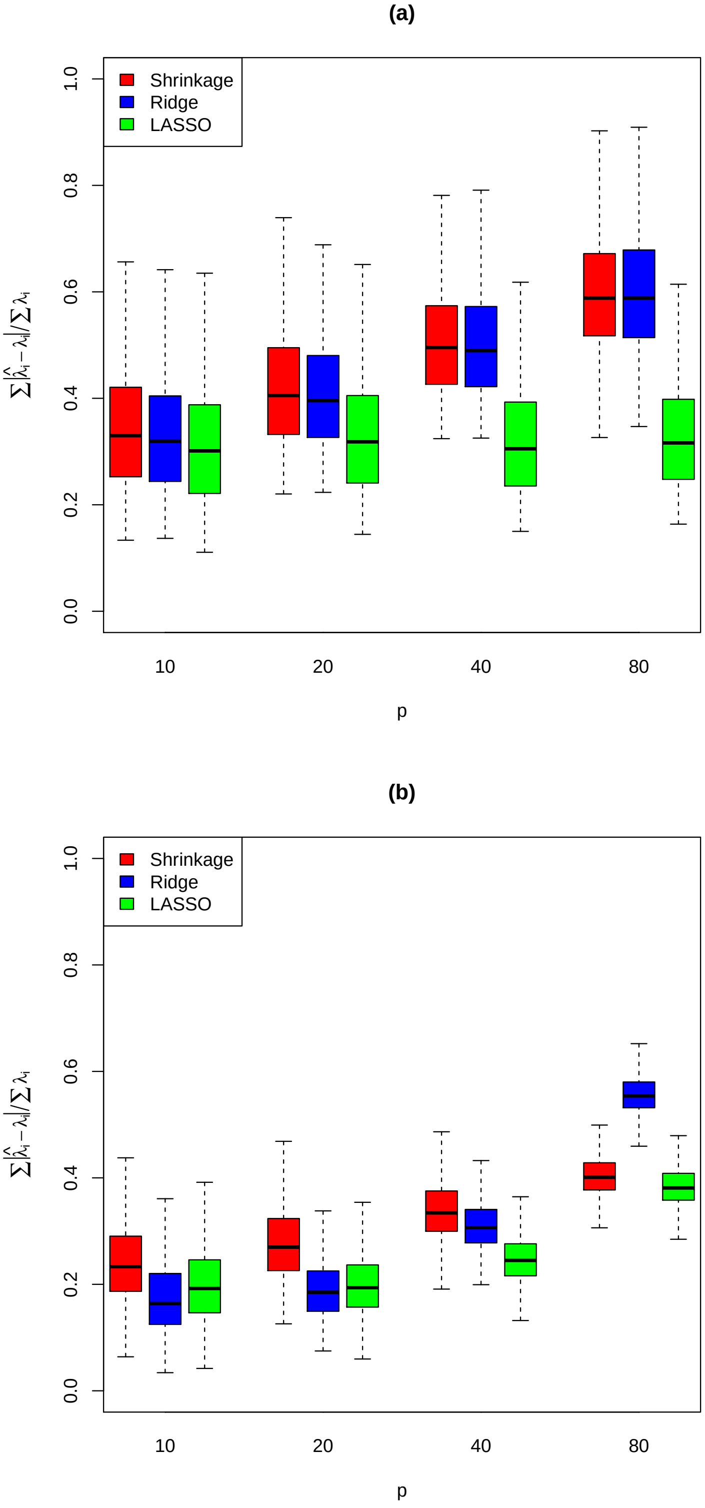


FIGURE 4.1: Distribution of the sum of absolute errors ($\sum_{i=1}^{p} \left|\hat{\lambda}_i - \lambda_i\right| / \sum_{i=1}^{p} \lambda_i$) in the estimated eigenvalues under (a) exchangeable and (b) AR(1) covariance structures both with $b = 0.6$. For each value of $p \in \{10, 20, 40, 80\}$, 1000 samples of size 20 are simulated from a multivariate normal distribution. Eigenvalues of the true covariance matrix are estimated using shrinkage, ridge, and lasso regularization and the sum of absolute errors are calculated for each of the 1000 samples. Note that, the estimation error increases as we increase p and the lasso regularization maintains the highest accuracy.

Table 4.1: Values of the shift parameter, δ, used in the simulation study to obtain a moderate power.

Orientation of δ	Covariance structure			
	AR(1)		Exchangeable	
	$b = 0.4$	$b = 0.8$	$b = 0.4$	$b = 0.7$
along 1^{st} eigenvector	3.40	4.00	4.20	4.00
along 2^{nd} eigenvector	1.90	4.00	1.90	2.00
along p^{th} eigenvector	1.30	1.00	1.30	1.00
along all eigenvectors	0.43	0.35	0.42	0.30

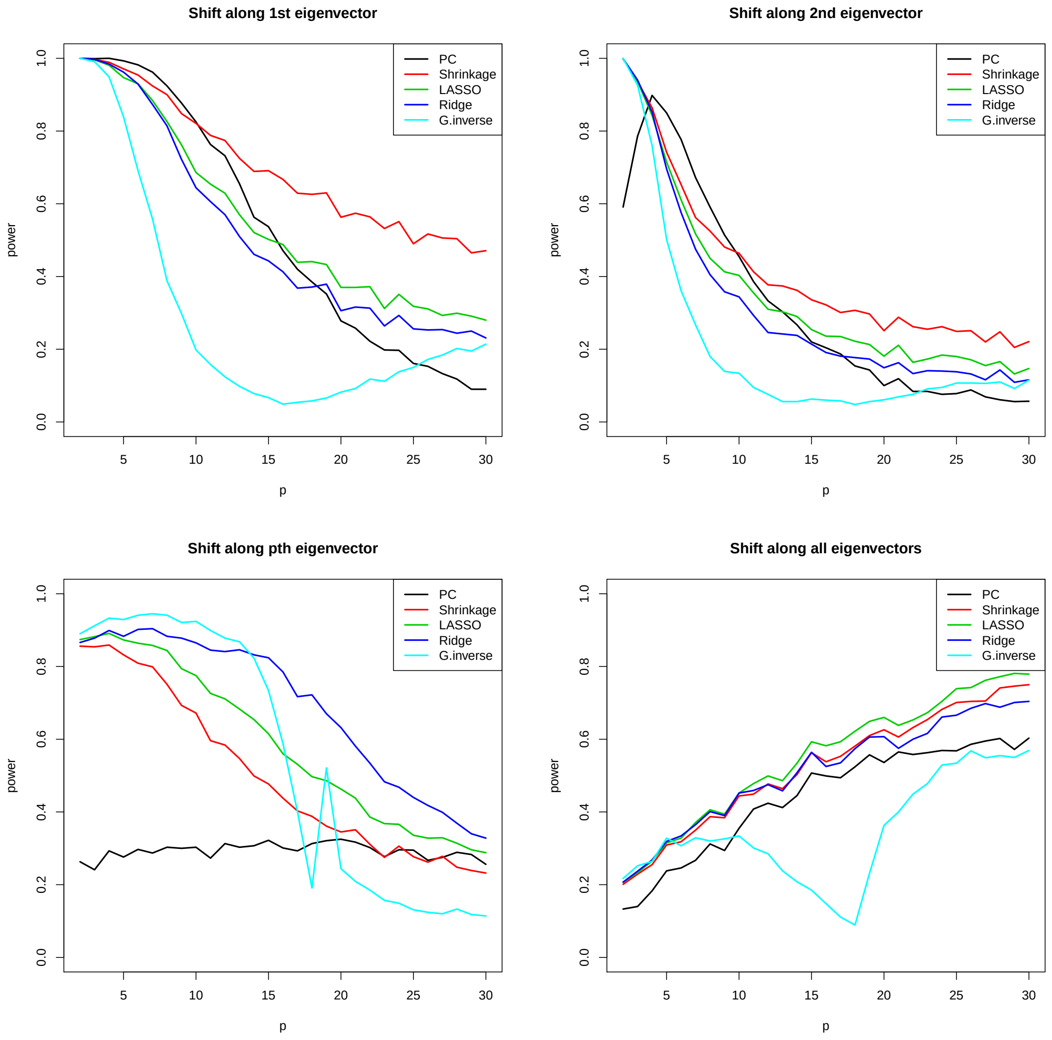


Figure 4.2: Power comparison of MANOVA test based on 5 competing procedures under AR(1) covariance structure with $b = 0.4$. For each value of $p \in [2, 30]$, the point on a power curve is the average of 1000 experiments. The significance level is 0.05.

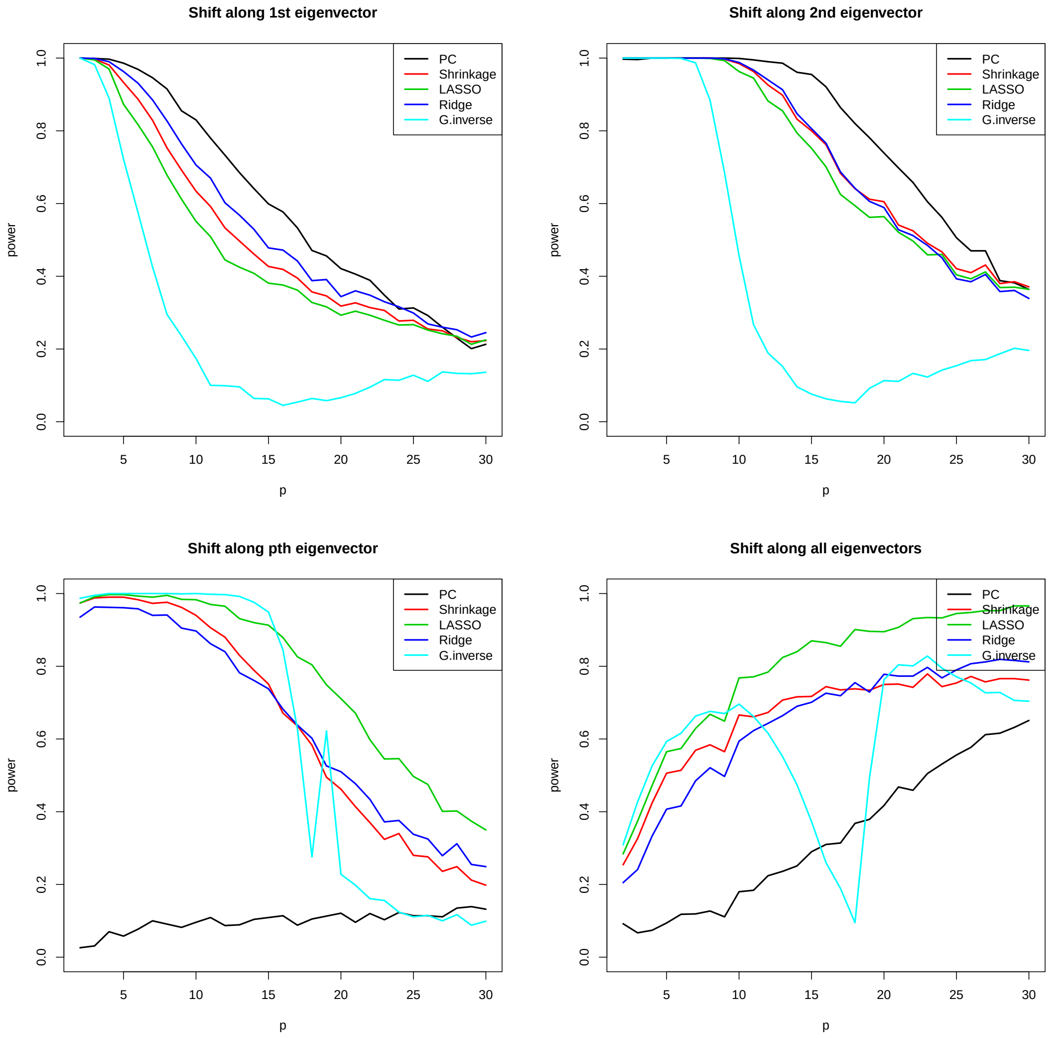


FIGURE 4.3: Power comparison of MANOVA test based on 5 competing procedures under AR(1) covariance structure with $b = 0.8$. For each value of $p \in [2, 30]$, the point on a power curve is the average of 1000 experiments. The significance level is 0.05.

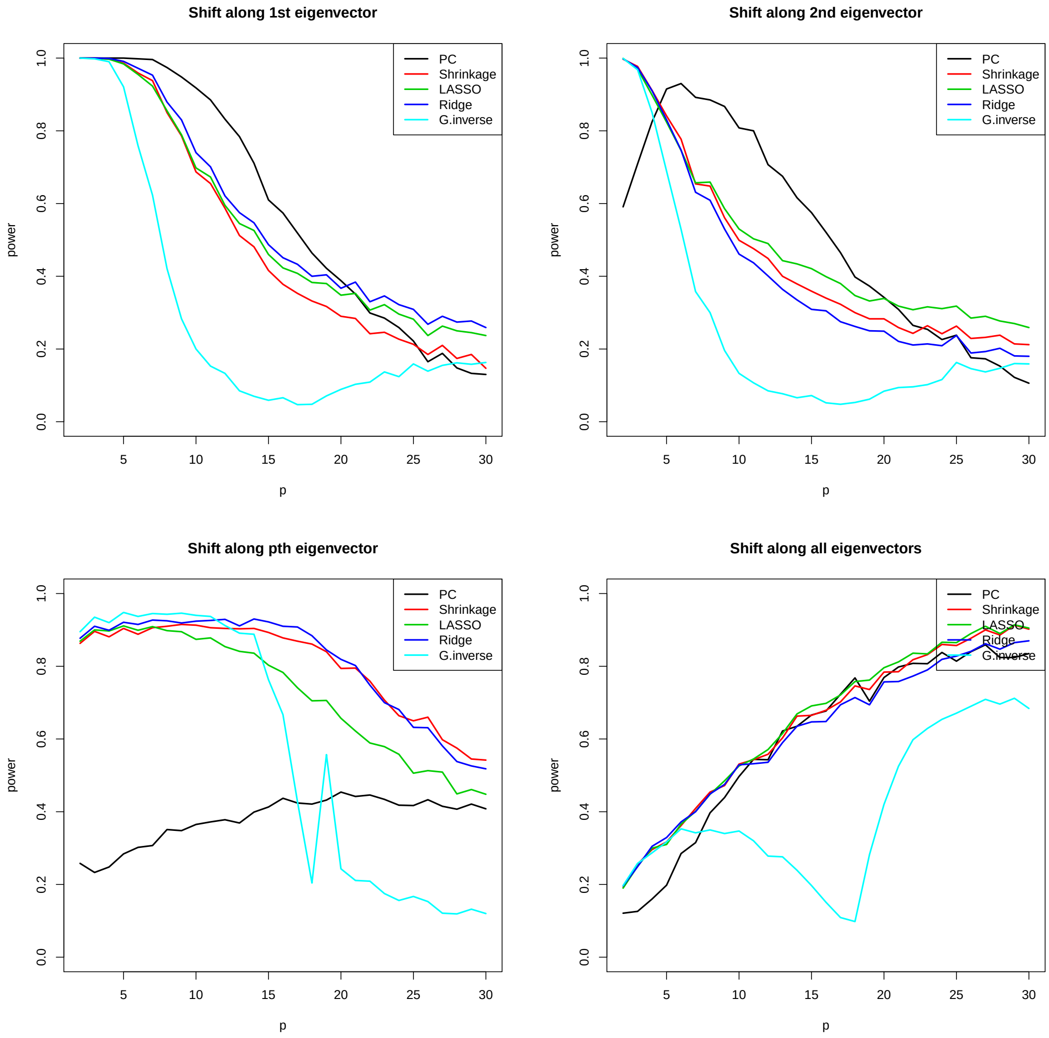


FIGURE 4.4: Power comparison of MANOVA test based on 5 competing procedures under exchangeable covariance structure with $b = 0.4$. For each value of $p \in [2, 30]$, the point on a power curve is the average of 1000 experiments. The significance level is 0.05.

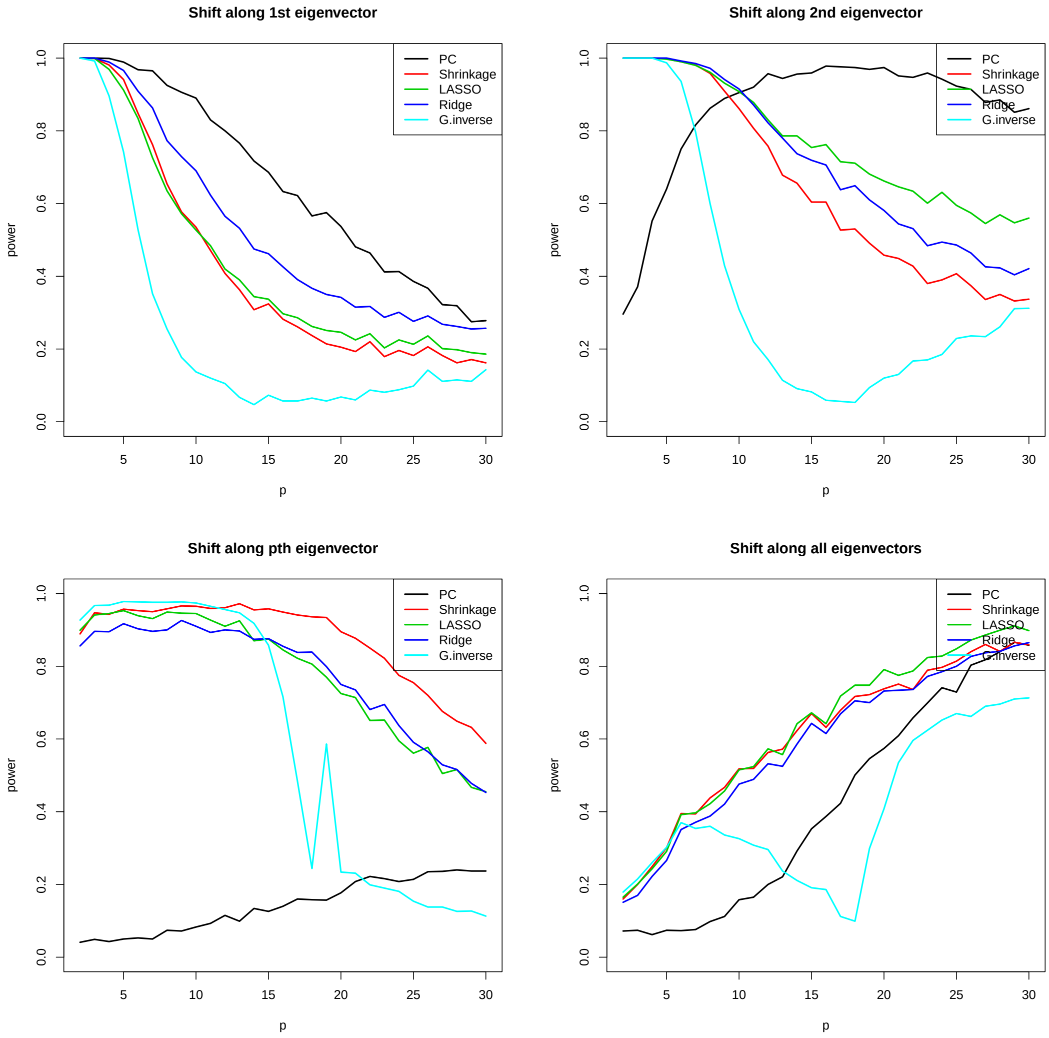


Figure 4.5: Power comparison of MANOVA test based on 5 competing procedures under exchangeable covariance structure with $b = 0.7$. For each value of $p \in [2, 30]$, the point on a power curve is the average of 1000 experiments. The significance level is 0.05.

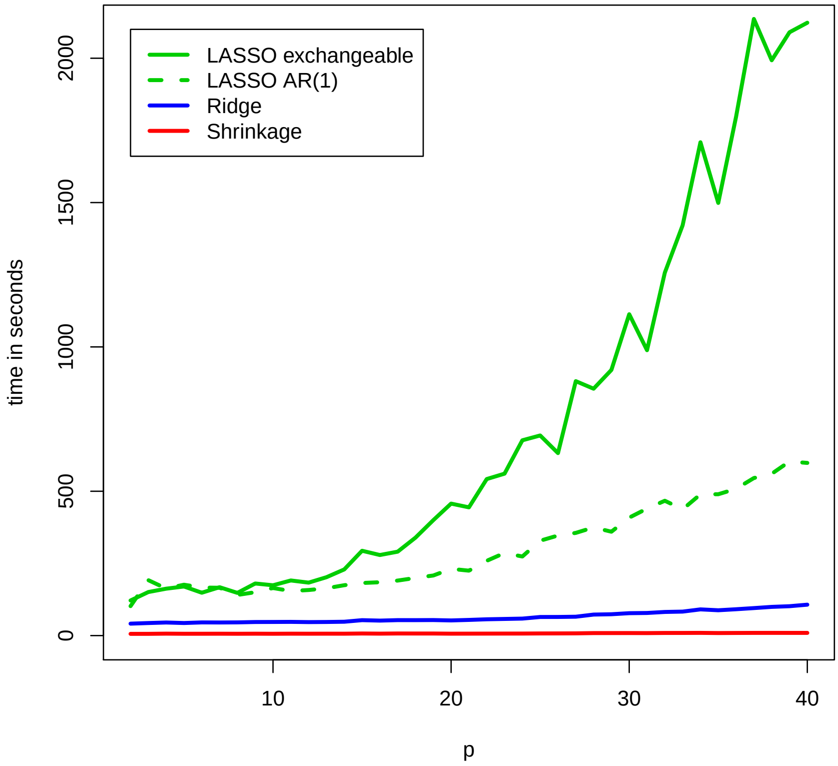


FIGURE 4.6: Time comparison of 3 competing regularization procedures under two different covariance structures: exchangeable and AR(1). Each point in the graph is averaged over 10 replicates. The covariance structure does not make big difference in computational time for ridge and shrinkage therefore only shown for exchangeable. The computational time for principal components and and generalized inverse is not shown but lie below the shrinkage estimate.

4.5 Practical application

In this section, we considered soil compaction profiles from three different positions (high, medium and low) along sunny ridge slopes. The data is from 7 different ridges in the eastern Qilian Mountans (China), a rangeland habitat. The measurements are of soil compaction at 18 different depths, ranging from 2.5 cm to 45 cm with an interval of 2.5 cm, so there are 18 variables, each variable corresponding to a particular depth, and 21 observations per variable. These 21 observations were divided into three groups of seven, corresponding to the position relative to the ridge top (high, medium and low). Initial investigations indicated the blocking structure (the seven ridges) was unimportant and we have ignored it. The attractive property of this data set, for our purposes, is that the number of variables are comparable to the number of observations.

After centering each group, the measurements at adjacent depths are highly correlated, but with the correlation falling away with more intervening distance (Figure 4.7), suggesting our AR(1), $b = 0.8$ simulations (Figure 4.3) are the best reference for selecting a regularization method. The plot of the data projected onto the first two principal components of the within group correlation matrix (Figure 4.8) does not even hint at a separation between the elevation groups, so we will concentrate on the lower left panel, where the shift was in the direction of the pth component. This suggests lasso will produce the best results, followed fairly closely by ridge and shrinkage, with principal components the worst procedure.

Our results are largely consistent with these predictions, (although ridge and shrinkage reverse their rankings). The p-values for each of the five techniques are provided in the top row of Table 4.2. Using all 21 observations with 7 observation in each group, the effect turns out to be non-significant at a 5% level of significance for principal components, ridge, and generalized inverse with p-values 0.156, 0.058, and 0.096 respectively, while it is significant for shrinkage and lasso regularization with p-value 0.023 for both of them.

To check the stability of the results, we repeated the analysis with the data from six ridges rather than seven, deleted each ridge in turn to get 7 new data sets with 18 observations each (3 groups of 6). The p-values for all 7 replicates are arranged in Table 4.2 with significant effects at $\alpha = 0.05$ shown in bold. The shrinkage approach and lasso regularization gave similar results in most cases, with ridge not far behind (but typically just missing the significance threshold). The performance of the generalized inverse became superior by reducing the sample size from 21 (where the covariance estimate is just the standard MLE) to 18. This is in close agreement with the power bump seen in our simulation for the generalized inverse procedure just after p exceeds the degrees of freedom for estimating the covariance. As predicted principal components is the least able to detect differences.

TABLE 4.2: Table of p-values for five competing procedures. Significant effects at $\alpha = 0.05$ are shown in bold.

deleted	regularization procedure				
observation	G.inverse	PC	ridge	shrinkage	lasso
none	0.096	0.156	0.058	**0.023**	**0.023**
1st	**0.001**	**0.013**	**0.009**	**0.002**	**0.011**
2nd	**0.042**	0.142	0.190	0.063	0.128
3rd	0.101	0.232	0.112	0.061	0.060
4th	**0.018**	0.171	0.108	**0.039**	**0.037**
5th	**0.000**	0.126	0.054	**0.015**	**0.030**
6th	**0.008**	0.618	0.093	0.079	0.074
7th	**0.005**	0.390	0.080	0.057	**0.046**

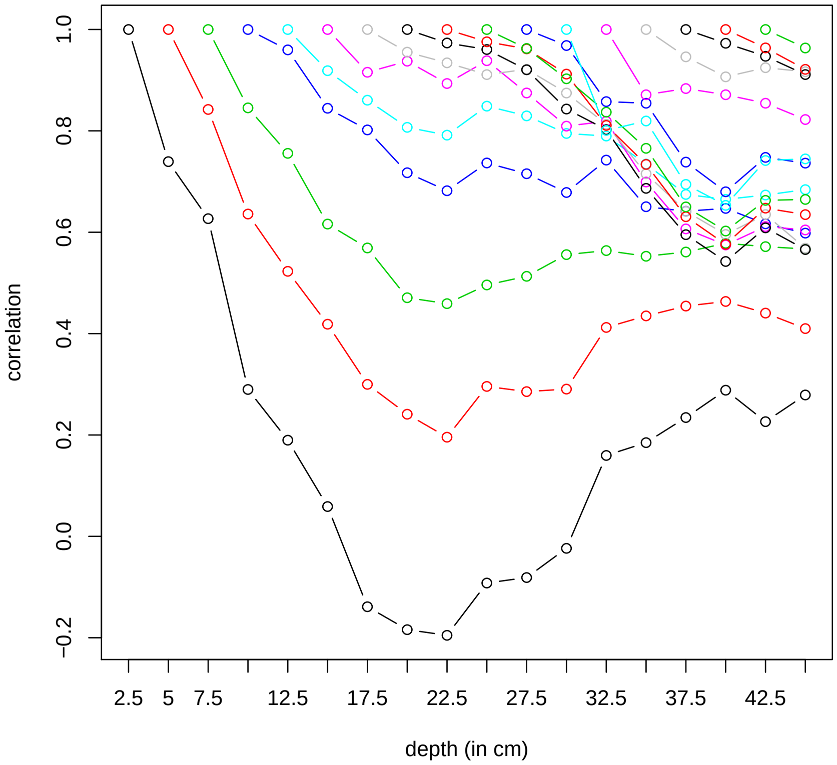


FIGURE 4.7: Serial correlation coefficients of soil compaction at shallower depths with soil compaction at all the deeper depths after adjusting for group means. Each sequence of joined points of the same color represents the correlation of the measurements at a certain depth with the measurements at deeper levels (with the depth value given on the x axis). Note that for each of the 18 depths (variables) we have 21 observations.

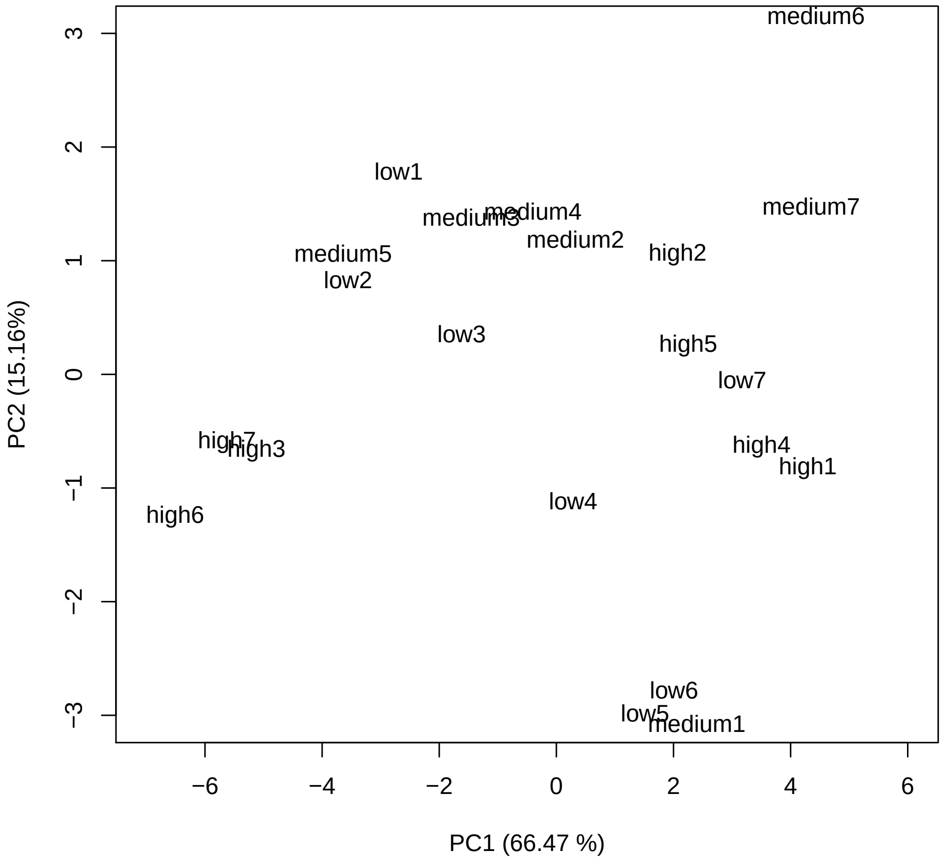


Figure 4.8: Projection of the data (variables are standardized to have zero means and unite standard deviations) onto the first two principal components of the data after adjusting for group means.

4.6 Conclusion and discussion

In this Chapter, we have used normal likelihood with a lasso penalty to estimate the inverse covariance matrix in the context of high-dimensional MANOVA. This was originally motivated by the relatively good performance of the lasso estimator in other areas like Gaussian graphical models. The method was tested by extensive simulations with two different covariance structures: exchangeable and AR(1). The approach based on lasso regularization is also compared with some other competing procedures including shrinkage approach of Schäfer & Strimmer (2005b) and ridge regularization approach used by Warton (2008). All three of these regularization procedure perform well, and perform better than the more conventional generalized inverse and principal component approach when the shift is not along the first few eigenvectors. Lasso regularization is always the best when the shift is along all the eigenvectors but the difference is not dramatic. The difference, however, increases as we increase the number of variables. We know that the lasso does well at recovering all p eigenvalues. The fact that the lasso is always the best method when the shift is in all directions suggests that accurate estimation of all eigenvalues is important in this case. The connection between accurately estimating the eigenstructure and increased power is a topic for further investigation.

In terms of computational time, however, there are big differences. The lasso approach is much more expensive than its other counterparts, because lasso regularization is an iterative procedure even before considering the computationally cumbersome tasks of cross-validation to choose a penalty parameter, and computation of the permutation based reference distribution. While not computationally prohibitive its computational time increases dramatically with the number of variables for a fixed sample size. Generally, the adaptive lasso (Zou, 2006) has been preferred over lasso for recovery of eigenvalues but it further increases computational time. Ridge regularization is comparatively less expensive than lasso but its implementation also involves cross-validation. The principal component approach, generalized inverse and shrinkage approach are all very fast. For principal components and the generalized inverse, this comes at the cost of poor performance in many cases, particularly when p is close to or exceeds n. However, shrinkage was typically at or near the top in power, and would be our recommendation whenever both power and computational time must be considered.

4.7 Contributions of the Chapter

Recently the shrinkage estimator and ridge regularization of the covariance matrices have been used to allow the use of MANOVA test in high-dimensional situations. While lasso regularization is a natural alternative, a lasso based MANOVA procedure has not yet been described. We provide the outline of such a procedure. The comparative performance of different approaches (lasso, ridge, shrinkage, and the more traditional generalized inverse and principal component approaches) is then explored via an extensive simulation study. The methods are also applied to real data set of soil compaction profiles at various elevation ranges. Note that this also provides a head-to-head comparison of the ridge and shrinkage approaches, which has not previously existed. The MANOVA test based on the lasso regularization performs better in terms of power of the test in some cases. We also see that the shrinkage intensity chosen by the closed form shrinkage approach is typically competitive with the cross validation based ridge procedure.

Chapter 5

Monitoring future observations in the high-dimensional setting

5.1 Introduction

Multivariate control charts are often used to detect unusual behavior in a process from which several quality characteristics are simultaneously measured in discrete sampling stages. The Hotelling T^2 statistic, which measures the distance of each sub-group mean from the process mean, is used as a charting statistic. The procedure is divided into two phases: Phase-I and Phase-II. In Phase-I analysis, historical data is tested retrospectively to establish the behavior of the process when it is in-control. The parameters are estimated using the in-control historical data and control limits are calculated. In Phase-II analysis, the continuation of the process is monitored for out-of-control signals using the control limits determined in Phase-I. For each future sub-group, the Hotelling T^2 statistic is calculated and plotted on the control chart. An out-of-control warning is issued when there is a departure from the limits established in Phase-I analysis (see Bersimis et al. (2007) and the references in there for details about multivariate control charts).

In those applications where the data generating process is too slow or it is impossible to make natural sub-groups, individual observations are monitored. A number of papers have been published on multivariate control charts for individual observations (for example, see Tracy et al. (1992), Lowry & Montgomery (1995) and Chou et al. (1999)). The methods described in these papers do not allow the monitoring process to start until the sample size, n, is more than the number of variables, p. Moreover, these methods are unreliable in high-dimensional problems unless we use a large sample size (see Champ et al. (2005) and Lowry & Montgomery (1995) for sample size recommendations). Here, we provide a more reliable and practical method for monitoring individual high-dimensional observations.

Consider an $n \times p$ matrix $\mathbf{X}$ of n individual baseline observations and assume that the p variables in $\mathbf{X}$ follow a multivariate normal distribution with a mean vector $\boldsymbol{\mu}^t = (\mu_1, \mu_2, ..., \mu_p)$ and a $p \times p$ covariance matrix $\boldsymbol{\Sigma}$, i.e. $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Denote the ith row of $\mathbf{X}$ by $\mathbf{X}_i$; then the squared Mahalanobis distance of a point, $\mathbf{X}_i$, from the process mean in a multidimensional space, is generally used as a charting statistic for multivariate control charts. If $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known, then the statistic

$$\chi_p^2 = (\mathbf{X}_i - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{X}_i - \boldsymbol{\mu}) \tag{5.1}$$

follows a chi-square distribution with p degrees of freedom and the corresponding control chart is usually called a chi-square control chart. In multivariate control charts, it is common to set the lower control limit to zero because the value of charting statistic in (5.1) is always positive and a shift in the mean vector always results in an increase in the value of the statistic. The upper control limit of the chi-square control chart is the $(1-\alpha)$th percentile of the chi-square distribution with p degrees of freedom, that is

$$UCL = \chi^2(\alpha; p). \tag{5.2}$$

Generally, the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown and are estimated from the retrospective data in Phase-I analysis. It is assumed that the process is in Phase-I stage, and $\bar{\mathbf{X}}$ and $\mathbf{S}$ are, respectively, the estimated baseline mean and unbiased sample covariance matrix. Then, the Hotelling T^2 statistic

$$T_i^2 = c_1 (\mathbf{X}_i - \bar{\mathbf{X}})^t \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}) \tag{5.3}$$

where $c_1 = n/(n-1)^2$, follows a beta distribution with $p/2$ and $(n-p-1)/2$ as shape parameters. The upper control limit for Phase-I analysis is

$$UCL = beta(\alpha; p/2, n-p-1/2), \tag{5.4}$$

where $beta(\alpha; a, b)$ is the $(1-\alpha)$th percentile of beta distribution with parameter

a and b (see Chou et al. (1999) for the proof of (5.4)). Denote a future observation by $\mathbf{X}_f$. Since $\mathbf{X}_f$ is independent of $\bar{\mathbf{X}}$ and $\mathbf{S}$ (because $\bar{\mathbf{X}}$ and $\mathbf{S}$ are estimated from baseline set of data), the statistic

$$T^2 = c_2(\mathbf{X}_f - \bar{\mathbf{X}})^t\mathbf{S}^{-1}(\mathbf{X}_f - \bar{\mathbf{X}}), \tag{5.5}$$

where $c_2 = n(n-p)/p(n-1)(n+1)$, follows an F-distribution with p and $n-p$ degrees of freedom. The upper control limit for future observations is therefore the $(1-\alpha)$th percentile of F distribution with p and $n-p$ degrees of freedom i.e.

$$UCL = F(\alpha; p, n-p) \tag{5.6}$$

(see Tracy et al. (1992) for the detailed proof of (5.6)). Generally, an out-of-control signal is issued if the calculated statistic in the above three cases goes beyond the corresponding upper control limits. The control charts based on the statistics in (5.3) for the Phase-I analysis and in (5.5) for the Phase-II analysis are termed as Hotelling T^2 control charts.

Estimation of the true parameters ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) from the baseline data causes the T^2 for $\mathbf{X}_i$ and $\mathbf{X}_f$ to follow different distributions, even though $\mathbf{X}_i$ and $\mathbf{X}_f$ follow the same multivariate normal distribution. The reason for this difference is that the T^2 statistics calculated for $\mathbf{X}_i$ are not independent of the estimated mean vector and sample covariance matrix, whereas the T^2 for $\mathbf{X}_f$ are independent (Tracy et al., 1992). Note that this difference exists only under a limited sample size, n. Asymptotically, as n approaches infinity, the distribution of T^2 for both baseline and future observations converges to a chi-square distribution with p degrees of freedom.

The Hotelling T^2 control charts are reasonably effective as long as the number of process variable, p, is small. Difficulties arise in finding an upper control limit when the number of process variables, p, is comparable to n. Firstly, as p approaches n, the estimate $\mathbf{S}$ (of $\boldsymbol{\Sigma}$) becomes unstable. As a result, the power to detect out-of-control signals progressively decreases. Secondly, the estimated covariance matrix $\mathbf{S}$ is rank-deficient if p exceeds n. Consequently, $\mathbf{S}^{-1}$ cannot be calculated and the Hotelling T^2 procedure fails to work. Data sets where p is large relative to n arise in many fields. We will consider a case where many gene expression measurements are taken on a limited number of patients.

In this Chapter, we propose a novel scheme for detecting one-off unusual changes in the process mean vector in a high-dimensional setting. We estimate the variance covariance matrix using a distribution free shrinkage estimator (lasso regularization is avoided because it assumes multivariate normality). We use a leave-one-out re-sampling procedure, with n baseline samples, to create n independent T^2 statistics. The empirical distribution of the T^2 statistics for future observations and the UCL for Phase-II analysis is estimated using a kernel smoothing technique.

5.2 Shrinkage estimate of a covariance matrix

The sample estimate of a covariance matrix is known to be unstable when p and n are comparable (Johnstone, 2001). Procedures have been proposed to regularize the sample covariance matrix. These procedures include the Steinian-class of shrinkage estimators that shrinks an estimator towards a pre-specified target to get a corresponding regularized version. Ledoit & Wolf (2004) take the idea of James & Stein (1961) and propose a shrinkage estimator of a covariance matrix that is the convex linear combination of the sample covariance and a target matrix. Consider the unbiased sample covariance, $\mathbf{S}$, of a high-dimensional data and let $\mathbf{T} = (t_{ij})_{1\leq i,j\leq p}$ be a target estimate towards which we want to shrink our sample covariance. The target estimator, $\mathbf{T}$, is required to be positive definite and its specification needs some assumptions about the structure of the true covariance matrix, $\mathbf{\Sigma}$. For example, Ledoit & Wolf (2004) uses a diagonal matrix as a structured target estimate (presuming that all variances are of the same size and all covariances are zero which make sense in high-dimensional scenarios) which is also positive definite. A Steinian-class of shrinkage estimators is obtained by the convex linear combination of $\mathbf{S}$ and $\mathbf{T}$, given by

$$\widehat{\mathbf{\Sigma}}_\rho = \rho\mathbf{T} + (1-\rho)\mathbf{S} \tag{5.7}$$

where $\rho \in [0,1]$ is the shrinkage intensity. Note that, for $\rho = 0$ we get $\widehat{\mathbf{\Sigma}}_\rho = \widehat{\mathbf{\Sigma}}$ while for $\rho = 1$, we have $\widehat{\mathbf{\Sigma}}_\rho = \mathbf{T}$. The regularized estimate, $\widehat{\mathbf{\Sigma}}_\rho$, obtained in this way is more accurate and statistically efficient than the estimators $\widehat{\mathbf{\Sigma}}$ and $\mathbf{T}$ in problems with n comparable to p (see Ledoit & Wolf (2004)).

Ledoit & Wolf (2004) provided a procedure for finding the optimal shrinkage intensity which asymptotically minimizes the expected quadratic loss function,

$E\|\widehat{\mathbf{\Sigma}}_\rho - \mathbf{\Sigma}\|^2$, where $\|\mathbf{A}\|^2$ is the squared matrix norm of $\mathbf{A}$. The expected quadratic loss function measure the mean-squared error and an estimator with minimal mean-squared error is desired. Schäfer & Strimmer (2005b) applies the Ledoit & Wolf (2003) formulation to different target estimators and derive closed-form analytical expressions for computing the value of ρ. To be more specific, they minimize the risk function

$$R(\rho) = E\|\widehat{\mathbf{\Sigma}}_\rho - \mathbf{\Sigma}\|^2 \tag{5.8}$$

to compute the value of ρ. Minimizing (5.8) with respect to ρ, the following expression has been obtained for the optimal value of ρ

$$\hat{\rho}^* = \frac{\sum_{i=1}^p \sum_{j=1}^p Var(s_{ij}) - Cov(t_{ij}, s_{ij})}{\sum_{i=1}^p \sum_{j=1}^p E\left[(t_{ij} - s_{ij})^2\right]}. \tag{5.9}$$

It is possible from the above expression to obtain a value of $\hat{\rho}^*$ that is either greater than 1 (over shrinkage) or even negative. This is avoided by using $\hat{\rho} = max(0, min(1, \hat{\rho}^*))$. It is worth noting at this point, that the shrinkage intensity varies as we change the target estimator. Schäfer & Strimmer (2005b) provide a detailed discussion about the six commonly used targets. A natural choice for $\mathbf{T}$ is $\mathbf{I}$, the identity matrix used by Ledoit & Wolf (2003) or its scalar multiple. This choice not only assume sparsity which is more intuitive in high-dimensional applications but also remarkably simple because it require no or one parameter to be estimated. Using identity matrix as a target estimate reduces the expression in (5.9) to

$$\hat{\rho}^* = \frac{\sum_{i=1}^p Var(s_{ij}, i \neq j) + \sum_{i=1}^p Var(s_{ii})}{\sum_{i=1}^p (s_{ij}^2, i \neq j)}. \tag{5.10}$$

Since (5.7) becomes a convex linear combination of positive definite target matrix, $\mathbf{T}$, and positive semidefinite matrix $\mathbf{S}$, therefore the obtained shrinkage estimate $\widehat{\mathbf{\Sigma}}_\rho$ is guaranteed to be positive definite.

In this Thesis, we use the function *cov.shrink()* with the default options, available in contributed R package “corpcor” (Schaefer et al., 2010), to calculate shrinkage estimate of the covariance matrix. The *cov.shrink()* function shrink the sample correlation matrix, $\mathbf{R} = (r_{ij})_{1 \leq i,j \leq p}$, towards the identity target estimate, $\mathbf{I}$. Replacing $\mathbf{S}$ by $\mathbf{R}$ and $\mathbf{T}$ by $\mathbf{I}$ in (5.7), the expression for shrinkage intensity in (5.9) solves down to

$$\hat{\rho}^* = \frac{\sum_{i=1}^p Var(r_{ij}, i \neq j)}{\sum_{i=1}^p (r_{ij}^2, i \neq j)}. \tag{5.11}$$

The shrunken covariance is then obtained using the equation

$$\widehat{\boldsymbol{\Sigma}}_{\rho} = \mathbf{S}_d^{1/2}\widehat{\mathbf{R}}_{\rho}\mathbf{S}_d^{1/2}, \tag{5.12}$$

where $\widehat{\mathbf{R}}_{\rho}$ is the regularized version of $\widehat{\mathbf{R}}$. This formulation is more appropriate when variables are measured on different scales. Note that, the *cov.shrink()* function also allows us to shrink the diagonal elements (this is the default option) with separate shrinkage intensity calculated by

$$\hat{\rho}^{*} = \frac{\sum_{i=1}^{p} Var(s_{ii})}{\sum_{i=1}^{p}(s_{ii}^{2} - median(s))^{2}}, \tag{5.13}$$

where $median(s)$ is the median of sample variances.

Figure 5.1 shows the ordered eigenvalues of a true covariance matrix (AR(1) with $b = 0.5$) in comparison with the sample covariance and shrinkage estimate, calculated for a sample of size 25 with dimension $p = 20$, drawn from a multivariate normal distribution (this is concordant with the results from Schäfer & Strimmer (2005b)). The eigenvalues for the sample covariance and shrinkage estimate are averaged over 1000 realizations. The figure illustrates that the sample covariance is likely to overestimate the larger eigenvalues and underestimate the smaller eigenvalues when p is large relative to the sample size. The shrinkage estimate deflates the large eigenvalues downwards and inflates the small eigenvalues upward, thereby overcoming the problem posed by dimensionality. Warton (2008) gives a schematic like that of Figure 5.2 to demonstrate the implications of using the shrinkage estimate of the covariance matrix. Geometrically, the shrinkage estimate reduces the eccentricity of the $100(1-\alpha)\%$ confidence ellipse. This increases the power if the shift is along the dominant eigenvector as illustrated by Figure 5.2(a). The power, however, will decrease if the shift is along the non-dominant eigenvector as shown in Figure 5.2(b). The diagram also shows that a larger shift is required along the dominant eigenvector to obtain a particular level of power as compared to if the shift is along non-dominant eigenvector.

It is important to note that the shrinkage estimator does not make any distributional assumptions about the underlying distribution of the data and its performance advantages are, therefore, not restricted to Gaussian assumptions. This is also the main reason why we avoided other regularization procedures (ridge and lasso) discussed in Chapter 2 because they use normal likelihood to select the tuning parameter, ρ.

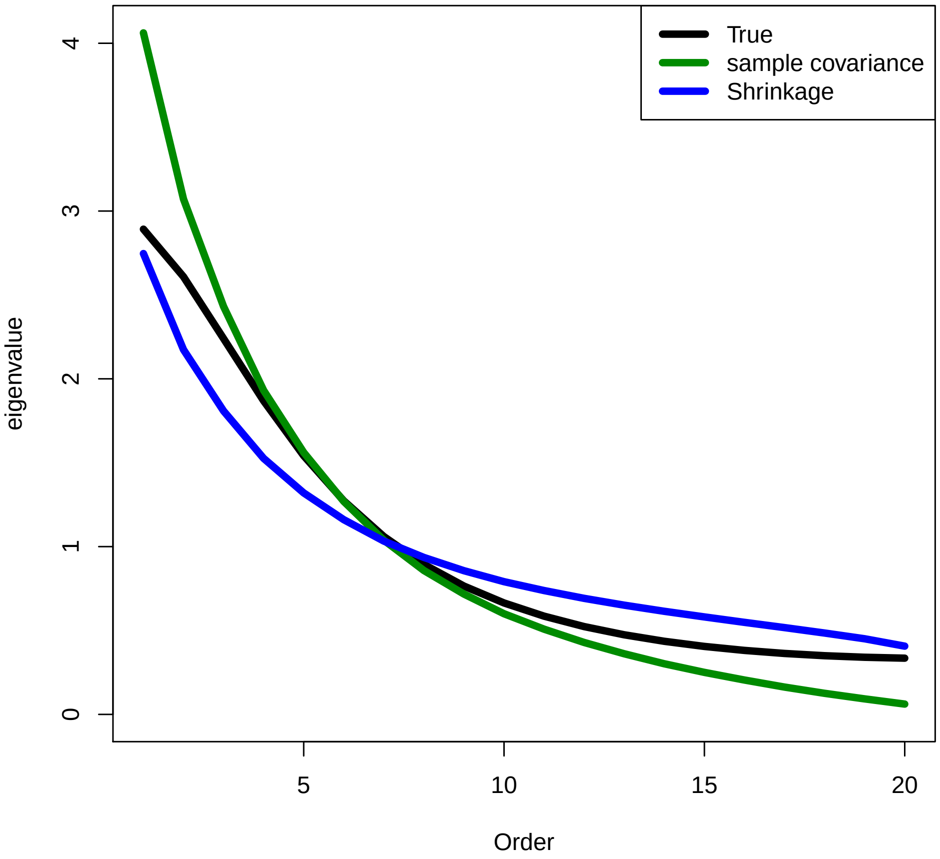


FIGURE 5.1: Ordered eigenvalues of a shrinkage estimate, $\widehat{\mathbf{\Sigma}}_\rho$, in comparison with the eigenvalues of a true covariance, $\mathbf{\Sigma}$, and the sample covariance matrix, $\mathbf{S}$. $\mathbf{\Sigma}$ is of AR(1) structure with $b = 0.5$, and $\mathbf{S}$ and $\widehat{\mathbf{\Sigma}}_\rho$ are calculated from a sample of size $n = 25$ drawn from a multivariate ($p = 20$) normal distribution with $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$.

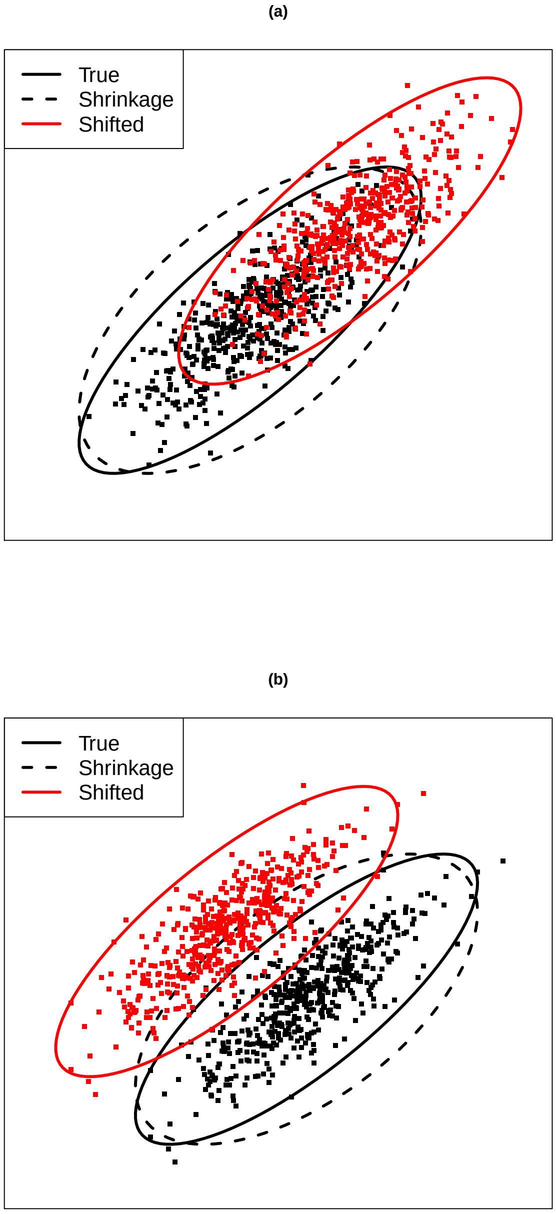


Figure 5.2: A hypothetical diagram illustrating the shrinkage effect for bivariate data (Warton, 2008). The lines are the 99% probability contours based on the true covariance (the solid lines) and shrunken covariance (the dashed lines). The shrinkage estimate reduces the eccentricity of the ellipse (the ellipse represented by solid black line is squashed along the major axis and make the ellipse represented by dashed line). The diagram also illustrates how the shift in different orientation can effect the power of a method to detect it. The red ellipse shows the shifted true distribution. It is shifted along the (a) first eigenvector (b) along second eigenvector. A larger shift is required along the first eigenvector to be detected as compared to the shift along second eigenvector. Note that "detected" means those red points that are outside the black ellipses.

5.3 Proposed procedure

Assume that we have a set of in-control baseline data, and also that p is large relative to n. The true parameters ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) are unknown and are required to be estimated from the baseline data. Since the sample covariance, $\mathbf{S}$, is either rank deficient (if $p > n$) or unstable (if p is comparable to n), it can be replaced by the shrinkage estimate described in the previous section (it makes the calculation of T^2 possible even when $p > n$). The distribution of T^2 when calculated using a shrinkage estimate of the covariance matrix is, however, unknown. The classical reference distributions (Beta distribution for Phase-I analysis and F distribution for Phase-II analysis) discussed in section 1, are no longer applicable because of the replacement of the sample covariance by its regularized counterpart. An obvious alternative is to use an empirical distribution of the T^2 values as an estimate of the reference distribution, and $(1-\alpha)th$ quantile as an estimate of the UCL for Phase-I observations. However, this is possible for Phase-I analysis but cannot be extended to Phase-II analysis because we do not have future observations. An alternative procedure is therefore required to obtain the UCL for Phase-II monitoring, as outlined below.

The T^2 for baseline observations, $\mathbf{X}_i$, and future observations, $\mathbf{X}_f$, follow different distributions. The reason for this difference is that the T^2 statistics calculated for the baseline observations are not independent of the estimated mean vector and sample covariance matrix while T^2 for future observations are independent (Tracy et al., 1992). This difference between the T^2 for baseline observations and T^2 for future observations can be exploited to generate T^2 values from the n available baseline observations, similar in properties to the T^2 values for future observations. This can be done by leaving one observation out at a time and using the rest of the $n-1$ baseline observations to estimate the mean vector and the covariance matrix. The T^2 statistic is calculated for the holdout observation using the mean vector and regularized estimate of covariance matrix calculated from the $n-1$ observations. The process is repeated, keeping each observation as a holdout in turn, and using the rest as a baseline data. The T^2 values obtained in this way are now independent of the estimated mean vector and sample covariance matrix and have the same properties as those of the T^2 values for future observations given that the process is in-control. Once we have n independent T^2 values, we use the empirical distribution of these T^2 values as an estimate of the reference distribution for Phase-II analysis. The $(1-\alpha)th$ quantile of this empirical distribution is the estimate of UCL for Phase-II monitoring.

Note that the empirical distribution obtained in this way is of a discrete nature. This can sometimes inflate or deflate the type-I error. A continuous distribution is desirable, especially if one is interested in the tail probabilities. This can be achieved using kernel density estimation (Polansky & Baker, 2000). We estimate the kernel distribution function using a built-in R function *kcde()* provided in contributed R package "ks" with default options (Duong, 2014).

Assuming that we have n p-dimensional observations and $\mathbf{X}_f$ denotes a future observation, then the step-by-step procedure is as follows:

1. Choose an observation in turn from the baseline data and call it $\mathbf{X}_f$, the future observation.

2. Calculate the mean vector and shrinkage estimate of a covariance matrix based on the remaining $n-1$ baseline observations excluding $\mathbf{X}_f$.

3. Calculate the T^2 statistic for $\mathbf{X}_f$ using the parameter estimates calculated in step 2.

4. Repeat steps 1-3 for each baseline observation to get n independent T^2 values, one for each observation.

5. Apply the kernel smoothing method to the T^2 values obtained in step 4 and estimate the distribution function.

6. The UCL of the control chart, for monitoring future observations, is the $(1-\alpha)th$ quantile of the kernel smoothed distribution function.

Note that we expect $\alpha\%$ T^2 values for future observations to be above the UCL, because the UCL is determined using an empirical distribution function. An examination of the algorithm reveals that we can, in fact, use it for Phase-I analysis as well, and we do this in a simulation study in the next section. We note that the procedure does not anywhere explicitly rely on normality of the data, although our assessments at this stage focus on multivariate normal data. Our reference distribution is based on an empirically generated distribution, it is valid even if the data are not normal. The shrinkage approach is also distribution free. It seems a promising non-parametric alternative and might work well for non-normal data.

5.4 Simulation study

In this section, we present a simulation study conducted to quantify the performance of the proposed method. We considered a number of different factors that could potentially affect the performance of the method: dimension, sample size, covariance structure, significance level α, and the orientation and size of the shift parameter δ for out-of-control observations.

In each simulation, we drew n baseline observations from a p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We used $n = 30$ and $p \in [2, 60]$ for baseline observations in order to show the characteristics of the proposed method for a range of p in relation to n. For the n baseline observations, we set $\boldsymbol{\mu}$ as $\mathbf{0}$ and considered two different covariance structures: the exchangeable structure given by

$$\sigma_{ij} \quad = \quad \begin{cases} 1 & \text{when } i = j \\ b & \text{when } i \neq j \end{cases} \qquad \text{for} \quad 1 \leq i, j \leq p \tag{5.14}$$

and the AR(1) structure given as

$$\sigma_{ij} = b^{|i-j|} \qquad \text{for} \quad 1 \leq i, j \leq p. \tag{5.15}$$

For both AR(1) and exchangeable structures, we did the experiments with $b \in \{0.3, 0.4, 0.6, 0.7\}$. We show the results for $b \in \{.3, .7\}$ because of the similar pattern of results across different values of b (the results for $b \in \{.4, .6\}$ are given in the AppendixB of the Thesis). The UCL obtained from the empirical distribution was in the far right tail and was sensitive to the largest T^2 values. To assess how the tail behavior affects the false alarm rate, we did the experiments with $\alpha \in \{.01, .05\}$. Other important aspects to consider were the size and orientation of shift parameter, δ. We did the experiment while making shift along the first, second, and last eigenvectors, and a shift along all eigenvectors. We varied the values of δ across different scenarios to examine cases of intermediate power that discriminate between the proposed and the classical methods. For the AR(1) structure, intermediate power was achieved with δ values 1, 7, 7, and 5 along all, 1st, 2nd, and pth eigenvector, respectively. For the exchangeable structure this was achieved with delta values 1, 14, 4, and 4 along all, 1st, 2nd, and pth eigenvector, respectively. Since most of the noise (error variation) lay along the dominant eigenvectors, a relatively larger shift was required to obtain the moderate power. On the other hand, a smaller shift was required to obtain a moderate power if it was along non-dominant eigenvectors.

Note that, for the AR(1) structure, the required δ value when looking at a shift along the second eigenvector was the same as that for shifting along the first eigenvector; for the exchangeable structure, the second eigenvector used the same δ as the pth eigenvector. This grouping of examples (shift along second eigenvector similar to first eigenvector for AR(1), similar to pth eigenvector for exchangeable) persisted in the other features of the power curves, so the second eigenvector simulations have been diverted to the supplementary material in Appendix B.

We used the proposed method and estimate the UCL using the baseline set of data to monitor future observations. For comparison, we present the corresponding results with T^2 values calculated using the true parameters as in (5.1). These are the best possible results one could hope to achieve. In practice, however, these parameters are unknown and are estimated from the baseline data. We also provide the results for the Hotelling T^2 method wherever possible (for $p < n$) using (5.5) with parameters estimated from baseline data. Note that the UCL for these two classical procedures are obtained from their respective known distributions discussed in section 1 (see (5.4) and (5.6)). A set of 5000 in-control future observations were generated to estimate α, the false alarm rate. A shift of size δ was created along different orientations using the same set of 5000 observations to make the data out-of-control. The false alarm rate and power were calculated, respectively, as the percentage of 5000 T^2 values for in-control observations that exceeded the UCL and the percentage of 5000 T^2 values for out-of-control observations that exceeded the UCL. This was repeated 5000 times, each time with a new set of baseline observations, to obtain 5000 values of false alarm rate and 5000 values of power. The averages over 5000 values of false alarm rate and power were examined and the typical results in different setups are presented in Figures 5.3 and 5.4.

The results were consistent across the two covariance structures. The power decreased as we increased p relative to n except if the shift was along all eigenvectors. This is not surprising because the shift cumulates across p dimensions and will be easily detected for larger value of p. On the other hand, a shift along a single dimension is masked by dimensionality and is harder to detect. The power of

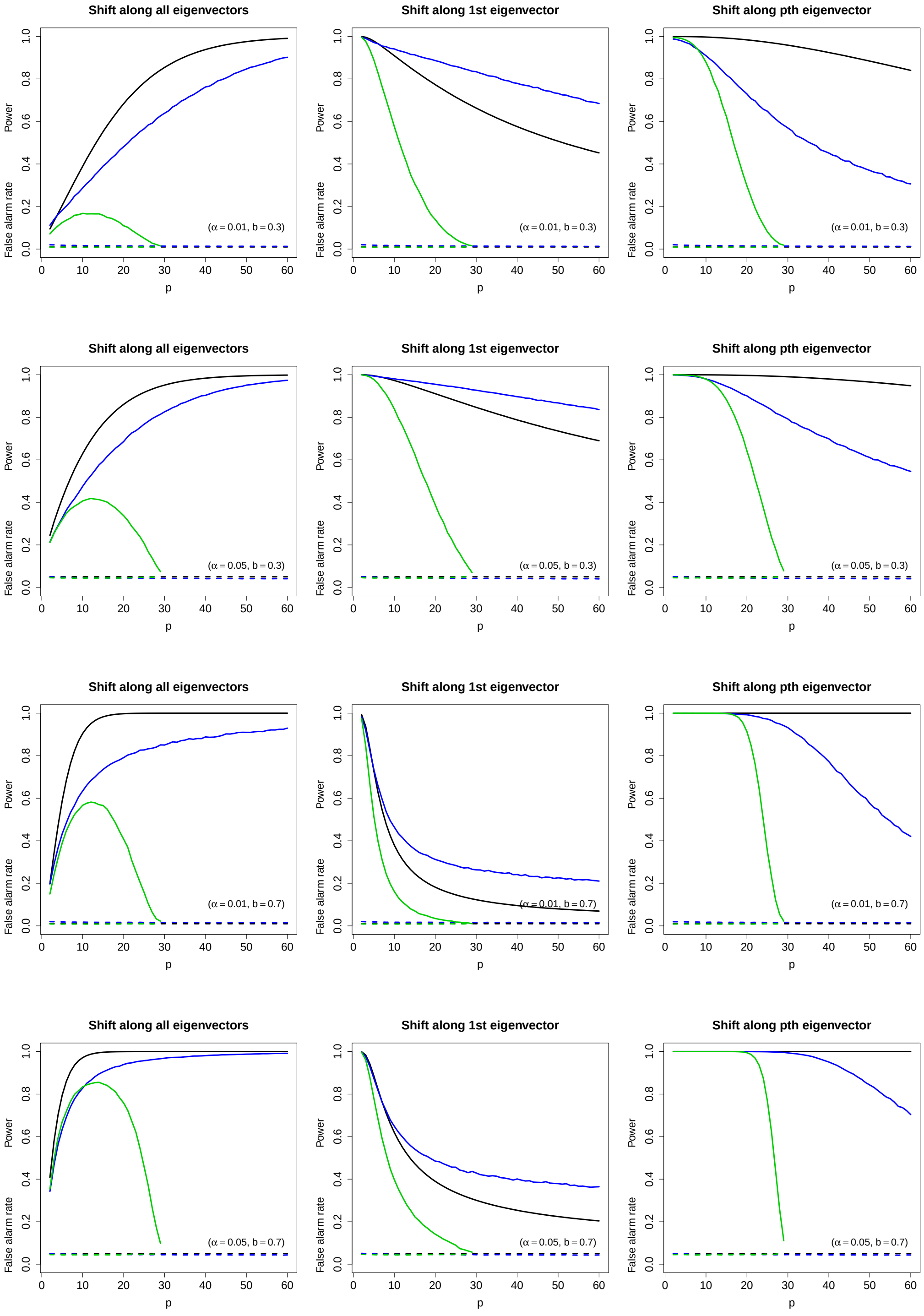


FIGURE 5.3: Power (solid lines) and false alarm rate (dashed lines) for AR(1) covariance structure. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for the Hotelling T^2 control chart (uses sample covariance matrix) and the blue lines are the results from new method.

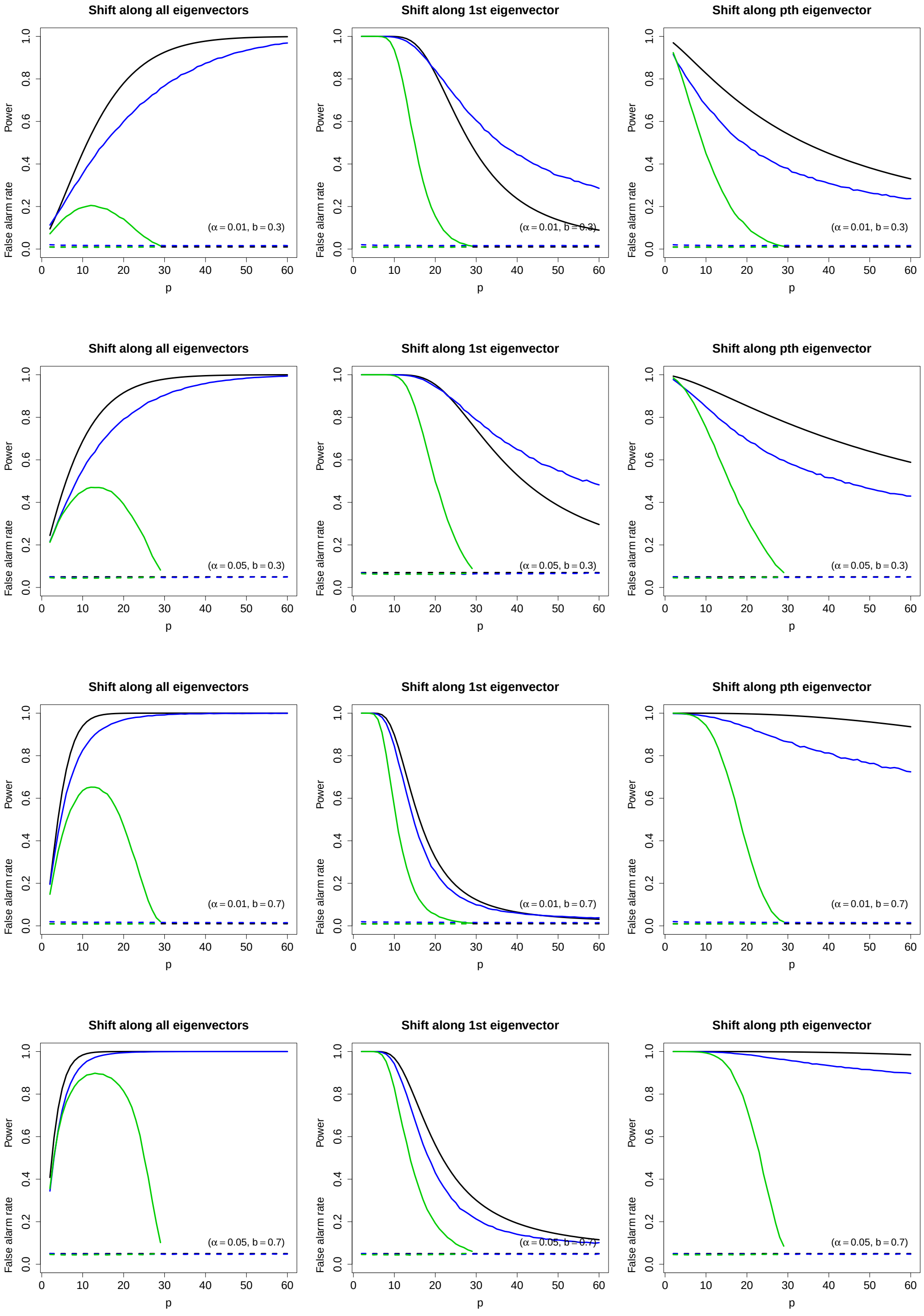


FIGURE 5.4: Power (solid lines) and false alarm rate (dashed lines) for exchangeable covariance structure. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for the Hotelling T^2 control chart (uses sample covariance matrix) and the blue lines are the results from new method.

the standard method deteriorated faster with increasing p compared to the new method and was not applicable for $p > n$.

The proposed method performed well throughout simulation study and was comparable to the results based on the true parameters. In fact, the power was higher than the results based on the true parameters if the shift is along first eigenvector. This was expected because of the shrinkage effect that is explained in Figure 5.2. The false alarm rate (Type I error) was slightly lower than the nominal value on the average for larger value of α because the extreme values in the far tail of the empirical distribution are less likely to occur and UCL is sensitive to these extreme values. This problem, however, was not very serious for moderate values of α (see results for $\alpha = 0.01$ and $\alpha = 0.05$).

5.5 The in-control run length performance

An average run length (ARL) is a measure to describe the performance of the control charts. In the context where the individual observations are monitored, the run length is the number of in-control observations that must be collected before an out-of-control signal appears (Montgomery & Woodall, 1999).

The ARL loses much of its attractiveness as a summary because it follows a geometric distribution which is positively skewed (Montgomery, 2007). Instead a median run length (MRL) is used to quantify the performance of control charts. We conducted simulation experiments to calculate the MRL of the proposed procedure. A random sample of size $n \in \{50, 100\}$ were simulated from a multivariate normal distribution, $N_p(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is either an AR(1) structure or exchangeable structure with $b = 0.6$. This was used to estimate the control limits for monitoring the future observations using the proposed method. Another in-control sample of size 100,000 was simulated to calculate MRL. This process was repeated 5,000 times, each time the control limit was estimated using a new set of baseline data. The median MRL over 5000 replications with their respective lower and upper quartiles are presented in Table 5.1 for $n = 50$ and in Table 5.2 for $n = 100$. The experiments are conducted for different choices of p and α (the nominal false alarm rate), $p \in \{50, 100, 150\}$ and $\alpha \in \{0.01, 0.02, ..., 0.10\}$.

The covariance structure does not have a noticeable effect on the MRL performance. The MRL values are closer to the desired MRL values (with lower interquartile range) for the higher values of alpha. For lower values of α the in-control MRL is higher than the desired MRL and the degree of departure from the desired MRL increases as we decrease the value of α. The higher median MRL is better in the sense that it checks less on false alarms than the nominal one, but there is a lot of variability, with more much frequent false positives for a substantial number of cases (see the lower quartile). The MRL values becomes closer to the desired MRL if we increase the sample size from 50 to 100. The variability reflects the varying quality of the covariance estimates based on the baseline observations.

5.6 Practical applications

In this section, we illustrate our method using two different real data sets. The first data set is a high-dimensional gene expression application (271 variables). The second data set, although not high-dimensional (4 variables), has been analyzed in previous literature.

5.6.1 Example 1: Gene expression data

We tested the method on a gene expression data set constructed from two studies (Pawitan et al. (2005), Miller et al. (2005)). Both studies took population samples of breast cancers from women treated in Sweden, but in different time periods, 1987-1989 and 1995-1996 respectively. Measurements of low quality or without associated survival data were eliminated, leaving 232 cases for the earlier data collection period and 159 for the later one.

The data are measurements of gene expression in breast cancer tumors using H133 A affymetrics chips. The composition of the samples is similar in terms of the tumour molecular subtypes and estrogen receptor status (Chisquared tests of association p=0.593 and 0.394 respectively). There are some differences in the descriptions of the protocols used to collect the tissues; the earlier samples (232 patients) are described as “frozen” while the later ones (159 samples) are “frozen immediately on dry ice or in liquid nitrogen and stored in -70C freezers.” We treated

Table 5.1: In-control median run length (MRL) under both the AR(1) structure of covariance and the exchangeable structure of covariance with $b = .6$ (the lower and upper quartiles are shown inside the parentheses). The desired MRL is the median of the geometric distribution with parameter α. The size of the baseline set of data is 50.

α	Desired MRL	AR(1)			Exchangeable		
		$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
0.01	68.00	78.0 (39.0, 192.6)	82.0 (41.0, 204.0)	83.0 (42.5, 200.0)	75.0 (38.0, 178.5)	74.0 (38.0, 173.0)	81.0 (39.0, 197.5)
0.02	34.00	45.5 (26.0, 90.0)	46.0 (26.0, 92.0)	48.0 (27.0, 90.0)	43.0 (24.0, 87.0)	43.0 (25.0, 84.0)	40.0 (23.0, 80.0)
0.03	22.00	28.0 (17.0, 50.0)	29.0 (18.0, 49.1)	29.0 (18.0, 49.0)	28.0 (17.0, 51.0)	27.0 17.0, 47.0)	26.0 (17.0, 44.0)
0.04	16.00	21.0 (14.0, 35.0)	22.0 (14.0, 35.0)	22.0 (15.0, 36.0)	21.0 (13.0, 35.0)	20.0 (13.0, 32.0)	19.0 (12.0, 30.0)
0.05	13.00	17.0 (11.0, 26.0)	17.0 (12.0, 27.0)	17.0 (12.0, 27.0)	16.0 (11.0, 26.0)	16.0 (10.0, 24.0)	15.0 (10.0, 22.0)
0.06	11.00	13.0 (9.0, 20.0)	14.0 (9.0, 21.0)	14.0 (10.0,21.0)	13.0 (9.0, 20.0)	13.0 (9.0, 19.0)	12.0 (8.0, 17.0)
0.07	9.00	11.0 (8.0, 16.0)	12.0 (8.0, 17.0)	12.0 (8.0,17.0)	11.0 (8.0, 17.0)	11.0 (7.0, 16.0)	10.0 (7.0, 14.0)
0.08	8.00	10.0 (7.0, 14.0)	10.0 (7.0, 14.0)	10.0 (7.0, 14.0)	10.0 (7.0, 14.0)	9.0 (6.0, 13.0)	9.0 (6.0, 12.0)
0.09	7.00	8.0 (6.0, 12.0)	9.0 (6.0, 12.0)	9.0 (6.0, 12.0)	8.0 (6.0, 12.0)	8.0 (6.0, 11.0)	7.0 (5.0, 10.0)
0.10	6.00	7.0 (5.0, 10.0)	8.0 (6.0, 11.0)	8.0 (6.0, 11.0)	7.0 (5.0, 11.0)	7.0 (5.0, 10.0)	6.0 (5.0, 9.0)

Table 5.2: In-control median run length (MRL) under both the AR(1) structure of covariance and the exchangeable structure of covariance with $b = .6$ (the lower and upper quartiles are shown inside the parentheses). The desired MRL is the median of the geometric distribution with parameter α. The size of the baseline set of data is 100.

α	Desired MRL	AR(1)			Exchangeable		
		$p = 50$	$p = 100$	$p = 150$	$p = 50$	$p = 100$	$p = 150$
0.01	68.00	84.0 (48.0, 168.0)	81.0 (48.0, 156.0)	80.0 (47.0, 157.0)	84.0 (47.0, 162.0)	83.0 (49.0, 163.0)	86.0 (50.0, 160.0)
0.02	34.00	40.0 (26.0, 66.6)	41.0 (28.0, 64.0)	41.0 (27.0, 65.0)	40.0 (26.0, 63.0)	42.0 (27.0, 69.0)	41.0 (28.0, 66.0)
0.03	22.00	27.0 (19.0, 42.0)	27.0 (19.0, 40.0)	26.0 (18.0, 40.0)	26.0 (19.0, 38.0)	27.0 18.0, 42.0)	27.0 (19.0, 39.0)
0.04	16.00	19.0 (14.0, 28.0)	19.0 (14.0, 27.0)	20.0 (14.0, 28.0)	19.0 (14.0, 27.0)	20.0 (14.0, 29.0)	20.0 (15.0, 28.0)
0.05	13.00	15.0 (11.0, 22.0)	15.0 (11.0, 21.0)	15.0 (11.0, 21.0)	15.0 (11.0, 21.0)	16.0 (12.0, 22.0)	15.0 (12.0, 21.0)
0.06	11.00	13.0 (9.0, 17.0)	12.0 (10.0, 17.0)	13.0 (10.0,17.0)	12.0 (9.0, 16.0)	13.0 (10.0, 18.0)	13.0 (10.0, 17.0)
0.07	9.00	11.0 (8.0, 14.0)	11.0 (8.0, 14.0)	11.0 (8.0,14.0)	10.0 (8.0, 14.0)	11.0 (8.0, 15.0)	11.0 (8.0, 14.0)
0.08	8.00	9.0 (7.0, 12.0)	9.0 (7.0, 12.0)	9.0 (7.0, 12.0)	9.0 (7.0, 11.0)	9.0 (7.0, 12.0)	9.0 (7.0, 12.0)
0.09	7.00	8.0 (6.0, 10.0)	8.0 (6.0, 10.0)	8.0 (6.0, 10.0)	8.0 (6.0, 10.0)	8.0 (6.0, 11.0)	8.0 (6.0, 10.0)
0.10	6.00	7.0 (6.0, 9.0)	7.0 (6.0, 9.0)	7.0 (5.0, 9.0)	7.0 (5.9, 9.0)	7.0 (6.0, 9.1)	7.0 (6.0, 9.0)

the differences between the groups as examples of differences that might arise if a specified protocol was deviated from or a different protocol was mistakenly used.

We did not use all the gene expression measurements, but rather a subset of scientific interest, 271 genes associated with the ERBB2 pathway. These were measured on 391 patients in total. A plot of the data on the first two principle components of the 271 variables is shown in Figure 5.5. The data from the different studies are represented by numbers 1 (1987-1989) and 2 (1995-1996) in the plot. The PCA plot shows a clear separation between the two studies.

We analyzed the data using our method while considering the data from study 1 as the baseline data and the one from study 2 as hypothetical out-of-control observations, as might result from a change in laboratory conditions or a mistake in protocol. The multivariate control chart in Figure 5.6 shows both the baseline data (black) and future data (red) with 842.4466 as the UCL at 5% level of significance. There are 12 points out of total 232 points, in the baseline data above the UCL. Unlike the classical procedure, where the UCL is obtained from a known distribution, the UCL here is obtained from the empirical distribution, and 12 points are constrained to be above the UCL at $\alpha = 0.05$. These 12 points (encircled 1's in Figure 5.5) should be investigated to ascertain reasons for their departure from the normal state. However, none of the 12 points were found to be largely influential and we retained them in the sample.

As shown in Figure 5.6, 72% of study 2 observations were identified as out-of-control. These out-of-control observations are encircled in the PCA plot in Figure 5.5. We identify some points that are different in ways not obvious by looking at first 2 PC's in Figure 5.5. Note that the classical control charts methods could not be applied here because there are more genes (variables) than patients (observations) in each group, so the sample covariance matrix is not invertible.

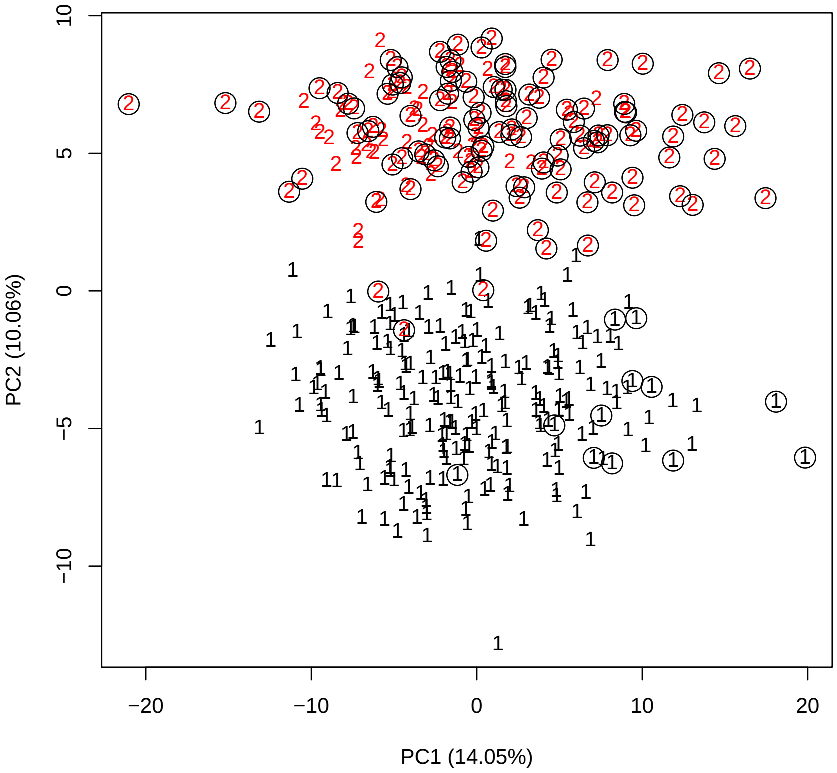


FIGURE 5.5: First two principle components of the gene expression data. All out-of-control points in Figure 5.6 (including 12 out-of control points from baseline set of data) are encircled.

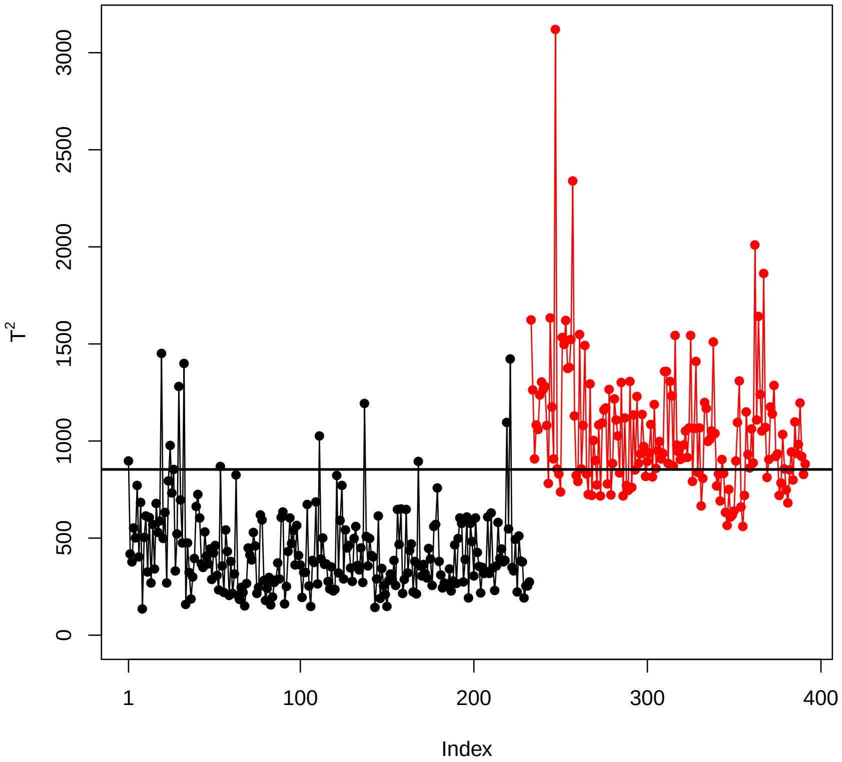


FIGURE 5.6: Multivariate control chart using study 1 data as a baseline set of data (black) and study 2 as a future set of data (red). The solid line at $T^2 = 853.504$ represent the UCL at 5% level of significance.

5.6.2 Example 2: Chemical process data

Here we apply the method to chemical process data with four variables and 30 observations. The data set is given in Table 5.3 and was originally used by Montgomery (2007) to demonstrate the method of principal component analysis for multivariate process monitoring. Following Montgomery (2007), the first 20 observations were used as Phase-I data and the last ten observations were used for testing and monitoring (Phase-II). The scatter plot of the first two principal component scores computed using the Phase-I observations is shown in Figure 5.7. The Phase-II observations are projected onto the PCA plot of Phase-I observations. The Phase-II observations are represented by a different plotting symbols together with the numbers from 21 to 30, to show the sequence of points. The observations outside the 99% confidence ellipse are considered the out-of-control signals and indicates that there has been a shift in the process mean. We analyze the data using our method and present the control chart in Figure 5.8. It is clear that our method shows more sensitivity and detects the shift one point in time earlier than the principal component trajectory plot.

To further assess our method in comparison with the principal components approach under a smaller sample size, we drop the first 10 observations of the baseline data and re-analyze the data using the middle 10 observations as a baseline set of data. The resultant PCA plot shown in Figure 5.9 while the control chart produced by our method is in Figure 5.10. The PCA performance is poor – because it fails to identify most of the out-of-control observations as only three observations (24rd, 26th, and 29th) are outside the 99% confidence ellipse. On the other hand, our method seems to be robust to the change in the sample size. It has detected 23rd the out-of-control signals even with the smaller Phase-I sample size ($n = 10$).

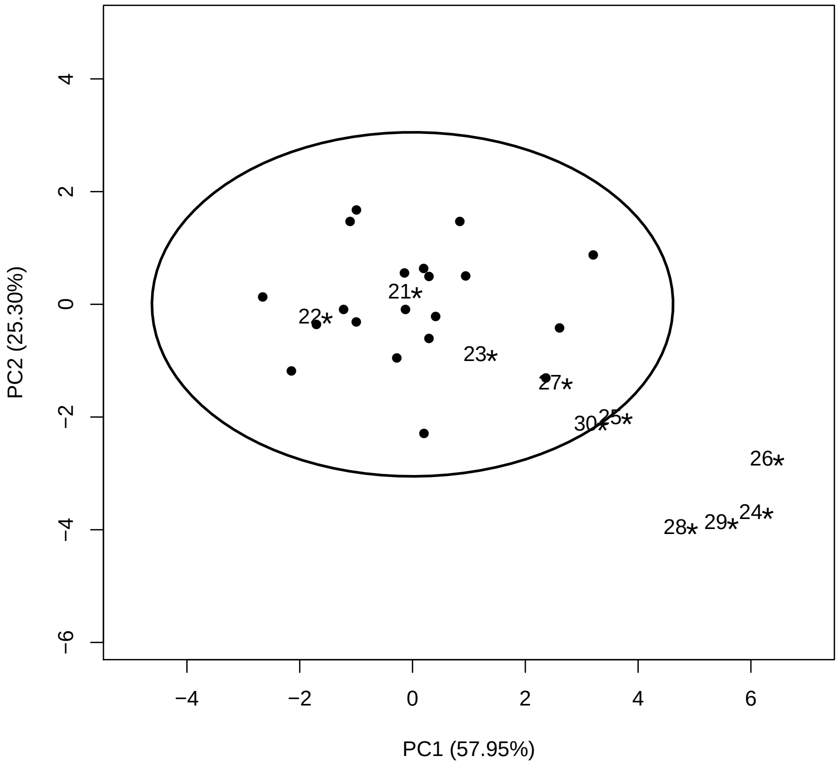


FIGURE 5.7: Principal components trajectory plot for the chemical process data with 99% confidence ellipse. Only the first 20 baseline observations are used to compute the principal components. The 10 future observation are plotted with star symbols and are numbered from 21 to 30 in order to show the natural sequence of the points.

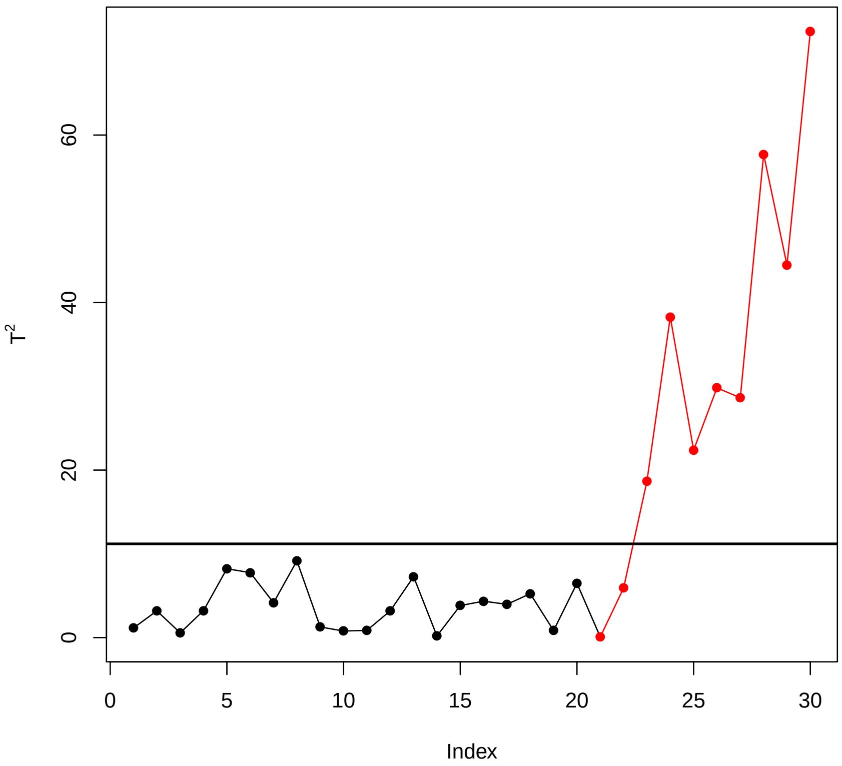


FIGURE 5.8: Control chart produced by the proposed method for monitoring the chemical process data shown in Figure 5.7. The first 20 observations are used to estimate the control limits. The solid line at $T^2 = 11.1887$ indicates the control limit of the chart at 1% level of significance.

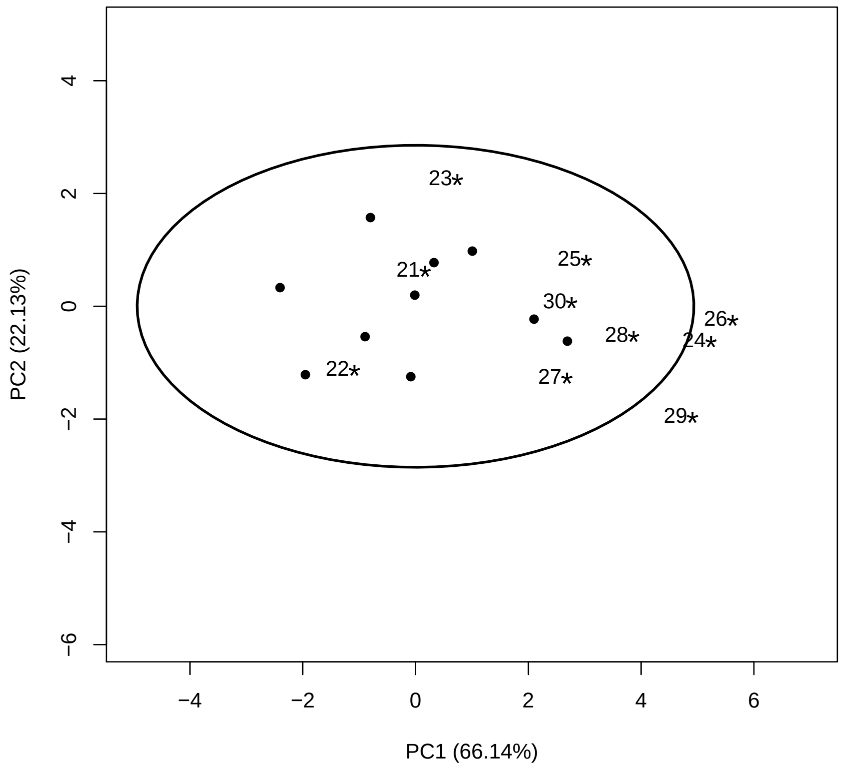


Figure 5.9: Principal components trajectory plot for the chemical process data with 99% confidence ellipse. Note that the first 10 baseline observations are dropped from the analysis and the principal components are computed from the middle 10 observations. The last 10 observations are plotted with star symbols and are numbered from 21 to 30 in order to show the natural sequence of the points.

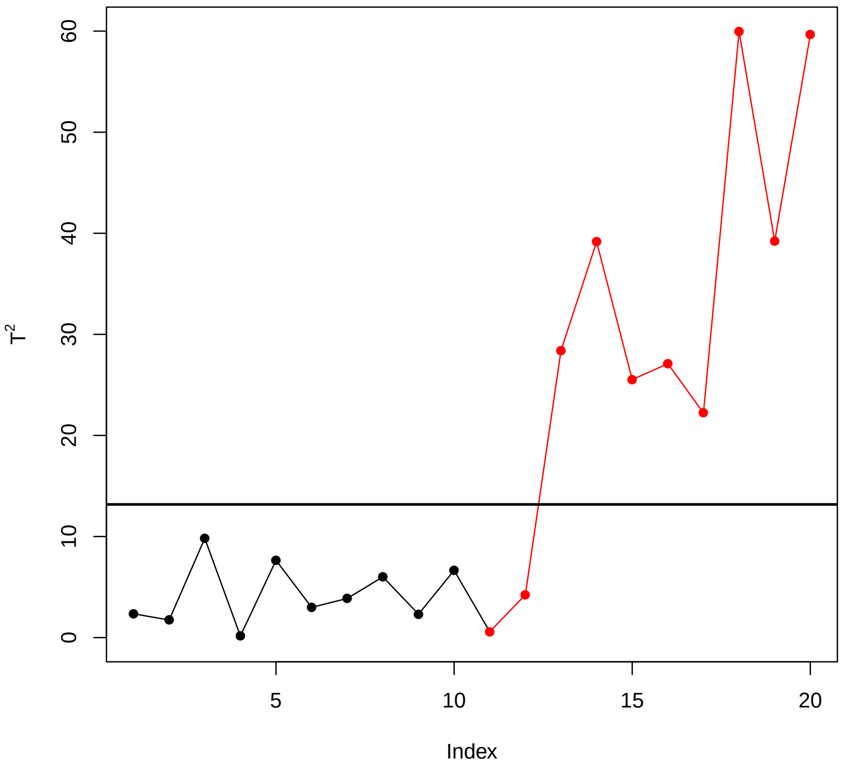


FIGURE 5.10: Control chart produced by the proposed method for monitoring the chemical process data shown in Figure 5.9. Only the middle 10 observations are used to estimate the empirical reference distribution. The solid line at $T^2 = 13.1733$ indicates the control limit of the chart at at 1% level of significance.

5.7 Summary and Conclusion

In this Chapter, we propose a method to detect unusual observations or changes in a high-dimensional stochastic process. In high-dimensional problems, the T^2 statistic is impossible to calculate if $p > n$ and even when $p < n$ the standard theory (Hotelling T^2 control charts) leads to a procedure with low power. We chose the shrinkage estimation of the covariance matrix to not only make the calculation of T^2 possible, but also reliable. Other regularization techniques for the covariance matrix (e.g. ridge or lasso regularization) could be used, but they are computationally expensive, typically relying on cross validation and use normal likelihood to select the value of the penalty parameter. One reason for choosing the shrinkage estimate is that it is based on a fully automated and computationally inexpensive data driven method. Other reason for choosing shrinkage estimator is that it does not make any distributional assumptions about the underlying distribution of the data and its performance advantages are potentially not restricted to Gaussian assumptions. This was combined with a leave-one-out re-sampling procedure to obtain n independent T^2 values. To estimate the distribution of T^2 for future observations, we used kernel smoothing. The UCL for monitoring future observations was the $(1-\alpha)$th quantile of the kernel smoothed distribution.

The performance of the procedure was evaluated using extensive Monte Carlo simulation. The method was compared with the standard procedure wherever possible and also with a hypothetically best case, based on the true parameters. We showed numerically that the new procedure competes well and had considerable power to detect signals under various simulations. The ability of the procedure to accurately characterize the shift in mean vector is also shown by applying it to a gene expression data, where we use data from two studies with different protocols to demonstrate detecting a protocol change or error. One natural competitor of the proposed method is the principal component approach. We use the chemical process data (previously used to demonstrate the principal component approach) to show the advantages of the proposed method.

The method may perform well in multivariate non-normal data, as it does not assume the data to be normally distributed. Further simulation, however, would be required to investigate the performance of the method in this more general setting. We leave assessment of this to future work.

The control charts generally use much lower false alarm rates. Our procedure works well for larger significance values (e.g., 5%). Therefore, this approach will be useful in situations where there is a high priority of quickly detecting any shift in the process, or where the follow-up to a detected outlier is easy and inexpensive, so that a higher false alarm rate can be tolerated. Note also that the proposed procedure relies on kernel estimation, it is sensitive to the quality and size of the training sample.

5.8 Contributions of the Chapter

The Hotelling T^2 control chart becomes unreliable and even impractical when $n < p$. In this Chapter, we propose a procedure to improve process monitoring in the high-dimensional setting. We use a shrinkage estimate of the covariance matrix as an estimate of the baseline parameter. A leave-one-out re-sampling procedure is used to obtain independent T^2 values. The upper control limit for monitoring the future observations is then calculated from kernel smoothed empirical distribution of the independent T^2 values. The performance of the proposed approach is tested, and compared to the Hotelling T^2 and the hypothetically "best possible" results, via an extensive simulation study. The procedure outperforms the standard Hotelling T^2 method and gives comparable results to the one based on true parameters. The procedure is also applied to a real gene expression data set and a chemical process data that has been analyzed in literature to demonstrate the principal component approach for process monitoring.

Table 5.3: Chemical process data. There are total 30 observations. The first 20 observations constitute the baseline set of data and the last 10 observations are the new observations used for testing and monitoring.

Observation	x1	x2	x3	x4
1	10.00	20.70	13.60	15.50
2	10.50	19.90	18.10	14.80
3	9.70	20.00	16.10	16.50
4	9.80	20.20	19.10	17.10
5	11.70	21.50	19.80	18.30
6	11.00	20.90	10.30	13.80
7	8.70	18.80	16.90	16.80
8	9.50	19.30	15.30	12.20
9	10.10	19.40	16.20	15.80
10	9.50	19.60	13.60	14.50
11	10.50	20.30	17.00	16.50
12	9.20	19.00	11.50	16.30
13	11.30	21.60	14.00	18.70
14	10.00	19.80	14.00	15.90
15	8.50	19.20	17.40	15.80
16	9.70	20.10	10.00	16.60
17	8.30	18.40	12.50	14.20
18	11.90	21.80	14.10	16.20
19	10.30	20.50	15.60	15.10
20	8.90	19.00	8.50	14.70
21	9.90	20.00	15.40	15.90
22	8.70	19.00	9.90	16.80
23	11.50	21.80	19.30	12.10
24	15.90	24.60	14.70	15.30
25	12.60	23.90	17.10	14.20
26	14.90	25.00	16.30	16.60
27	9.90	23.70	11.90	18.10
28	12.80	26.30	13.50	13.70
29	13.10	26.10	10.90	16.80
30	9.80	25.80	14.80	15.00

Chapter 6

General Conclusions

In recent years, there has been intense activity in improving estimation of the covariance matrix when $n < p$, or they are of comparable size. Despite this, the regularized estimators can be rarely seen in practice. One reason is because the covariance is merely an ingredient of the common multivariate procedures. Replacing the sample covariance by its regularized alternatives does not guarantee that the procedure still works. There is a need to adapt procedures to allow the use of regularized estimators, and assess these new techniques. When regularized estimators have found their way into multivariate techniques, their assessment is frequently restricted to the applications that were the original motivation for its development. For example, lasso-type regularization has been used primarily when a sparse covariance is expected, while the shrinkage estimate is common in portfolio optimization problems. There is a need for a more comprehensive comparison, to broadly characterize the situations where each method performs well.

The main objective of this Thesis was to explore some of the important regularization techniques for estimation of high-dimensional covariance matrices and their properties in different high-dimensional multivariate methods. We have investigated the behavior of the regularized alternatives to the sample covariance in three different multivariate techniques. These are summarized in Sections 6.2, 6.3, and 6.4 with common themes and future directions highlighted in Section 6.5.

6.1 Regularized estimation of the high-dimensional covariance matrices

The behavior of the sample covariance in high-dimensional problems is well-known to be poor. In recent years various regularization techniques have been developed to improve over the sample covariance in a high-dimensional setting. Chapter 2 reviews some of the important recent regularization techniques. These techniques include the shrinkage estimation of the sample covariance matrix, the ridge regularization, the lasso regularization, and weighted versions of the lasso regularization. To assess how well different regularized estimators perform to estimate the true covariance matrix, a simulation study was conducted. The lasso estimator, although computationally intensive and assuming multivariate normality, has the competitive accuracy. The shrinkage estimator, on the other hand, is computationally inexpensive and does not make distributional assumption about the underlying set of data. The ridge estimator differs from the shrinkage estimator in that it chooses the shrinkage parameter using cross-validation. In our simulations, we found no advantage of the cross-validation over the computationally fast closed form expression used to estimate the shrinkage intensity for the shrinkage estimator. Our simulations show that the covariance matrix can be more accurately estimated using lasso penalty (if diagonal elements are not penalized) rather than using adaptive lasso and SCAD penalties that are shown to be superior to the lasso penalty in model selection.

6.2 Hierarchical covariance estimation

Multivariate random effect models need proper estimates of the within-group and the between-group covariance matrices. The computation of a between-group covariance involves the difference of two mean-square matrices and often results negative elements on the diagonal. The probability of negative elements increases as we increase the number of variables. This makes it difficult, in the high-dimensional setting, to obtain a proper between-group covariance matrix. In Chapter 3, a hierarchical model is proposed based on the EM-algorithm. The Lasso-regularized estimates of the covariance matrices are embedded in the algorithm to ensure estimates of covariance matrices are positive definite. Our simulation study showed that the proposed model performs well and returns a positive definite estimate of the between-group covariance. We applied the method to a glass chemical composition data. In this data set, we are interested in the between-group covariance and the standard estimate is not positive definite. We show that using the proposed method one can obtain a proper between-group covariance. In our simulation experiments, the algorithm needed few iterations to converge (especially in cases where the between-group variation is dominant) and is a good replacement in the situations where the traditional analysis of variance technique fails to work.

6.3 Regularized MANOVA for high-dimensional data

MANOVA is used to test hypothesis of group effects on multiple response variables simultaneously. High-dimensionality poses a serious problem to MANOVA tests. We have shown that the estimation error in terms of eigenvalues of covariance matrices is the least, on average, for the lasso-regularized covariance matrix in comparison with ridge and shrinkage estimates. In Chapter 4 of the Thesis, we propose an approach based on the lasso regularization. We investigate the behavior of the novel approach via extensive simulations taking into account a number of different factors. The new approach is also compared with other existing approaches.

In our simulation experiments, the three recent high-dimensional regularization procedures of the sample covariance: the shrinkage, the ridge, and the lasso regularization perform well, and in many cases perform better than the more conventional generalized inverse and principal component approach. None of the shrinkage, the ridge, and lasso is universally superior across scenarios considered in our simulation study. Lasso regularization is always the best when the shift is along all the eigenvectors but the difference is not dramatic. The difference, however, increases as we increase the number of variables. The better performance of lasso regularization seems to be because of its better recovery of the true eigenvalues.

Lasso regularization, however, is computationally expensive than the other competing procedures we have considered in this study. This is because lasso regularization by itself is an iterative procedure. The two computationally cumbersome jobs, cross-validation to choose penalty parameter and permutation test, add substantially to its computational time. While not computationally prohibitive its computational time is increasing dramatically with the number of variables for a fixed sample size. Ridge regularization is comparatively less expensive than lasso but its implementation also involves cross-validation. Principal component approach, generalize inverse and shrinkage approach are computationally simple and very fast. However, the performance of principal component approach have been very poor both in our simulation experiments and the real data. The performance of generalized inverse has been very poor in the zone where p is large but close to n. Taking both things — power and computational time— into consideration the shrinkage approach has an excellent balance.

6.4 Monitoring individual future observations in the high-dimensional setup

The Hotelling T^2 control chart is used to detect unusual changes in the mean vector of a stochastic process while monitoring individual future observations. In high-dimensional problems the T^2 statistic is impossible to calculate if $n < p$ and even for $n > p$, the UCL of the Hotelling T^2 control chart for monitoring future observations is unknown. In Chapter 5, a new method is proposed for monitoring future observations in high-dimensional setting. The method is superior to the standard Hotelling T^2 control chart in practice.

We evaluate the performance of the novel procedure using extensive Monte Carlo simulation. The method is also compared with the standard procedure wherever possible and also with a hypothetical "best case" that is based on the true parameters. We show numerically that the new procedure competes well and gains considerable power to detect signals under various simulations. The ability of the procedure to accurately characterize the shift in mean vector is also shown by applying it to a gene expression data, where we use data from two studies with different protocols to demonstrate detecting a protocol change or error. One natural competitor of the proposed method is the principal component approach. We use the chemical process data (previously used to demonstrate the principal component approach) to show the out-performance of the proposed method.

It is important to note that we do not make any particular assumption about the distribution of the data. The method may perform well in multivariate non-normal data. Further simulation, however, would be required to investigate the performance of the method in multivariate non-normal data. Although the new method is proposed to monitor future observations, it can be used for Phase-I analysis as well.

6.5 Commonalities and future work

Lasso regularization is being used for identifying zero elements in the inverse covariance matrix (model selection in the context of Gaussian graphical models). It can be used to obtain a regularized estimate of a covariance matrix. The shrinkage and ridge regularization are being used to estimate high-dimensional covariance matrices but always produce dense estimate of the inverse covariance matrix. In Chapter 2 and Chapter 4, the lasso regularization was expected to have a bigger advantage on the AR(1) structure (because its inverse is sparse) as compared to the exchangeable structure but it is not true in our simulation. The performance of the regularized covariance matrix obtained using lasso regularization is better even if the inverse covariance matrix is not sparse. The computational time of lasso regularization is, however, less for the covariance structure whose inverse is sparse.

Chapter 4 and 5 both involve creating reference distributions. In Chapter 5, we drop one observation from the baseline set of data at a time and estimate the parameters from the rest. The Hotelling T^2 statistic is calculated for the observation that was dropped. This gives us only n different T^2 values. The empirical distribution of the T^2 values is more of a discrete nature (we avoid this using kernel density estimate) and is not suitable for estimating the p-value. This is totally different from the permutation test we have used in Chapter 5 because there are many possible permutations even for a small sample size that produce many values of the test statistic. The empirical distribution in this case is smoother and is suitable to estimate p-value.

The first problem that needs to be further investigated is related to the two dimensionality reduction techniques we have considered in Chapter 4, namely, Moore-Penrose generalized inverse and principal component approach. The generalized inverse exhibit rather surprisingly good behavior outside the critical zone (when n is just less than p) relative to other methods we have considered in Chapter 4. On the other hand, the performance of the approach based on principal components has been poor. Both these methods reduces the dimension of the data. The generalized inverse reduces the dimension from p to the $rank(\widehat{\mathbf{\Sigma}})$. It reduces to standard matrix inverse when $rank(\widehat{\mathbf{\Sigma}}) \geq p$. The performance of the standard inverse covariance matrix is already known to be poor in high-dimensional problems. The principle component approach reduces the dimension of the data to first q principal components (q is the number of principal components chosen using the eigenvalue-greater-than-one rule). The two methods becomes similar and are expected to provide similar results when the $rank(\widehat{\mathbf{\Sigma}}) = q$. The performance of generalized inverse might improve in the critical zone if we use it with fewer dimensions than the $rank(\widehat{\mathbf{\Sigma}})$ and do not allow it to reduce to the standard matrix inverse. A different criteria to choose an intermediate number of dimensions would be required rather than the $rank(\widehat{\mathbf{\Sigma}})$ or eigenvalue-greater-than-one rule.

Another problem that needs to be investigated is related to the method we presented in Chapter 5 for monitoring future observations. To obtain the UCL, a kernel density approach is used that is good to estimate central quantiles. Another possibility is to estimate the quantiles using extreme value distribution (Beirlant et al., 2006). This approach is preferred to estimate the extreme quantiles and may perform better than the kernel density approach. It needs to be explored.

Finally, real data is frequently non-normal. The Hotelling T^2 control chart method for detecting unusual future observations assume that the observations follow a multivariate normal distribution. In the procedure presented in Chapter 5, our reference distribution is based on an empirically generated distribution, it is valid even if the data are not normal. The shrinkage approach is also distribution free. It seems a promising non-parametric alternative and might work well for non-normal data. Further investigation would be required to study the procedure for different flavors of non-normality.

Appendix A

Supplementary Figure for Chapter 4

Table A.1: Values of the shift parameter, δ, used in the simulation experiments whose results are presented in Figure A.1 and Figure A.2.

Orientation	Covariance structure	
of δ	AR(1)	Exchangeable
along 1^{st} eigenvector	4.080	5.040
along 2^{nd} eigenvector	2.280	2.280
along p^{th} eigenvector	1.560	1.560
along all eigenvectors	0.516	0.504

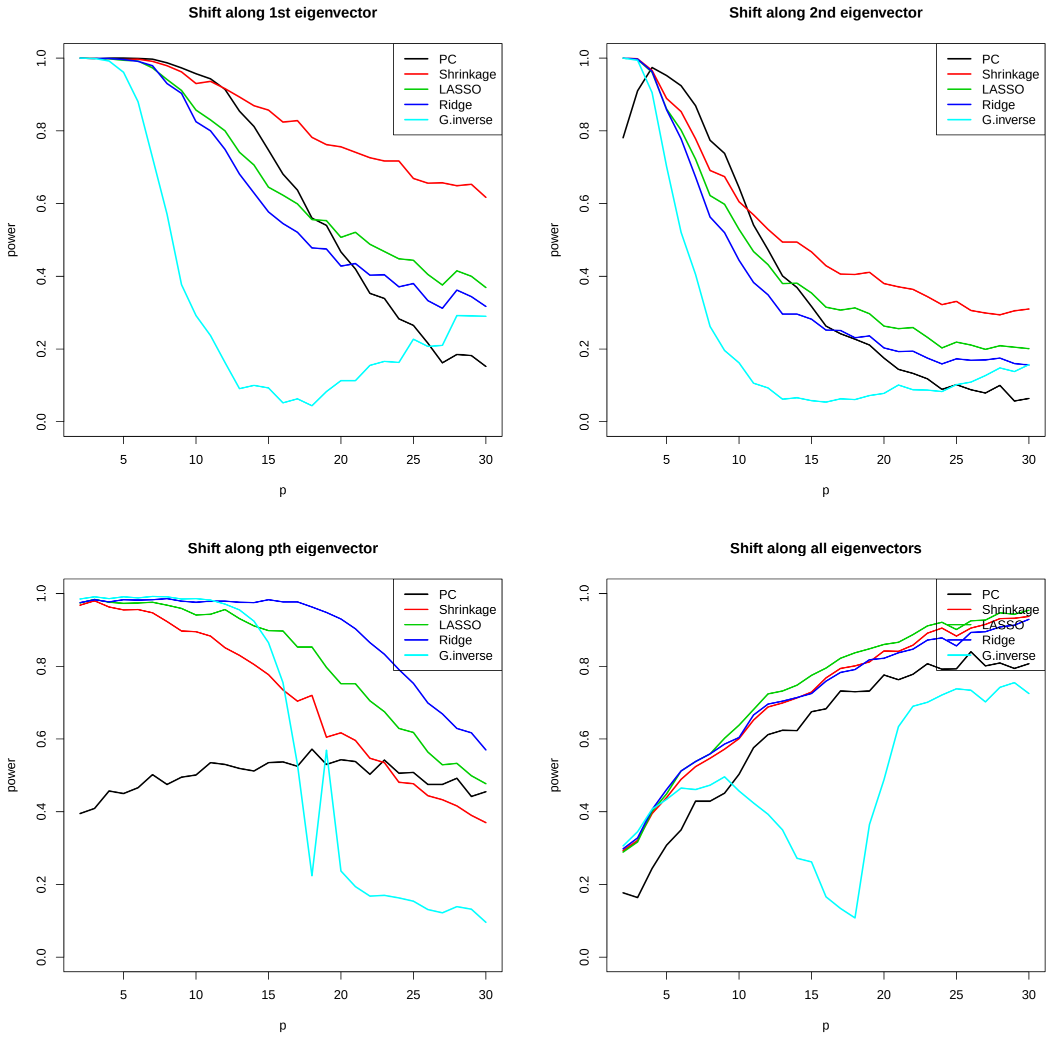


Figure A.1: Power comparison of MANOVA test based on 5 competing procedures under AR(1) covariance structure with $b = 0.4$. For each value of $p \in [2, 30]$, the power is estimated using 1000 samples and the significance level is kept as 0.05.

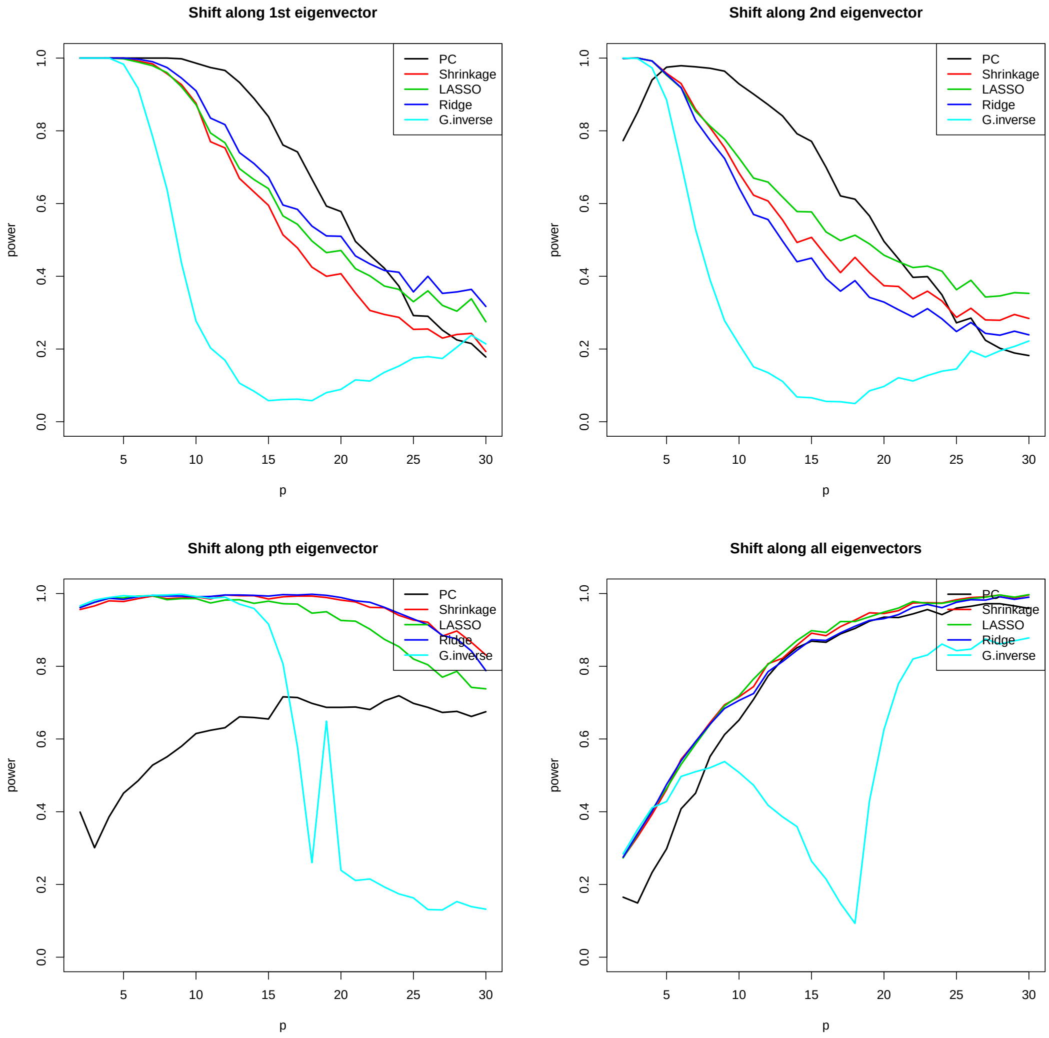


FIGURE A.2: Power comparison of MANOVA test based on 5 competing procedures under exchangeable covariance structure with $b = 0.4$. For each value of $p \in [2, 30]$, the power is estimated using 1000 samples and the significance level is kept as 0.05.

Appendix B

Supplementary Figure for Chapter 5

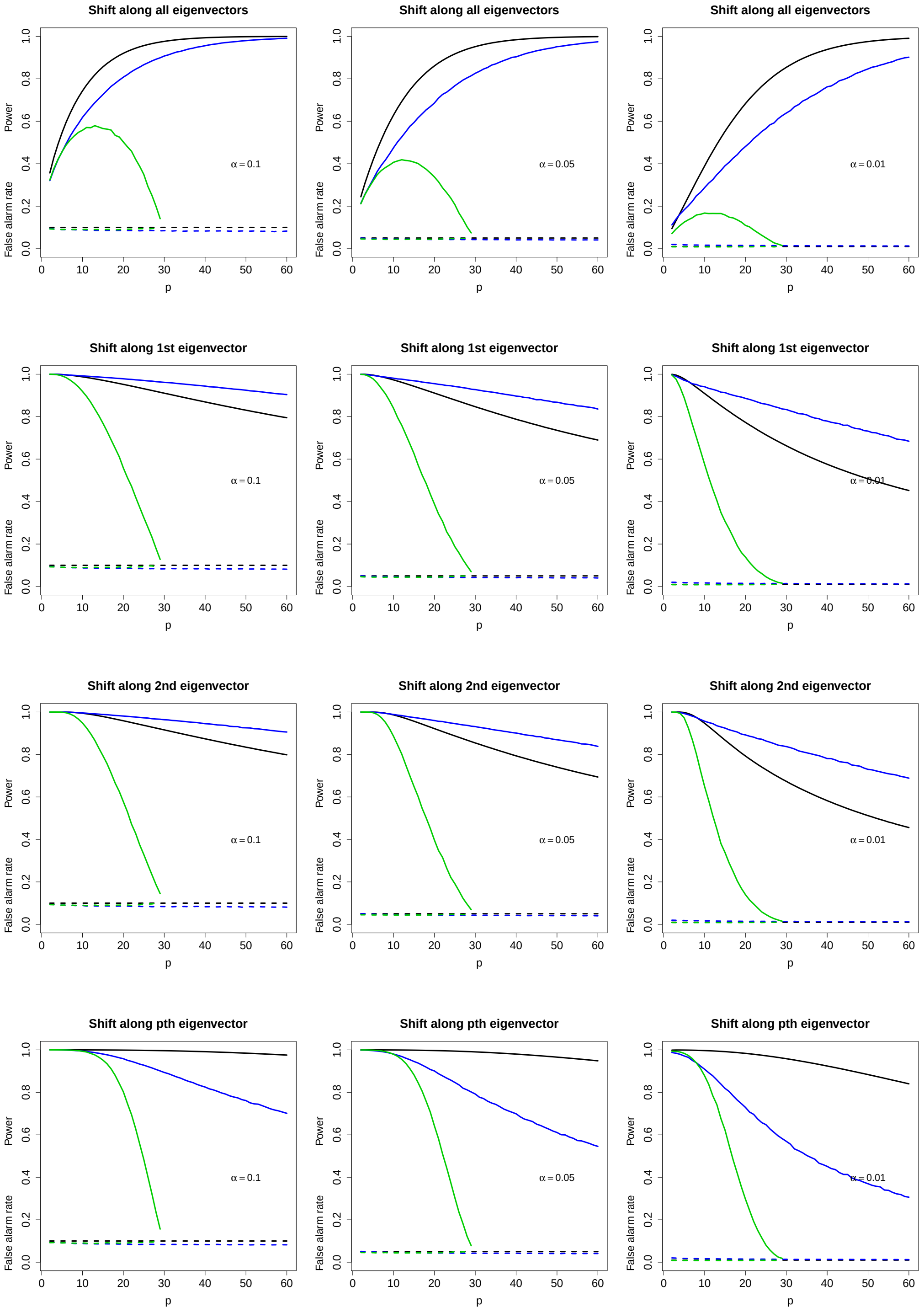


FIGURE B.1: Power (solid lines) and false alarm rate (dashed lines) for AR(1) covariance structure with $b = 0.3$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.

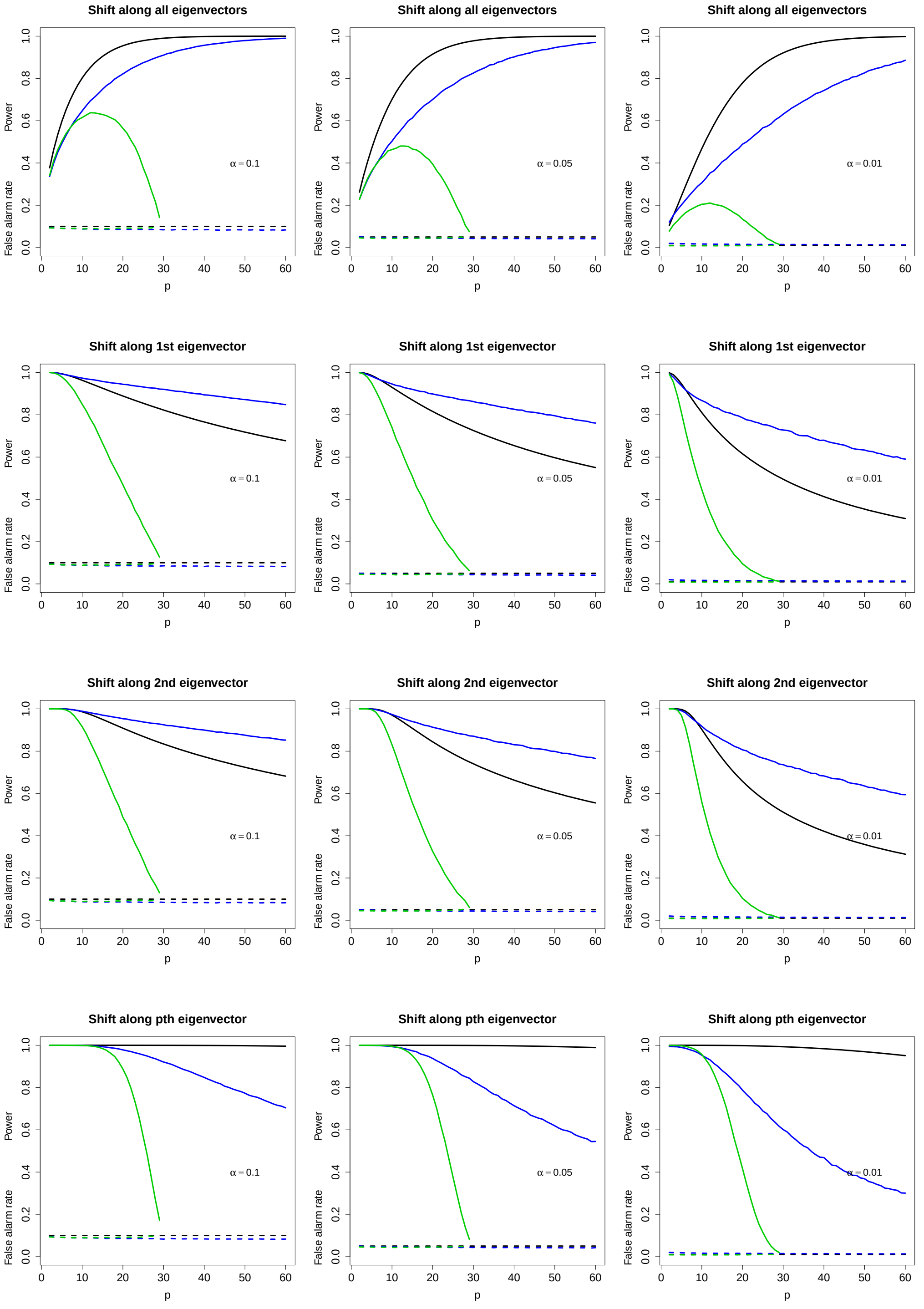


Figure B.2: Power (solid lines) and false alarm rate (dashed lines) for AR(1) covariance structure with $b = 0.4$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.

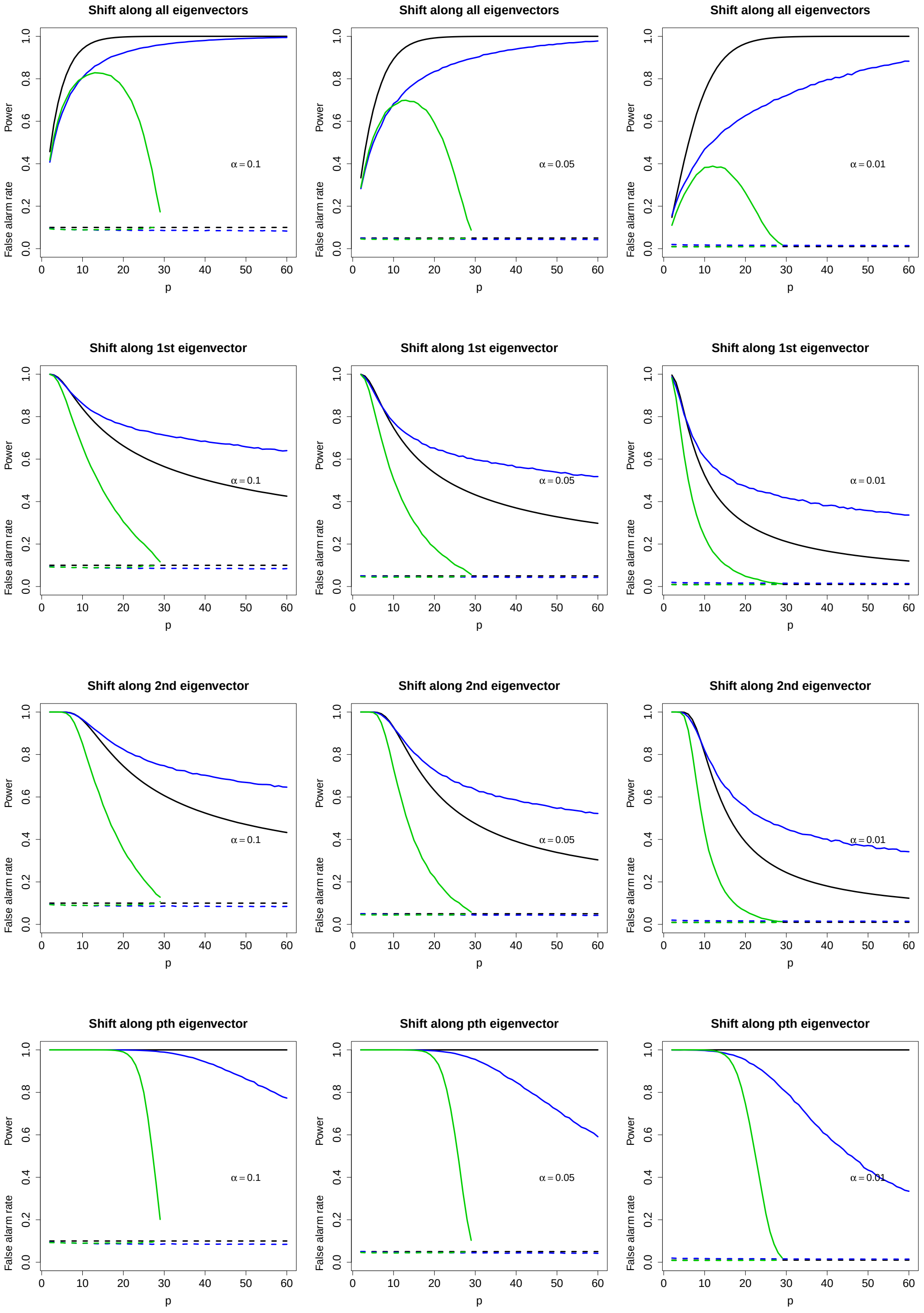


FIGURE B.3: Power (solid lines) and false alarm rate (dashed lines) for AR(1) covariance structure with $b = 0.6$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.

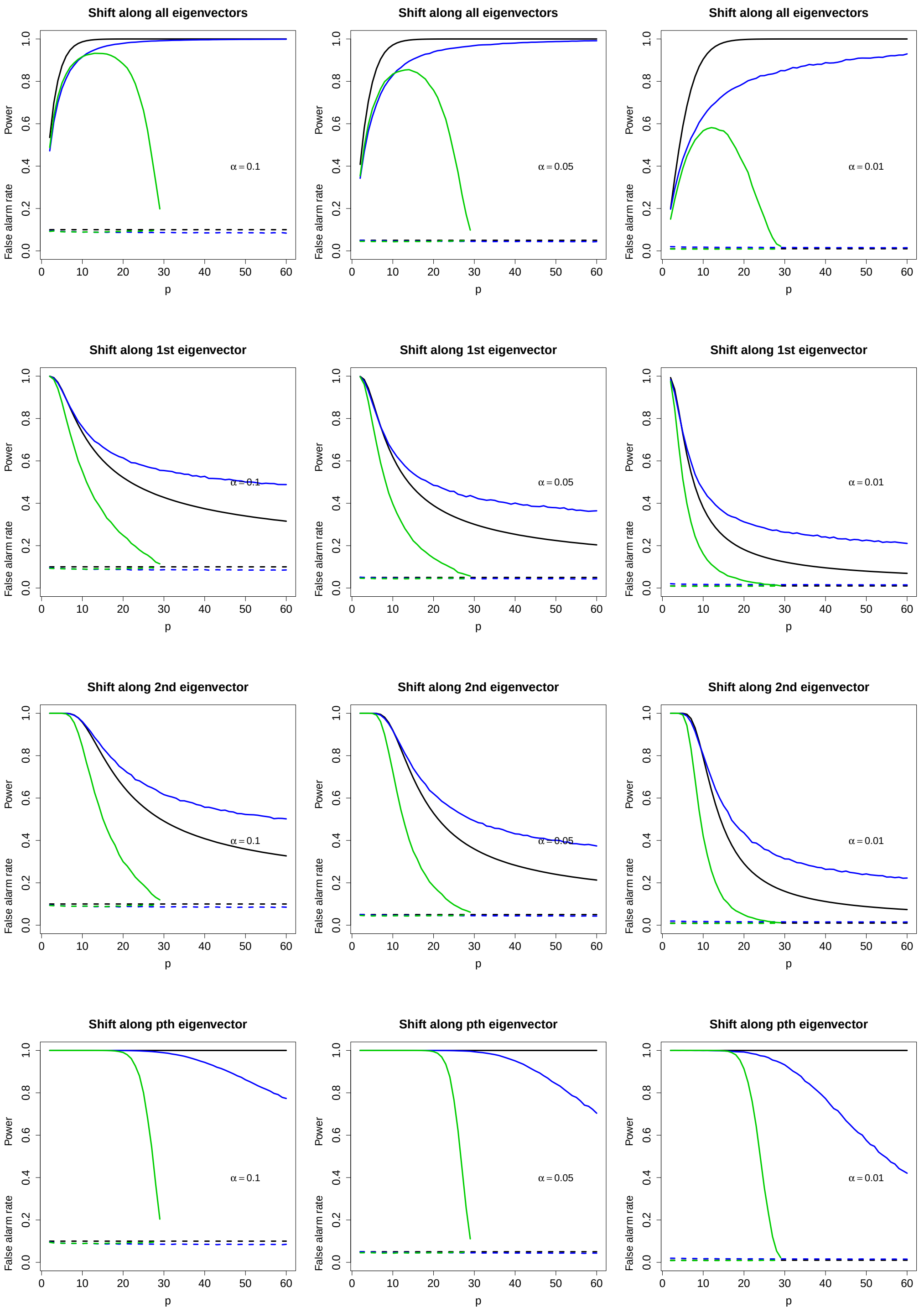


FIGURE B.4: Power (solid lines) and false alarm rate (dashed lines) for AR(1) covariance structure with $b = 0.7$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.

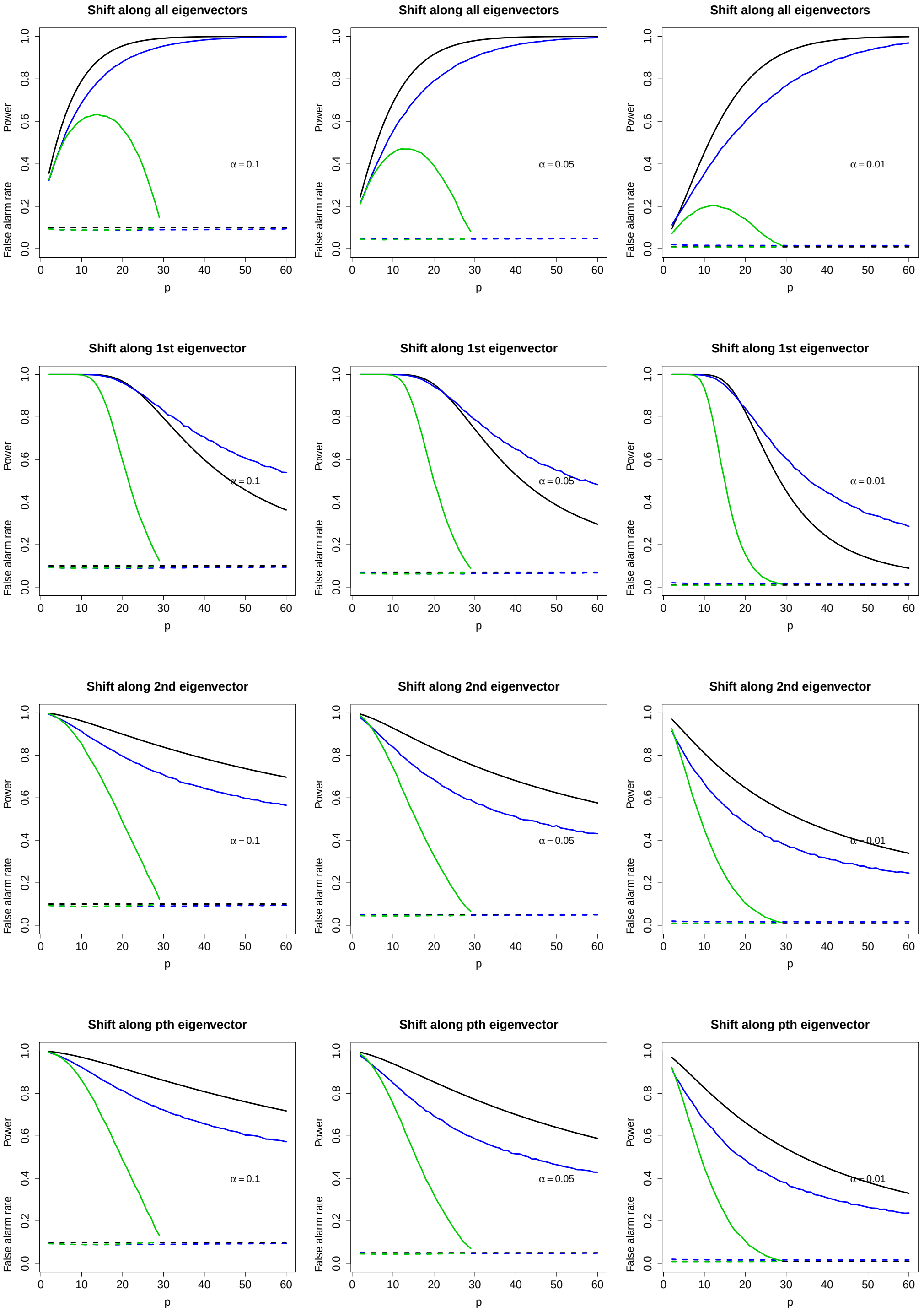
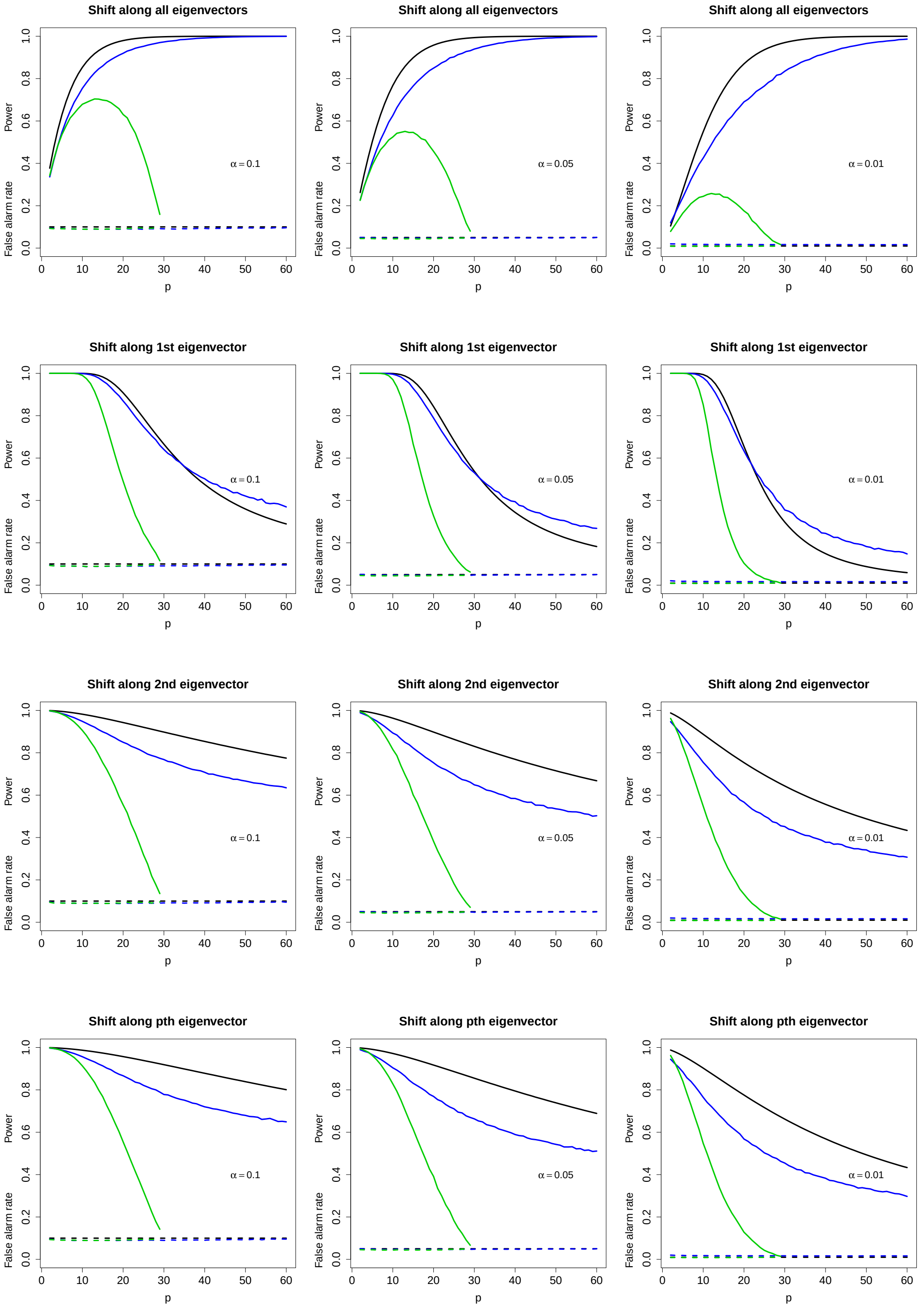


FIGURE B.5: Power (solid lines) and false alarm rate (dashed lines) for exchangeable covariance structure with $b = 0.3$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.



Shift along all eigenvectors

Shift along 1st eigenvector

Shift along 2nd eigenvector

Shift along pth eigenvector

FIGURE B.6: Power (solid lines) and false alarm rate (dashed lines) for exchangeable covariance structure with $b = 0.4$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.

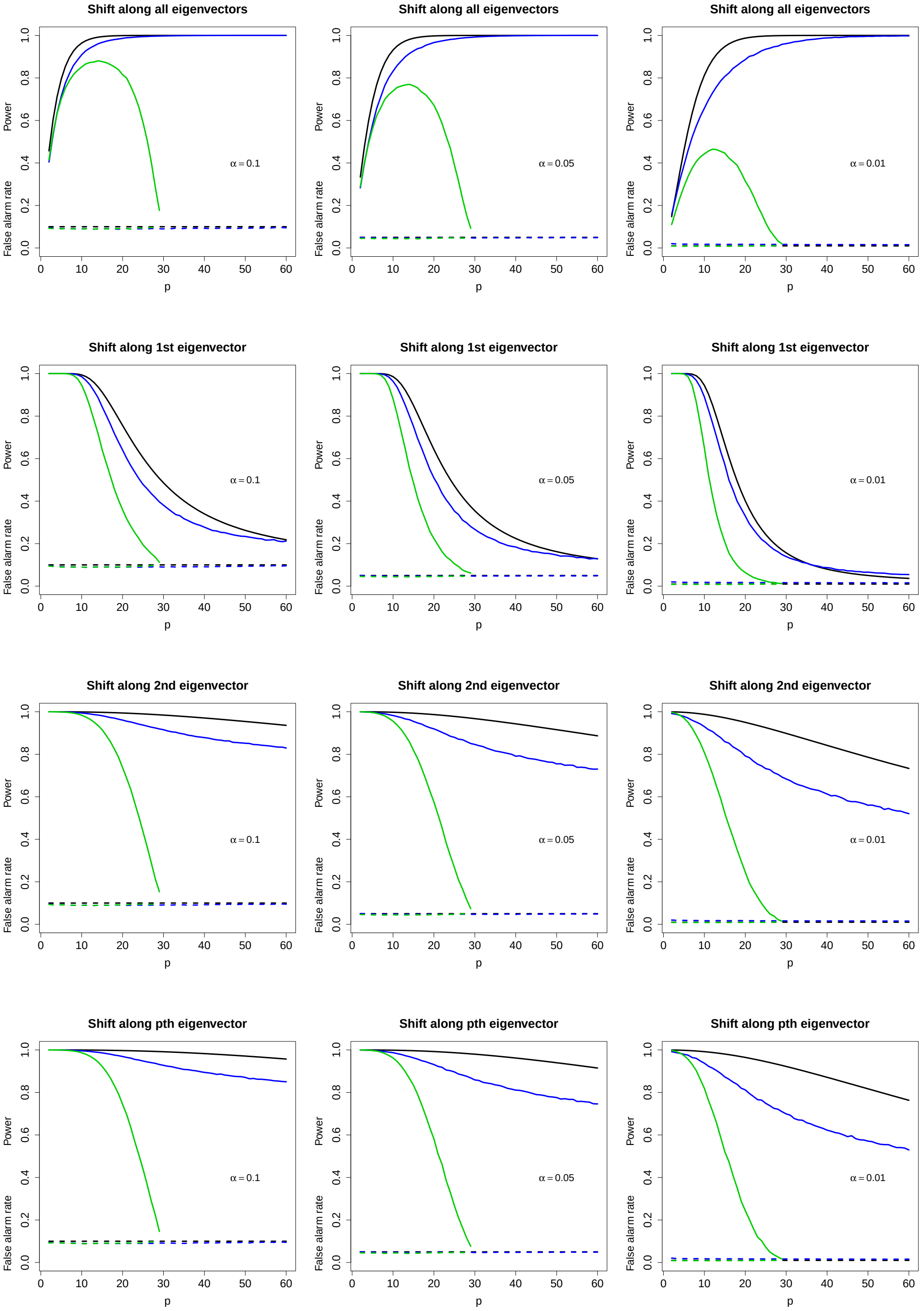


FIGURE B.7: Power (solid lines) and false alarm rate (dashed lines) for exchangeable covariance structure with $b = 0.6$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.

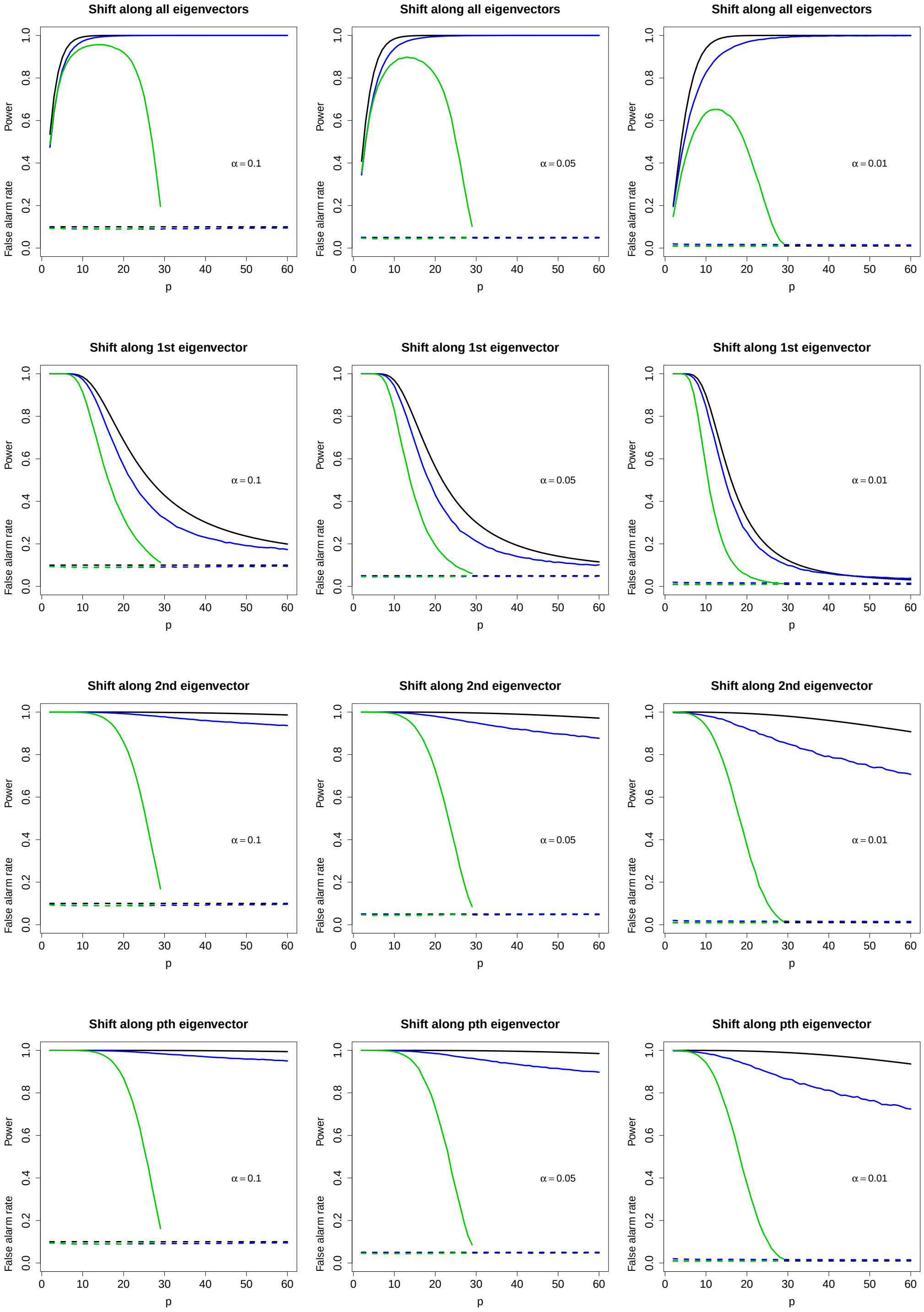


FIGURE B.8: Power (solid lines) and false alarm rate (dashed lines) for exchangeable covariance structure with $b = 0.7$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.

Appendix C

DRC forms

The signed statement of contribution to doctoral thesis containing publications is attached immediately followed this page.

References

Aitken, C. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*(1), 109–122.

Amemiya, Y. (1985). What should be done when an estimated between-group covariance matrix is not nonnegative definite? *The American Statistician*, *39*(2), 112–117.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis.* Wiley New York.

Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. (2006). *Statistics of extremes: theory and applications.* John Wiley & Sons.

Bennett, R. L. (2002). *Aspects of the analysis and interpretation of glass trace evidence.* Unpublished master's thesis, New Zealand.

Bersimis, S., Psarakis, S., & Panaretos, J. (2007). Multivariate statistical process control charts: an overview. *Quality and Reliability Engineering International*, *23*(5), 517–543.

Bickel, J., & Levina, E. (2007). *Covariance regularization by thresholding* (Tech. Rep.).

Calvin, J. A. (1993). Reml estimation in unbalanced multivariate variance components models using an em algorithm. *Biometrics*, 691–701.

Calvin, J. A., & Dykstra, R. L. (1991). Maximum likelihood estimation of a set of covariance matrices under lowner order restrictions with applications to balanced multivariate variance components models. *The Annals of Statistics*, 850–869.

Champ, C. W., Jones-Farmer, L. A., & Rigdon, S. E. (2005). Properties of the T^2 control chart when parameters are estimated. *Technometrics*, *47*(4).

Chou, Y., Mason, R. L., & Young, J. C. (1999). Power comparisons for a Hotelling's T^2 statistic. *Communications in Statistics - Simulation and Computation*, *28*(4), 1031-1050.

Cox, D. R., & Wermuth, N. (1996). *Multivariate dependencies: Models, analysis and interpretation* (Vol. 67). CRC Press.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*(1), pp. 157-175.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Dey, D. K., & Srinivasan, C. (1985). Estimation of a covariance matrix under stein's loss. *The Annals of Statistics*, 1581–1591.

Duong, T. (2014). ks: Kernel smoothing [Computer software manual]. Available from http://CRAN.R-project.org/package=ks (R package version 1.9.2)

Edgington, E., & Onghena, P. (2007). *Randomization tests.* CRC Press.

Fan, J., Feng, Y., & Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, *3*(2), 521.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, J., Peng, H., et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*(3), 928–961.

Fitch, A. M., Jones, M. B., & Massam, H. (2014). The performance of covariance selection methods that consider decomposable models only. *Bayesian Analysis*, *9*(3), 659–684.

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

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

Friedman, J., Hastie, T., & Tibshirani, R. (2013). glasso: Graphical lasso- estimation of gaussian graphical models [Computer software manual]. Available from http://www-stat.stanford.edu/~tibs/glasso (R package version 1.7)

Gao, X., Pu, D. Q., Wu, Y., & Xu, H. (2009). Tuning parameter selection for penalized likelihood estimation of inverse covariance matrix. *ArXiv e-prints*.

Golub, G., & Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial & Applied Mathematics, Series B: Numerical Analysis*, *2*(2), 205–224.

Haff, L. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, 586–597.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*(358), 320–338.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning* (Vol. 2) (No. 1). Springer.

Hill, W. G., & Thompson, R. (1978). Probabilities of non-positive definite between-group or genetic covariance matrices. *Biometrics*, *34*(3), pp. 429-439. Available from http://www.jstor.org/stable/2530605

Huang, J. Z., Liu, N., Pourahmadi, M., & Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, *93*(1), 85-98. Available from http://biomet.oxfordjournals.org/content/93/1/85.abstract

James, W., & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, *1*(1961), 361–379.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, 295–327.

Kong, S. W., Pu, W. T., & Park, P. J. (2006). A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, *22*(19), 2373-2380.

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

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365 - 411.

Ledoit, O., Wolf, M., et al. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, *40*(2), 1024–1060.

Lowry, C. A., & Montgomery, D. C. (1995). A review of multivariate control charts. *IIE transactions*, *27*(6), 800–810.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., et al. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(38), 13550–13555.

Montgomery, D. C. (2007). *Introduction to statistical quality control* (Sixth ed.). John Wiley & Sons.

Montgomery, D. C., & Woodall, W. (1999). Research issues and and ideas in statistical process control. *Journal of Quality Technology*, *31*(4), 376–387.

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.-L., Egyhazi, S., Hall, P., et al. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, *7*(6), R953.

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, *51*(03), 406–413.

Phillips, P. C., & Arnold, S. J. (1999). Hierarchical comparison of genetic variance-covariance matrices. i. using the Flury hierarchy. *Evolution*, *53*(5), 1506–1515.

Polansky, A. M., & Baker, E. R. (2000). Multistage plugin bandwidth selection for kernel distribution function estimates. *Journal of Statistical Computation and Simulation*, *65*(1-4), 63–80.

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

Schaefer, J., Opgen-Rhein, R., , & Strimmer., K. (2010). corpcor: Efficient estimation of covariance and (partial) correlation [Computer software manual]. Available from http://CRAN.R-project.org/package=corpcor (R package version 1.5.7)

Schäfer, J., & Strimmer, K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, *21*(6), 754-764. Available from http://bioinformatics.oxfordjournals.org/content/21/6/754.abstract

Schäfer, J., & Strimmer, K. (2005b). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology, The Berkeley Electronic Press*, *4*(1). Available from http://www.bepress.com/sagmb/vol4/iss1/art32/

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1).

Seber, G. A. (2009). *Multivariate observations* (Vol. 252). John Wiley & Sons.

Shaw, R. G. (1991). The comparison of quantitative genetic parameters between populations. *Evolution*, 143–151.

Shen, Y., Lin, Z., & Zhu, J. (2011). Shrinkage-based regularization tests for high-dimensional data with application to gene set analysis. *Computational Statistics & Data Analysis*, *55*(7), 2221–2233.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, *1*(399), 197–206.

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

Tomfohr, J., Lu, J., & Kepler, T. (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, *6*(1), 1-11.

Tracy, N. D., Young, J. C., & Mason, R. L. (1992). Multivariate control charts for individual observations. *Journal of Quality Technology*, *24*(2), 88-95.

Tsai, C.-A., & Chen, J. J. (2009). Multivariate analysis of variance test for gene set analysis. *Bioinformatics*, *25*(7), 897-903. Available from http://bioinformatics.oxfordjournals.org/content/25/7/897.abstract

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth ed.). New York: Springer. Available from http://www.stats.ox.ac.uk/pub/MASS4 (ISBN 0-387-95457-0)

Wang, H., Li, R., & Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, *94*(3), 553–568.

Wang, Y., Naumann, U., Wright, S. T., & Warton, D. I. (2012). mvabund–an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, *3*(3), 471–474.

Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association*, *103*(481), 340-349.

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

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

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Insha Ullah

Name/Title of Principal Supervisor: Dr. Beatrix Jones

Name of Published Research Output and full reference:

Ullah, I. and Jones, B. (submitted). "Regularized MANOVA for high-dimensional data", Australian & New Zealand Journal of Statistics.

In which Chapter is the Published Work: Chapter 4

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate: 90%

  and / or

- Describe the contribution that the candidate has made to the Published Work:

Candidate's Signature

11/02/2015
Date

Principal Supervisor's signature

11/2/2015
Date