Sample Integrity in High Dimensional Data

by

Jung Ae Lee

(Under the direction of Jeongyoun Ahn)

Abstract

This dissertation consists of two parts for the topic of sample integrity in high dimensional data. The first part focuses on batch effect in gene expression data. Batch bias has been found in many microarray studies that involve multiple batches of samples. Currently available methods for batch effect removal are mainly based on gene-by-gene analysis. There has been relatively little development on multivariate approaches to batch adjustment, mainly because of the analytical difficulty that originates from the high dimensional nature of gene expression data. We propose a multivariate batch adjustment method that effectively eliminates inter-gene batch effects. The proposed method utilizes high dimensional sparse covariance estimation based on a factor model and a hard-thresholding technique. We study theoretical properties of the proposed estimator. Another important aspect of the proposed method is that if there exists an ideally obtained batch, other batches can be adjusted so that they resemble the target batch. We demonstrate the effectiveness of the proposed method with real data as well as simulation study. Our method is compared with other approaches in terms of both homogeneity of adjusted batches and cross-batch prediction performance.

The second part deals with outlier identification for high dimension, low sample size (HDLSS) data. The outlier detection problem has been hardly addressed in spite of the enormous popularity of high dimensional data analysis. We introduce three types of distances in order to measure the "outlyingness" of each observation to the other data points: centroid

distance, ridge Mahalanobis distance, and maximal data piling distance. Some asymptotic properties of the distances are studied related to the outlier detection problem. Based on these distance measures, we propose an outlier detection method utilizing the parametric bootstrap. The proposed method also can be regarded as an HDLSS version of quantile-quantile plot. Furthermore, the masking phenomenon, which might be caused by multiple outliers, is discussed under HDLSS situation.

INDEX WORDS: Batch effect; Centroid distance; Factor model; Gene expression data; High dimensional covariance estimation; Masking effect; Maximal data piling distance; Outlier detection; Ridge Mahalanobis distance

Sample Integrity in High Dimensional Data

by

Jung Ae Lee

B.A., Ewha Womans University, 1999

M.A., Ewha Womans University, 2004

M.S., University of Georgia, 2009

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2013

Sample Integrity in High Dimensional Data

by

Jung Ae Lee

Approved:

Major Professor: Jeongyoun Ahn

Committee: Kevin K. Dobbin
Liang Liu
Jaxk Reeves
Paul Schliekelman

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2013

Dedication

To my parents, two sisters, and Andy

Acknowledgments

This dissertation is gratefully dedicated to my family members and some special friends. First of all, I thank my parents for their love and understanding during the Ph.D study. I also thank my two sisters, Jung Eui and Jung Su, who have been constant supporters and sincere friends in my life.

Second of all, I would like to express my deepest gratitude to my advisor, Dr. Jeongyoun Ahn, for her advice and help during the research years. Our meetings every week were productive and enjoyable thanks to her enthusiastic help. Without her leadership and encouragement, I would not have successfully finished my dissertation. I also thank her for her recommendation letter for the IIRG award and a job. Moreover, I am grateful to my committee members, especially to Dr. Kevin Dobbin and Dr. Jaxk Reeves who wrote recommendation letters for my first job in the United States.

Third of all, I would like to thank my special friends, Marty and Mitch, who have been the sweetest friends while I stayed in Athens. I appreciate their love and care like American parents and their prayers for my life. I also give thanks to my friend, Jungsoon in Korea, who has prayed for my graduation and job. I also appreciate the Hemerlein Family and Delaney for giving me a lot of good memories in Athens such as game nights, strawberry picking, steak dinners with mashed potatoes, and wedding pictures.

Last but not least, I give thanks to my husband, Andy Bartlett, for his care and support while I am studying. Thanks to him, I could enjoy the journey toward the Ph.D. degree. In particular, I appreciate his full support and understanding when I had to complete this dissertation while we were preparing for our wedding ceremony in July.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the last few decades, our society has produced all kinds of databases in many scientific and business areas. Collecting and processing massive data sets have entailed the development of numerous new statistical methodologies and theories. In statistical data analysis, *observation* is the experiment unit or data collection unit in which particular phenomena of the object (e.g., human being) are watched and recorded. An observation typically contains multiple variables (e.g., height, weight, age, etc.), and the number of variables is referred to as the *dimension* of the data. Many traditional statistical theories are built on the assumption that the number of observations (n) is much larger than the dimension (p). Also most asymptotic studies are regarding situations in which n tends to infinity while p is a fixed constant.

However, data nowadays often come with an extremely large number of variables. The radically increasing number of variables is due to the growth of information technology that collects, processes, and manages massive data sets. We are now in a situation where many important data analysis problems are high-dimensional. Gene expression data, financial data, hyperspectral imagery data, and internet commerce data are well-known examples. Commonly, high dimensional data sets have a much smaller number of observations than variables. For example, a typical microarray data set has gene expression measurements with tens of thousands of genes for only up to a few hundred patients with a certain disease. We refer to such data as high dimension, low sample size (HDLSS) data in this dissertation.

Unfortunately, classical statistical methods are not designed to cope with this kind of explosive growth of dimensionality that the current data sets face. The analytic difficulty

associated with adding extra dimensions is mentioned as "curse of dimensionality" by many authors (e.g., Bellman (1961), Hastie et al. (2001)). It generally refers to the problem that the previously well-fitting model is not valid or feasible in higher dimension even when p is still less than n. For this reason, there have been continuous efforts to deal with a few "well-chosen variables" from a large set of variables. These efforts are often categorized as variable or feature selection or dimension reduction methods, which commonly involve the process of selecting a subset of relevant variables for model construction such as regression and classification.

Once a successful dimension reduction method transforms a data set to be a manageable one by decreasing the number of variables below the sample size, conventional statistical tools can be useful for the analysis of the data. Although practitioners search for a tractable size of database in various ways, the attempt to reduce the number of variables is not always successful. Often the variables should remain greater than the sample size to provide greater details and better explanations, resulting in the failure of many statistical results.

Some solutions to these challenges have been offered, singularly or in combination with the existing methodologies, for example, factor models (Bai, 2003; Fan et al., 2008), regularization methods (Friedman, 1989), and sure independent screening methods (Fan and Lv, 2007). During the process of creating alternative approaches for HDLSS data, researchers discovered that the HDLSS problems require new or different mathematics along with novel computational and modeling issues. For example, for the theoretical development of HDLSS studies, it is general to have the asymptotic situation that p approaches infinity instead of a fixed p; the sample size n grows along with the dimension or sometimes remains fixed. In another example, Hall et al. (2005) pointed out that the high dimensional data has a quite different geometric representation from the low dimensional data.

While we mention "curse" due to the practical difficulty of high dimensional data, there are benefits of high-dimensionality, called "blessing of dimensionality" (Donoho, 2000). This view is often found in probability theory such as concentration of measure. In this view,

increases in dimensionality can be helpful since certain random fluctuations are well controlled in high dimension rather than moderate dimension. Not only theoretical benefits, there are also great opportunities in high dimension on the practical side; for example, in gene expression data, the more measurements are taken on each individual, the more informative details are available to characterize a certain disease.

Among emerging issues in HDLSS data analysis, this dissertation deals with the so called "sample integrity" issue. Sample integrity means a sound, uncontaminated, or ideal condition of a data set which leads to consistent and reliable results in the statistical analysis. Our concern for high dimensional data lies in how to ensure the quality of a sample while previous trends focus on the quality of variables. Actually, a sample, as a subset of population, is impossible to be perfect unless it includes all objects in the population. Statistical models based on a sample always carry some sample biases. Such biases challenge estimation, inference, and prediction, especially in high dimension. HDLSS data require extra efforts to make the sample the best condition, not necessarily perfect but "unbiased" enough to extract important patterns and trends in populations. This dissertation consists of two parts for the topic of sample integrity in high dimensional data, which are introduced in Chapters 2 and 3, respectively.

In Chapter 2, we discuss eliminating batch bias found in many microarray studies in which multiple batches of samples are inevitably used. Sometimes, researchers attempt to integrate several data sets for the sake of prediction power. But they often find a problem that analyzing the large combined data sets is not free from "batch effect," which is possibly caused by different experimental conditions. If the batch effect is not well controlled, the statistical results can be erroneous. As a solution, we propose a novel batch effect removal method regarding inter-gene relationships as well as location adjustment.

In Chapter 3, we suggest an outlier identification method for HDLSS data. Detecting outliers is a common first step in the exploratory data analysis, but it has been hardly addressed for HDLSS data. Conventional methods, which utilize mean and covariance, are not

effective in characterizing the outlying observation in high dimensional space. This chapter proposes an outlier detection method in HDLSS setting based on newly defined distance measures and a parametric bootstrap.

Chapter 2

Covariance Adjustment for Batch Effect in Gene Expression Data

2.1 Introduction

2.1.1 Microarray and Batch Biases

Since the mid 1990s, DNA microarray, also commonly known as gene chip, technologies have become enormously popular in biological and medical research. This technology enables one to monitor the expression levels of many genes (usually up to 20,000s) simultaneously. Careful and efficient statistical analyses of microarray data can lead some important scientific discoveries, for example gaining an understanding in the pathogenesis of a disease, identifying clinical biomarkers, and ultimately personalized medicine (Böttinger et al., 2002; Heidecker and Hare, 2007; Lee and Macgregor, 2004).

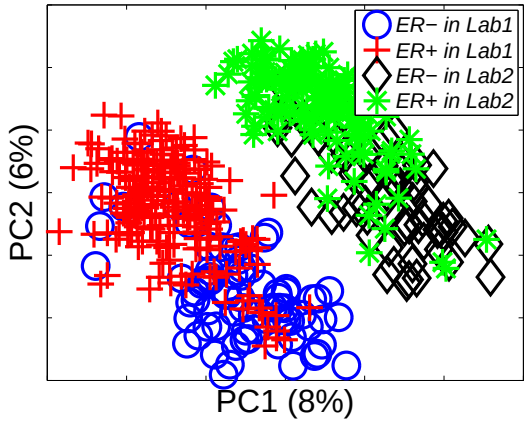
An important aspect of microarray studies is the management of biases (Scherer, 2009). Unfortunately, microarray technology is susceptible to measurement variability due to the complexity of the experiment process (Altman, 2009). The biases that are induced by technical factors can mislead results of statistical analysis and thus threaten the validity of the entire study. Although proper employment of a well-planned experimental design can reduce biases, the costly and time-consuming procedure of a microarray experiment makes it difficult to perform the "optimal" experiment. In reality, most microarray studies are faced with not only random technical errors in the analytic phase, but also substantial systematic biases.

One of the biggest challenges in statistical analysis of microarray data is how to deal with "batch effect," a systematic bias caused by the samples collected at different times or sites.

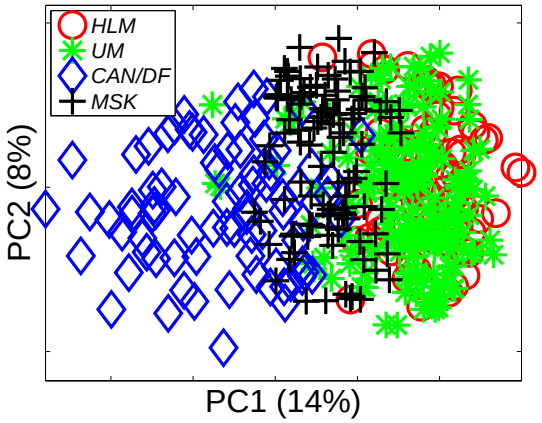
Often researchers use multiple batches of samples in order to increase the sample size for the sake of a potential benefit of increased statistical power. For example a predictive model from combined sample is more robust than one from individual studies (Vachani et al., 2007; Xu et al., 2005). But sometimes it is not clear whether the advantage of increased samples size outweighs the disadvantage of higher heterogeneity of merged data sets. In order to make the combining serve its purpose, there is a pressing need to find a "batch effect" removal method that can create a merged data set without any batch bias.

Some experiments, such as a cancer study, require sufficiently large samples to enhance the prediction performance (Ein-Dor et al., 2006). In the following example we consider two different breast cancer data sets collected at different laboratories. Analyzing these data sets individually may limit the scope of the study due to relatively small sample sizes compared to tens of thousands of genes; the sample sizes are 286 and 198, respectively. With a goal of predicting the estrogen receptor (ER) status, we want to create a combined data set in order to increase the statistical power. Figure 2.1-(a) displays projections of the data onto the first two principle component directions of the whole data set. We can see that the two batches are clearly separated. In fact, this batch difference dominates biological difference of ER+ and ER−. One naturally needs to narrow or even close this gap between the two batches of samples before any classification analysis. Both data sets are preprocessed by MAS5.0 for the Affymetrix platform. The detailed description of these data sets can be found in Section 2.7.

Another example of batch effect can be found in Figure 2.1-(b), where four lung cancer microarray data sets from four different laboratories are shown. Shedden et al. (2008) used these data sets to perform a gene-expression-based survival prediction study. The detailed description of the data set can be found in Section 2.7. In the figure, four different symbols represent four different laboratories. We notice that there are visible gaps among the batches, even after all four samples are preprocessed by RMA together for the Affymetrix platform.



(a) Breast cancer data



(b) Lung cancer data

Figure 2.1. Illustration of a batch effect in microarray data. In (a), breast cancer data sets from two different laboratories are projected on the first two PC directions. It is clear that the batch effect dominates the biological (ER+, ER−) signal. In (b), the projections of four lung cancer data sets are shown. We can see strong batch bias especially in the first PC direction.

Batch effects exist not only in microarray but also in other newer technologies. Recently, researchers have found that there exist significant batch effects in mass spectrometry data, copy number abnormality data, methylation array data, and DNA sequencing data (Leek and Storey, 2007). In particular, recently research transition from microarrays to next-generation sequencing is notable; e.g., RNA-sequencing is gaining more popularity since it provides greater accuracy than microarray and a dynamic range of gene expression values. Even though most of the existing methods including the proposed work have been developed for microarray, the experience with microarray technologies may lead to the future success for these new technologies.

2.1.2 Preprocessing

Preprocessing is an important step for gene expression data analysis since this step reduces technical variation across arrays (samples) ahead of any statistical analyses such as clustering, classification and regression analysis. Here we use the term "array" to refer to a sample which includes information of individual probe sets. The choice of pre-processing method affects both the experiment's sensitivity and its bias (Scherer, 2009).

Preprocessing of microarray data generally involves the following steps: background correction, normalization, and filtering. In particular, the normalization step puts each array on a common scale, which allows for the comparison of expression levels across different arrays. There are two methods frequently used to normalize the microarray; one is global median normalization, which forces each array to have equal median expression value, and the other is quantile normalization, proposed by Bolstad et al. (2003). The quantile normalization method transforms each array so that it has the same empirical distribution of gene expression values as the other arrays. The common empirical distribution which each array will have is driven by mean or median over all arrays. See the example in Figure 2.2; the sorted gene expression values for each array is replaced by the average value of all arrays in

each position, and consequently all arrays come to have the same distribution of intensity values.

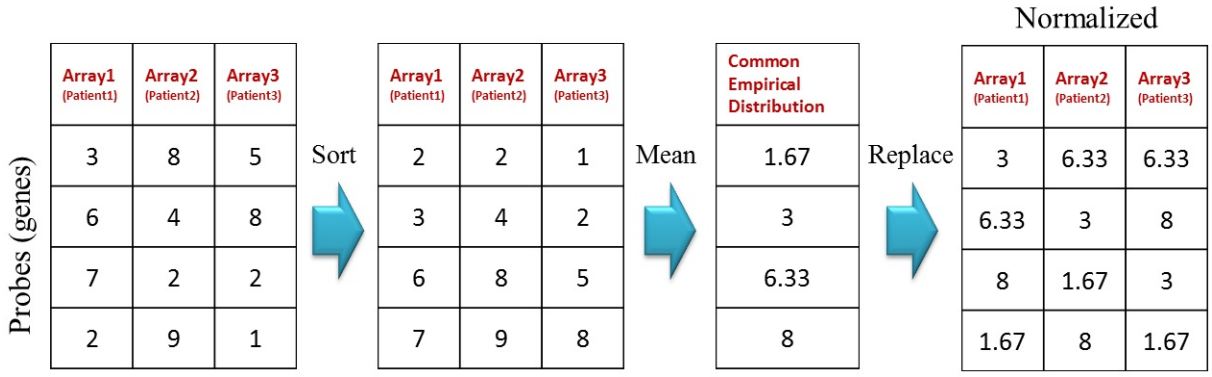


Figure 2.2. Illustration of quantile normalization. The sorted gene expression values for each array is replaced by the average value of all arrays in each position, so that each array has the same distribution of expression values.

Practically, a preprocessed data set is obtained by using software packages. Common choices are MAS5.0, robust multiarray analysis (RMA), or dChip for Affymetrix platform, and locally weighted scattered smoothing (LOWESS) method for Agilent platform and others. In our work, MAS5.0 is used for the breast cancer data, and RMA is used for the lung cancer data. The softwares are publicly available at http://www.bioconductor.org and http://www.dChip.org. The BioConductor package also includes a variety of graphical tools. Figure 2.3, for example, displays the box plots for the first 19 arrays of a lung cancer data set before and after RMA preprocessing. It is clear that the scale of probe intensities became similar across arrays after preprocessing. For the our lung cancer data, RMA has been applied to four data sets simultaneously because by using one unified reference we can avoid extra bias between batches. For this matter, there are more recent methods such as frozen robust multiarray analysis (fRMA) (McCall et al., 2010), which is known to be effective when preprocessing multiple batches.

Preprocessing methods contribute to reducing array-specific technical variation such as dye-balance, spatial dependency induced by experimental procedures. In many cases, however, systematic biases still remain after normalization. There are also many other issues in

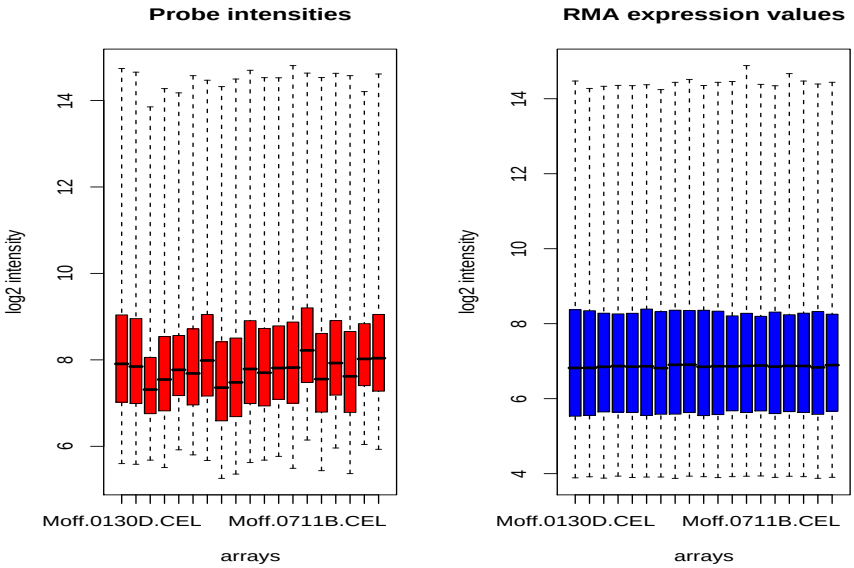


Figure 2.3. Box plots before and after RMA preprocessing for the first 19 arrays of a lung cancer data set. The left panel displays the original intensity of arrays before RMA preprocessing. The intensities' scale becomes similar among arrays in the right panel after preprocessing.

preprocessing, but we restrict our attention to "batch effect," which is occurred in a combined data set from different sources.

2.2 Existing Batch Adjustment Methods

In this section we introduce some currently available batch adjustment methods: mean-centering, standardization, empirical bayes method, discrimination-based methods, and cross-platform normalization. Let Y_{ijg} represent the expression value of sample j ($j = 1, \ldots, n_i$), for gene g ($g = 1, \ldots, p$), in batch i ($i = 1, \ldots, K$). Here the gene index g is also called "variable" or "dimension" in this dissertation.

2.2.1 Mean Centering and Standardization

The mean-centering (or median-centering) method sets the mean (or median) of each gene to zero for each batch. Let us define the sample mean for each batch i and gene g as

$\bar{Y}_{ig} = \sum_{j=1}^{n_i} Y_{ijg}/n_i$, and the mean-centering method replaces each observation by

$$Y^*_{ijg} \quad = \quad Y_{ijg} - \bar{Y}_{ig} \ .$$

Despite its simplicity, the mean-centering method works reasonably well in practice (Chen et al., 2011; Luo et al., 2010; Shabalin et al., 2008; Sims et al., 2008). This method is essentially optimal under the assumption that the batch bias only exists in the shift of the means. However, it has been found that the batch effect is more complex than mean shifts. This method is implemented in PAMR R software package (Prediction Analysis of Microarrays for R).

Standardization makes each gene within each batch have a unit variance and zero mean (Johnson et al., 2007; Luo et al., 2010). Defining the sample variance for each batch i and gene g as $s^2_{ig} = \sum_{j=1}^{n_i}(Y_{ijg} - \bar{Y}_{ig})^2/(n_i - 1)$, the standardized gene expression is

$$Y^*_{ijg} \quad = \quad \frac{Y_{ijg} - \bar{Y}_{ig}}{s_{ig}}.$$

2.2.2 Empirical Bayes Method

The Empirical Bayes (EB) method is introduced by Johnson et al. (2007). The EB method is a gene-wise model-based approach for adjusting batch effects. This method regards the batch bias as a random block effect, and biological signals as fixed treatment effects, and uses analysis of variance (ANOVA) technique to estimate the bias for each gene. The batch bias is obtained by an empirical Bayes estimator. Let us define Y_{ijgc} as the expression value of sample j for gene g in batch i and biological covariates c ($c = 1, \ldots, C$). A two way ANOVA with batch effect can be generally expressed as

$$Y_{ijgc} = \alpha_g + \beta_{gc} + \gamma_{ig} + \delta_{ig}\epsilon_{ijgc} \ ,$$

where β_{gc} is biological effect (e.g., positive or negative for a disease, or treatment vs. control groups) for gene g, and also γ_{ig} and δ_{ig} represent the additive and multiplicative batch effects of batch i for gene g, respectively. Here both γ_{ig} and δ_{ig} are random effects. It is assumed that $\epsilon_{ijgc} \sim N(0, \sigma^2_g)$ for gene g.

In the EB method, all samples are standardized together to have zero mean and unit variance for each gene g. After this standardization, samples in each batch follow a normal distribution, $Z_{ijgc} \sim N(\gamma_{ig}, \delta_{ig}^2)$. By estimating $\hat{\gamma}_{ig}^*$ and $\hat{\delta}_{ig}^*$ using an empirical bayes estimator, the expression values are modified to

$$Y_{ijgc}^* = \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}^*}(Z_{ijgc} - \hat{\gamma}_{ig}^*) + \hat{\alpha}_g + \hat{\beta}_{gc} \ .$$

One advantage of this method is that the magnitude of the adjustments may vary from gene to gene, and it avoids over-adjusting due to the outliers (Johnson et al., 2007). In particular, the EB method adjusts the variances between two batch of samples as well as the mean location by estimating appropriate amount of multiplicative batch effect in ANOVA model.

However, there are two potential drawbacks of the EB method. First we notice the fact that the modified expression values by this method includes the biological effect in the model. Thus it may be subject to the so-called "double dipping" problem because the transformed data will be used later again to analyze the biological feature. Another drawback is that it is a gene-wise approach. Since the batch effect estimation is performed on the individual ANOVA for each gene, the possible inter-gene batch effect is ignored.

The R software of the EB method (ComBat) is available at http://www.dchip.org.

A similar approach was attempted by Leek and Storey (2007), whose surrogate variable analysis (SVA) identifies the effect of the hidden factors that may be the sources of data heterogeneity, and recovers it in a subsequent regression analysis.

2.2.3 Discrimination-based Approach

Benito et al. (2004) applied Distance Weighted Discrimination (DWD) (Marron et al., 2007) method for adjusting systematic microarray data biases. DWD is originally designed for a discrimination analysis in high dimension, low sample size (HDLSS) setting. For batch adjustment, this method first aims to find the optimal direction to maximize separation

between the batches, and then on the given direction each subpopulation moves until its mean reaches the (separating) hyperplane between the two batches.

Even though Benito et al. (2004) used DWD for the purpose of the batch adjustment, the core idea of their method can be applied with any discrimination method. They chose DWD over other discrimination methods, such as support vector machines (SVM), because they believe that DWD is better at finding the optimal separating hyperplane for HDLSS discrimination. They also claimed that the projected data onto the normal direction vector of the DWD hyperplane looks more Gaussian than other methods, which makes mean-shift adjustment reasonable.

In what follows, the general concept of a discrimination-based approach is explained with two batches case. Let $\mathbf{y}_{ij}$ be the p-dimensional vector of jth sample ($j = 1, \ldots, n_i$) for batch $i = 1, 2$, and $\bar{\mathbf{y}}_i = \sum_{j=1}^{n_i} \mathbf{y}_{ij}/n_i$. Suppose that we have found the optimal direction vector on which the separation between two batches is maximized. The hyperplane that efficiently separates the two batches in the p-dimensional space can be expressed as

$$H = \{\mathbf{y} | \mathbf{w}^{\mathrm{T}}\mathbf{y} = m\},$$

where $\mathbf{w}$ is the normal direction vector of the separating hyperplane, and m is the constant that indicates the middle point between the two batches on $\mathbf{w}$, i.e., $m = (\mathbf{w}^{\mathrm{T}}\bar{\mathbf{y}}_1 + \mathbf{w}^{\mathrm{T}}\bar{\mathbf{y}}_2)/2$. Then, the adjusted data for batch i can be obtained by

$$\mathbf{y}^*_{ij} = \mathbf{y}_{ij} + (m - \mathbf{w}^{\mathrm{T}}\bar{\mathbf{y}}_i)\mathbf{w}. \tag{2.1}$$

The equation (2.1) is obtained by the following steps. The mean vector $\bar{\mathbf{y}}_i$ for batch i approaches to the separating hyperplane along with the direction $\mathbf{w}$ and arrives at a point $\mathbf{h}_i$ on H. The moving path $\mathbf{h}_i - \bar{\mathbf{y}}_i$ can be expressed by the vector $k_i\mathbf{w}$ with a scalar k_i, i.e., $\mathbf{h}_i - \bar{\mathbf{y}}_i = k_i\mathbf{w}$. Since $\mathbf{w}^{\mathrm{T}}\mathbf{h}_i = m$, replacing it with $\mathbf{h}_i = \bar{\mathbf{y}}_i + k_i\mathbf{w}$ becomes $\mathbf{w}^{\mathrm{T}}\bar{\mathbf{y}}_i + k_i = m$ since $\mathbf{w}^{\mathrm{T}}\mathbf{w} = 1$. Thus, $k_i = m - \mathbf{w}^{\mathrm{T}}\bar{\mathbf{y}}_i$, and therefore the moving path $k_i\mathbf{w}$ is $(m - \mathbf{w}^{\mathrm{T}}\bar{\mathbf{y}}_i)\mathbf{w}$. Each sample, $\mathbf{y}_{ij}$, is shifted along the path by this amount as shown in (2.1).

This procedure can be extended in a natural way to handle more than two batches. Linear discrimination with K groups produces $K-1$ dimensional discriminant space. Note that we can set the middle point m to be any arbitrary number without changing the relative positions of the data vectors after adjustment. Suppose $K = 3$ and $\mathbf{w}_1$ and $\mathbf{w}_2$ are the two orthogonal directions that generate the 2-dimensional discriminant space. Then the three batches of samples can be adjusted by the following steps.

$$\begin{aligned} \mathbf{y}_{ij}^{*} &= \mathbf{y}_{ij} - \mathbf{w}_1^{\mathrm{T}}\bar{\mathbf{y}}_i\mathbf{w}_1, \\ \mathbf{y}_{ij}^{**} &= \mathbf{y}_{ij}^{*} - \mathbf{w}_2^{\mathrm{T}}\bar{\mathbf{y}}_i\mathbf{w}_2. \end{aligned}$$

2.2.4 Cross Platform Normalization

Shabalin et al. (2008) proposed the cross platform normalization (XPN) method for the problem of combining data from different array platforms. The XPN procedure is based on a simple block-linear model. In other words, under the assumption that the samples of each platform fall into one of L homogeneous groups within each of M groups of similar genes, the expression value Y_{ijg} is written as block mean plus noise.

$$Y_{ijg} = A_{i,\alpha(g),\beta_i(j)} \cdot b_{ig} + c_{ig} + \sigma_{ig}\epsilon_{ijg} \ . \tag{2.2}$$

The functions $\alpha : \{1,\ldots,p\} \mapsto \{1,\ldots,M\}$ and $\beta_i : \{1,\ldots,n_i\} \mapsto \{1,\ldots,L\}$ define the linked groups of genes and samples, respectively. The A_{iml} are block means, and b_{ig}, c_{ig} are slope and offset parameter, respectively, where $m = 1,\ldots,M$ and $l = 1,\ldots,L$. It is assumed that $\epsilon_{ijg} \sim \mathcal{N}(0,1)$.

Prior to the estimation of the model in (2.2), k-means clustering is performed independently to the rows and columns of a p by n data matrix, to identify homogeneous group of genes and samples. From the mapping $\alpha(g)$ and $\beta_i(j)$, model parameters $\hat{A}_{iml}$, $\hat{b}_{ig}$, $\hat{c}_{ig}$ and $\hat{\sigma}_{ig}$ are estimated by using standard maximum likelihood methods. Common parameters $\hat{\theta}_g = (\hat{b}_g,\ \hat{c}_g,\ \hat{\sigma}_g^2)$ and $\hat{A}_{ml}$ are then obtained by weighted averages of the parameters from

the two batches, i.e.,

$$\hat{\theta}_g = \frac{n_1\hat{\theta}_{1,g} + n_2\hat{\theta}_{2,g}}{n_1 + n_2} \quad \text{and} \quad \hat{A}_{ml} = \frac{n_{1,l}\hat{A}_{1,m,l} + n_{2,l}\hat{A}_{2,m,l}}{n_{1,l} + n_{2,l}},$$

where $n_{i,l}$ is the number of samples in the lth sample group of batch i. Finally, the expression values are modified as

$$Y^*_{ijg} = \hat{A}_{\alpha(g),\beta_i(j)} \cdot \hat{b}_g + \hat{c}_g + \hat{\sigma}_g \left(\frac{Y_{ijg} - \hat{A}_{i,\alpha(g),\beta_i(j)} \cdot \hat{b}_{ig} + \hat{c}_{ig}}{\hat{\sigma}_{ig}} \right).$$

An advantage of the XPN method is taking the relationship of genes into account, by the row and column clustering, followed by estimating block means in block linear model, but block means assumption is arbitrary and may not be justified in the biological context.

2.2.5 Comparison of the Existing Methods

Mean-centering, standardization and the EM method are gene-wise approaches. Essentially, they assume that the batch effect applies to each gene independently. These methods are relatively easy to implement compared to a multivariate approach because the possible inter-gene batch effect is not taken into account. However, it has been noted that some batch effects exist in a multivariate space, i.e., the covariance structure among genes may be different from a batch to a batch (Leek et al., 2010).

On the other hand, the discrimination-based batch adjustment attempts to remove the batch effect in the multivariate space. These methods seem more efficient in the sense that they use fewer direction vectors ($K - 1$ discriminant directions for K batches) than the univariate approaches which remove batch effect in each of the p dimensions. However, finding the discriminant direction such as the DWD can be computationally intensive. Furthermore, as Proposition 1 below implies, the discrimination-based approach is essentially a gene-wise approach, and it can even be regarded as an "incomplete" mean-centering.

Proposition 1. *Applying the mean-centering adjustment to the data that have been adjusted with any discrimination-based method is equivalent to the mean-centering of the original data.*

Proof. Suppose $\mathbf{Y}_i$ is $p \times n_i$ data matrix for batch i $(i = 1, 2)$, and $\bar{\mathbf{y}}_i$ is $p \times 1$ mean vector, defined by $\mathbf{Y}_i\mathbf{1}/n_i$, where $\mathbf{1}$ is a $p \times 1$ vector that all elements are 1. Adjusted data $\mathbf{Y}_i^{mc}$ by mean-centering method can be written as

$$\mathbf{Y}_i^{mc} = \mathbf{Y}_i - [\bar{\mathbf{y}}_i, \ldots, \bar{\mathbf{y}}_i]_{p\times n_i}.$$

Adjusted data by a discrimination methods can be expressed by

$$\mathbf{Y}_i^{dm} = \mathbf{Y}_i - \mathbf{w}_1'\bar{\mathbf{y}}_i[\mathbf{w}_1, \ldots, \mathbf{w}_1]_{p\times n_i},$$

where $\mathbf{w}_1$ denotes a $p \times 1$ discriminant direction of two batches with unit length. Mean-centering on data $\mathbf{Y}^{dm}$ is

$$\mathbf{Y}_i^{dm+mc} = \mathbf{Y}_i^{dm} - [\bar{\mathbf{y}}_i^{dm}, \ldots, \bar{\mathbf{y}}_i^{dm}]_{p\times n_i}.$$

Note that the mean-centering result can be also obtained from an iteration process, that is, p sequential mean-shift to zero in orthogonal vectors, $\mathbf{w}_1$, $\mathbf{w}_2$, $\ldots$, $\mathbf{w}_p$, where $\mathbf{w}_j'\mathbf{w}_j = 1$ and $\mathbf{w}_j'\mathbf{w}_k = 0$ for $j \neq k$. We can formulate this process as below

$$\begin{aligned}
\mathbf{Y}_i^{dm+mc(1)} &= \mathbf{Y}_i - \mathbf{w}_1'\bar{\mathbf{y}}_i[\mathbf{w}_1, \ldots, \mathbf{w}_1]_{p\times n_i} = \mathbf{Y}_i^{(1)}(say),\\
\mathbf{Y}_i^{dm+mc(2)} &= \mathbf{Y}_i^{(1)} - \mathbf{w}_2'\bar{\mathbf{y}}_i^{(1)}[\mathbf{w}_2, \ldots, \mathbf{w}_2]_{p\times n_i}\\
&= \mathbf{Y}_i^{(1)} - \mathbf{w}_2'\left(\frac{1}{n_i}\mathbf{Y}_i^{(1)}\mathbf{1}\right)[\mathbf{w}_2, \ldots, \mathbf{w}_2]_{p\times n_i}\\
&= \mathbf{Y}_i^{(1)} - \mathbf{w}_2'\left(\frac{1}{n_i}\{\mathbf{Y}_i\mathbf{1} - \mathbf{w}_1'\bar{\mathbf{y}}_i[\mathbf{w}_1, \ldots, \mathbf{w}_1]_{p\times n_i}\mathbf{1}\}\right)[\mathbf{w}_2, \ldots, \mathbf{w}_2]_{p\times n_i}\\
&= \mathbf{Y}_i^{(1)} - \mathbf{w}_2'\left(\bar{\mathbf{y}}_i - \frac{1}{n_i}n_i\mathbf{w}_1'\bar{\mathbf{y}}_i\mathbf{w}_1\right)[\mathbf{w}_2, \ldots, \mathbf{w}_2]_{p\times n_i}\\
&= \mathbf{Y} - \mathbf{w}_1'\bar{\mathbf{y}}_i[\mathbf{w}_1, \ldots, \mathbf{w}_1]_{p\times n_i} - \mathbf{w}_2'\bar{\mathbf{y}}_i[\mathbf{w}_2, \ldots, \mathbf{w}_2]_{p\times n_i} - (\mathbf{w}_1'\bar{\mathbf{y}}_i)\mathbf{w}_2'\mathbf{w}_1[\mathbf{w}_2, \ldots, \mathbf{w}_2]_{p\times n_i}\\
&= \mathbf{Y}_i - \mathbf{w}_1'\bar{\mathbf{y}}_i[\mathbf{w}_1, \ldots, \mathbf{w}_1]_{p\times n_i} - \mathbf{w}_2'\bar{\mathbf{y}}_i[\mathbf{w}_2, \ldots, \mathbf{w}_2]_{p\times n_i},
\end{aligned}$$

finally, the adjusted data becomes

$$\mathbf{Y}_i^{dm+mc(p)} = \mathbf{Y}_i - \mathbf{w}_1'\bar{\mathbf{y}}_i[\mathbf{w}_1, \ldots, \mathbf{w}_1]_{p\times n_i} - \mathbf{w}_2'\bar{\mathbf{y}}_i[\mathbf{w}_2, \ldots, \mathbf{w}_2]_{p\times n_i} - \cdots - \mathbf{w}_p'\bar{\mathbf{y}}_i[\mathbf{w}_p, \ldots, \mathbf{w}_p]_{p\times n_i}.$$

Here we can replace $\mathbf{w}_1$, $\mathbf{w}_2$, ..., $\mathbf{w}_p$ with $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_p$, where $\mathbf{e}_1 = (1, 0, \ldots, 0)'$, $\mathbf{e}_2 = (0, 1, \ldots, 0)'$ and so on. This is because, for any other orthogonal basis $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_p$, it is true that

$$\begin{aligned} \bar{\mathbf{y}} &= \mathbf{w}_1'\bar{\mathbf{y}}\mathbf{w}_1 + \cdots + \mathbf{w}_p'\bar{\mathbf{y}}\mathbf{w}_p \\ &= \mathbf{v}_1'\bar{\mathbf{y}}\mathbf{v}_1 + \cdots + \mathbf{v}_p'\bar{\mathbf{y}}\mathbf{v}_p. \end{aligned}$$

Since $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_p$ form an orthogonal basis in p dimensions, replacing $\mathbf{w}$'s by these gives the same result. Therefore,

$$\begin{aligned} \mathbf{Y}_i^{dm+mc} &= \mathbf{Y}_i - \mathbf{e}_1'\bar{\mathbf{y}}_i[\mathbf{e}_1, \ldots, \mathbf{e}_1]_{p\times n_i} - \mathbf{e}_2'\bar{\mathbf{y}}_i[\mathbf{e}_2, \ldots, \mathbf{e}_2]_{p\times n_i} - \cdots - \mathbf{e}_p'\bar{\mathbf{y}}_i[\mathbf{e}_p, \ldots, \mathbf{e}_p]_{p\times n_i} \\ &= \mathbf{Y}_i - [\bar{\mathbf{y}}_i, \bar{\mathbf{y}}_i, \ldots, \bar{\mathbf{y}}_i]_{p\times n_i} \\ &= \mathbf{Y}_i^{mc}. \end{aligned}$$

□

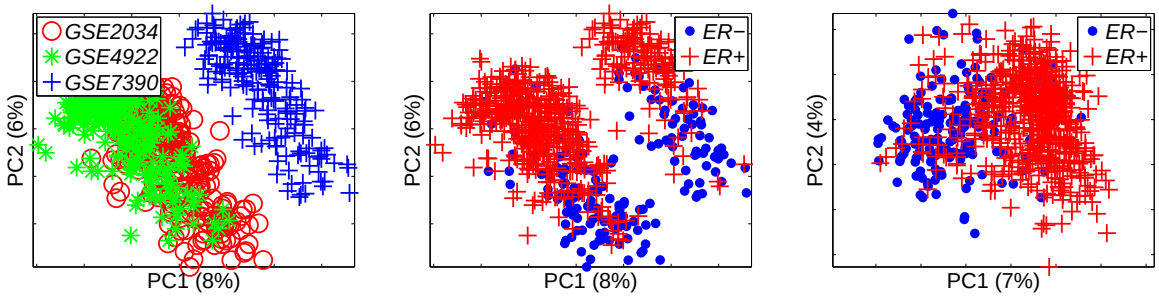
2.3 Evaluation of Batch Effect Adjustment

Another important aspect of batch adjustment is how to justify or evaluate a given batch adjustment method. Success of batch effect removal has been typically judged in the following two aspects: 1) whether the method would enhance prediction performance of biological class in the combined data set, 2) how homogeneous batches would become after adjustment.

2.3.1 Prediction Performance

Successful batch adjustment is often evaluated based on the prediction performance in a combined data set. Some microarray studies are attempted with the purpose of exploring predictors; e.g., clinically useful prognostic markers for cancer. Unfortunately, there are few overlaps among individual studies due to the limited number of patients (Michiels et al., 2005). Naturally a goal of integrating data sets is increasing the sample sizes, thereby discovering more reliable predictors and increasing prediction accuracy (Cheng et al., 2009; Tan et al., 2003; Vachani et al., 2007; Xu et al., 2005).

One simple way to check the improved performance of merged data sets is through the graphical technique such as principal component plot. Suppose we consider combining three different breast cancer data sets to predict estrogen receptor (ER) status. Figure 2.4-(a) displays three batches of sample projected on the first two principal component directions. In (b), before batch adjustment, the plot shows that batch difference somewhat dominates biological difference (ER+ and ER−). Meanwhile, in (c), after the batch adjustment by mean-centering, biological signal became apparent without batch effect; the difference between ER+ and ER− is well separated in the first principal component direction.



(a) Three batches (b) Before batch correction (c) After batch correction

Figure 2.4. PC scatter plot of breast cancer data sets by bio-label before and after batch adjustment. In (a), three batches of samples are displayed on the first two principal component directions, labeled by batch membership. In (b), the same data sets are labeled with biological classes. One can see that batch difference dominates biological difference (ER+, ER−). In (c), after batch correction by mean-centering, biological signal became more clear with no apparent batch bias.

Beyond graphical measures, there have been some efforts to see the improved performance of merged data sets based on classification performance (Huang et al., 2011; Luo et al., 2010; Xu et al., 2005), or survival prediction (Cheng et al., 2009; Yasrebi et al., 2009). In particular, Luo et al. (2010) compared several batch bias removal methods based on cross-batch prediction, i.e., establishing a prediction model in one batch and applying to another. The Matthews correlation coefficient (MCC) is used as the evaluation measure for binary prediction that is known to be useful for unbalanced sample sizes. The MCC can be calculated directly from the confusion matrix using the formula

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively.

The principle of evaluating batch adjustment methods is that if applying a batch adjustment method yields a better MCC score, batch biases would be regarded as effectively removed and the method worked well. However, we notice that removing batch effect is not necessarily followed by good prediction performance, regardless of the choice of batch effect removal method (Luo et al., 2010; Yasrebi et al., 2009). There are more factors affecting predictive power other than a batch adjustment technique, such as classification methods, classifiers, sample size, biological natures, and others. For example, poor prediction performance can be introduced by clinical nature; some biological endpoints (e.g.,overall survival) are notoriously hard to predict while ER status is relatively easy to predict.

Therefore, even though a batch effect removal method has a positive impact on prediction performance in a combined data set or validation data set, it may not necessarily imply that the batch adjustment is effectively done. Rather, we should evaluate how the batch adjustment contributes to homogenizing independently generated data sets.

2.3.2 Similarity of Adjusted Batches

The ideal way to evaluate the success of batch effect removal is to focus on the homogeneity of sample batches, that is, how "similar" the batches have become after the adjustment. There are several ways to look at the similarity of the batches. A simple approach is to use a visualization technique. Benito et al. (2004) employed PC plot to justify their DWD batch adjustment, which provides evidence of well-mixed samples in an informative subspace as shown in Figure 2.5. A heatmap is also often used to visualize the difference between batches (Johnson et al., 2007).

If we regard "similarity" as closeness in terms of location, the performance of methods can be assessed by the difference between means or medians of batches. Chen et al. (2011) compared several batch effect methods based on the presence of batch effect within the

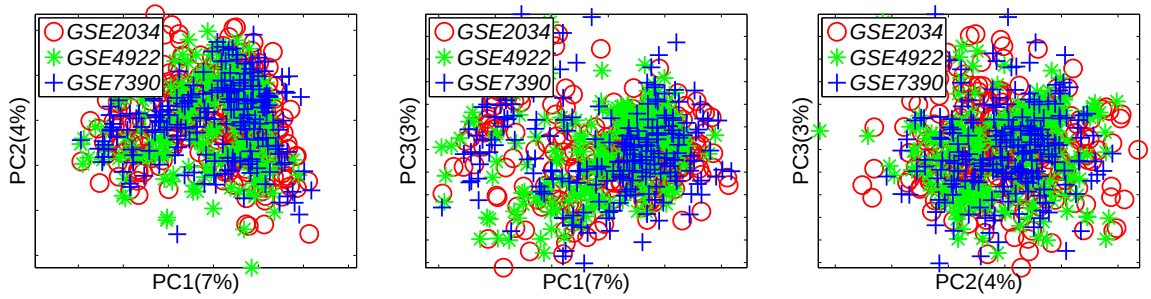


Figure 2.5. PC scatter plot of breast cancer data sets after adjustment, labeled by batch membership. In the previous Figure 2.4-(a), three batches (groups) of samples are separated on the first two PC directions before batch adjustment. In this figure, the batches appear to be well mixed in the PC plots after the adjustment by mean-centering.

analysis of variance (ANOVA) model; i.e., testing the existence of mean difference between batches.

However, evaluating only the mean's location of batches is too simple and even meaningless since most methods accomplish mean-centering in some way (Chen et al., 2011). We will need to consider diverse aspects of the homogeneity of samples more than closeness of the location. Shabalin et al. (2008) suggested various measures for a validation of the XPN methods, such as average distance to the nearest array in a different batch, correlation between genes in different batches, correlation of t-statistics and preservation of significant genes, and other measures.

In this dissertation we consider the closeness of covariance structures of batches. Test statistics for equal covariance for high dimensional Gaussian data are proposed by Schott (2007) and Srivastava and Yanagihara (2010). We choose the Q_K^2 test statistic proposed by Srivastava and Yanagihara (2010) because it shows better performance in both power and attained significant level especially when the variables are correlated; these facts will be shown in the simulations in Section 2.3.3. In the equal covariance test, the null hypothesis is $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_K = \mathbf{\Sigma}$, where K is the number of batches. It is shown that the test statistic Q_K^2 asymptotically follows χ^2_{K-1} under the null hypothesis as both sample size

and dimension grow. For $K = 2$, their test statistic Q_K^2 is based on the difference between $tr(\hat{\mathbf{\Sigma}}_1^2)/\{tr(\hat{\mathbf{\Sigma}}_1)\}^2$ and $tr(\hat{\mathbf{\Sigma}}_2^2)/\{tr(\hat{\mathbf{\Sigma}}_2)\}^2$. A smaller value of the test statistic indicates greater similarity of two population covariance matrices. This test is also available when there are more than two batches. The general form of the test statistic is given by

$$Q_K^2 = \sum_{i=1}^{K} \frac{(\hat{\gamma}_i - \bar{\hat{\gamma}})^2}{\hat{\xi}_i^2}, \tag{2.3}$$

where

$$\begin{aligned}
\bar{\hat{\gamma}} &= \frac{\sum_{i=1}^K \hat{\gamma}_i/\hat{\xi}_i^2}{\sum_{i=1}^K 1/\hat{\xi}_i^2}, \\
\hat{\gamma}_i &= \frac{\hat{a}_{2i}}{\hat{a}_{1i}^2},\ i = 1, \ldots, K\ , \\
\hat{\xi}_i^2 &= \frac{4}{n_i^2}\left\{\frac{\hat{a}_2^2}{\hat{a}_1^4} + \frac{2n_i}{p}\left(\frac{\hat{a}_2^3}{\hat{a}_1^6} - \frac{2\hat{a}_2\hat{a}_3}{\hat{a}_1^5} + \frac{\hat{a}_4}{\hat{a}_1^4}\right)\right\},\ i = 1, \ldots, K.
\end{aligned}$$

Here $\hat{a}_i$ is a consistent estimator of $a_i = tr(\mathbf{\Sigma}^i)/p$ for $i = 1, \ldots, 4$. Let us denote $\mathbf{S}_i$ the sample covariance matrix and n_i is the degree of freedom for each batch. Also define that $\mathbf{V}_i = n_i\mathbf{S}_i$, $\mathbf{V} = \sum_{i=1}^K \mathbf{V}_i$ and $n = \sum_{i=1}^K n_i$. Then,

$$\begin{aligned}
\hat{a}_{1i} &= \frac{1}{pn_i} tr(\mathbf{V}_i), \quad \text{and} \quad \hat{a}_1 = \frac{1}{pn} tr(\mathbf{V}), \\
\hat{a}_{2i} &= \frac{1}{p(n_i - 1)(n_i + 2)}\left\{tr(\mathbf{V}_i^2) - \frac{1}{n_i}(tr\mathbf{V}_i)^2\right\}, \\
\hat{a}_2 &= \frac{1}{(n-1)(n+2)p}\left\{tr(\mathbf{V}^2) - \frac{1}{n}(tr\mathbf{V})^2\right\}, \\
\hat{a}_3 &= \frac{n}{(n-1)(n-2)(n+2)(n+4)}\left\{\frac{1}{p}tr(\mathbf{V}^3) - 3(n+2)(n-1)\hat{a}_2\hat{a}_1 - np^2\hat{a}_1^3\right\}, \\
\hat{a}_4 &= \frac{1}{c_0}\left(\frac{1}{p}tr(\mathbf{V}^4) - pc_1\hat{a}_1 - p^2c_2\hat{a}_1^2\hat{a}_2 - pc_3\hat{a}_2^2 - np^3\hat{a}_1^4\right),
\end{aligned}$$

where

$$c_0 = n(n^3 + 6n^2 + 21n + 18),\ c_1 = 2n(2n^2 + 6n + 9),$$

$$c_2 = 2n(3n + 2),\ \text{and}\ c_3 = n(2n^2 + 5n + 7).$$

In the following subsection, we carry out some simulations to see the performance of the Q_K^2 test statistic as well as other competing test methods.

2.3.3 Tests for the Equality of High Dimensional Covariance Matrices

The high dimensional covariance test is first introduced in Schott (2007). The null hypothesis is

$$H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_K = \mathbf{\Sigma}, \tag{2.4}$$

where $\mathbf{\Sigma}_i, i = 1, \ldots, K$ is a covariance matrix of a p-variate normal population. For the test of H_0 when sample size n is less than dimension p, Schott (2007) suggests the test statistic,

$$\begin{aligned} J_K &= \sum_{i<j} \Big[tr(\mathbf{S}_i - \mathbf{S}_j)^2 - (n_i\eta_i)^{-1}\{n_i(n_i-2)tr(\mathbf{S}_i^2) + n_i^2 tr(\mathbf{S}_i)^2\} \\ &- (n_j\eta_j)^{-1}\{n_j(n_j-2)tr(\mathbf{S}_j^2) + n_j^2 tr(\mathbf{S}_j)^2\} \Big], \end{aligned} \tag{2.5}$$

where n_i denotes degrees of freedom (sample size $-$ 1) for ith group, and η_i is $(n_i+2)(n_i-1)$. $\mathbf{S}_i$ denotes sample covariance matrix. Note that (2.5) is based on the trace of the sample covariance matrices. This is a major difference from the conventional equal covariance test when $n > p$ that uses the determinants of matrices (Muirhead, 1982).

The basic idea of deriving (2.5) is to find an unbiased estimator of $\sum_{i<j} tr(\mathbf{\Sigma}_i - \mathbf{\Sigma}_j)^2$. Thus, (2.5) has the following property

$$\mathrm{E}(J_K) = \sum_{i<j} tr(\mathbf{\Sigma}_i - \mathbf{\Sigma}_j)^2. \tag{2.6}$$

If the null hypothesis (2.4) holds, (2.6) will be zero. Specifically, Schott (2007) proved that J_K approximately follows normal distribution with zero mean and a constance variance, say θ^2, as $(n_i, p) \to \infty$ under H_0. Unfortunately, however, they admit that this test is somewhat limited in that the convergence of null distribution to normal distribution is slower when $\mathbf{\Sigma} \neq \mathbf{I}_p$.

On the other hand, Srivastava and Yanagihara (2010) proposed the statistic T_K^2 and Q_K^2. The T_K^2 statistic is based on the differences of $tr(\mathbf{\Sigma}_1^2)$ and $tr(\mathbf{\Sigma}_2^2)$, and the Q_K^2 is based on

the differences of $tr(\mathbf{\Sigma}_1^2)/(tr\mathbf{\Sigma}_1)^2$ and $tr(\mathbf{\Sigma}_1^2)/(tr\mathbf{\Sigma}_2)^2$. So a smaller value of these two test statistics implies greater similarity of two covariances. Under the null hypothesis, both T_K^2 and Q_K^2 are asymptotically distributed as χ_{K-1}^2 as (n_i, p) approach infinity. But Srivastava and Yanagihara (2010) showed that Q_K^2 performs better than T_K^2 in terms of the power of the test.

In what follows we provide some simulations to compare the performance of the high dimensional covariance tests: J_K, T_K^2 and Q_K^2.

a. **Simulation Study 1**: Power of the test for different covariance structures.

We carry out a simulation in order to see the power of the tests in diverse situations. With a goal of rejecting the null hypothesis ($\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$), the alternatives ($\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$) are constructed with the following four different scenarios.

- Diagonal signal: $\mathbf{\Sigma}_1 = \mathbf{I}_p$ vs. $\mathbf{\Sigma}_2 = 2\mathbf{I}_p$.
- Off-diagonal signal: $\mathbf{\Sigma}_1 = (1-0.5)\mathbf{I} + 0.5\mathbf{1}\mathbf{1}^{\mathrm{T}}$ vs. $\mathbf{\Sigma}_2 = (1-0.3)\mathbf{I} + 0.3\mathbf{1}\mathbf{1}^{\mathrm{T}}$.
- Diagonal signal*: $\mathbf{\Sigma}_1 = \delta_i\mathbf{I}_p$ vs. $\mathbf{\Sigma}_2 = \delta\mathbf{I}_p$, where $\delta = \{\sum_{i=1}^p \delta_i^2/p\}^{1/2}$.
- Both diagonal and off-diagonal signal*: $\mathbf{\Sigma}_1 = (1-0.5)\mathbf{I} + 0.5\mathbf{1}\mathbf{1}^{\mathrm{T}}$ vs. $\mathbf{\Sigma}_2 = (\delta - 0.3)\mathbf{I} + 0.3\mathbf{1}\mathbf{1}^{\mathrm{T}}$, where $\delta = \{1 + (p-1)0.5^2 - (p-1)0.3^2\}^{1/2}$.

Note that in the last two scenarios with the asterisk notation, a constant δ is determined under the condition that $tr(\mathbf{\Sigma}_1^2) = tr(\mathbf{\Sigma}_2^2)$, which avoids the rejection by the size of trace square. Three test statistics are calculated based on two samples from $\mathcal{N}_{200}(\mathbf{0}, \mathbf{\Sigma}_1)$ and $\mathcal{N}_{200}(\mathbf{0}, \mathbf{\Sigma}_2)$, where the sample size is 50 for each data. We repeat this 100 times. Then we observe how many times the tests reject the null hypothesis at 0.05 significance level. In these scenarios, we intend to reject H_0 since data sets are generated from different covariance structures. Thus, high frequency of the rejection out of 100 repetitions indicates that the tests work well.

The results are summarized in Table 2.1, which shows the percentage of the rejections. A high percentage indicates that a test successfully detects the difference between two

covariance matrices, whereas a low percentage indicates that the test does not always detect the difference of the covariances. The Q_K^2 has the best results among the three tests in that it shows high rejection percentages in all cases. However, the J_K cannot detect off-diagonal signal well, and the T_K^2 test only detects the diagonal signal well.

Table 2.1

Comparison of three test statistics. The percentage indicates the number of rejections out of 100 repetitions for the null hypothesis of equal covariance. A high percentage indicates that the test works well. The Q_K^2 generally performs well among the three tests.

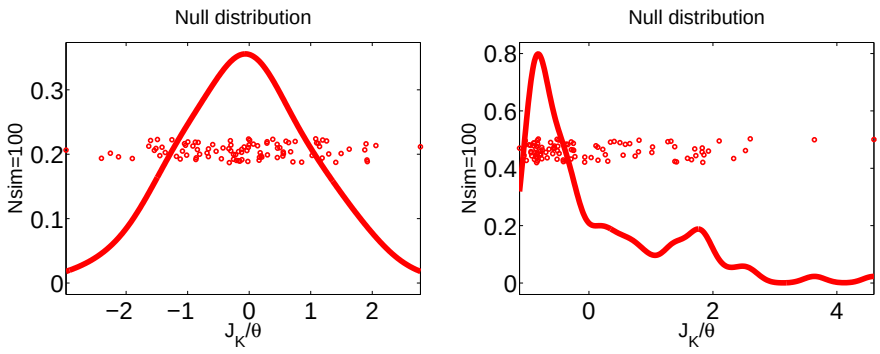
	J_K^2	T_K^2	Q_K^2
Diagonal signal	100%	100%	87%
Off-diagonal signal	47%	36%	81%
Diagonal signal*	100%	20%	100%
Both diagonal and off-diagonal signal*	100%	8%	100%

b. **Simulation Study 2**: Null distribution of J_K

In order to take a closer look at the J_K statistic (since it shows weak performance for the detection of off-diagonal signal in Table 2.1), we run a simulation and observe the performance of the null distribution. Two groups of samples are generated from $\mathcal{N}_{200}(0, \mathbf{\Sigma})$ with the sample size 50. We compute J_K from those samples, and repeat this 100 times. Since two data sets are from the equal covariance population, the distribution of J_K is supposed to be approximately normal. The results are shown in Figure 2.6. When $\mathbf{\Sigma}$ is set to $\mathbf{I}_{200}$ in (a), the null distribution of J_K/θ is close to standard normal. Meanwhile, when $\mathbf{\Sigma}$ is set to $0.5\mathbf{I}+0.5\mathbf{1}\mathbf{1}^{\mathrm{T}}$ in (b), the null distribution is not close to standard normal. These results say that using the J_K may be a problematic approach especially when a data set includes many correlated variables.

c. **Simulation Study 3:** Attained significant level and power

The performance of the test statistics can be seen through the attained significant level (ASL) and power. In Table 2.1, the Q_K^2 outperforms the J_K and T_K^2, showing that the J_K



(a) $\mathbf{\Sigma} = \mathbf{I}_p$ (b) $\mathbf{\Sigma} \neq \mathbf{I}_p$

Figure 2.6. Convergence of null distribution of J_K/θ. In (a), $\mathbf{\Sigma} = \mathbf{I}_p$, the null distribution is close to standard normal. Meanwhile, in (b), $\mathbf{\Sigma} \neq \mathbf{I}_p$, the null distribution is not close to normal.

and T_K^2 are not responsive to off-diagonal signal of covariance matrix. To generalize this argument, we carry out a simulation with ASL and power for the third case in Table 2.1, which is the case when the covariance matrices are different in their off diagonals. This simulation is reproduced from those of Srivastava and Yanagihara (2010) with a slight modification for the null and the alternative hypothesis as below.

$$H \quad : \quad \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = (1 - 0.5)\mathbf{I} + 0.5\mathbf{1}\mathbf{1}^{\mathrm{T}}, \tag{2.7}$$

$$A \quad : \quad \mathbf{\Sigma}_1 = 0.5\mathbf{I} + 0.5\mathbf{1}\mathbf{1}' \text{ and } \mathbf{\Sigma}_2 = (\delta - 0.3)\mathbf{I} + 0.3\mathbf{1}\mathbf{1}^{\mathrm{T}}, \tag{2.8}$$

where $\delta = \{\sum_{i=1}^{p} \delta_i^2/p\}^{1/2}$. Following Srivastava and Yanagihara (2010), the ASL and the attained power are defined as

$$\hat{\alpha} = \frac{\sum_{i=1}^{Nsim} I(QH > \chi^2_{1,\alpha})}{Nsim}, \quad \hat{\beta} = \frac{\sum_{i=1}^{Nsim} I(QA > \hat{\chi}^2_{1,\alpha})}{Nsim},$$

respectively, where $Nsim$ is the number of replications, QH are the values of the test statistic computed from data simulated under the null hypothesis, and $\chi^2_{1,\alpha}$ is the upper 100α percentile of the chi-square distribution with 1 degree of freedom. The QA are the values of the test statistic computed from data simulated under the alternative hypothesis, and $\hat{\chi}^2_{1,\alpha}$ is the estimated upper 100α percentile from the empirical chi-square distribution of QH.

In our simulation, $Nsim$ is 1000, α is set to 0.05, and p = 10, 40, 60, 100, 200 and n_i = 10, 20, 40, 60. The ASL and power for the three test statistics are calculated under (2.7) and (2.8). The results are presented in Table 2.2. In ASL, T_K^2 and Q_K^2 are showing decent performance, converging to 0.05 as n_i and p increase. Meanwhile, J_K is supposed to reach 0.025 in both the left and right tail under H_0 by the two-tail test, but it is not converging to 0.025 especially in the left tail. In power, the Q_K^2 is converging to 1 as n_i and p increase. The J_K is also approaching to 1 although the convergence of J_K is slower than that of Q_K^2. However, the T_K^2 is not approaching to 1. Also note that the T_K^2 test is not robust for different covariance structures in Table 2.1. This weak power is because the T_K^2 test statistic only depends on the difference of $tr(\mathbf{\Sigma}_1^2)$ and $tr(\mathbf{\Sigma}_2^2)$, but there are many different covariance matrices that have the same trace size. In conclusion, Q_K^2 shows the best performance in both ASL and power while J_K is not stable in ASL and T_K^2 shows weakness in power. Similar conclusions are derived when $K > 2$ as well.

Table 2.2

Attained significance level (ASL) and power. For the Q_K^2 test statistic, the convergence of ASL to 0.05 is stable and the attained power tends to converge to 1 as both n_i and p increase. Meanwhile, the ASL of the J_K is not converging to 0.025 especially in the left tail, and the attained power of the T_K^2 is not converging to 1.

		ASL				Power		
$K = 2$		J_K		T_K^2	Q_K^2	J_K	T_K^2	Q_K^2
p	n_i	(L)	(R)					
20	10	0	0.0740	0.0250	0.2130	0.0270	0.0540	0.3690
20	20	0	0.0750	0.0260	0.0770	0.1350	0.0800	0.9450
20	40	0	0.0700	0.0380	0.0410	0.3840	0.0770	0.9980
20	60	0	0.0850	0.0460	0.0520	0.7110	0.0700	1.0000
40	10	0	0.0860	0.0240	0.1870	0.0290	0.0240	0.7660
40	20	0	0.0890	0.0360	0.0730	0.1340	0.0620	0.9960
40	40	0	0.0790	0.0400	0.0540	0.6680	0.0690	1.0000
40	60	0	0.0650	0.0260	0.0400	0.9870	0.0850	1.0000
60	10	0	0.0890	0.0160	0.1660	0.0410	0.0340	0.7610
60	20	0	0.0620	0.0250	0.0510	0.2000	0.0810	0.9980
60	40	0	0.0940	0.0390	0.0550	0.7650	0.0720	1.0000
60	60	0	0.0830	0.0460	0.0540	0.9890	0.0880	1.0000
100	10	0	0.0660	0.0130	0.1580	0.0350	0.0220	0.6180
100	20	0	0.0840	0.0320	0.0720	0.1350	0.0520	0.9960
100	40	0	0.0840	0.0430	0.0610	0.8290	0.0620	1.0000
100	60	0	0.1000	0.0510	0.0630	0.9830	0.0600	1.0000
200	10	0	0.0700	0.0110	0.1480	0.0050	0.0130	0.2090
200	20	0	0.0900	0.0470	0.0790	0.0510	0.0220	0.9530
200	40	0	0.0920	0.0410	0.0550	0.8100	0.0550	1.0000
200	60	0	0.0660	0.0330	0.0440	1.0000	0.0720	1.0000

2.4 Batch Effect Removal by Covariance Adjustment

In this section we propose a novel batch effect removal method that adjusts the covariance structure across batches. This method utilizes a factor model and a hard-thresholding idea for the high dimensional covariance estimation, which are described in Section 2.4.2 and 2.4.3, respectively.

2.4.1 Multivariate Batch Adjustment

Let us assume an imaginary situation where batch effect does not exist, or that all current and future data are from the same batch. Define the (unobservable) random vector of gene expression values $\mathbf{Y}^* = (Y_1^*, \ldots, Y_p^*)^{\mathrm{T}}$. The p-dimensional $\mathbf{Y}^*$ is assumed to be from a multivariate distribution with mean vector $\mathrm{E}(\mathbf{Y}^*) = \boldsymbol{\mu}^*$ and nonsingular covariance matrix $\mathrm{Var}(\mathbf{Y}^*) = \boldsymbol{\Sigma}^*$, where p is the number of genes. We also assume that $\mathbf{Z} = \boldsymbol{\Sigma}^{*-1/2}(\mathbf{Y}^* - \boldsymbol{\mu}^*)$ has $\mathrm{E}(\mathbf{Z}) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{Z}) = \mathbf{I}$, which is a common assumption in high dimensional analysis (Ahn et al., 2007; Yata and Aoshima, 2010).

In a more realistic scenario, we observe array vectors $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijp})^{\mathrm{T}}$ from batch i, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$. We assume that each sample array from the ith batch follows a multivariate distribution with mean vector $\mathrm{E}(\mathbf{Y}_{ij}) = \boldsymbol{\mu}_i$ and nonsingular covariance matrix $\mathrm{Var}(\mathbf{Y}_{ij}) = \boldsymbol{\Sigma}_i$. Then we can express

$$\begin{aligned}
\mathbf{Y}_{ij} &= \boldsymbol{\Sigma}_i^{1/2}\mathbf{Z}_j + \boldsymbol{\mu}_i \\
&= \boldsymbol{\Sigma}_i^{1/2}(\boldsymbol{\Sigma}^{*-1/2}(\mathbf{Y}_j^* - \boldsymbol{\mu}^*)) + \boldsymbol{\mu}_i \\
&= \boldsymbol{\Sigma}_i^{1/2}\boldsymbol{\Sigma}^{*-1/2}\mathbf{Y}_j^* - \boldsymbol{\Sigma}_i^{1/2}\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\mu}^* + \boldsymbol{\mu}_i \\
&= f_i(\mathbf{Y}_j^*),
\end{aligned}$$

where a function f_i represents the ith batch effect and $\mathbf{Y}_j^*$ is a realization of $\mathbf{Y}^*$. Thus, the function f_i is an affine transformation of the unobservable $\mathbf{Y}^*$, i.e.,

$$f_i(\mathbf{Y}^*) = \mathbf{A}_i\mathbf{Y}^* + \mathbf{b}_i,$$

where $\mathbf{A}_i = \boldsymbol{\Sigma}_i^{1/2}\boldsymbol{\Sigma}^{*-1/2}$ and $\mathbf{b}_i = -\boldsymbol{\Sigma}_i^{1/2}\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\mu}^* + \boldsymbol{\mu}_i$. Now it can be seen that the batch effect can be adjusted by applying the inverse function f_i^{-1} such that

$$f_i^{-1}(\mathbf{Y}) = \mathbf{A}_i^{-1}(\mathbf{Y} - \mathbf{b}_i). \tag{2.9}$$

Note that in this way we could adjust for the batch effect that possibly distorts inter-gene relationship as well as mean and variance effect for individual genes.

Then the critical question is how to estimate the matrix $\mathbf{A}_i^{-1} = \boldsymbol{\Sigma}^{*1/2}\boldsymbol{\Sigma}_i^{-1/2}$, which is a function of the two population covariance matrices. Note that $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}^*$ implies that the covariance of ith batch is the same as the one under the assumption of no batch effect. It is clear that in this case $\mathbf{A}_i^{-1}$ is identity, so there is no need to correct the multiplicative effect in (2.9). In general, we need to suggest how to obtain the estimates of $\hat{\boldsymbol{\Sigma}}_i^{-1}$ and $\hat{\boldsymbol{\Sigma}}^*$. It is noted that using the usual sample covariance is not acceptable because of the singularity from $p > n_i$.

2.4.2 High Dimensional Covariance Estimation Based on Factor Model

Estimating high dimensional covariance matrices has been gaining much importance over the recent years. The classical sample covariance estimator is not directly applicable to many multivariate studies when the dimension is large relative to the sample size. According to Bickel and Levina (2008b), most problems related to the high dimensional covariance estimation can be solved by two main approaches. One is the estimation of eigenstructure of the covariance matrix, which is useful for some extended research of the principal component analysis. The other approach is the estimation of the inverse, usually called precision matrix, which relates to linear discriminant analysis, regression, conditional independence analysis in graphical model and many others.

We note that in the ideal batch adjustment suggested in the previous section, the multiplicative factor matrix $\mathbf{A}_i^{-1}$ has two components. The first part is related to the covariance of the *true batch*, i.e., data without batch bias. The second component is the inverse of the covariance of an observed batch. Therefore we need a unified framework under which both

covariance and precision matrices are estimated. To achieve this goal, we employ the factor model proposed by Fan et al. (2008).

One advantage of using the factor model is that gene categorization can be taken into account, by assuming that the behaviors of the many numbers of observed variables are determined by a much smaller number of "factors." A common belief in gene expression analysis is that there exist groups of genes within which the genes "act" together (Dettling and Buehlmann, 2004; The Gene Ontology Consortium, 2008). There are a few different approaches for defining factors for genes. Most commonly, one can use the Gene Ontology (GO), grouping genes that belong to the same pathway. The so-called "gene set enrichment" analysis (Efron and Tibshirani, 2007; Subramanian et al., 2005) is based on this approach. A possible drawback is that at the current moment the pathway information is not complete and can be inaccurate (Montaner et al., 2009). The DNA sequence information of the genes can also be used to define factors, that is, the genes with similar DNA sequences are expected to form a group (Claesson et al., 2009; Davey et al., 2007). This approach utilizes the evolutionary information. Another approach is to use the data set at hand to create clusters of the genes, using a clustering algorithm such as k-means. In this work, we use the pathway approach, provided at http://go.princeton.edu/cgi-bin/GOTermMapper.

Under this presumption of factors, we briefly introduce the framework of the factor model. The factor model is a multiple regression model on each gene expression level Y_i, $i = 1, \ldots, p$,

$$Y_i = \beta_{i1} X_1 + \cdots + \beta_{iq} X_q + \varepsilon_i, \tag{2.10}$$

where $X_1, \ldots, X_q$ are q known factors and β_{ij}, $j = 1, \ldots, q$, are regression coefficients. Following Fan et al. (2008), let $\mathbf{y} = (Y_1, \ldots, Y_p)^{\mathrm{T}}$, $\mathbf{x} = (X_1, \ldots, X_q)^{\mathrm{T}}$ and $\mathbf{e} = (\varepsilon_1, \ldots, \varepsilon_p)^{\mathrm{T}}$, and rewrite (2.10) in a matrix form

$$\mathbf{y} = \mathbf{B}\mathbf{x} + \mathbf{e}. \tag{2.11}$$

Note that $\mathbf{B} = \{\beta_{ij}\}$ is a $p \times q$ regression coefficient matrix. For microarray data, a random vector $\mathbf{y}$ indicates p gene expression values for an individual subject. Each of p variables has

a linear relationship with known q factors, denoted by a vector $\mathbf{x}$. In this study, we utilize 21 categories of Gene Ontology as the q factors, and X_j represents the group mean of expression values that belong to the jth category. We make the common assumption that $\mathrm{E}(\mathbf{e}|\mathbf{x}) = \mathbf{0}$, and $\mathrm{cov}(\mathbf{e}|\mathbf{x})$ is diagonal (Fan et al., 2008).

Fan et al. (2008) estimated covariance of $\mathbf{y}$ based on (2.11). They define two data matrices $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]_{p\times n}$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]_{q\times n}$ from n i.i.d sample of random vector $\mathbf{y}$ and $\mathbf{x}$, respectively. Then the least squares estimator of $\mathbf{B}$ in (2.11) is given by $\hat{\mathbf{B}} = \mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}$. Let $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{y})$ and $\mathbf{\Sigma}_0 = \mathrm{cov}(\mathbf{e}|\mathbf{x})$. The estimation of the covariance matrix of $\mathbf{y}$ from n samples can be derived from the model (2.11):

$$\hat{\mathbf{\Sigma}} \;=\; \hat{\mathbf{B}}\widehat{\mathrm{cov}}(\mathbf{x})\hat{\mathbf{B}}^{\mathrm{T}} + \hat{\mathbf{\Sigma}}_0, \tag{2.12}$$

where $\widehat{\mathrm{cov}}(\mathbf{x})$ is a $q \times q$ nonsingular sample covariance matrix from the given factor matrix $\mathbf{X}$. Lastly, $\hat{\mathbf{\Sigma}}_0$ is obtained by $\mathrm{diag}(n^{-1}\hat{\mathbf{E}}\hat{\mathbf{E}}^{\mathrm{T}})$ where $\hat{\mathbf{E}}$ is the residual matrix, $\mathbf{Y} - \hat{\mathbf{B}}\mathbf{X}$. The estimator $\hat{\mathbf{\Sigma}}$ in (2.12) is always invertible even when the dimension p exceeds n. Fan et al. (2008) has shown that $\hat{\mathbf{\Sigma}}^{-1}$ performs better than the inverse of the sample covariance matrix as $(p, q, n) \to \infty$; see Fan et al. (2008, 2011) for detailed discussions on the convergence rate of the estimate.

Back to the batch problem, we can use the estimator in (2.12) for a covariance estimator for each batch $\hat{\mathbf{\Sigma}}_i$, $i = 1, \ldots, K$. Another matrix to estimate, $\hat{\mathbf{\Sigma}}^*$, is the covariance of the "true batch." If one of the observed batches, say i^*th batch, can be regarded as close to the true batch, one can use $\hat{\mathbf{\Sigma}}_{i*}$. This batch may have been produced under the best conditions. Specifically, it may be that the facility that produced these measurements had the most experience with the technology. The batch might show the best quality metrics, or the best reproducibility on technical replicates, such as Strategene Universal Human Reference samples. In this case the proposed adjustment method transforms the data so that all the batches mimic the ideal batch.

When it is difficult to pinpoint a better batch, we can pool the covariance estimates for each batch as following

$$\hat{\mathbf{\Sigma}}^* = \frac{(n_1 - 1)\hat{\mathbf{\Sigma}}_1 + \cdots + (n_K - 1)\hat{\mathbf{\Sigma}}_K}{n_1 + \cdots + n_K - K}. \tag{2.13}$$

Note that this assumes that $\mathbf{\Sigma}^*$ is reasonably close to $K^{-1}\sum_{i=1}^{K}\mathbf{\Sigma}_i$.

2.4.3 Sparse Estimation through hard thresholding

In practice, the suggested covariance adjustment estimator $\hat{\mathbf{A}}_i^{-1}$ in the previous section can induce a substantial amount of uncertainty since the estimation involves a multiplication of high-dimensional covariance estimates, one of which is inverted. In high dimensional data analysis, it is a common assumption that not all variables are signal variables. Thus some degree of sparsity is usually imposed in the estimation process to achieve a more stable estimator, especially when high dimensional covariance matrix is being estimated. See for example Bickel and Levina (2008b), Friedman et al. (2008), and Cai and Liu (2011).

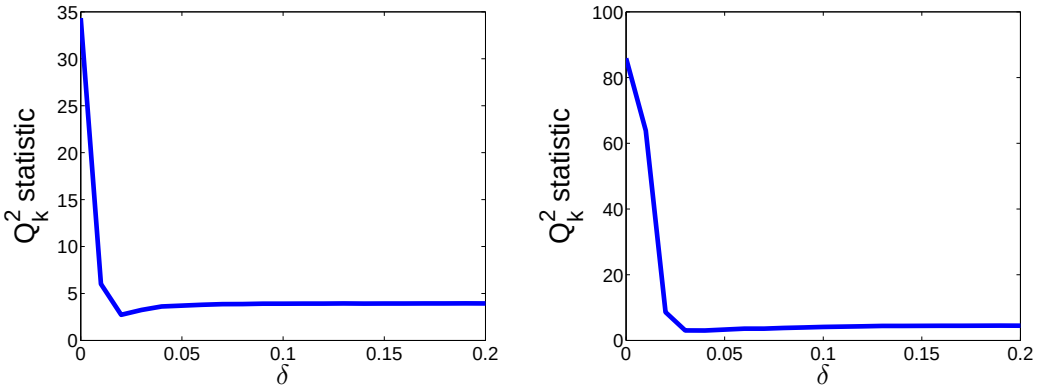
In this work we use a hard-thresholding idea (Bickel and Levina, 2008a; Shao et al., 2011), i.e., entries that are smaller than some tuning parameter, say δ, in the estimated matrix are forced to be zero. Let us define $\hat{\mathbf{A}}_i^{-1}(\delta)$ to be a sparse estimate of $\mathbf{A}_i^{-1}$ by hard-thresholding with an appropriately chosen δ, i.e., the (j,k)-th off-diagonal element a_{jk} of $\mathbf{A}_i^{-1}$ is

$$a_{jk}(\delta) = a_{jk}I(|a_{jk}| > \delta), \quad j \neq k.$$

In order to choose δ, we consider similarity between covariances of the adjusted batches. Let $\mathbf{S}_i$ and $\mathbf{S}_i^\delta$ be the sample covariance matrices of the ith batch before and after the adjustment, respectively. Note that $\mathbf{S}_i^\delta = \hat{\mathbf{A}}_i^{-1}(\delta)\mathbf{S}_i(\hat{\mathbf{A}}_i^{-1}(\delta))^{\mathrm{T}}$. We propose to choose δ that makes $\mathbf{S}_i^\delta$ as similar to each other as possible. In particular, we consider the equal covariance test statistic for high dimensional data proposed by Srivastava and Yanagihara (2010). Their test statistic Q_K^2 is based on the difference between $tr\{(\mathbf{S}_i^\delta)^2\}/\{tr(\mathbf{S}_i^\delta)\}^2$ and $tr\{(\mathbf{S}_j^\delta)^2\}/\{tr(\mathbf{S}_j^\delta)\}^2$. A smaller value of the test statistic indicates greater similarity of two population covariance

matrices. This test is also applicable for comparison of more than two batches. The general form of the Q_K^2 test statistic is given in Section 2.3.2.

Figure 2.7 displays the test statistic Q_K^2 for both the breast cancer data and the lung cancer data in Section 2.7 for a range of $\hat{\delta}$. It can be seen that $\hat{\delta} = .02$ and $\hat{\delta} = .03$ are the best choices for respective data sets. In Section 2.7, we separately choose the level of sparsity for each batch. For computational efficiency, the search is performed around the common $\hat{\delta}$ found in Figure 2.7. As a result, $\hat{\delta} = (0.01, 0.03, 0.02)$ are used for the breast cancer data, and $\hat{\delta} = (0.03, 0.02, 0.06, 0.05)$ for the lung cancer data.



(a) Breast cancer data (b) Lung cancer data

Figure 2.7. Change of the test statistic Q_k^2 over $\hat{\delta}$. The lowest value is obtained at $\hat{\delta} = .02$ and $\hat{\delta} = .03$ in (a) and (b), respectively.

2.5 Theoretical Properties

In this section we study some theoretical properties of $\hat{\mathbf{A}}_i^{-1} = \hat{\mathbf{\Sigma}}^{*1/2}\hat{\mathbf{\Sigma}}_i^{-1/2}$, $i = 1, \ldots, K$, with growing dimensionality (p), number of factors (q), and sample size (n_i). The rate of convergence is studied in terms of Frobenius norm. For any matrix $\mathbf{C} = \{c_{ij}\}$, its Frobenius norm is given by

$$\|\mathbf{C}\| = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |c_{ij}|^2 \right)^{1/2} = \{tr(\mathbf{C}\mathbf{C}^{\mathrm{T}})\}^{1/2}.$$

For the sake of simplicity, we impose the same set of assumptions for each batch so that we can omit the subscript i when discussing the estimation of $\hat{\mathbf{\Sigma}}_i$. Also we assume that

the sample size n_i is all equal to n for all batches. In the following, we repeat some basic assumptions in Fan et al. (2008) for readers. Let $b_n = \mathrm{E}\|\mathbf{y}\|^2$, $c_n = \max_{1\le i\le q}\mathrm{E}(X_i^4)$, and $d_n = \max_{1\le i\le p}\mathrm{E}(\varepsilon_i^4)$.

(i) $(\mathbf{y}_1, \mathbf{x}_1), \ldots, (\mathbf{y}_n, \mathbf{x}_n)$ are i.i.d samples of $(\mathbf{y}, \mathbf{x})$. $\mathrm{E}(\mathbf{e}|\mathbf{x}) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{e}|\mathbf{x}) = \mathbf{\Sigma}_0$ is diagonal. Also the distribution of $\mathbf{x}$ is continuous and the number of factor q is less than dimension p.

(ii) $b_n = O(p)$ and the sequences c_n and d_n are bounded. Also, there exists a constant $\sigma_1 > 0$ such that $\lambda_q(\mathrm{cov}(\mathbf{x})) \ge \sigma_1$ for all n.

(iii) There exists a constant $\sigma_2 > 0$ such that $\lambda_p(\mathbf{\Sigma}_0) \ge \sigma_2$ for all n.

In this paper we further assume that

(iv) Denote the eigenvector and the corresponding eigenvalue of $\mathbf{\Sigma}$ as $(\mathbf{u}_j, \lambda_j)$, and those of $\hat{\mathbf{\Sigma}}$'s as $(\hat{\mathbf{u}}_j, \hat{\lambda}_j)$, $j = 1, \ldots, p$. The conditional expectation $\mathrm{E}(\hat{\mathbf{u}}_j\hat{\mathbf{u}}_j^{\mathrm{T}} \mid \hat{\lambda}_j) = \mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}$ for all j, and $P\{\sum_{j=1}^p(\hat{\lambda}_j - \lambda_j)^2 \ge \sum_{j=1}^p(\hat{\lambda}_j^{1/2} - \lambda_j^{1/2})^2\} = 1$.

Note that the last assumption, which can be re-written as $\sum_{j=1}^p \hat{\lambda}_j(\hat{\lambda}_j - 1) + \lambda_j(\lambda_j - 1) - 2\hat{\lambda}_j^{1/2}\lambda_j^{1/2}(\hat{\lambda}_j^{1/2}\lambda_j^{1/2} - 1) \ge 0$, is easily met in general unless most eigenvalues are less than one. In what follows we list our theoretical findings as well as the proofs.

Lemma 1 (Convergence rate for pooled covariance matrix)**.** *Under the assumptions (i) and (ii), we have*

$$\|\hat{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^*\| = O_p(n^{-1/2}pq).$$

Proof of Lemma 1. Suppose that we use the pooled covariance for the ideal batch and the number of batch K is finite.

$$\begin{aligned}\|\hat{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^*\| &= \|\Big\{\frac{(n-1)\hat{\mathbf{\Sigma}}_1 + (n-1)\hat{\mathbf{\Sigma}}_2 + \cdots + (n-1)\hat{\mathbf{\Sigma}}_K}{(Kn-K)}\Big\} - \Big\{\frac{\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 + \cdots + \mathbf{\Sigma}_K}{K}\Big\}\| \\ &= \frac{1}{K}\|\big(\hat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}_1\big) + \big(\hat{\mathbf{\Sigma}}_2 - \mathbf{\Sigma}_2\big) + \cdots + \big(\hat{\mathbf{\Sigma}}_K - \mathbf{\Sigma}_K\big)\|\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{K}\sum_{i=1}^{K}\|\hat{\mathbf{\Sigma}}_i - \mathbf{\Sigma}_i\| \\
&= O_p(n^{-1/2}pq).
\end{aligned}
$$

The proof of $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = O_p(n^{-1/2}pq)$ has been shown in Fan et al. (2008). □

Lemma 2 (Inequality for the rates of convergence)**.** *Under the assumptions (i), (ii) and (iv), we have*

$$
\mathrm{E}\|\hat{\mathbf{\Sigma}}^{1/2} - \mathbf{\Sigma}^{1/2}\|^2 \leq \mathrm{E}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2.
$$

Proof of Lemma 2.

$$
\begin{aligned}
\mathrm{E}\|\hat{\mathbf{\Sigma}}^{1/2} - \mathbf{\Sigma}^{1/2}\|^2 &= \mathrm{E}tr(\hat{\mathbf{\Sigma}}^{1/2} - \mathbf{\Sigma}^{1/2})^2 \\
&= \mathrm{E}tr\big(\hat{\mathbf{\Sigma}} + \mathbf{\Sigma} - 2\hat{\mathbf{\Sigma}}^{1/2}\mathbf{\Sigma}^{1/2}\big) \\
&= \mathrm{E}\big\{tr(\hat{\Lambda}) + tr(\Lambda) - 2tr(\hat{\mathbf{U}}\hat{\Lambda}^{1/2}\hat{\mathbf{U}}^{\mathrm{T}}\mathbf{U}\Lambda^{1/2}\mathbf{U}^{\mathrm{T}})\big\},
\end{aligned}
$$

where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p]_{p\times p}$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. From the condition (iv), we know that $\mathrm{E}(\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^{\mathrm{T}} \mid \hat{\lambda}_i) = \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$. Then we show that

$$
\begin{aligned}
&\mathrm{E}tr(\hat{\mathbf{U}}\hat{\Lambda}^{1/2}\hat{\mathbf{U}}^{\mathrm{T}}\mathbf{U}\Lambda^{1/2}\mathbf{U}^{\mathrm{T}}) \\
&= \mathrm{E}tr\{(\hat{\Lambda}^{1/2}\hat{\mathbf{U}}^{\mathrm{T}}\mathbf{U})(\Lambda^{1/2}\mathbf{U}^{\mathrm{T}}\hat{\mathbf{U}})\} \\
&= \mathrm{E}\{\textstyle\sum_{i=1}^{p}(\hat{\mathbf{u}}_i^{\mathrm{T}}\mathbf{u}_i)^2\hat{\lambda}_i^{1/2}\lambda_i^{1/2} + \sum_{i\neq j}^{p}(\hat{\mathbf{u}}_i^{\mathrm{T}}\mathbf{u}_j)^2\hat{\lambda}_i^{1/2}\lambda_j^{1/2}\} \\
&= \mathrm{E}\{\textstyle\sum_{i=1}^{p}(\mathbf{u}_i^{\mathrm{T}}\hat{\mathbf{u}}_i)(\hat{\mathbf{u}}_i^{\mathrm{T}}\mathbf{u}_i)\hat{\lambda}_i^{1/2}\lambda_i^{1/2} + \sum_{i\neq j}^{p}(\mathbf{u}_j^{\mathrm{T}}\hat{\mathbf{u}}_i)(\hat{\mathbf{u}}_i^{\mathrm{T}}\mathbf{u}_j)\hat{\lambda}_i^{1/2}\lambda_j^{1/2}\} \\
&= \mathrm{E}\{\textstyle\sum_{i=1}^{p}\mathbf{u}_i^{\mathrm{T}}\mathrm{E}(\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^{\mathrm{T}} \mid \hat{\lambda}_i)\mathbf{u}_i\hat{\lambda}_i^{1/2}\lambda_i^{1/2} + \sum_{i\neq j}^{p}\mathbf{u}_j^{\mathrm{T}}\mathrm{E}(\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^{\mathrm{T}} \mid \hat{\lambda}_i)\mathbf{u}_j\hat{\lambda}_i^{1/2}\lambda_j^{1/2}\} \\
&= \mathrm{E}\{\textstyle\sum_{i=1}^{p}(\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_i)^2\hat{\lambda}_i^{1/2}\lambda_i^{1/2} + \sum_{i\neq j}^{p}(\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j)^2\hat{\lambda}_i^{1/2}\lambda_j^{1/2}\} \\
&= \mathrm{E}\big(\textstyle\sum_{i=1}^{p}\hat{\lambda}_i^{1/2}\lambda_i^{1/2}\big).
\end{aligned}
$$

Similarly, we can show that $\mathrm{E}tr(\hat{\mathbf{U}}\hat{\Lambda}\hat{\mathbf{U}}^{\mathrm{T}}\mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}}) = \mathrm{E}(\sum_{i=1}^{p}\hat{\lambda}_i\lambda_i)$. Note that $tr(\Lambda) = \sum_{i=1}^{p}\lambda_i$. Therefore, we compare two equations

$$
\mathrm{E}\|\hat{\mathbf{\Sigma}}^{1/2} - \mathbf{\Sigma}^{1/2}\|^2 - \mathrm{E}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2
$$

$$
\begin{aligned}
&= \mathrm{E}\big\{ \textstyle\sum_{i=1}^{p} \hat{\lambda}_i + \sum_{i=1}^{p} \lambda_i - 2\sum_{i=1}^{p} \hat{\lambda}_i^{1/2}\lambda_i^{1/2} \big\} - \mathrm{E}\big\{ \textstyle\sum_{i=1}^{p} \hat{\lambda}_i^2 + \sum_{i=1}^{p} \lambda_i^2 - 2\sum_{i=1}^{p} \hat{\lambda}_i\lambda_i \big\} \\
&= \mathrm{E}\big\{ \textstyle\sum_{i=1}^{p} (\hat{\lambda}_i^{1/2} - \lambda_i^{1/2})^2 - \sum_{i=1}^{p} (\hat{\lambda}_i - \lambda_i)^2 \big\} \\
&\leq 0,
\end{aligned}
$$

under the condition (iv). □

Lemma 3 (Convergence rate for the inverse of square root covariance estimator)**.** *Suppose that $q = O(n^{\alpha_1})$ and $p = O(n^{\alpha})$ where $\alpha_1, \alpha \geq 0$. Under the assumptions (i)–(iv), we have*

$$
\|\hat{\mathbf{\Sigma}}^{-1/2} - \mathbf{\Sigma}^{-1/2}\| = o_p\left((p^3 q^4 \log(n)/n)^{1/2}\right).
$$

Proof of Lemma 3. The basic idea is the same as the proof of Theorem 3 in Fan et al. (2008), which showed the weak convergence of $\hat{\mathbf{\Sigma}}^{-1}$. We also follow the same steps for $\hat{\mathbf{\Sigma}}^{-1/2}$. Firstly, in Fan et al. (2008), they applied Sherman-Morrison-Woodbury formular in (2.12) to get

$$
\hat{\mathbf{\Sigma}}^{-1} = \hat{\mathbf{\Sigma}}_0^{-1} - \hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big[\widehat{\mathrm{cov}}(\mathbf{x})^{-1} + \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big]^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_0^{-1}. \tag{2.14}
$$

Here we modify (2.14) by multiplying $\hat{\mathbf{\Sigma}}^{1/2}$ to both sides, and we will get

$$
\hat{\mathbf{\Sigma}}^{-1/2} = \hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{\Sigma}}^{1/2} - \hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big[\widehat{\mathrm{cov}}(\mathbf{x})^{-1} + \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big]^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{\Sigma}}^{1/2}.
$$

Secondly, like Fan et al. (2008), we evaluate the estimation error of each term of $\hat{\mathbf{\Sigma}}^{-1/2}$ as below

$$
\begin{aligned}
\|\hat{\mathbf{\Sigma}}^{-1/2} - \mathbf{\Sigma}^{-1/2}\| \;\leq\; & \|\hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{\Sigma}}^{1/2} - \mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}^{1/2}\| \\
+\; & \|\big(\hat{\mathbf{\Sigma}}_0^{-1} - \mathbf{\Sigma}_0^{-1}\big)\hat{\mathbf{B}}\big[\widehat{\mathrm{cov}}(\mathbf{x})^{-1} + \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big]^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{\Sigma}}^{1/2}\| \\
+\; & \|\mathbf{\Sigma}_0^{-1}\hat{\mathbf{B}}\big[\widehat{\mathrm{cov}}(\mathbf{x})^{-1} + \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big]^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\big(\hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{\Sigma}}^{1/2} - \mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}^{1/2}\big)\| \\
+\; & \|\mathbf{\Sigma}_0^{-1}\big(\hat{\mathbf{B}} - \mathbf{B}\big)\big[\widehat{\mathrm{cov}}(\mathbf{x})^{-1} + \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big]^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}^{1/2}\| \\
+\; & \|\mathbf{\Sigma}_0^{-1}\mathbf{B}\big[\widehat{\mathrm{cov}}(\mathbf{x})^{-1} + \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big]^{-1}\big(\hat{\mathbf{B}}^{\mathrm{T}} - \mathbf{B}^{\mathrm{T}}\big)\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}^{1/2}\| \\
+\; & \|\mathbf{\Sigma}_0^{-1}\mathbf{B}\big\{\big[\widehat{\mathrm{cov}}(\mathbf{x})^{-1} + \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big]^{-1} - \big[\mathrm{cov}(\mathbf{x})^{-1} + \mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}_0^{-1}\mathbf{B}\big]^{-1}\big\}\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}^{1/2}\| \\
\widehat{=}\; & \mathcal{H}_1 + \mathcal{H}_2 + \mathcal{H}_3 + \mathcal{H}_4 + \mathcal{H}_5 + \mathcal{H}_6.
\end{aligned}
$$

Examining each of $\mathcal{H}_1$ to $\mathcal{H}_6$ is quite tedious, and many details are identical to Fan et al. (2006, 2008), so here we only provide some additional proofs with the outline of the remains.

Let us look at the first term $\mathcal{H}_1$.

$$\begin{aligned}
\mathcal{H}_1 &= \|\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}^{1/2}\| \\
&= \|\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}^{1/2} + \boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}^{1/2}\| \\
&\leq \|\hat{\boldsymbol{\Sigma}}_0^{-1} - \boldsymbol{\Sigma}_0^{-1}\|\|\hat{\boldsymbol{\Sigma}}^{1/2}\| + \|\boldsymbol{\Sigma}_0^{-1}(\hat{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}^{1/2})\| \\
&= O_p(n^{-1/2}p^{1/2})O_p(p^{1/2}q^{1/2}) + \mathcal{J}_1 \\
&= O_p(n^{-1/2}pq). \qquad (2.15)
\end{aligned}$$

$\|\hat{\boldsymbol{\Sigma}}_0^{-1} - \boldsymbol{\Sigma}_0^{-1}\|$ is given in Fan et al. (2008). To check $\|\hat{\boldsymbol{\Sigma}}^{1/2}\| = O_p(p^{1/2}q^{1/2})$, we consider an ideal situation in (2.11). We have q factors $X_1, \ldots, X_q$, and each factor has an equal variance, i.e., $\sigma^2 = \text{var}(X_j)$ for all j, as well as the factors are independent each other. Thus let us say $\text{cov}(\mathbf{x}) = \sigma^2\mathbf{I}_q$. Also assume that the factor loadings are all equal over the response variable Y_i, $i = 1, \ldots, p$; for example let $\mathbf{B} = [\mathbf{1}, \ldots, \mathbf{1}]_{p\times q}$. In addition we suppose that $\text{cov}(\mathbf{e}|\mathbf{x}) = \mathbf{I}_p$. Then (2.12) is

$$\hat{\boldsymbol{\Sigma}} = \hat{\sigma}^2 q\mathbf{1}\mathbf{1}^{\mathrm{T}} + \mathbf{I}_p.$$

We now have that $\mathrm{E}\|\hat{\boldsymbol{\Sigma}}^{1/2}\|^2 = \mathrm{E}tr(\hat{\boldsymbol{\Sigma}}) = tr\mathrm{E}(\hat{\sigma}^2 q\mathbf{1}\mathbf{1}^{\mathrm{T}} + \mathbf{I}_p) = tr(\sigma^2 q\mathbf{1}\mathbf{1}^{\mathrm{T}}) + tr(\mathbf{I}_p)$. It turns out to be the sum of eigenvalues. Therefore, $\mathrm{E}\|\hat{\boldsymbol{\Sigma}}^{1/2}\|^2 = \sigma^2 qp + p = O(qp)$. Lastly, we consider the term $\mathcal{J}_1$. From Lemma 2, we see that $\|\hat{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}^{1/2}\| = O_p(n^{-1/2}pq)$. Also note that from the assumption (iii), we know that $\boldsymbol{\Sigma}_0$ is diagonal in which entries are a positive constant. Thus, we have

$$\mathcal{J}_1 = O_p(n^{-1/2}pq).$$

Next, we look at the second term $\mathcal{H}_2$.

$$\begin{aligned}
\mathcal{H}_2 &= \|\big(\hat{\boldsymbol{\Sigma}}_0^{-1} - \boldsymbol{\Sigma}_0^{-1}\big)\hat{\mathbf{B}}\big[\widehat{\text{cov}}(\mathbf{x})^{-1} + \hat{\mathbf{B}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big]^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Sigma}}^{1/2}\| \\
&\leq \|\big(\hat{\boldsymbol{\Sigma}}_0^{-1} - \boldsymbol{\Sigma}_0^{-1}\big)\hat{\boldsymbol{\Sigma}}_0^{1/2}\|\|\hat{\boldsymbol{\Sigma}}_0^{-1/2}\hat{\mathbf{B}}\big[\widehat{\text{cov}}(\mathbf{x})^{-1} + \hat{\mathbf{B}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big]^{-1}\hat{\mathbf{B}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_0^{-1/2}\|\|\hat{\boldsymbol{\Sigma}}_0^{-1/2}\hat{\boldsymbol{\Sigma}}^{1/2}\| \\
&\widehat{=} \mathcal{L}_1\mathcal{L}_2\mathcal{L}_3
\end{aligned}$$

$$= \quad O_p(n^{-1/2}pq). \tag{2.16}$$

$\mathcal{L}_1$ and $\mathcal{L}_2$ have been proved in Fan et al. (2008). There is some change in the term $\mathcal{L}_3$, but it contains the same calculation as in $\mathcal{H}_1$. Thus, the result above is combined from

$$\mathcal{L}_1 = O_p(n^{-1/2}p^{1/2}), \quad \mathcal{L}_2 = O(q^{1/2}), \quad \text{and} \quad \mathcal{L}_3 = O_p(p^{1/2}q^{1/2}).$$

Here we notice that both term $\mathcal{H}_1$ and $\mathcal{H}_2$ will approach zero as $n \to \infty$ since $p^{3/2}q = o((n/log(n))^{1/2})$. This fact relies on the range of α_1 and α from our initial assumption $p = O(n^{\alpha})$ and $q = O(n^{\alpha_1})$. In other words, $n^{-\beta/2}$ consistency is obtained when $\beta = 1-3\alpha-2\alpha_1 \geq 0$. Similar argument about the range of α_1 and α can be found in Theorem 2 in Fan et al. (2008). Similarly, we can bound $\mathcal{H}_3$

$$\begin{aligned}
\mathcal{H}_3 &= \|\boldsymbol{\Sigma}_0^{-1}\hat{\mathbf{B}}\big[\widehat{\text{cov}}(\mathbf{x})^{-1} + \hat{\mathbf{B}}^{\text{T}}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big]^{-1}\hat{\mathbf{B}}^{\text{T}}\big(\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}^{1/2}\big)\| \\
&\leq \|\boldsymbol{\Sigma}_0^{-1/2}\|\|\boldsymbol{\Sigma}_0^{-1/2}\hat{\mathbf{B}}\big[\widehat{\text{cov}}(\mathbf{x})^{-1} + \hat{\mathbf{B}}^{\text{T}}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\mathbf{B}}\big]^{-1}\hat{\mathbf{B}}^{\text{T}}\boldsymbol{\Sigma}_0^{-1/2}\|\|\boldsymbol{\Sigma}_0^{1/2}\big(\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}^{1/2}\big)\| \\
&= O(p^{1/2})O(q^{1/2})o_p((n/log(n))^{-1/2}pq) \\
&= o_p((n/log(n))^{-1/2}p^{3/2}q^{3/2}).
\end{aligned} \tag{2.17}$$

Finally, we consider terms $\mathcal{H}_4$, $\mathcal{H}_5$ and $\mathcal{H}_6$. All of these terms have a slight modification in Fan et al. (2008) by the term $\|\boldsymbol{\Sigma}^{1/2}\|$.

$$\mathcal{H}_4 = O_p(n^{-1/2}p^{3/2}q), \quad \mathcal{H}_5 = O_p(n^{-1/2}p^{3/2}q) \quad \text{and} \quad \mathcal{H}_6 = o_p((n/log(n))^{-1/2}p^{3/2}q^2). \tag{2.18}$$

In conclusion, the next result follows from (2.15) - (2.18) that

$$\sqrt{np^{-3}q^{-4}/log(n)}\ \|\hat{\boldsymbol{\Sigma}}^{-1/2} - \boldsymbol{\Sigma}^{-1/2}\| \xrightarrow{P} 0 \quad \text{as } n \to \infty.$$

□

In Lemma 1, the convergence rate of pooled covariance estimator is determined by the term $n^{-1/2}pq$. Note that this rate is the same for the individual covariance estimator $\hat{\boldsymbol{\Sigma}}_i$, as shown Fan et al. (2008). From Lemma 2, the convergence rate of $\hat{\boldsymbol{\Sigma}}^{1/2}$ is bounded by that of $\hat{\boldsymbol{\Sigma}}$. Furthermore, in Lemma 3, we show the weak convergence of the estimator $\hat{\boldsymbol{\Sigma}}^{-1/2}$ as

$(p, q, n) \to \infty$. Note that p and q increase as n increases, thus the impact of dimensionality can considerably slow down the convergence rate. The idea of Lemma 3 is originated from Fan et al. (2008), who obtained the convergence rate of $\hat{\mathbf{\Sigma}}^{-1}$. In a comparison to the convergence rate for $\hat{\mathbf{\Sigma}}^{-1}$: $\|\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\| = o_p\{(p^2q^4 \log(n)/n)^{1/2}\}$, it can be seen that $\hat{\mathbf{\Sigma}}^{-1/2}$ converges slightly slower than $\hat{\mathbf{\Sigma}}^{-1}$ by the order of $p^{1/2}$. From Lemma 1, 2 and 3 above, the next theorem follows.

Theorem 1 (Convergence rate for the covariance adjustment estimator)**.** *Under assumptions (i)–(iv), we have*

$$\|\hat{\mathbf{\Sigma}}^{*1/2}\hat{\mathbf{\Sigma}}_i^{-1/2} - \mathbf{\Sigma}^{*1/2}\mathbf{\Sigma}_i^{-1/2}\| = o_p\left((p^4q^5 log(n)/n)^{1/2}\right).$$

Proof of Theorem 1. For easy notation, let $\mathbf{A} = \mathbf{\Sigma}^{*1/2}$, $\hat{\mathbf{A}} = \hat{\mathbf{\Sigma}}^{*1/2}$, $\mathbf{B} = \mathbf{\Sigma}_i^{-1/2}$ and $\hat{\mathbf{B}} = \hat{\mathbf{\Sigma}}_i^{-1/2}$ and consider the following problem

$$\begin{aligned}
\|\hat{\mathbf{A}}\hat{\mathbf{B}} - \mathbf{AB}\| &= \|\hat{\mathbf{A}}\hat{\mathbf{B}} - \hat{\mathbf{A}}\mathbf{B} + \hat{\mathbf{A}}\mathbf{B} - \mathbf{AB}\| \\
&= \|\hat{\mathbf{A}}(\hat{\mathbf{B}} - \mathbf{B}) + (\hat{\mathbf{A}} - \mathbf{A})\mathbf{B}\| \\
&\leq \|\hat{\mathbf{A}}\|\|\hat{\mathbf{B}} - \mathbf{B}\| + \|\hat{\mathbf{A}} - \mathbf{A}\|\|\mathbf{B}\|.
\end{aligned}$$

We need to examine each of the four terms. Note that $\|\hat{\mathbf{A}}\| = \|\hat{\mathbf{\Sigma}}^{*1/2}\| = \|\hat{\mathbf{\Sigma}}^{1/2}\| = O_p(p^{1/2}q^{1/2})$ when $n_i = n$, and also we know that $\|\mathbf{\Sigma}^{-1/2}\| = O(p^{1/2})$ from the assumption (iii). Furthermore, by Lemmas 1 and 2,

$$\|\hat{\mathbf{A}} - \mathbf{A}\| = \|\hat{\mathbf{\Sigma}}^{*1/2} - \mathbf{\Sigma}^{*1/2}\| = O_p(n^{-1/2}pq).$$

Combining the result of $\|\hat{\mathbf{B}} - \mathbf{B}\|$ in Lemma 3,

$$\|\hat{\mathbf{A}}\hat{\mathbf{B}} - \mathbf{AB}\| = O_p(p^{1/2}q^{1/2})o_p\{(p^3q^4 log(n)/n)^{1/2}\} + O_p(n^{-1/2}pq)O(p^{1/2}).$$

Since $p^{3/2}q = o((n/log(n))^{1/2})$, we therefore have the following result

$$\sqrt{np^{-4}q^{-5}/log(n)}\ \|\hat{\mathbf{\Sigma}}^{*1/2}\hat{\mathbf{\Sigma}}_i^{-1/2} - \mathbf{\Sigma}^{*1/2}\mathbf{\Sigma}_i^{-1/2}\| \xrightarrow{P} 0 \quad \text{as } n \to \infty.$$

□

The following corollary shows that the after-adjustment covariance estimator has the same convergence rate as the pooled estimator $\hat{\mathbf{\Sigma}}^*$.

Corollary 1. *Under the same conditions as in Theorem 1, we have*

$$\|\widehat{\text{cov}}(\mathbf{Y}_i^*) - \mathbf{\Sigma}^*\| = O_p(n^{-1/2}pq),$$

where $\mathbf{Y}_i^*$ *is the adjusted data in the ith batch.*

Proof of Corollary1. We denote the adjusted data in each batch as $\mathbf{Y}_i^* = \hat{\mathbf{\Sigma}}^{*1/2}\hat{\mathbf{\Sigma}}_i^{-1/2}(\mathbf{Y}_i - \hat{\mathbf{b}}_i)$ in (2.9) Thus, we have

$$\begin{aligned}
\|\widehat{\text{cov}}(\mathbf{Y}_i^*) - \mathbf{\Sigma}^*\| &= \|\hat{\mathbf{\Sigma}}^{*1/2}\hat{\mathbf{\Sigma}}_i^{-1/2}\widehat{\text{cov}}(\mathbf{Y}_i)\hat{\mathbf{\Sigma}}_i^{-1/2}\hat{\mathbf{\Sigma}}^{*1/2} - \mathbf{\Sigma}^*\| \\
&= \|\hat{\mathbf{\Sigma}}^{*1/2}\hat{\mathbf{\Sigma}}_i^{-1/2}\hat{\mathbf{\Sigma}}_i\hat{\mathbf{\Sigma}}_i^{-1/2}\hat{\mathbf{\Sigma}}^{*1/2} - \mathbf{\Sigma}^*\| \\
&= \|\hat{\mathbf{\Sigma}}^* - \mathbf{\Sigma}^*\| \\
&= O_p(n^{-1/2}pq)
\end{aligned}$$

by Lemma 1. □

Theorem 1 shows that the dimensionality $p^2q^{5/2}$ slows down the convergence rate of $\hat{\mathbf{\Sigma}}^{*1/2}\hat{\mathbf{\Sigma}}_i^{-1/2}$ as $n \to \infty$. Suppose both the dimension p and the number of factor q is a fixed small number, then the estimator is slightly slower than $\hat{\mathbf{\Sigma}}_i^{-1/2}$ in Lemma 3. But its performance becomes considerably slower when q increases along with p. In Corollary 1, the convergence rate of $\widehat{\text{cov}}(\mathbf{Y}_i^*)$ is determined by the order of pq, and this rate is the same as that of $\hat{\mathbf{\Sigma}}_i$. Therefore the covariance estimator after batch adjustment achieves the same performance as before batch adjustment. Corollary 1 is the expected consequence due to the fact that $\widehat{\text{cov}}(\mathbf{Y}_i^*)$ pursuits to be $\hat{\mathbf{\Sigma}}^*$ in order to obtain the homogeneity of covariances among batches.

2.6 Simulation Study

In this section, we carry out some simulations to see the performance of the proposed method as well as other existing methods. As the first step, we generate two heterogeneous data

sets in both location and covariance. Then, six methods are attempted to adjust two data sets to make them homogeneous. The methods are mean-centering (MC), distance-weighted discrimination (DWD) method, standardization (Z-score), empirical Bayes (EB) method, cross-platform normalization (XPN) and our proposed multi-batch covariance adjustment (MBCA) method. We also add MBCA(B1) and MBCA(B2), which consider adjusting data with a target batch 1 and 2, respectively.

For the evaluation of the successful adjustment, we investigate data similarity in various aspects. The evaluation items are some graphical measures, the test for the equal covariance (Q_2^2), average distance of nearest samples between batch, comparison of within-batch pair distance versus between-batch pair distance.

2.6.1 Simulation Data

Let us fix the number of batch $K = 2$, dimension $p = 800$, and sample size $n_i = 50$ $(i = 1, 2)$. Data sets are generated from two normal populations. Let us define data matrices $\mathbf{Y}_1 = [\mathbf{y}_{11}, \ldots, \mathbf{y}_{1n}]_{(p\times n_1)}$ and $\mathbf{Y}_2 = [\mathbf{y}_{21}, \ldots, \mathbf{y}_{2n}]_{(p\times n_2)}$, where $\mathbf{y}_{ij}$ is a p-dimensional data vector j for batch i. In other words,

$$\begin{aligned} \mathbf{y}_{1j} &\sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad j = 1, \ldots, n_1 \ , \\ \mathbf{y}_{2j} &\sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \quad j = 1, \ldots, n_2 \ . \end{aligned}$$

Here $\boldsymbol{\mu}_1$ is a p-dimensional mean vector, and the p elements are randomly generated from $unif(0, 1.4)$. Similarly, $\boldsymbol{\mu}_2$'s elements are generated from $unif(-1.4, 0)$. For the covariance, $\boldsymbol{\Sigma}_i$ $(i = 1, 2)$ is determined by $\mathbf{U}_i\mathbf{D}_i\mathbf{U}_i^{\mathrm{T}}$, where $\mathbf{U}_i$ is a $p \times p$ random orthonormal matrix and $\mathbf{D}_i$ is a diagonal matrix that contains eigenvalues. The diagonal elements of $\mathbf{D}_1$ are set to $px^{-1}/3$, $x = 1, \ldots, p$, and the diagonal elements of $\mathbf{D}_2$ are set to $p\exp(-0.042x)/6$. In Figure 2.8, two sets of diagonal elements (eigenvalues) are displayed on the range [1,100] of dimension. $\mathbf{D}_1$ is decreasing in polynomial order and $\mathbf{D}_2$ is exponentially decreasing.

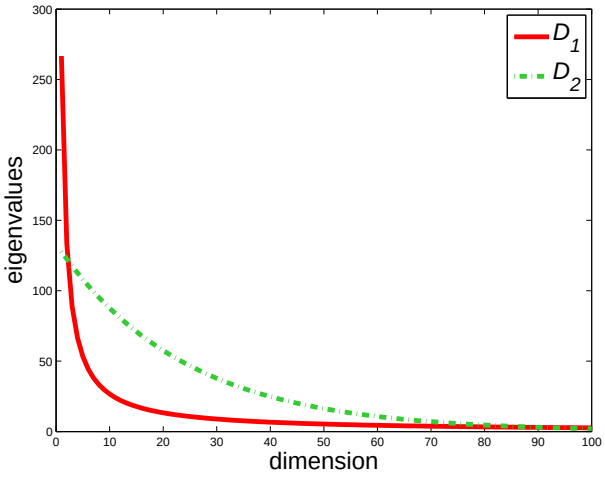


Figure 2.8. Eigenvalues of two population covariance matrices in the simulation setting.

For example, two generated data sets are displayed by the principal components in Figure 2.9. Since we intended to generate heterogeneous data, it appears that two sample batches are different in both location and dispersion.

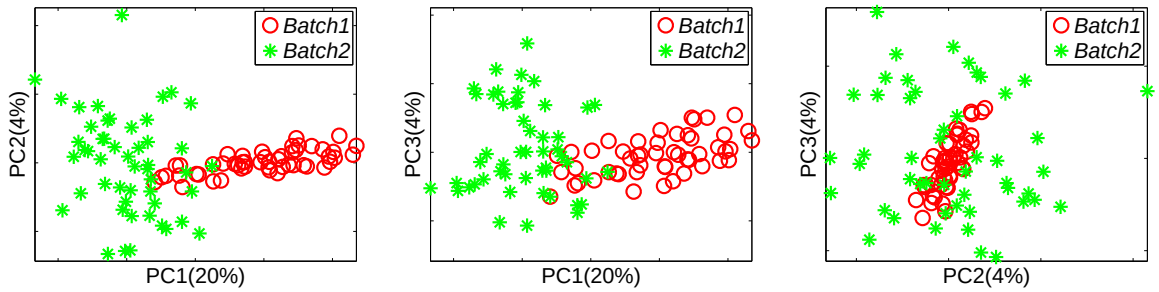


Figure 2.9. PC scatter plot of two sample batches under our simulation settings. We can see the data sets are different in both location and shape.

For the simulated data sets above, we apply several batch adjustment methods to create a homogeneous data set. For our method MBCA, the factor matrix $\mathbf{X}_i$ $(i = 1, 2)$, which is a $k \times n_i$ matrix where $k < p$, is needed corresponding to the data matrix $\mathbf{Y}_i$ $(i = 1, 2)$. The k-means clustering is used to construct the factor matrix with $k = 5$. In real data analysis one can utilize an important factor information if any such as gene pathway for microarray data. For applying the EB and XPN, the software from http://www.bioconductor.org and

their providing software are used under their default setting. These simulations are repeated one hundred times ($Nsim = 100$). The results are following.

2.6.2 Simulation Results

a. Eigenvalues

The similarity of eigenvalues is observed after batch adjustment across methods. In Figure 2.10, coincidence of two eigenvalue curves indicates similarity of data sets. For MC and DWD, the eigenvalues do not change from the Before. It is not surprising because those two methods correct only the mean difference but not the variance-covariance. Meanwhile, Z-score, EB and XPN adjust the variance of data sets, leading to some change in eigenvalue structure. Although the eigenvalue curves come to similar as dimension increases, these are not quite similar in the first few PC directions. In MBCA, MBCA(B1), two eigenvalue curves show the best similarity.

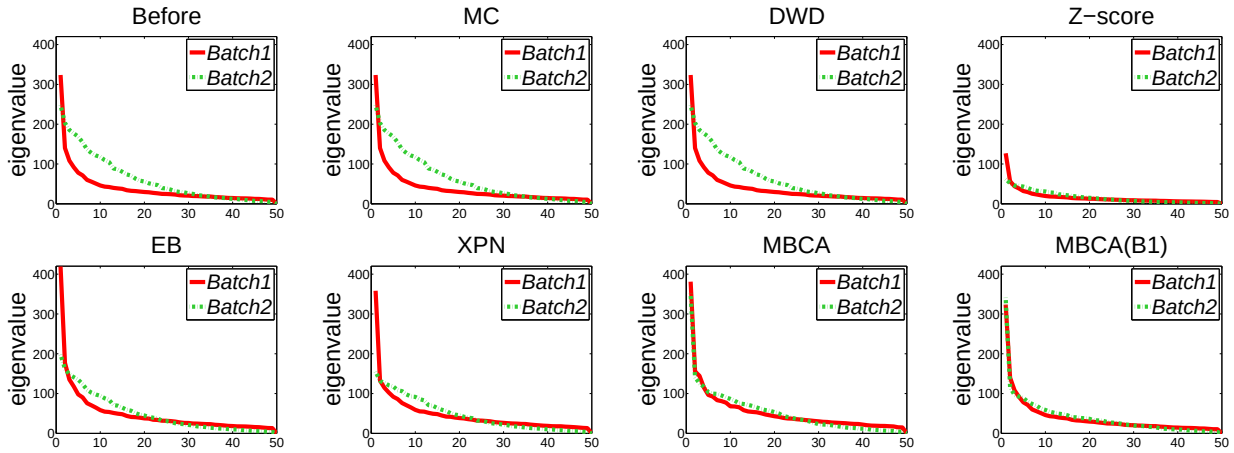


Figure 2.10. Eigenvalues after batch adjustment across methods. Coincidence of two eigenvalue curves indicates similarity of data sets. Among seven methods, our proposed method shows the best similarity.

b. Principal component scatter plot

The principal component (PC) scatter plot is widely used in many statistical analysis. Unlike the usual scatter plot which displays data based on canonical basis, the PC plot

projects data onto the space spanned by eigenvectors (also called principal component direction vectors). This makes it possible that one investigates only a few direction vectors in order to understand the feature of data. Due to its dimensional efficiency, using the principal components is almost essential in high dimensional data analysis.

In this work we use the PC scatter plot to see the data homogeneity after the adjustment. In the previous Figure 2.9, we could see the dissimilarity of the two data sets. After the batch adjustment, it is expected that the two data sets are homogeneous. In Figure 2.11, adjusted data by different methods are shown on the first two PC directions. Although PC1 and 2 contain limited information, these are still useful to see the adjusted results. For instance, the MC and DWD adjust the location of samples but not the shape. Other methods adjust the variance and therefore show well-mixed shape compared to the MC and DWD. Among these, XPN, MBCA and MBCA(B1) are better in the homogeneity of the two batches than Z-score and EB method.

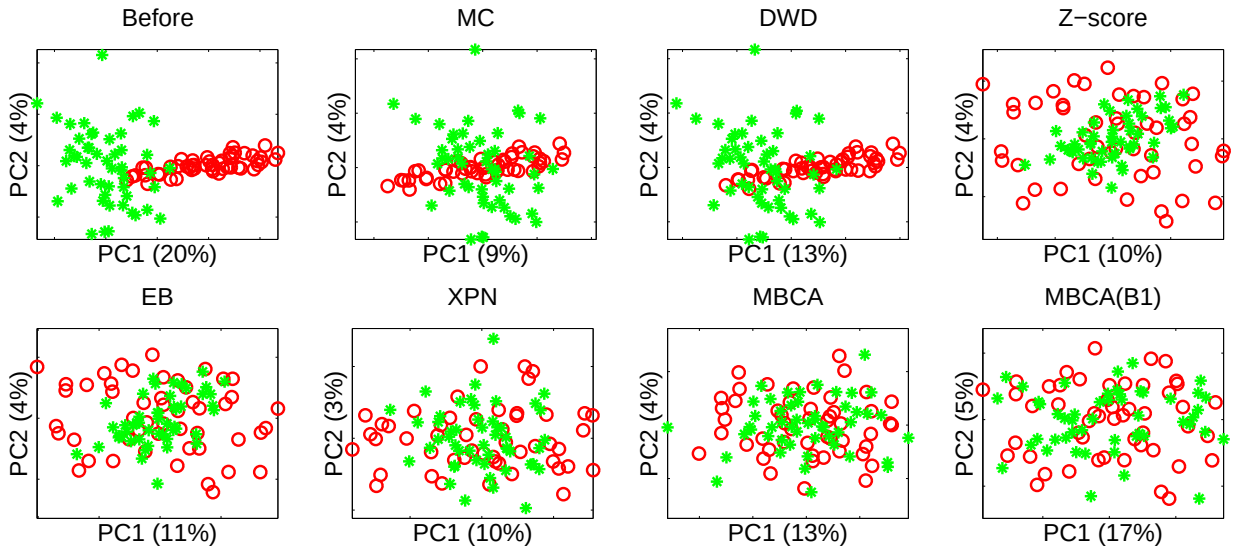


Figure 2.11. PC scatter plot after adjustment across methods. the MC and DWD adjust the location of samples but not the shape. Other methods adjust the variance and thus show well-mixed shape compared to the MC and DWD.

But note that this criterion provides limited knowledge about data feature since a few principal components explain only partial variance of whole data. In some case, Z-score,

EB, XPN, and MBCA produce indistinguishably similar results in the PC plot. In other case, PC3 and 4 provide more useful information. Therefore, we only use this PC plot to gain general idea before and after adjustment. To reach more reliable conclusion, we further investigate numerical measurements based on repetitions.

c. Equal covariance test (Q_2^2)

The adjusted data sets can be tested in the equality of covariance matrices, using Q_2^2 statistic. The Q_K^2 statistic is proposed by Srivastava and Yanagihara (2010) to test the equality for K high dimensional covariance matrices. For the null hypothesis of equal covariance, small p-value indicates dissimilarity of covariances. After adjustment, it is expected to have higher p-values indicating that data sets are similar in terms of covariance. In Figure 2.12, the Before has the small p-values, which is intended from the initial simulation setting. During 100 iterations, the MC and DWD do not change the p-values meaning that these methods do not alter the covariance. Meanwhile, Z-score, EB and XPN methods adjust the covariance in a small range. However, in fact, if our rejection criterion is 0.1, the covariance equality is not guaranteed in these methods. On the other hand, the MBCA, MBCA(B1) and MBCA(B2) achieve the higher p-values during all iterations meaning that our proposed method successfully adjusts the covariance.

d. Average distance of nearest samples between batch

Another way to measure the similarity of data sets can be using the pairwise L_2 distances between arrays after adjustment. As similarly done by Shabalin et al. (2008), we calculate the distance of an array in one batch to the nearest sample in another batch. At first, we standardize the whole data before calculating the distances for the sake of fair comparison across methods. Once the calculation of the distance is done for all samples in one batch, the same thing is repeated for the other batch. Then we average the distances that are produced from both batches. After batch adjustment, it is expected that the nearest sample distance becomes smaller than the Before. In Figure 2.13-(a), during 100 iterations, the average distances of nearest samples between batch are declined by all methods. In

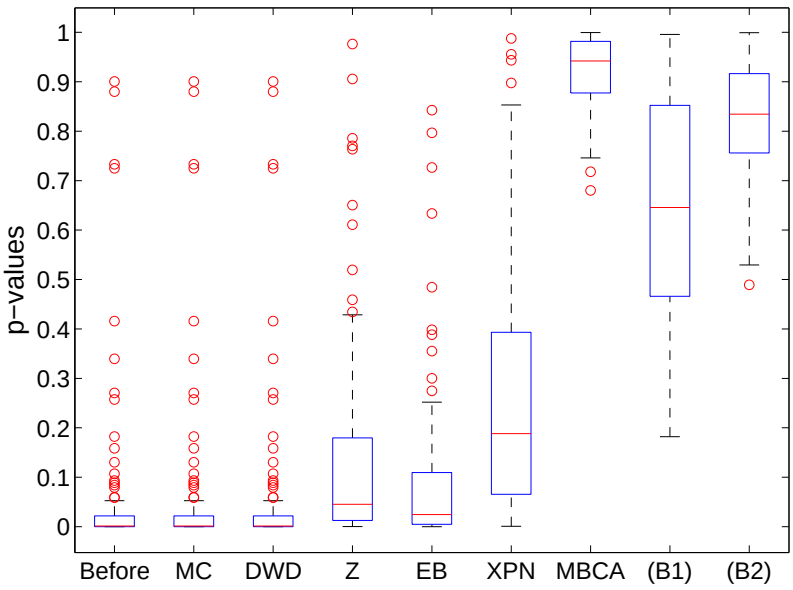
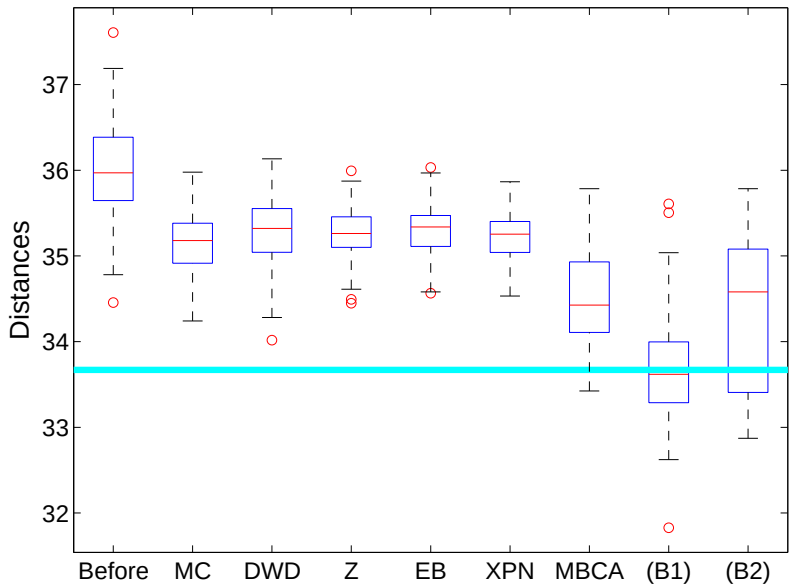


Figure 2.12. Equal covariance test (Q_2^2) after adjustment across the methods. Higher p-values, e.g., over 0.1 significant level, indicates that the adjustment achieves the similarity of data sets in terms of covariance. During 100 simulations, the MC and DWD do not change the p-values. Our proposed method achieves the higher p-values compared to other methods.

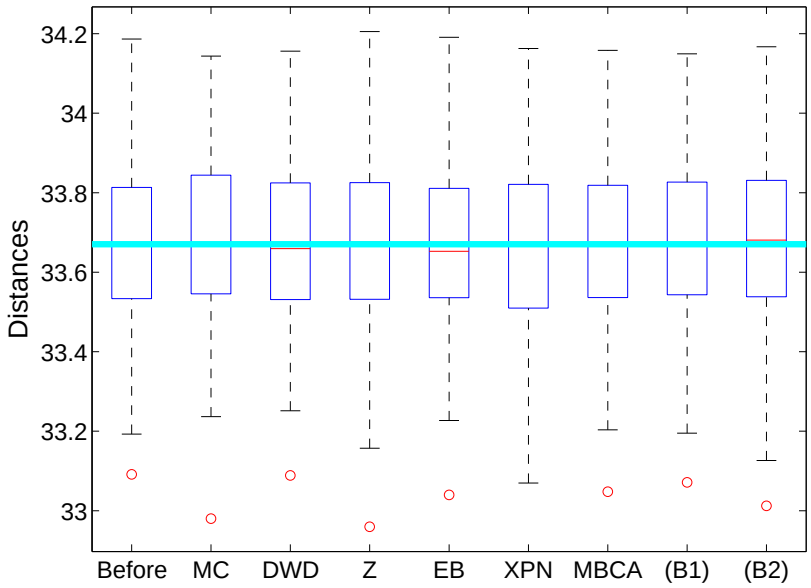
particular, the MBCA method reduces the distances close to the baseline, which is referred from 2.13-(b). The baseline is calculated by the same way of (a) except the fact that two data groups are from random split within batch. This split is repeated 10 times in each batch. The baseline drawn in (a) and (b) are from the median of the 100 baselines from the Before. In comparison to the baseline, the proposed method shows the best performance.

e. Within batch pair distance versus between batch pair distance

We compare the distribution of the pairwise distances between samples in the same batch with those in different batches. Note that we standardize the whole data before calculating the distances. In each panel of Figure 2.14, we overlay two estimated density curves of pairwise distances, which are from within and between batches, respectively for a simulated data. Coincidence of two density curves indicates similarity of the data sets. MC and DWD change the location of two density curves but not the shape of the density.



(a) Average distance of nearest samples between batch



(b) Baseline: the distance from random split within batch

Figure 2.13. Average distance of nearest samples between batch. After batch adjustment, it is expected to decrease the nearest sample distance. In (a), during 100 iterations, the distances are declined by all methods. In particular, the MBCA method reduces the distances close to the baseline. The baseline is calculated based on random split within batch.

Note that the DWD sometimes shows incomplete mean-centering since it only shifts the data sets in a selected direction; see Proposition 1 in Section 2.2.5. The other methods

show better coincidence of two density curves than MC and DWD. In particular, it can be seen that MBCA has the most similar density curves.

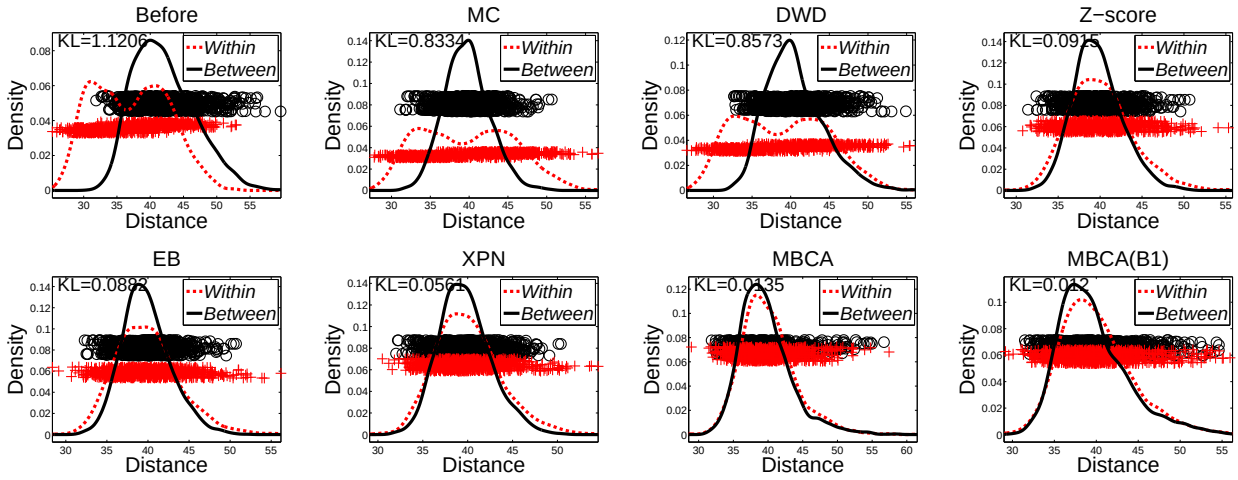


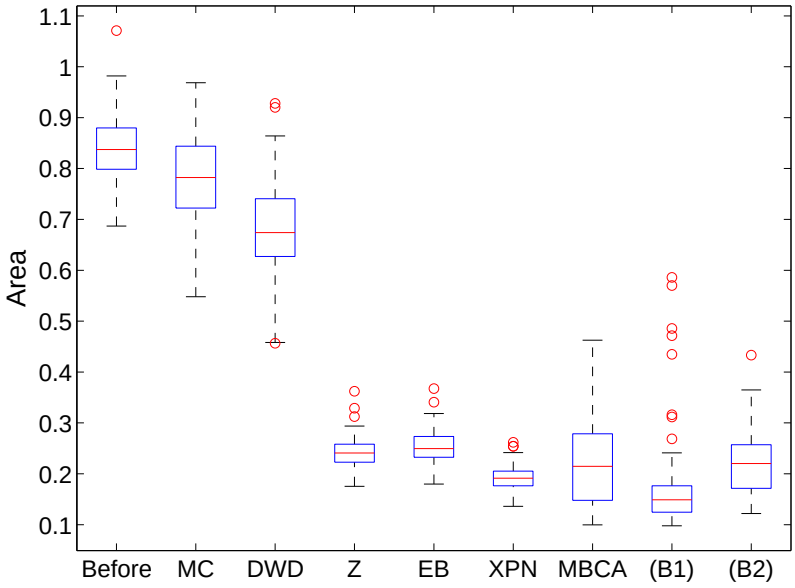
Figure 2.14. Estimated density curves of within and between batch pairwise distances in the simulated data. Coincidence of two density curves indicates similarity of the data sets. Kullback-Leibler (KL) divergence in the left-up corner measures the difference between two density curve.

Furthermore in order to compare the methods numerically, we calculate the absolute difference between the areas under each density curve. After successful batch adjustment, the difference of the areas will be small since the two density curves be more coincident. The results are shown in Figure 2.15-(a). All methods decrease the area between two density curves. Rather than MC and DWD, the other methods considerably reduce the the area. Similarly, we calculate Kullback-Leibler (KL) divergence, which also measures the difference between two density curve. The KL divergence is a nonsymmetric measure of the difference between two probability distribution. Let us say P is a true probability and Q is an approximate probability. The KL divergence of Q from P is defined by

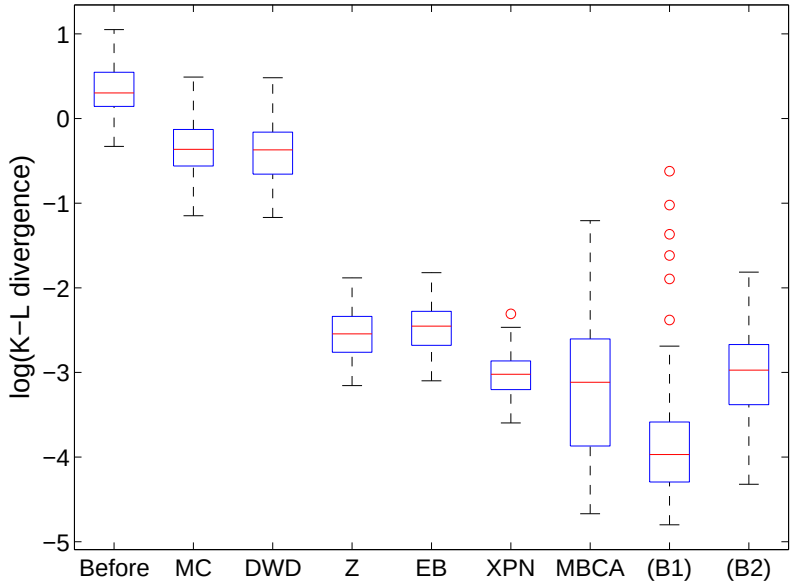
$$KL(P\|Q) = \int_{-\infty}^{\infty} \ln\left(\frac{p(x)}{q(x)}\right) p(x)\, dx.$$

We first set the distribution of within batch pair distance as P and the distribution of between batch pair distance as Q and change the role of P and Q, then two KL values are averaged. Figure 2.15-(b) shows box plots of the log divergence values from 100 repetitions.

The last three box plots, the original MBCA and the two targeted versions, are the lowest, indicating that the batches are merged well.



(a) Absolute difference between the areas under each density curve.



(b) Kullback-Leibler (KL) divergence.

Figure 2.15. Difference between two density curves in Figure 2.14. In (a), absolute difference between the areas under each density curve is computed for 100 repetitions. In (b), log Kullback-Leibler divergence is calculated. Smaller value indicates the similarity of two density curves, therefore indicating successful adjustment.

2.7 Real Data Analysis

2.7.1 Data

We use two cancer microarray gene expression data sets: Breast cancer data and Lung cancer data. Originally two data sets have 22,283 genes, and this number is down to 12,526 genes by averaging the replicates that are identified by the UniGene cluster. Then we select 4,364 genes that are assigned by Gene Ontology (GO); mapping of the GO annotations for genes can be found at http://go.princeton.edu/cgi-bin/GOTermMapper.

The breast cancer data set consists of three batches that were independently collected in various places in Europe at different times. All three batches were preprocessed by MAS5.0 for Affymetrix U133a GeneChip. The data are available at the http://www.ncbi.nlm.nih.gov/projects/geo/ with the GEO accession number : GSE2034, GSE4922 and GSE7390. As for the biological signal, we have a estrogen status (ER+, ER−) for each subject. A brief summary of the three batches is shown in Table 2.3.

Four batches in the lung cancer data set were collected at four different laboratories: Moffitt Cancer Center (HLM), University of Michigan Cancer Center (UM), the Dana-Farber Cancer Institute (CAN/DF), and Memorial Sloan-Kettering Cancer Center (MSK). These four institutes formed a consortium with the US National Cancer Institute to develop and validate gene expression signatures of lung adenocarcinomas. This experiment was designed to characterize the performance of several prognostic models across different subjects and different laboratories. A relevant study has been published by Shedden et al. (2008). The CEL files are available at http://caarraydb.nci.nih.gov/caarray/. The data sets for our analysis are preprocessed by the robust multi-array average (RMA) method. As for the biological signal, we use overall survival status (dead, alive) reported at last follow-up time. Table 2.3 displays a brief summary of the data.

Table 2.3
A brief summary of the two gene expression data sets.

Breast cancer data				Lung cancer data			
Batch	Sample size	ER−	ER+	Batch	Sample size	Live	Die
GSE2034	286	77	209	HLM	79	19	60
GSE4922	245	34	211	UM	178	76	102
GSE7379	198	64	134	CAN/DF	82	47	35
				MSK	104	65	39

2.7.2 Comparison of Methods

In this section we compare the proposed multi-batch covariance adjustment (MBCA) approach with several other existing methods including mean-centering (MC), distance weighted discrimination (DWD), standardization, the empirical Bayes (EB) and the cross-platform normalization (XPN) methods. The criteria that we use for comparison are homogeneity of covariance, cross-batch prediction performance, pairwise distances for between-batch samples and within-batch samples, correlation of t-statistics and a preservation score of important genes.

a. Homogeneity of Covariance

We measure the inter-batch homogeneity in terms of their covariance matrices. Note that most methods adjust for the means so the adjusted batch means are equal. We obtained p-values based on the Q_k^2 test statistic ($H_0 : \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_K$) discussed in Section 2.3.2. A higher p-value indicates greater similarity of covariance matrices among batches. Table 2.4 shows the results for both data sets. As expected from the proposition 1, the MC and DWD methods yield the same p-value before and after adjustment. Meanwhile, the EB, XPN and MBCA have higher p-values, which indicates these methods alter covariances and homogenize them among batches. The proposed MBCA method has the highest p-value. The results for the lung cancer data sets also suggest a similar conclusion. In particular, the sample standardization has the p-value (0.0002) that is smaller than the

before p-value (0.0456), which indicates that the gene-wise approach can even make the batches more dissimilar. We acknowledge that our method is designed to achieve the best covariance similarity as much as possible, so this criterion alone does not necessarily justify the superior performance. However, it is still useful to see how other methods measure up one another.

The p-values of the EB, XPN, and MBCA methods in Table 2.4 suggest that for both data sets the methods transform the batches in such a way that they have a common covariance. However, this does not necessarily imply that the adjusted data by the three methods indeed have the identical covariance. In Table 2.5, we have results from pairwise tests for the equality of covariance matrices produced by each method. We can see that the three methods homogenize the covariances differently; while the EB and XPN are similar each other, both are different from the MBCA.

Table 2.4

Comparison of test statistics for equality of covariance and corresponding p-values. Higher p-values indicate greater similarity of covariance matrices among the batches.

	Breast cancer data		Lung cancer data	
Method	Q_K^2	p-value	Q_K^2	p-value
Before	7.7697	0.0206	8.0182	0.0456
MC	7.7697	0.0206	8.0182	0.0456
DWD	7.7697	0.0206	8.0182	0.0456
Z-score	4.8767	0.0873	19.4446	0.0002
EB	0.7093	0.7014	3.3300	0.3435
XPN	0.2802	0.8693	1.8651	0.6009
MBCA	0.2421	0.8860	0.1585	0.9840

Table 2.5

Pairwise covariance-equality test for different methods.

	Breast cancer data		Lung cancer data	
Method	Q_K^2	p-value	Q_K^2	p-value
MBCA vs. EB	10.9304	0.0009	8.0243	0.0046
MBCA vs. XPN	14.6737	0.0001	2.3989	0.0004
EB vs.XPN	0.2734	0.6011	0.4285	0.5127

b. Target Batch Consideration

As discussed in Section 2.4.2, the proposed MBCA is able to transform the batches to resemble an "ideal" batch. For the data sets analyzed in this section, there is no known ideal batch to target at. Therefore we set each batch as the target batch and compare the covariance homogeneity between the target and the other batches ($H_0 : \mathbf{\Sigma}_i = \mathbf{\Sigma}_{(-i)}$, where i is the target batch and $\mathbf{\Sigma}_{(-i)}$ is the common covariance of all the other batches) in Table 2.6. It is not surprising that MBCA successfully makes the whole data mimic the target batch each time. From the table, on the other hand, it is observed that few hypotheses are rejected for EB and XPN. For example, for the breast cancer data, XPN yields batches 1 and 3 that are different from batch 2 (p-value = 0.0069), and for the lung cancer data, EB yields batches 1, 2, and 4 that are different from batch 3 (p-value = .0011). Furthermore, relatively low p-values of batch 3 in the lung cancer data may indicate that this batch is more difficult to be homogenized with others. This finding is also supported by Shedden et al. (2008), where arrays from batch 3 were found to cluster separately from the other batches, perhaps related to the observation that these arrays were also dimmer; this difference could reflect a departure from protocol or a technical issue with the batch.

c. Cross-batch Prediction

Since the batch bias usually interferes with the biological signal, a successful adjustment can improve separability of the biological classes, making a classifier produce better prediction performance. However, one should use caution when using the strengthened biological signal as a criterion for a successful batch effect adjustment. If the adjustment is overly done, it may force the data to be spuriously more separable than what the underlying population can allow. A method that uses the biological signal in the adjustment process, such as the EB method, can be prone to this problem.

In this dissertation, rather than the performance itself, we use cross-batch prediction, i.e., we see if one can build a classification rule based on one batch, that can be effectively

Table 2.6

Equal covariance test after adjusting data for a target batch.

	Breast cancer data		Lung cancer data	
Batch 1 vs. Others	Q_K^2	p-value	Q_K^2	p-value
MBCA(B1)	0.0553	0.8141	0.2697	0.6036
EB	0.8670	0.3518	0.6437	0.4224
XPN	2.4537	0.1172	0.1310	0.7174
Batch 2 vs. Others	Q_K^2	p-value	Q_K^2	p-value
MBCA(B2)	0.0611	0.8047	0.1239	0.7249
EB	5.5035	0.0190	1.2084	0.2717
XPN	7.2880	0.0069	3.1155	0.0776
Batch 3 vs. Others	Q_K^2	p-value	Q_K^2	p-value
MBCA(B3)	1.1364	0.2864	0.6530	0.4191
EB	1.4948	0.2215	6.3813	0.0115
XPN	0.5219	0.4700	10.6768	0.0011
Batch 4 vs. Others	-	-	Q_K^2	p-value
MBCA(B4)	-	-	0.2609	0.6081
EB	-	-	6.7489	0.0094
XPN	-	-	7.5614	0.0060

applied to other batches. If the adjustment is reasonable, we expect that prediction performance of a classifier would be similar from batch to batch. As for the classification method, we choose the regularized linear discriminant analysis Guo et al. (2005), because of its known superior performance for high dimensional data such as gene expression. Since each batch can have different proportions of biological signals, we use Matthews correlation coefficient (MCC), which is known to be useful for unbalanced sample sizes (Luo et al., 2010), to measure prediction performance.

For the breast cancer data, we use two batches as training data to determine the discrimination rule, with which the tuning parameter is chosen with five-fold cross-validation. Then we predict ER status in the left-out batch, and report the MCC. The results are shown in Table 2.7. It is evidenced that cross-batch prediction performance has been generally improved by all methods and the proposed method again shows competitive performance. For the lung cancer data, we train the classifier with three batches and test

Table 2.7

MCC from cross-batch prediction by ridge linear discriminant analysis. The numbers in the parentheses indicate batches used for training the classifier.

	Breast cancer data			Lung cancer data			
Method	$(23) \to 1$	$(13) \to 2$	$(12) \to 3$	$(234) \to 1$	$(134) \to 2$	$(124) \to 3$	$(123) \to 4$
Before	0.6267	0.5597	0.7191	0.0492	-0.0554	0.0416	0.1933
MC	0.6955	0.5794	0.7317	0.1268	0.1431	0.0812	0.2334
DWD	0.7133	0.6140	0.6941	0.0654	0.1421	0.1183	-0.0201
Z-score	0.6711	0.6125	0.7178	0.0367	0.1658	0.1656	0.1300
EB	0.6623	0.5159	0.7317	0.0479	0.1298	0.1183	0.0852
XPN	0.7080	0.5962	0.7209	0.1410	0.1502	0.0175	0.1654
MBCA	0.7106	0.5877	0.7317	0.1696	0.1334	0.0812	0.1442

it on the left-out batch. The results for the lung cancer data with overall survival (OS) as the biological signal are shown in the table as well. Note that the MCC values for the lung cancer data are much smaller than for the breast cancer data since the prediction of the overall survival is notoriously difficult (Luo et al., 2010) while ER status is relatively easy to predict.

Table 2.8 shows the MCC values when MBCA is applied with a target batch for the lung cancer data. The results suggest that some batches work better as a target batch. In particular, MBCA (B1), with batch 1 as the target, excels any results in Table 2.7, which is consonant with the fact that this site was a high volume facility experienced with microarrays.

d. Within and Between Batch Pairwise Distance

In order to measure the similarity of data sets after adjusting batches, we compare the distributions of the pairwise distances between arrays in the same batch with those in different batches. In each panel of Figure 2.16, 2.17, we overlay two estimated density curves of pairwise distances, which are from within and between batches, respectively. Coincidence of two density curves indicates similarity of the data sets. Furthermore, in order to compare the methods numerically, we also calculate Kullback-Leibler (KL) divergence

Table 2.8

MCC from cross-batch prediction for the lung cancer data with MBCA with a target batch.

	Lung cancer data			
Method	(234) → 1	(134) → 2	(124) → 3	(123) → 4
Before	0.0492	-0.0554	0.0416	0.1933
MBCA(B1)	0.1410	0.1693	0.1027	0.2636
MBCA(B2)	0.0654	0.1658	0.1027	0.1788
MBCA(B3)	0.0675	0.1817	0.0386	0.1689
MBCA(B4)	0.0533	0.1466	0.0745	0.1992

which measures the difference between two density curve, shown in the left-up corner of Figure 2.16, 2.17. The smaller value indicates the better coincidence of two density curves. For the breast cancer data in Figure 2.16, the standardization, EB, XPN and MBCA methods show similarly good performance while these methods are better than the MC and DWD. For the lung cancer data in Figure 2.17, there is no substantial difference among the compared methods.

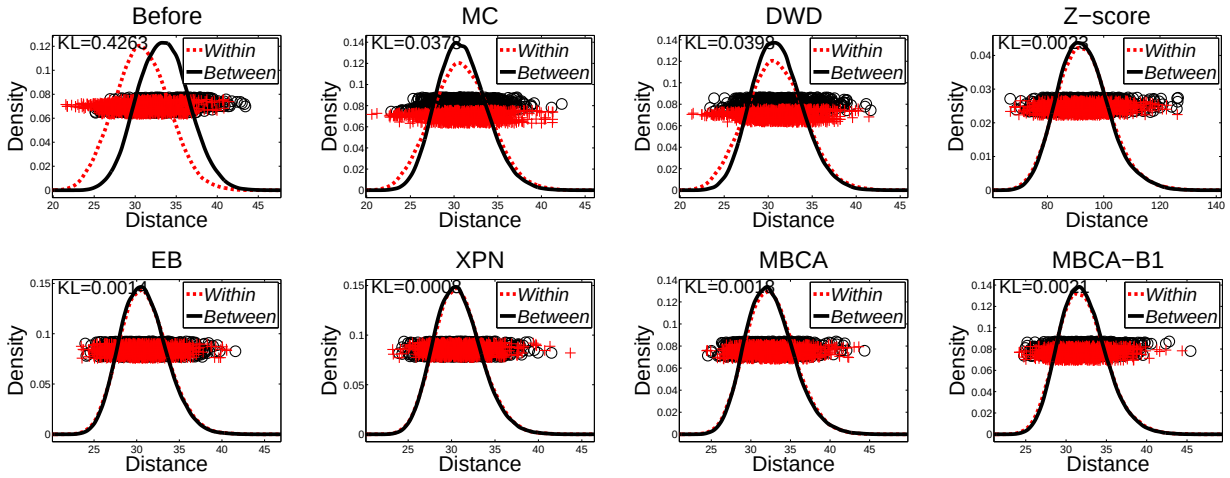


Figure 2.16. (Breast cancer data) Estimated density curves of within and between batch pairwise distances in the breast cancer data. Coincidence of two density curves indicates similarity of the data sets.

e. Correlation of t-statistic

Another evidence of batch bias is disagreement of t-statistics of biological class among

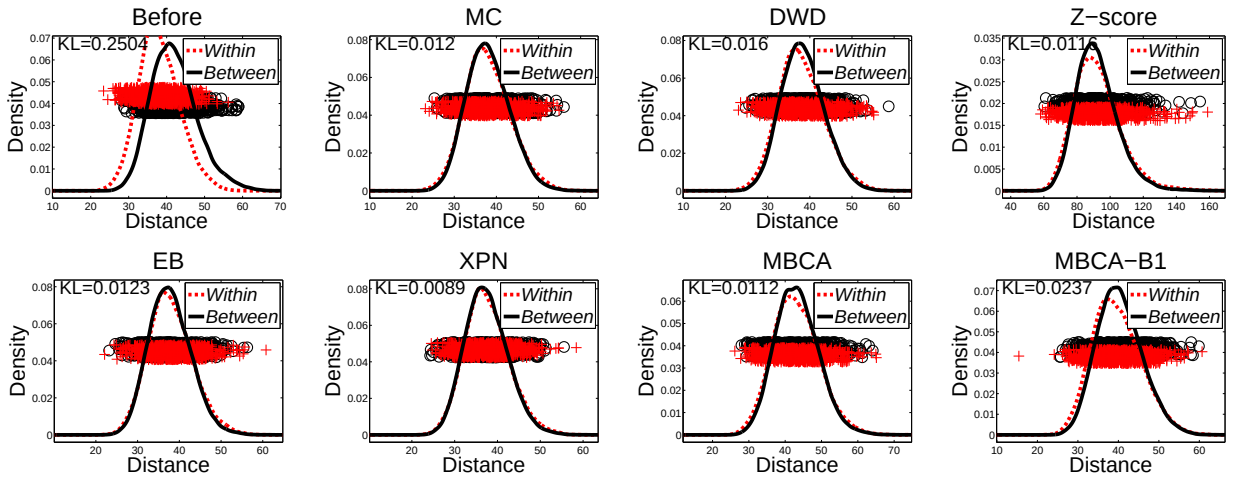


Figure 2.17. (Lung cancer data) Estimated density curves of within and between batch pairwise distances in the lung cancer data. Coincidence of two density curves indicates similarity of the data sets.

different batches. After removing batch bias, it is expected to see more concordant result in t-statistics across batches. In Shabalin et al. (2008), this concordance has been measured by Pearson correlation in a pair of t-statistics for ER status in the breast cancer data. We applies this measure for our two types of endpoints: ER status in the breast cancer data and OS in the lung cancer data. For three batches, three Pearson correlations are calculated from three pairs of t-statistics and averaged. The results are shown in Table 2.9. Higher value indicates better concordance of t-statistic after the batch adjustment. In both data sets, the greatest increase is achieved by the XPN. Our method MBCA also increases the correlation of t-statistics in the breast cancer data.

f. Preservation of Important Genes

As pointed out by Shabalin et al. (2008), excessive homogenization of batches can result in a loss of biological information. Thus it is important to see whether a given method keeps the biological signal in the individual batches after adjustment. In particular, Shabalin et al. (2008) use the sets of genes that are found to be important before adjustment and check whether those genes remain important. We borrow their approach in this study.

Table 2.9

Correlation of t-statistic. Higher value indicates better concordance of t-statistic across batches after the batch adjustment

	Breast cancer data	Lung cancer data
Biological class	ER status	Overall survival
Method	Correlations	Correlations
Before	0.8031	0.1412
MC	0.8031	0.1412
DWD	0.8031	0.1412
Z-score	0.8031	0.1412
EB	0.7885	0.1239
XPN	0.8525	0.1434
MBCA	0.8122	0.1381

Let L_i $(i = 1, \ldots, K)$ be the set of genes that have p-value less than .1 for the two-sample t-test in the ith batch before adjustment. Also let L^a be the gene list in the combined data set after adjustment. Ideally many genes in L_i should appear in L^a. We evaluate the preservation of significant genes by the following measures.

$$V_1 = |(L_1 \cap \cdots \cap L_K) \cap L^a| / |L_1 \cap \cdots \cap L_K|,$$

$$V_2 = |(L_1 \cup \cdots \cup L_K) \cap L^a| / |L_1 \cup \cdots \cup L_K|.$$

Higher value of these measures (closer to one) indicates better preservation of significant genes. The results are presented in Table 2.10, where we can see that XPN has the lowest V_2 for the breast cancer data and EB has the highest V_2 for the lung cancer data.

2.8 Discussion

In this dissertation we propose a novel multivariate batch adjustment method. The approach taken is one step advanced to the gene-wise approach in the sense that we directly estimate and adjust for correlations between genes. It is shown to be effective to obtain best homogeneity among batches through our data analysis. Furthermore, the proposed MBCA method

Table 2.10
Preservation of Significant genes.

	Breast cancer data		Lung cancer data	
Biological class	*ER* status		Overall survival	
Method	V_1	V_2	V_1	V_2
MC	0.9991	0.8185	0.3333	0.4310
DWD	0.9991	0.8260	0.3333	0.4367
Z-score	0.9991	0.8201	0.3333	0.4305
EB	0.9991	0.8289	0.3333	0.4672
XPN	0.9991	0.8066	0.3333	0.4202
MBCA	0.9983	0.8237	0.3333	0.4243

is greatly useful when there exists an ideal batch which is obtained in the best experimental conditions since the other batches can mimic the ideal one.

There are some practical issues with the proposed MBCA method. First is the choice of factors. One can use data-driven clustering results with genes, for example k-means. However, choosing k is another non-trivial problem. In this work, we use the gene ontology to avoid such argument and it is reasonable in the biological sense. Second is computational burden because the MBCA method involves calculation of high dimensional covariances. If this were a concern, one can use singular value decomposition as discussed in Hastie and Tibshirani (2004), which would reduce the computing time from $O(p^3)$ to $O(pn^2)$.

Chapter 3

Outlier Detection in High Dimension, Low Sample Size Data

3.1 Introduction

3.1.1 Outliers In HDLSS Data

Identifying outliers has been important in many statistical analyses since outlying observation may distort parameter estimation and mislead the experimental result. Several authors have defined outliers as observations far away from the main mass of data, or unlike observation from the baseline distribution (Barnett and Lewis, 1994; Becker and Gatheer, 1999; Hawkins, 1980). As the definition implies, classifying outliers from the rest of data is somewhat relative; in other words, it depends on various factors such as distribution assumption, distance measure, covariance structure, and existence of group outliers.

Our study focuses on identifying outliers in high dimension, low sample size (HDLSS) data. In spite of the enormous popularity of high dimensional data analysis nowadays, the outlier detection problem has been hardly addressed. Classical outlier detection methods that are based on estimated mean and covariance can be useless when the dimension is relatively large compared to the sample size. Therefore, developing a more suitable outlier detection method that can handle the growth of dimensionality in many real data is imperative.

Outliers are commonly viewed as unusual observations that aberrantly behave relative to other observations. As dimension increases, it is not clear whether the growth of dimensionality hides the unusual observations, or reversely, falsely claims outliers that belong to the ordinary pattern of high-dimensional samples. The important principle for detecting outliers is how to define the ordinary pattern of a group of observations, which is hard for HDLSS data since the distributions of data are difficult to infer with a small sample size.

Another important aspect in outlier detection for HDLSS data is how to measure the "outlyingness," specifically, the distance between a potential outlier and the rest of the data points. In this dissertation we investigate centroid distance, maximal data piling distance (Ahn and Marron, 2010) and ridge Mahalanobis distance. Before introducing the proposed outlier method for HDLSS data, in the following subsection we briefly introduce existing approaches for traditional low-dimensional multivariate data.

3.1.2 Existing Approaches For Low-Dimensional Data

There are two main approaches to the outlier detection problem. First is a diagnostic approach often used in regression. In this context, outliers refer to regression outliers, which are observations deviating from the linear relationship between explanatory variables (X-space) and the response variable (Y-space). For example, a residual-based method considers the extremeness in Y-space, whereas a hat matrix is used to detect a high leverage point in X-space. Let us consider the following regression model for a given data set $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^n$:

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta} + e_i, \quad i = 1, \ldots, n \ ,$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$. The most popular estimator for the regression coefficients $\boldsymbol{\theta}$ is least squares (LS) estimator that solves the following minimization problem:

$$\min_{\hat{\boldsymbol{\theta}}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}})^2.$$

Unfortunately, in LS estimation, even a single outlier can easily inflate or deflate parameter estimation $\hat{\boldsymbol{\theta}}$. The mathematical formula of this concept is introduced as the breakdown point in Donoho and Huber (1983) and Rousseeuw (1987), which is defined by the proportion of contaminants that leads the estimates to arbitrarily large aberrant values. For the LS estimator, the breakdown point is $1/n$, which is called 0 % breakdown point because of the limit value as $n \to \infty$, indicating that the estimation is very sensitive to outliers. Therefore, many estimators have been developed for robust regression such as L_1 regression,

M-estimator in Huber (1981), Least Median of Squares (LMS) and Least Trimmed Squares (LTS) in Rousseeuw (1984), S-estimator in Rousseeuw and Yohar (1984) and many others. Among these, LMS and LTS regression that have a 50% breakdown point have gained popularity due to its robustness to outliers in both X- and Y-space, and these estimates inspired the proposal of weighted least squares (WLS), which has been popularly used until now. Further, these model-diagnostic ideas for outliers can be extended to other linear models such as designed experiments, non-linear regression, and time series.

The second approach for the outlier detection problem is to see the "extremeness" of an observation relative to the data mass, which requires one to estimate the location and the scale of the data mass. In the view of regression, this approach focuses on only X-space regardless of the response variable. As a regression model includes more explanatory variables, there is a higher possibility that outliers appear in X-space, and those outliers are often hard to recognize in the multivariate space. Certain outliers are not identifiable in individual dimensions as shown with a toy example in Figure 3.1.

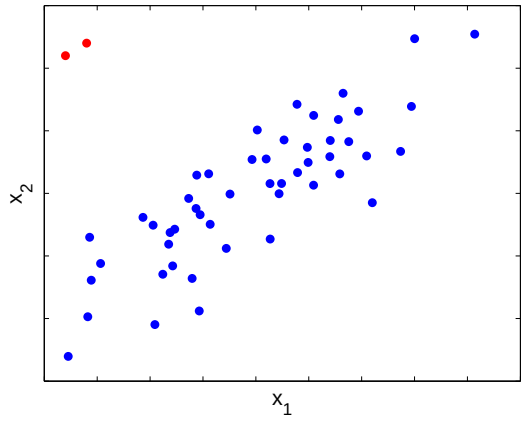


Figure 3.1. An example of multivariate outliers. Two data points in the up left corner are separated from the other observations, but those outliers are not detectable in individual dimensions, $\mathbf{x}_1$ or $\mathbf{x}_2$.

It is commonly preferred to use both location and shape parameters in order to estimate the distance of a data point to data mass in the multivariate space. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be multivariate data vectors in $\mathbb{R}^p$. Also let d_i^2 denote the squared distance of the ith data point to

the center of data considering dispersion of data clouds. Then,

$$d_i^2 = d^2(\mathbf{x}_i; \mathbf{t}, \mathbf{V}) = (\mathbf{x}_i - \mathbf{t})^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{x}_i - \mathbf{t}), \quad i = 1, \ldots, n \ , \tag{3.1}$$

where $\mathbf{t}$ is the location parameter that represents the center of data and $\mathbf{V}$ is the shape parameter that represents the dispersion of data. Intrinsic estimators for $\mathbf{t}$ and $\mathbf{V}$ are sample mean vector and sample covariance matrix; in that case, d_i^2 is called the squared Mahalanobis distance. Unfortunately, the Mahalanobis distance is known to suffer from masking or swamping phenomena as shown in Barnett and Lewis (1994) and many other literatures. The masking effect is caused by multiple outliers, one of which hides the other. Conversely the swamping effect occurs when the outlier, say $\mathbf{x}_{(n)}$, carries the less outlying observation $\mathbf{x}_{(n-1)}$ with it even though it may not be an outlier, resulting in a false judgement in regard to $\mathbf{x}_{(n-1)}$. For this reason, it seems natural to replace $\mathbf{t}$ and $\mathbf{V}$ with robust estimators such as median and median absolute deviation so that the distance in (3.1) is not vulnerable to masking or swamping, and thus rightly declares outliers. Some sophisticated methods for the estimation of robust location and scatter parameters have been studied regarding high breakdown point. (Davies, 1992, 1987; Lopuhaä, 1989; Maronna et al., 2006; Rousseeuw, 1985, 1987; Rousseeuw and van Zomeren, 1990)

However, when dimension becomes higher, even at the range of tens, the robust estimators for $\mathbf{t}$ and $\mathbf{V}$ that have the high breakdown points become computationally intensive and often infeasible in practice . Thus, some researchers proposed the fast algorithm for the robust estimators along with the outlier detection (Maronna and Zamar, 2002; Maronna and Youhai, 1995; Peña and Prieto, 2001; Rousseeuw and van Driessen, 1999). The main rule of those methods is selecting useful dimensions, such as principal component directions, that effectively estimate the location and scatter parameters or that explicitly represent the outlier. For example, Peña and Prieto (2001) suggested to search for up to $2p$ directions, by maximizing or minimizing the kurtosis coefficient of the projections. In another example, Filzmoser et al. (2008) proposed the PCout for outlier detection. They found suitable prin-

cipal component directions of data on which the outliers are readily apparent and thus one downweights those directions to obtain a robust estimator.

At times, researchers focus on the outliers themselves for the purpose of deleting them rather than developing a robust method for outliers. This approach requires knowledge of the distribution of the data since we decide a cut-off value, rejection point for outliers, based on that distribution. There are many proposals for discordancy test of multivariate samples from a multivariate distribution in (Barnett and Lewis, 1994). A conventional method is a chi-square quantile-quantile plot. This method is using the fact that the squared Mahalanobis distance calculated from Gaussian data follows a chi-square distribution. On the other side, Hardin and Rocke (2005) proposed the scaled F-distribution substituting chi-square distribution for some robust estimation of $\mathbf{t}$ and $\mathbf{V}$ in (3.1). In the next subsection, we describe the details of the chi-square quantile-quantile plot.

3.1.3 The Chi-square Quantile-Quantile Plot

The chi-square quantile-quantile plot (qq plot) is originally designed to check the multivariate normality assumption of a data set. The principle of the plot lies in the linearity between the χ^2_p quantiles and the sample quantiles. Let us denote a sample vector $\mathbf{x}_j$, $j = 1, \ldots, n$, which comes from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The squared Mahalanobis distance of each data point, $\mathbf{x}_j \in \mathbb{R}^p$, is defined by

$$D_j^2 = (\mathbf{x}_j - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu}),\ j = 1, \ldots, n\ . \tag{3.2}$$

It is well known that (3.2) follows χ^2_p distribution. Since $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are usually unknown, we would estimate (3.2) from the data as

$$\hat{D}_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}),\ j = 1, \ldots, n\ , \tag{3.3}$$

where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean vector and the sample covariance matrix, respectively. Note that, as $n \to \infty$, the squared Mahalanobis distance from data in (3.3) will behave as if it is a chi-square random variable. This relationship can be seen as a straight line having

slope 1 over the plot with respect to the pairs $\left(\chi^2_p\big((j-0.5)/n\big),\ \hat{D}^2_{(j)}\right)$, $j = 1, \ldots, n$, where $\chi^2_p(\cdot)$ is the lower quantiles of chi-square distribution and $\hat{D}^2_{(j)}$ is the jth order statistics of $\hat{D}^2_1, \ldots, \hat{D}^2_n$. In the outlier context, the observation deviating from the straight line is more likely an outlier.

However, the chi-square qq plot has suffered from some drawbacks. Only large samples consistently produce statistically significant results. Another problem can stem from the Mahalanobis distance's masking effect. Even though there are many efforts to overcome such drawbacks by, for example, finding robust estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the Mahalanobis distance including its robust versions are hardly useful when $p > n$.

3.1.4 New Approach With Parametric Bootstrap

In this dissertation, we propose a novel outlier detection method using a parametric bootstrap, which can be interpreted as an HDLSS version of qq plot. For the outlier detection problem for HDLSS data, we employ different distance measures instead of the Mahalanobis distance, which are introduced in Section 3.2. Some high dimensional asymptotic properties of these distances are studied in Section 3.3. A problem in the usage of the newly defined distances is that the distribution of the distances is unknown and difficult to infer due to the small sample size. Our solution is to utilize a bootstrap method, which is known to be useful for estimating the sampling distribution when the theoretical distribution is unknown and the sample size is insufficient. Specifically, we use the parametric bootstrap based on the mean and covariance of Gaussian distribution. The details of the parametric bootstrap method for the outlier detection are discussed in Section 3.4.

3.2 Distance for Outlier detection

In this section, we introduce three distance measures for outlier detection in HDLSS data: namely, centroid distance, maximal data piling distance proposed by Ahn and Marron (2010), and ridge Mahalanobis distance. We will use these distances to measure the oulyingness of

a data point relative to others. This approach is often called the leave-one-out procedure. In the following subsection, we investigate the property of each distance measure as well as the relationships among the distances.

3.2.1 Centroid distance

Suppose there is an HDLSS data set denoted by $\mathbf{X}_{(p\times n)} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ with $p > n$. First, we consider the distance between an observation and the center of the rest. The centroid distance (CD) can be calculated by the Euclidean distance between the jth data point in $\mathbb{R}^p$ and the mean (centroid) of the other $n-1$ data points for all n data points; i.e.,

$$D_c(j) \;=\; \sqrt{(\mathbf{x}_j - \bar{\mathbf{x}}_{(-j)})^{\mathrm{T}}(\mathbf{x}_j - \bar{\mathbf{x}}_{(-j)})},\ j = 1, \ldots, n\ , \tag{3.4}$$

where $\bar{\mathbf{x}}_{(-j)}$ is the sample mean vector without the jth observation.

Stating outlier as an unusual observation from the ordinary pattern of the others often implies that the observation might be from a different distribution. Suppose one data point $\mathbf{x}_1$ came from a different distribution than the other $n-1$ data points. This fact also determines the distribution of the distance between two data points: $\mathbf{x}_1$ vs $\bar{\mathbf{x}}_{(-1)}$. For example, assume that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently generated p-variate Gaussian data. Note that D_c^2 denotes the squared centroid distance between a data point $\mathbf{x}_1$ and the mean of the others; i.e., $D_c^2 = (\mathbf{x}_1 - \bar{\mathbf{x}}_{(-1)})^{\mathrm{T}}(\mathbf{x}_1 - \bar{\mathbf{x}}_{(-1)})$. The following proposition states that D_c^2 is a noncentral chi-squared distribution.

Proposition 2. *Suppose there are n independent sample vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. While a data vector $\mathbf{x}_1$ is from $\mathcal{N}_p(\boldsymbol{\theta}, \tau^2\mathbf{I}_p)$, the other data vectors $\mathbf{x}_2, \ldots, \mathbf{x}_n$ are from $\mathcal{N}_p(\boldsymbol{\mu}, \sigma^2\mathbf{I}_p)$. Then, D_c^2 follows a noncentral chi-squared distribution; $\big(\tau^2 + \sigma^2/(n-1)\big)\chi_p^2(\lambda)$, where $\lambda = \big(\tau^2 + \sigma^2/(n-1)\big)^{-1}\|\boldsymbol{\theta} - \boldsymbol{\mu}\|^2$.*

Proof. Since $\mathbf{x}_2, \ldots, \mathbf{x}_n$ are independent, the mean vector denoted by $\bar{\mathbf{x}}_{(-1)} = \sum_{j=2}^n \mathbf{x}_j/(n-1)$ is $\mathcal{N}_p\big(\boldsymbol{\mu}, \sigma^2/(n-1)\mathbf{I}_p\big)$. Also, $\mathbf{x}_1$ and $\bar{\mathbf{x}}_{(-1)}$ are independent, and

$$\big(\tau^2 + \sigma^2/(n-1)\big)^{-1/2}(\mathbf{x}_1 - \bar{\mathbf{x}}_{(-1)}) \sim \mathcal{N}_p\big(\boldsymbol{\theta} - \boldsymbol{\mu}, \mathbf{I}_p\big). \tag{3.5}$$

The squared form of (3.5) can be seen as a noncentral chi-squared distribution with the degree of freedom p and the noncentrality parameter $\lambda = \big(\tau^2 + \sigma^2/(n-1)\big)^{-1}\|\boldsymbol{\theta} - \boldsymbol{\mu}\|^2$. $\square$

Remark 1 states a special case of Proposition 2 with $\boldsymbol{\theta} = \boldsymbol{\mu}$ and $\tau^2 = \sigma^2$, which represents a situation where all data vectors are from the same normal distribution with spherical covariance.

Remark 1. *Suppose there are n i.i.d. sample vectors from $\mathcal{N}_p(\mathbf{0}, \sigma^2\mathbf{I}_p)$. Then, D_c^2 between an observation and the others follows $\sigma^2\big(1 + 1/(n-1)\big)\chi_p^2$.*

3.2.2 Maximal Data Piling Distance

In the previous section, the CD addresses the proximity between two centers of data, where one of two groups has a single observation. The maximal data piling (MDP) distance focuses on the proximity between two affine subspaces. More precisely, the MDP distance is the orthogonal distance between two affine subspaces that each of two data groups generates. In the outlier context, it is the orthogonal distance from an observation to the affine subspace generated by the other observations as seen in Figure 3.2.

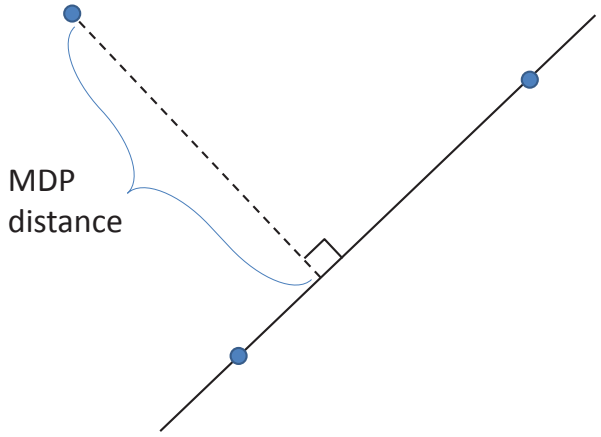


Figure 3.2. Illustration of MDP distance. MDP distance is the orthogonal distance from an observation to the affine subspace generated by the other observations.

It is suggested in Ahn et al.(2011) that the MDP distance is an appropriate distance measure for high dimensional clustering problem. Originally the MDP direction vector is

proposed by Ahn and Marron (2010) for a discrimination problem for HDLSS data. Let us assume a binary discrimination situation. Ahn and Marron (2010) showed that there exists a unique direction vector that maximizes the distance between projections of each class as well as the amount of data piling within class. Data piling is a phenomenon that projection values of data vectors are piled at a point. In fact, the data piling onto this direction are two distinct values, one for each class. This optimization problem resembles the Fisher's linear discriminant (FLD) in low-dimensional setting in the sense that FLD also maximizes the distance of two class means and minimizes the within-class variance. But note the MDP direction vector is defined only for HDLSS data since in low-dimensional data there does not exist a direction vector that yields the complete data piling. The MDP direction vector has been found to work well for HDLSS data classification problem especially when variables are correlated (Ahn and Marron, 2010).

In what follows, we introduce the mathematics of the MDP distance along with the MDP direction vector. Regarding the outlier problem, we assume a special case of binary classification. Suppose there are two classes where the first class contains a single observation $\mathbf{x}_1$ and the second class contains $n-1$ observations $\mathbf{x}_2, \ldots, \mathbf{x}_n$. Then the mean vector of the second class is $\bar{\mathbf{x}}_{(-1)}$. Let $\mathbf{w} = \mathbf{x}_1 - \bar{\mathbf{x}}_{(-1)}$ denote the mean difference vector of two classes. Also let $\mathbf{C}$ denote the centered data matrix by subtracting the mean vector in the second class, i.e., $\mathbf{C} = [\mathbf{x}_2 - \bar{\mathbf{x}}_{(-1)}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}_{(-1)}]$. Further, let $\mathbf{P} = \mathbf{C}\mathbf{C}^\dagger$ be the projection matrix to the column space of $\mathbf{C}$, where $\mathbf{A}^\dagger$ is the Moore-Penrose generalized inverse of a matrix $\mathbf{A}$. Ahn and Marron (2010) showed that the maximal data piling vector $\mathbf{v}_{\text{MDP}}$ is obtained from the following optimization problem:

$$\text{finding } \mathbf{v} \text{ that maximizes } |\mathbf{v}^{\mathrm{T}}\mathbf{w}| \text{ subject to } \mathbf{C}^{\mathrm{T}}\mathbf{v} = \mathbf{0} \text{ and } \|\mathbf{v}\| = 1. \tag{3.6}$$

Here $\mathbf{C}^{\mathrm{T}}\mathbf{v} = \mathbf{0}$ is called the data piling constraint, also rewritten as $\mathbf{x}_i^{\mathrm{T}}\mathbf{v} = \bar{\mathbf{x}}_{(-1)}^{\mathrm{T}}\mathbf{v}$ for $i = 2, \ldots, n$, which implies that the projection of every data point in the second class onto $\mathbf{v}$ is the same as its class mean. Note that the first class is not included in this constraint since it has already a single data point. $\|\mathbf{v}\| = 1$ is a normalization constraint so that the vector

has unit length. The solution of (3.6) can be found by projecting $\mathbf{w}$ onto the orthogonal complement of the column space of $\mathbf{C}$,

$$\mathbf{v}_{\text{MDP}} \propto (\mathbf{I}_p - \mathbf{P})\mathbf{w}. \tag{3.7}$$

Then, the MDP distance is defined as the distance between two class projections onto $\mathbf{v}_{\text{MDP}}$, i.e.,

$$D_{\text{MDP}} = |(\mathbf{x}_1 - \bar{\mathbf{x}}_{(-1)})^{\text{T}}\mathbf{v}_{\text{MDP}}|. \tag{3.8}$$

In the following, properties of the MDP distance are studied under the spherical Gaussian data. Let D^2_{MDP} denote the squared maximal data piling distance. The following proposition states that D^2_{MDP} has a chi-squared distribution, which can be seen as a special case of Theorem 3 in Ahn et al. (2011).

Proposition 3. *Suppose there are n i.i.d. sample vectors from $\mathcal{N}_p(\mathbf{0}, \sigma^2\mathbf{I}_p)$. Then, D^2_{MDP} between an observation and the others follows $\sigma^2\big(1 + 1/(n-1)\big)\chi^2_{p-n+2}$.*

Back to our outlier detection problem, in order to identify outliers in a HDLSS data set $\mathbf{X}_{(p\times n)} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, we compute the MDP distance for all n data points as below

$$\begin{aligned} D_{\text{MDP}}(j) &= |(\mathbf{x}_j - \bar{\mathbf{x}}_{(-j)})^{\text{T}}\mathbf{v}_{\text{MDP}}|,\ j = 1, \ldots, n\ , & (3.9)\\ &= 2/\|\mathbf{Z}^{\text{T}\dagger}\ell(j)\|. & (3.10) \end{aligned}$$

For calculation purpose, we can use the equation (3.10), where $\mathbf{Z}$ is the centered data matrix obtained by subtracting the overall mean, i.e., $\mathbf{Z} = [\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}]$, and $\ell(j) = (1, \ldots, 1, -1, 1, \ldots, 1)^{\text{T}}$ is a label vector whose jth element is -1. Note that while centroid distance exists regardless of dimension and sample size, MDP distance exists only for HDLSS data, specifically, $p > n - 2$ ($df = p - n + 2$).

Here we would like to point out that we are not able to directly apply the chi-squared distribution for outlier detection. The first reason is that Propositions 2 and 3 are shown under the spherical case. However, this assumption is usually not true in real data setting since many variables are correlated. The second reason is that the leave-one-out distance

within a sample pool as in (3.4) and (3.9) does not fully meet the conditions in Propositions 2 and 3 in that $\mathbf{x}_j - \bar{\mathbf{x}}_{(-j)}$ and $\mathbf{x}_{j'} - \bar{\mathbf{x}}_{(-j')}$ are dependent each other and so on. Therefore, we rather generate the reference distribution using parametric bootstrap. This method will be introduced in Section 3.4.

3.2.3 Ridge Mahalanobis Distance

In this section, we introduce the ridge Mahalanobis (rMH) distance for HDLSS data, which measures the proximity from an observation to the others considering dispersion of data clouds. The squared Mahalanobis distance (3.3) in Section 3.1.3 is not applicable when $p > n$ due to the singularity of the sample covariance matrix $\mathbf{S}$. Thus, we define a modified version for HDLSS data. A popular solution for the singularity of $\mathbf{S}$ is utilizing ridge-type correction using α ($\alpha > 0$). Then, rMH distance of each data point to the rest for n data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is defined as

$$D_{\text{rMH}}(j) \;=\; \sqrt{(\mathbf{x}_j - \bar{\mathbf{x}}_{(-j)})^{\mathrm{T}}(\mathbf{S}_{(-j)} + \alpha\mathbf{I}_p)^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}_{(-j)})},\ j = 1, \ldots, n\ , \tag{3.11}$$

where $\bar{\mathbf{x}}_{(-j)}$ and $\mathbf{S}_{(-j)}$ are the sample mean vector and the sample covariance matrix without the jth observation, respectively, and α is the ridge parameter ($\alpha > 0$). The rMH distance shares some properties with both centroid distance and MDP distance depending on the value of α as shown in Propositon 4.

Proposition 4. *As α increases, the rMH distance becomes equivalent (up to a constant multiple) to the centroid distance. As α decreases, the rMH distance becomes equivalent (up to a constant multiple) to the MDP distance.*

Proof. First of all, as α increases, we can show that the rMH distance becomes proportional to the centroid distance. Let $\mathbf{w} = \mathbf{x}_1 - \bar{\mathbf{x}}_{(-1)}$. The squared rMH distance between $\mathbf{x}_1$ and the others is written as $D^2_{\text{rMH}} = \mathbf{w}^{\mathrm{T}}(\mathbf{S}_{(-1)} + \alpha\mathbf{I}_p)^{-1}\mathbf{w}$. Also the squared centroid distance can be rewritten as $D_c^2 = \mathbf{w}^{\mathrm{T}}\mathbf{w}$. For large value of α, we can see that the impact of correlation coefficient is weak, therefore $(\mathbf{S}_{(-1)} + \alpha\mathbf{I}_p)$ approximately becomes proportional to identity.

The inverse works similarly as shown below.

$$\begin{aligned}(\mathbf{S}_{(-1)}+\alpha\mathbf{I}_p)^{-1} &\propto \alpha(\mathbf{S}_{(-1)}+\alpha\mathbf{I}_p)^{-1} \quad \text{for } \alpha>0\\ &= (1/\alpha\mathbf{S}_{(-1)}+\mathbf{I}_p)^{-1}\\ &\to \mathbf{I}_p \quad as\ \alpha\to\infty.\end{aligned}$$

Therefore, $D^2_{\text{rMH}} \to c_1 D^2_c$ as $\alpha\to\infty$, where c_1 is a constant.

Second of all, as α decreases, we can show that the rMH distance becomes proportional to the MDP distance. Let $\mathbf{v}_\alpha = (\mathbf{S}_{(-1)}+\alpha\mathbf{I}_p)^{-1}\mathbf{w}$. The squared rMH distance can also be written as

$$D^2_{\text{rMH}} = \mathbf{w}^{\text{T}}\mathbf{v}_\alpha\ .$$

Similarly, the squared MDP distance is

$$\begin{aligned}D^2_{\text{MDP}} &= (\mathbf{w}^{\text{T}}\mathbf{v}_{\text{MDP}})^2 = \left(\frac{\mathbf{w}(\mathbf{I}_p-\mathbf{P})\mathbf{w}}{\|(\mathbf{I}_p-\mathbf{P})\mathbf{w}\|}\right)^2 = \mathbf{w}^{\text{T}}(\mathbf{I}_p-\mathbf{P})\mathbf{w}\\ &\propto \mathbf{w}^{\text{T}}\mathbf{v}_{\text{MDP}}\ .\end{aligned}$$

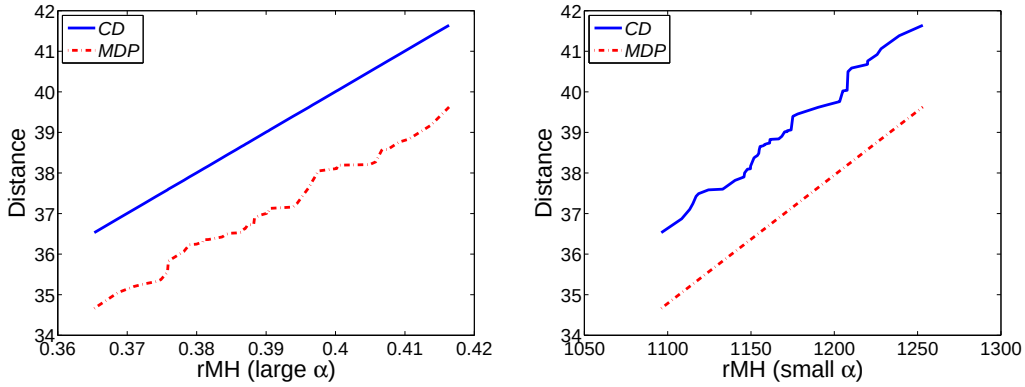
Thus the asymptotic equivalence of D^2_{rMH} and D^2_{MDP} can be shown through that of $\mathbf{v}_\alpha$ and $\mathbf{v}_{\text{MDP}}$ as $\alpha\to 0$.

$$\begin{aligned}\mathbf{v}_\alpha &= (\mathbf{S}_{(-1)}+\alpha\mathbf{I}_p)^{-1}\mathbf{w}\\ &\propto \alpha(\mathbf{C}\mathbf{C}^{\text{T}}+\alpha\mathbf{I}_p)^{-1}\mathbf{w}\\ &= \left[\mathbf{I}-\mathbf{C}\mathbf{C}^{\text{T}}(\mathbf{C}\mathbf{C}^{\text{T}}+\alpha\mathbf{I})^{-1}\right]\mathbf{w} \qquad (3.12)\\ &\to (\mathbf{I}-\mathbf{C}\mathbf{C}^{\dagger})\mathbf{w} \quad \text{as } \alpha\to 0. \qquad (3.13)\end{aligned}$$

The step from (3.12) to (3.13) is due to the fact that $\lim_{\alpha\to 0} = \mathbf{C}^{\text{T}}(\mathbf{C}\mathbf{C}^{\text{T}}+\alpha\mathbf{I})^{-1} = \mathbf{C}^{\dagger}$. Note that this limit exists even if $(\mathbf{C}\mathbf{C}^{\text{T}})^{-1}$ does not exist (Golub and Van Loan, 1996). Remind that $\mathbf{P} = \mathbf{C}\mathbf{C}^{\dagger}$. Therefore, $\mathbf{v}_\alpha \to c_2\mathbf{v}_{\text{MDP}}$ as $\alpha\to 0$, where c_2 is a constant. □

In addition to Proposition 4, we provide a numerical example to see the equivalence of rMH to either centroid distance or MDP distance. Let us generate 50 random vectors from

$\mathcal{N}_{500}(0,\, 3\mathbf{I}_p)$ and calculate CD, rMH and MDP distances. For the ridge parameter of rMH distance, two cases are computed: $\alpha = 10^4$, $\alpha = 10^{-3}$. In Figure 3.3, the rMH distance with the change of α is compared to the other two distances. In each panel, X-axis displays the order statistics of rMH distance, and Y-axis displays two other distances of the same observation. In (a), rMH with large α ($\alpha = 10^4$) is linear to CD. In (b), rMH with small α ($\alpha = 10^{-3}$) is linear MDP.



(a) rMH vs. CD (b) rMH vs. MDP

Figure 3.3. The ridge Mahalanobis (rMH) distance with α. In (a), rMH with a large α (10^4) is linear to the centroid distance. In (b), rMH with a small α (10^{-3}) is linear to the maximal data piling distance.

Regarding computational burden, rMH distance in (3.11) can be replaced with the computationally efficient version utilizing singular value decomposition. The singular value decomposition of $\mathbf{X}_{(p\times n)}$ is a factorization to be $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$. Let us denote $\mathbf{Y} = \mathbf{D}_0\mathbf{V}^{\mathrm{T}}$, where $\mathbf{D}_0$ is a $n \times n$ diagonal matrix keeping only non-zero diagonal elements of $\mathbf{D}$. We obtain the same result by replacing $\mathbf{X}_{(p\times n)}$ with $\mathbf{Y}_{(n\times n)} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$. Thus, the distance in (3.11) is equivalent to

$$D_{\mathrm{rMH}}(j) \;=\; \sqrt{(\mathbf{y}_j - \bar{\mathbf{y}}_{(-j)})^{\mathrm{T}}(\mathbf{S}_{y(-j)} + \alpha\mathbf{I}_n)^{-1}(\mathbf{y}_j - \bar{\mathbf{y}}_{(-j)})}\;.$$

3.3 HDLSS Asymptotics of Outlier Detection

In this section, we discuss some HDLSS asymptotic results of the outlier detection for MDP and CD methods. Theorems in the following subsections have been proved by Ahn et al. (2013).

3.3.1 The Maximal Data Piling Distance

In this section we study asymptotic properties of the MDP distance in the context of outlier by utilizing the asymptotic geometric representation of HDLSS data by Hall et al. (2005) and Ahn et al. (2007). The two papers established the representation under different distributional settings and later Jung and Marron (2009) did it in a unified framework. In this section we use the assumptions in Hall et al. (2005) since it is easier to discuss the geometry in their setting. Let $\mathbf{X}_{p\times n} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ be non-outliers data matrix with $p > n$ where $\mathbf{x}_j = (X_{1j}, \ldots, X_{pj})^{\mathrm{T}}$ are i.i.d. p-variate random vectors from a population. Suppose that we add $n_0 (< n)$ outliers to the data set and denote outlier vectors as $\mathbf{X}^0_{p\times n_0} = [\mathbf{x}^0_1, \ldots, \mathbf{x}^0_{n_0}]$, where $\mathbf{x}^0_j = (X^0_{1j}, \ldots, X^0_{pj})^{\mathrm{T}}$ are p-variate random vectors. Let N be the total number of sample size, $N = n + n_0$. It is not so realistic to assume that all $n_0 (> 1)$ outliers are coming from an identical distribution (which is different from the distribution of X). Yet, for the theoretical development of the method, we will view data with multiple outliers as a mixture model where majority of data vectors are from one population and relatively small fraction of data vectors come from another (but a common) population.

Assume that the population structure of the data satisfies the following conditions specified in Hall et al. (2005) for HDLSS asymptotics. (a) The fourth moments of the entries of the data vectors are uniformly bounded; (b) $\sum_{j=1}^{p} \mathrm{var}(X_j)/p$ and $\sum_{j=1}^{p} \mathrm{var}(X^0_j)/p$ converge to positive constants σ^2 and τ^2, respectively; (c) $\sum_{j=1}^{p} \{E(X_j) - \mathrm{E}(X^0_j)\}^2/p$ converges to nonnegative μ^2;(d) There exists a permutation of the entries of the data vectors such that the sequence of the variables are ρ-mixing for functions that are dominated by quadratics.

Under these conditions, as p tends to infinity, the data vectors approximately form an N-polyhedron while $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ forms a regular n-simplex with n and $\{\mathbf{x}_{n+1}^0, \ldots, \mathbf{x}_N^0\}$ with n_0 vertices, denoted by $\mathcal{X}$ and $\mathcal{X}_0$ respectively. The length of an edge connecting data vectors in $\mathcal{X}$ (or $\mathcal{X}_0$) is approximately $\sqrt{2}\sigma$ (or $\sqrt{2}\tau$) after scaled by $\sqrt{p}$. The length of an edge between a point in $\mathcal{X}$ and $\mathcal{X}_0$ is $\sqrt{\sigma^2 + \tau^2 + \mu^2}$ after scaled by $\sqrt{p}$.

First, we start with an easy case where there is only one outlier. The following proposition states a condition where the proposed detection method can identify a single outlier with probability one.

Proposition 5. *Assume that the above assumptions (a) - (d) are satisfied and there is a single outlier, i.e., $n_0 = 1$. Further, assume that $\mu^2 + \tau^2 > \sigma^2$. Then, the leave-one-out MDP distance for the outlier is bigger than the distances for non-outliers with probability 1.*

Note that in the large p limit, the condition $\mu^2 + \tau^2 > \sigma^2$ implies that the squared Euclidean distance between a data vector in $\mathcal{X}$ and the outlier is greater than all the pairwise distances between the data vectors in $\mathcal{X}$.

In the presence of multiple outliers, ($n_0 > 1$), the so-called masking effect can make the detection challenging. The following theorem, of which Proposition 5 is a special case, specifies situations where the proposed method can overcome the masking phenomenon.

Theorem 2. *[Multiple outliers.] Assume that the above assumptions (a) - (d) are satisfied and $n > n_0$. Let $\tau^2 \overset{let}{=} c\sigma^2$ for some constant $c > 0$. Under either of the following further assumptions,*

(i) If $c > 1$ and $\mu \geq 0$, or

(ii) if $c = 1$ and $\mu > 0$, or

(iii) if

$$\frac{n(n_0 - 1)}{n_0(n - 1)} < c < 1 \tag{3.14}$$

and

$$\mu^2 > f_{n,n_0}(c)\sigma^2, \tag{3.15}$$

where

$$f_{n,n_0}(c) = \frac{c^2(n-1) - c(n-n_0) - (n_0-1)}{n(n_0-1) - cn_0(n-1)},$$

then the $(N-1)$ *vs 1 splits- MDP outlier detection method detects outliers in the large p-limit in the sense that the MDP distance is bigger when an outlier is split from the rest than when non-outlier is split from the rest with probability 1.*

Remark 2. *(i): If* $c > 1$*, then the variation of outlier data vectors is larger than the that of non-outlier data vectors. As a result, in the large* $p-$*limit, the outliers tend to be so far away from each other that the masking effect becomes negligible.*

Remark 3. *(ii): If* $c = 1$*, then the variation of outlier data vectors is the same as that of non-outlier data vectors. Outlier detection is possible only if* μ *is strictly larger than 0. From condition (d) in the previous page, this implies that the squared mean difference between non-outliers and outliers should grow at least at the order of* $O(p)$ *as the dimension grows.*

Remark 4. *(iii): If the variation of the outliers is moderate, then successful outlier detection is possible only the mean difference is relatively large.*

It is evident that any outlier detection method will eventually struggle with too many outliers. But, how many outliers is too many? It is practically important and interesting to answer the question "how many outliers can MDP method tolerate in the large p-limit?". Theorem above under moderately isolated outliers case provides an immediate and intuitive answer to this question.

Corollary 2. *[How many outliers can MDP handle?] Assume that the above assumptions (a) - (d) are satisfied and* $n > n_0$*. Suppose that we have "moderately" isolated outliers from a population with* $\tau^2 = c\sigma^2$*, where* $0 < c < 1$ *is fixed. The MDP outlier method detects outliers successfully w.p. 1 as long as the the number of outliers does not exceed the upper bound,*

i.e., $n_0 < n/\{n(1-c)+c\}$ *and the mean difference exceeds the lower bound in the sense that* $\mu^2 > f_{n,n_0}(c)\sigma^2$.

Example 1. *For two different* c *values, shown in Figure 3.4 are the plots of the minimum required* μ^2 *from* (3.15) *for* n_0 *satisfying* (3.14), $c = 0.9$ *on top and* $c = 0.7$ *on bottom. Outlier detection is possible up to* $n_0 = 9$ *for* $c = 0.9$ *whereas we have the limit* $n_0 = 3$ *for* $c = .7$. *In both cases, the minimum mean difference is increasing as the number of outlier increases. In fact, we have a noticeable jump from* $n_0 = 8$ *to 9 (on top) and from* $n_0 = 2$ *to 3, and after that jump, the masking effect becomes so dominant that satisfactory outlier identification is impossible.*

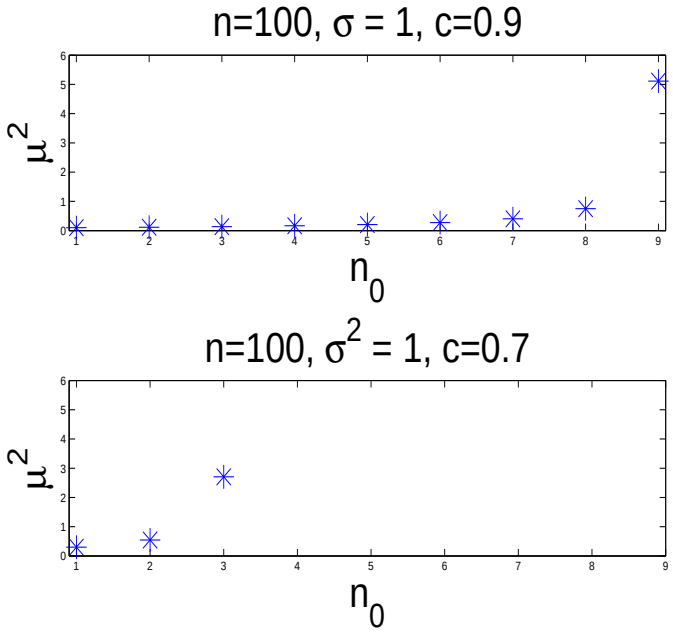


Figure 3.4. The number of outliers paired with the minimum μ^2 required for successful outlier detection. Top: $\tau^2 = .9\sigma^2$. Bottom: $\tau^2 = .7\sigma^2$. Outliers have to be farther away from the population mean as the number of outliers increases up to the limit, given by (3.14). If n_0 is beyond that limit, the detection methods break down.

As far as the proof goes, Proposition 5 is a special case of Theorem 2 when $n_0 = 1$. The proof of Theorem 2 shown in Ahn et al. (2013) is as follows:

Proof of Theorem 2

In this proof we will use 1 (non-outliers $\mathbf{x}$ distribution) and 0 (outliers $\mathbf{x}^0$ distribution) as the class labels for convenience's sake. In the HDLSS geometrical limit, the data vectors from Class 1 and Class 0 form two simplices, denoted by $\mathcal{X}$ and $\mathcal{X}^0$. Assume that the respective sample sizes are n and n_0. Since only the relative locations of the data vectors are of interest for all our purposes, we can express the data matrices for $\mathcal{X}$ and $\mathcal{X}^0$, after scaled by $\sqrt{p}$, as the following:

$$\mathcal{X} = \begin{bmatrix} \sigma \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{bmatrix} & \begin{bmatrix} \delta_{\mathrm{X}} & \cdots & \delta_{\mathrm{X}} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \delta_{\mathrm{X}} & \cdots & \delta_{\mathrm{X}} \end{bmatrix} & \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \end{bmatrix},$$

$$\mathcal{X}^0 = \begin{bmatrix} \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} & \begin{bmatrix} -\delta_{\mathrm{X}^0} & \cdots & -\delta_{\mathrm{X}^0} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ -\delta_{\mathrm{X}^0} & \cdots & -\delta_{\mathrm{X}^0} \end{bmatrix} & \tau \begin{bmatrix} 1-\frac{1}{n_0} & -\frac{1}{n_0} & \cdots & -\frac{1}{n_0} \\ -\frac{1}{n_0} & 1-\frac{1}{n_0} & \cdots & -\frac{1}{n_0} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n_0} & -\frac{1}{n_0} & \cdots & 1-\frac{1}{n_0} \end{bmatrix} \end{bmatrix},$$

where

$$\delta_{\mathrm{X}} = \frac{n_0}{N}\frac{\mu_0}{\sqrt{p-N}} \quad \text{and} \quad \delta_{\mathrm{X}^0} = \frac{n}{N}\frac{\mu_0}{\sqrt{p-N}}.$$

Note that this formulation ensures the same roles of σ^2, τ^2 as in the geometric representation and $\mu_0^2 = \mu^2 + \sigma^2/n + \tau^2/n_0$. For any split of data vector, $\ell \in \{1,0\}^N$, we formulate the resulting MDP distance. From the proof of the Theorem 1 in Ahn et al. (2011), we have an explicit expression for the MDP distance in terms of ℓ,

$$D^2_{\mathrm{MDP}}(\ell) = \mu_0^2 \left\{ \frac{\mu_0^2}{\sigma^2}\frac{n_{11}n_{10}}{n} + \frac{\mu_0^2}{\tau^2}\frac{n_{01}n_{00}}{n_0} + (\frac{n_{10}}{n} - \frac{n_{00}}{n_0})^2 \right\}^{-1}, \tag{3.16}$$

where n_{ij} denotes the number of samples that are that are actually from Class i, but classified into Class j. Note that $n_{11} + n_{10} = n$ and $n_{01} + n_{00} = n_0$. For outlier detection, we only

consider all possible $(N-1)$ vs 1 splits and there are only two possible scenarios; a non-outlier (a $\mathbf{x}$ vector) separated from the rest or an outlier (a $\mathbf{x}^0$ vector) separated from the rest. Thus, the pairs of consideration for (n_{11}, n_{00}) are $(n-1, 0)$ or $(n, 1)$. The former corresponds to the instances when one of the non-outliers is separated from the rest and the latter is the case when an outlier is separated from the rest.

In order to achieve successful outlier detection, we require that $D^2_{\text{MDP}}(\ell_0) \geq D^2_{\text{MDP}}(\ell_1)$, where ℓ_0 is the label assignments which splits an outlier from the rest and ℓ_1 is the label vector which separates a non-outlier from the rest. From the expression of MDP distance for any given label vector above, (3.16), let us define a bivariate function f as follows:

$$f(x, y) = \frac{\mu_0^2}{\sigma^2}\frac{x(n-x)}{n} + \frac{\mu_0^2}{\tau^2}\frac{(n_0-y)y}{n_0} + \left(1 - \frac{x}{n} - \frac{y}{n_0}\right)^2,$$

where (x, y) are the pair of legitimate values of (n_{11}, n_{00}), i.e., either $(n-1, 0)$ or $(n, 1)$. Successful outlier detection is possible if and only if

$$f(n-1, 0)^{-1} < f(n, 1)^{-1} \Leftrightarrow f(n-1, 0) > f(n, 1) \tag{3.17}$$

By plugging $\mu_0^2 = \mu^2 + \sigma^2/n + \tau^2/n_0$ and $\tau^2 = c\sigma^2$ into (3.17), we get equivalent condition,

$$g_{n,n_0}(c)\mu^2 + h_{n,n_0}(c)\sigma^2 > 0, \tag{3.18}$$

where the coefficients of μ^2 and σ^2 are

$$g_{n,n_0}(c) = cn_0(n-1) - n(n_0-1)$$

and

$$h_{n,n_0}(c) = \left(c^2(n-1) - c(n-n_0) - (n_0-1)\right).$$

Note that $g_{n,n_0}(c) < 0$ iff $0 < c < n(n_0-1)/\{n_0(n-1)\}$ and $h_{n,n_0}(c) < 0$ iff $0 < c < 1$. Depending on the signs of these coefficients, there are several possible scenarios:

1. If $c > 1$, then $g_{n,n_0}(c) > 0$ and $h_{n,n_0}(c) > 0$. Thus, the condition (3.18) holds for any values of $\mu \geq 0$.

2. If $c = 1$, then $g_{n,n_0}(c) > 0$ and $h_{n,n_0}(c) = 0$. The condition (3.18) holds for any values of $\mu > 0$.

3. If $n(n_0 - 1)/\{n_0(n-1)\} < c < 1$, then (3.18) holds if and only if $\mu^2 > -h_{n,n_0}(c)/g_{n,n_0}(c)\sigma^2$.

4. If $0 < c \leq n(n_0 - 1)/\{n_0(n-1)\}$, then $g_{n,n_0}(c) \leq 0$ and $h_{n,n_0}(c) < 0$ and the condition (3.18) does not holds for any values of $\mu \geq 0$.

□

3.3.2 Centroid Distance

In this section, we study the HDLSS asymptotic properties of outlier detection based on CD.

Theorem 3. *[Multiple outliers.] Assume that the above assumptions (a) - (e) are satisfied and $n > n_0$. Let $\tau^2 \stackrel{let}{=} c\sigma^2$ for some constant $c > 0$. Under either of the following further assumptions,*

(i) if $c > 1$ and $\mu \geq 0$, or

(ii) if $c = 1$ and $\mu > 0$, or

(iii) if $0 < c < 1$ and $\mu^2 > \sigma^2(1 - c)(N - 2)/(n - n_0)$,

then the $(N - 1)$ vs 1 splits- CD outlier detection method detects outliers in the large p-limit in the sense that the CD distance is bigger when an outlier is split from the rest than when non-outlier is split from the rest with probability 1.

Ahn et al. (2013) proved Theorem 3 as below.

Proof of Theorem 3

We start from the same data vector position as we used for MDP. The squared of the Centroid distance when one outlier separated from the rest is

$$D^{*2} = (\frac{n}{N-1})^2\mu_0^2 + \tau^2(\frac{N}{N-1})^2(\frac{n_0 - 1}{n_0}).$$

If a non-outlier data vector is separated from the rest, the role of n and n_0, σ^2 and $c\sigma^2$ are reversed from the expression above. We get the squared of the Centroid distance,

$$D^2 = (\frac{n_0}{N-1})^2(\mu^2 + \frac{\sigma^2}{n} + \frac{c\sigma^2}{n_0}) + (\frac{N}{N-1})^2(\frac{n-1}{n})\sigma^2.$$

For correct outlier detection, we require that $D^{*2} > D^2$, which is equivalent to

$$(n^2 - n_0^2)\mu^2 + (c-1)N(N-2)\sigma^2 > 0. \tag{3.19}$$

Let $p_{n,n_0}(c) = (c-1)N(N-2)$.

1. If $c > 1$, then $p_{n,n_0}(c) > 0$ and (3.19) is satisfied for any $\mu^2 \geq 0$.

2. If $c = 1$, then $p_{n,n_0}(c) = 0$ and (3.19) is satisfied for any $\mu^2 > 0$.

3. If $0 < c < 1$, then (3.19) can be written as $\mu^2 > \sigma^2(1-c)(N-2)/(n-n_0)$. □

For outliers with moderate variation ($c < 1$), the detection method for HDLSS data is successful if the outliers are deviated from the non-outliers more than minimum required quantity. Outlier detectable areas are determined by the number of parameters, μ^2, c, N and n_0. We fix $N = 50$ and change $n_0 = 1, 2$ or 5 and compare the detectable areas for CD and MDP. In the plot below, the x-axis is for $\log_2(1-c)$ and y-axis for $\log \mu^2$ and detectable areas by CD (MDP or both) are colored as blue (red or purple). In the top panel, when $n_0 = 1$, the CD and MDP ares are the same. With multiple outliers, $n_0 = 2$ (middle) and $n_0 = 5$, CD detectable area is slightly bigger than MDP area, however, the difference is noticeable only when $\log_2(1-c) > 2$. The difference in the areas with $c < 3/4$ is not much of an impact since the assumption of the variation within outliers less than 3/4 of variation within non-outliers seems to be unrealistic.

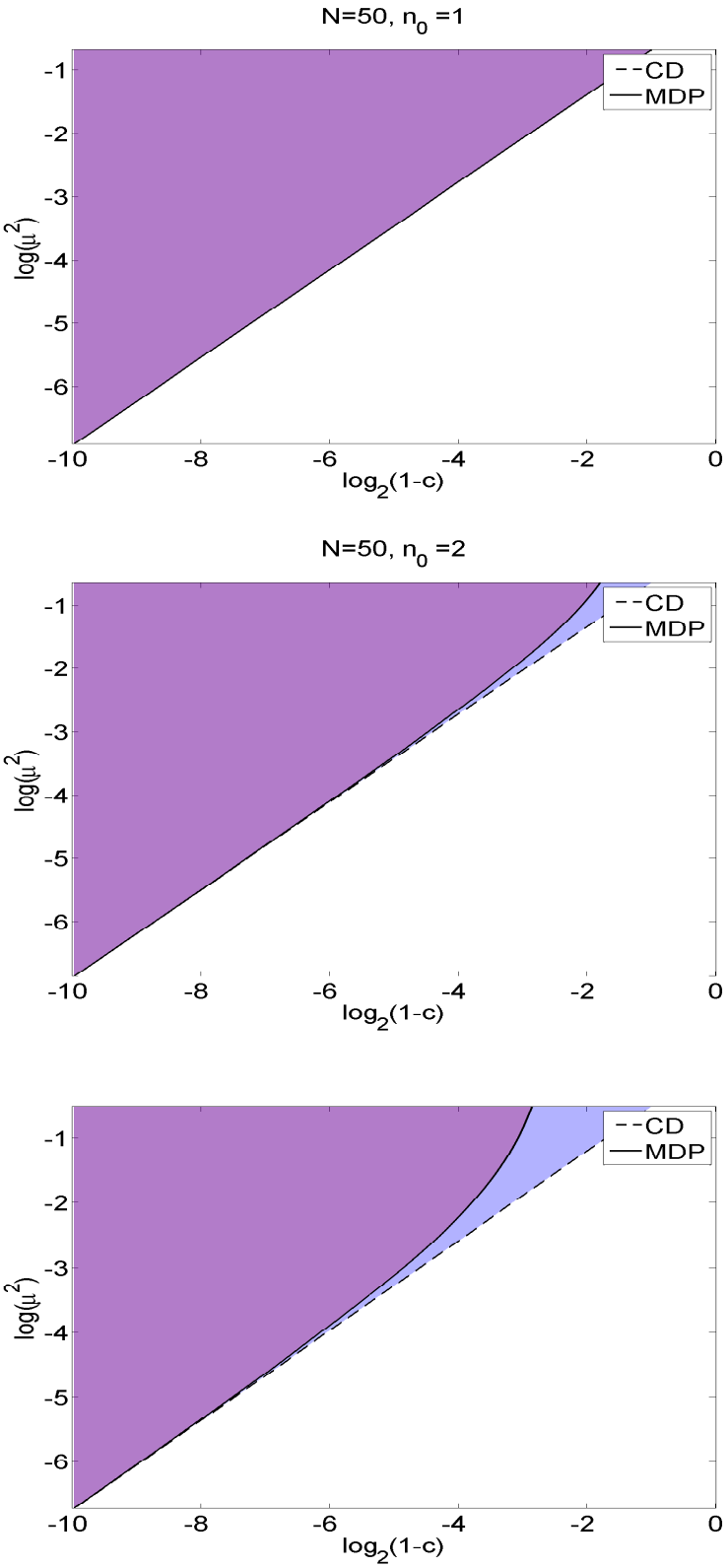


Figure 3.5. Detectable areas by CD and MDP.

3.4 Proposed Method For HDLSS Outlier Detection

In the previous section, we discussed three distance measures. The next question is how to judge the outlier, or how confidently one can say that the distance of the outlier is unusually large compared to those of other data points. In order to answer this question, one may like to see the "extremeness" of each data point in comparison with theoretical values that would be observed if there were no outliers. To see this, we need to obtain the distribution of the distances under the hypothesis of no outlier, i.e., the null distribution of the distance.

In this section we introduce a parametric bootstrap method in order to estimate the null distribution of the distance with which we can judge the outlier. In Section 3.4.1 we will describe the details about how to apply the parametric bootstrap approach for the estimation of the null distribution. In Section 3.4.2 we suggest the outlier detection method using a quantile-quantile plot based on the empirical null distribution. In Section 3.4.3 we discuss the consistency of the proposed method.

3.4.1 Estimation For the Null Distribution Of the Distance

Suppose there are n data points denoted by $\mathbf{x}_j$, $j = 1, \ldots, n$, in $\mathbb{R}^p$ and there is one possible outlier in the data. Let us denote $\hat{D}_j = d(\mathbf{x}_j, \mathbf{X}_{(-j)})$ the distance of the jth data point from the other observations, where $\mathbf{X}_{(-j)} = [\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_n]$, and sort these distance values to be $\hat{D}_{(1)} < \cdots < \hat{D}_{(n)}$. We can reasonably assume that the observation that has the maximum distance is a possible outlier. The question is when we can say the maximum is far enough to be an outlier. In order to judge whether a seemingly outlying data point is a true outlier, we like to compare the observed distances to the regular distances which are obtained under a non-outlier situation.

A parametric bootstrap approach is useful for the estimation of the null distribution of the distances, which would be observed under the situation that there are no possible outliers. In fact, we are interested in the n quantiles of this null distribution, denoted by $q_{(1)}, \ldots, q_{(n)}$. The parametric bootstrap estimates $q_{(1)}, \ldots, q_{(n)}$ with $\tilde{D}_{(1)}, \ldots, \tilde{D}_{(n)}$, which is the ordered

one-vs-others distances computed from a bootstrap sample of size n. Let us denote $\tilde{\mathcal{D}}_n$ a set of distances from the bootstrap samples.

At a first step, we estimate the mean and covariance $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ from the given data set for the preparation of re-sampling. It is important to assume that these two re-sampling parameters are robust to outliers. That way we can assure the bootstrap samples are free of outliers, from which we can reasonably estimate $\tilde{\mathcal{D}}_n$. One easy way to achieve the outlier-free condition is eliminating the potential outliers if any. It is natural to suspect the observation that is the farthest from the others. Suppose the kth observation is selected as a possible outlier based on a distance measure $d(\cdot, \cdot)$, i.e., $\hat{D}_k = \hat{D}_{(n)}$. We can think of a sample mean and a sample covariance estimation without the kth observation; say $\bar{\mathbf{x}}_{(-k)}$ and $\mathbf{S}_{(-k)}$. Apparently, $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}_{(-k)}$ and $\hat{\boldsymbol{\Sigma}} = \mathbf{S}_{(-k)}$ would be better estimators rather than $\bar{\mathbf{x}}$ and $\mathbf{S}$ if the kth observation were a true outlier.

The inference from $\tilde{\mathcal{D}}_n$ entails repetition for large B. For $b = 1, \ldots, B$, we draw n samples $\mathbf{y}_1^b, \ldots, \mathbf{y}_n^b$ from $\mathcal{N}_p(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, which provides us with B realizations $\tilde{\mathcal{D}}_n^{\,1}, \ldots, \tilde{\mathcal{D}}_n^{\,B}$ of $\tilde{\mathcal{D}}_n$. We sort the distances in each set $\tilde{\mathcal{D}}_n$, and average $\tilde{D}_{(j)}$ at each j for the estimation of $q_{(1)}, \ldots, q_{(n)}$. Notice the hat notation is used for the first generation estimators (samples), whereas the tilde notations are used for the second generation estimators (bootstrap samples).

When we generate a bootstrap sample of size n from $\mathcal{N}_p(\bar{\mathbf{x}}_{(-k)}, \mathbf{S}_{(-k)})$, we will encounter the singularity problem with the covariance $\mathbf{S}_{(-k)}$ due to $(n-1) < p$. In that case we add a small number $\alpha > 0$ to the diagonal elements of $\mathbf{S}_{(-k)}$, which produces a nonsingular matrix $(\mathbf{S}_{(-k)} + \alpha \mathbf{I}_p)$ that is still extremely close to $\mathbf{S}_{(-k)}$. By doing this, we ensure to draw samples from well-defined distribution. The small positive number α is arbitrary. We set α equal to $\sqrt{log(p)/(n-1)}$, following the asymptotic term $\sqrt{log(p)/n} \to 0$ which is found in some related works with covariance estimation (Bickel and Levina, 2008a,b; Cai and Liu, 2011).

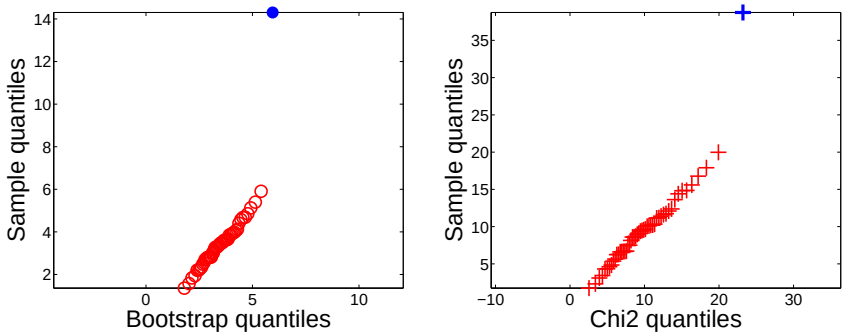
3.4.2 Comparison Of the Null Distribution and Sample Quantiles

Once we obtain the empirical null distribution from the bootstrap samples, we can create a quantile-quantile plot to see whether there is any aberrant pattern among data points. Specifically, we plot the ordered distances calculated from the sample, $\hat{D}_{(1)} < \cdots < \hat{D}_{(n)}$ in the y-axis (sample quantiles). Also we plot the average of the ordered distances based on B bootstrap samples, $\tilde{D}^*_{(1)} < \cdots < \tilde{D}^*_{(n)}$, where $\tilde{D}^*_{(j)} = \text{average}\,\{\tilde{D}^b_{(j)}\}_{b=1}^B$ for $j = 1, \ldots, n$, in the x-axis (bootstrap quantiles). If there is no outlier in the data set, it is expected that the pairs of the distances $(\hat{D}_{(j)}, \tilde{D}^*_{(j)})$, $j = 1, \ldots, n$, will be approximately linear, showing a straight line. If a certain data point deviates from the straight line, that can indicate the outlier.

Note that the proposed parametric bootstrap approach above has a similar concept as the so-called chi-square qq plot that is used for low-dimensional multivariate data. In the following example, we want to show that our proposed method can also work well for traditional low dimensional data. Suppose we have 50 random vectors from $\mathcal{N}_{10}(\mathbf{0}, \mathbf{\Sigma})$, and deliberately input one outlier. Details about how to input the outlying observation are described in Section 3.5. In Figure 3.6-(a), we implement our proposed parametric bootstrap approach. For the calculation of the distance, we use $d(\mathbf{x}_j, \mathbf{X}_{(-j)}) = \{(\mathbf{x}_j - \bar{\mathbf{x}}_{(-j)})^\top \mathbf{S}^{-1}_{(-j)} (\mathbf{x}_j - \bar{\mathbf{x}}_{(-j)})\}^{1/2}$. As expected, one observation is outlying from the rest in this figure. In addition, we draw the chi-square qq plot in (b) for the same data. The y-axis displays the order statistics of the squared Mahalanobis distances from the sample (sample quantiles). The x-axis displays the quantiles of the chi-square distribution with the degree of freedom 10. With the comparison to (b), the plot (a) looks visually equivalent to or even better than (b); it is fully functioning for distinguishing the outlier.

In the example of Figure 3.6 with low-dimensional data, both plots in (a) and (b) work well for the detection of outlier. However, the chi-square qq plot is not appropriate for the high-dimensional data since the estimated squared Mahalanobis distances in $n < p$ do not follow the chi-square distribution. Meanwhile, the great advantage of our proposed

method is the flexibility in higher dimension, which will be shown later in our simulation and real data analysis. Not only dimension, the method is flexible for different distance measures. This is because unlike the chi-square plot the our method does not require any distribution assumption of the distance. Instead, the parametric bootstrap method estimates the distribution of the distance empirically regardless of the type of distance function.



(a) Parametric bootstrap (b) Chi-square qq plot

Figure 3.6. An example of multivariate outlier identification in low dimension. In (a), the parametric bootstrap idea is applied for the qq plot with the bootstrap quantiles in the x-axis and the sample quantiles in the y-axis. In (b), the chi-square qq plot is displayed with χ^2 quantiles versus sample quantiles. In both plots, one observation is shown as an outlier.

Generally, suppose there are at most $n_0 \ll n$ potential outliers. We assume that removing all these n_0 outliers yields the remaining data free of outliers from which the parameter estimation is reasonable. In practice when we do not know how many outliers actually exist, we suggest to see all results from $n_0 = 1, \cdots, N_0$, where N_0 is a reasonable upper bound for the number of outliers. The procedures are performed sequentially. Assuming $n_0 = 1$, one potential outlier can be found by $\hat{D}_{k_1} = \hat{D}_{(n)}$. By deleting the k_1th observation from a data set, the qq plot using the parametric bootstrap is drawn. If there is no evidence that the k_1th observation is a true outlier, we stop the procedure and conclude there is no outlier. However, if the k_1th observation seems to be "real" outlier on the plot, we remove it and attempt to identify another potential outlier among $n-1$ observations. In this new step, we set $n_0 = 2$, which means there are two possible outliers, one of which are already removed in the previous step. The second potential outlier can be found by $\hat{D}_{k_2} = \hat{D}'_{(n-1)}$.

Note that there may be some change in the order statistics $\hat{D}'_{(1)}, \ldots, \hat{D}'_{(n-1)}$ after deleting k_1th observation. If there is no evidence that the k_2th observation is an outlier, we stop the procedure and conclude that there is only one outlier. If the plot shows the k_2th observation is also an outlier, we remove it and set $n_0 = 3$, and so forth.

This algorithm is illustrated in Figure 3.7. In each step of $n_0 = 1$, 2 and 3, we examine the extremeness of the observation that has the maximum distance. When the observation shows an outlying pattern, it is removed from the data set. The process is continued until we do not see any extreme pattern of the observations. In the step $n_0 = 3$, there is no enough evidence for the outlier. Thus we would conclude that there exist two outliers in this data set. Note that in this illustration we use MDP distance in (3.9) for the computation of the distances.

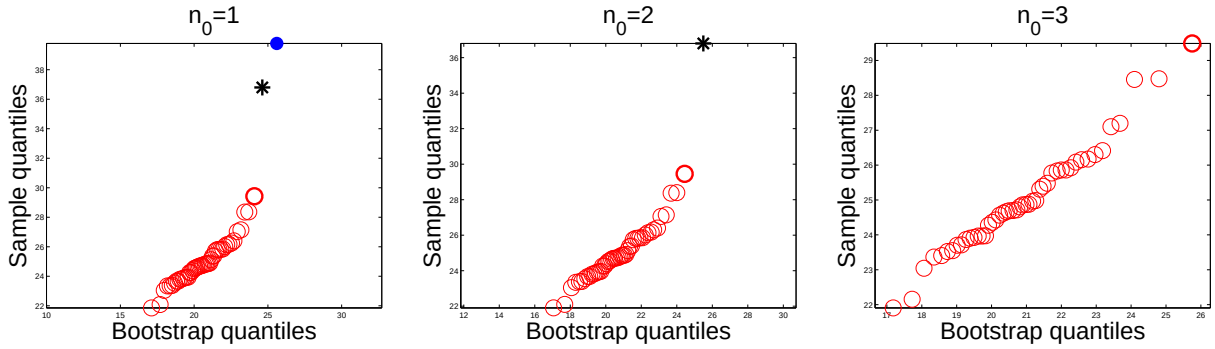


Figure 3.7. Illustration of the algorithm for outlier detection. In each step of n_0, we examine the extremeness of the observation that has the maximum distance. The process is continued until we do not see any evidence of outlying observations. Based on the plots, we would conclude that there exist two outliers.

Furthermore, we summarize the proposed algorithm in light of the computational steps as below. As for the distance measures, we can use the centroid, MDP, and rMH distances introduced in Section 3.2. From now on, we omit the hat and tilde notation for simplicity's sake.

Algorithm for Outlier Detection

Suppose $n_0 = 1$, which means there is one possible outlier.

(1) Calculate the distances of each observation (say D_j, $j = 1, \ldots, n$), and sort them by ascending order to be $D_{(1)} < \cdots < D_{(n)}$.

(2) Find the observation which has the largest distance, $D_k = D_{(n)}$, then delete the kth observation from the data set.

(3) Compute the sample mean vector $\bar{\mathbf{x}}_{(-k)}$ and covariance matrix $\mathbf{S}_{(-k)}$ without the kth observation.

(4) Generate n random vectors from $\mathcal{N}_p(\bar{\mathbf{x}}_{(-k)}, \mathbf{S}_{(-k)} + \alpha \mathbf{I}_p)$, where $\alpha = \sqrt{log(p)/(n-1)}$.

(5) Calculate the distances as in step (1) with the simulated data in step (4). Say D_j^1 $(j = 1, \ldots, n)$ and sort them to be $D_{(1)}^1 < D_{(2)}^1 < \cdots < D_{(n)}^1$.

(6) Repeat the steps (4) and (5) B times, and average the ordered distances in order to obtain $D_{(1)}^* < \cdots < D_{(n)}^*$, where $D_{(j)}^* = \text{average}\,\{D_{(j)}^b\}_{b=1}^B$ for $j = 1, \ldots, n$.

(7) Create a qq plot with $D_{(j)}^*$ in the x-axis and $D_{(j)}$ in the y-axis. Deviating from the straight line indicates an outlier.

This procedure will stop here if there is no evidence for the outlier in the plot (7). If the kth observation looks a true outlier, it is removed from the data set. In the next step, we set $n_0 = 2$, which means there are two possible outliers, one of which is removed and we search for the second one if any. The steps for $n_0 = 2$ is the same as those of $n_0 = 1$ but with $(n-1)$ observations.

(1′) Calculate the the distances $D'_{(1)} < \cdots < D'_{(n-1)}$.

$\vdots$

These procedures are continued until there is no outlying pattern of the observations, or a reasonable upper bound for the number of outliers is reached.

3.4.3 Consistency of the parametric bootstrap method

In what follows we discuss some theoretical reasoning of the proposed parametric bootstrap method. Suppose there are two Gaussian data sets. The first data set is $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, from which we calculate the leave-one-out distances, and denote a set of the distances by $\hat{\mathcal{D}}_n$. Also sort these distances to be $\{\hat{D}_{(1)}, \ldots, \hat{D}_{(n)}\}$. The second is a bootstrap sample of size n, i.e., $\mathbf{y}_1, \ldots, \mathbf{y}_n$ from $\mathcal{N}_p(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, from which we also find a set of the distances denoted by $\tilde{\mathcal{D}}_n$ with the sorted distances $\{\tilde{D}_{(1)}, \ldots, \tilde{D}_{(n)}\}$.

We are interested in the consistency of $\tilde{\mathcal{D}}_n$ and the performance of the parametric bootstrap when $p \gg n$. Suppose that the null distribution of the distances is known and we denote the quantiles of the distribution by $q_{(j)}$, $j = 1, \ldots, n$. Then, we have $P(\hat{D}_{(j)} = q_{(j)})$ converges to 1 as the sample size goes to infinity for all j. Also, because of the fact that $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ is uniformly consistent to $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as $n \to \infty$, it is true that $P(\hat{\mathcal{D}}_n = \tilde{\mathcal{D}}_n) \to 1$ as $n \to \infty$ since both $P(\hat{D}_{(j)} = q_{(j)}) \to 1$ and $P(\tilde{D}_{(j)} = q_{(j)}) \to 1$. However, if p is very large relative to n and we do not have this uniform consistency, the consistency of $\tilde{\mathcal{D}}_n$ to $\hat{\mathcal{D}}_n$ is unsure. Thus, an important thing is to establish the rate at which p can converge to infinity while still having the required uniform consistency.

The similar argument has been discussed in van der Lann and Bryan (2001) for their bootstrap approach. Let $p(n)$ be such that $n/log(p(n)) \to \infty$ as $n \to \infty$. In their Theorem 3.1, the consistency of $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ to $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has been shown as $n/log(p(n)) \to \infty$. The idea of the proof is using Bernstein's inequality (van der Vaart and Wellner, 1996). An application of Bernestein's inequality gives the result that $\max_{ij} P(|\hat{\boldsymbol{\Sigma}}_{ij} - \boldsymbol{\Sigma}_{ij}| > \epsilon) \leq C \exp(-n\epsilon)$ for some $C < \infty$. Thus,

$$P(\max_{i \leq j} |\hat{\boldsymbol{\Sigma}}_{ij} - \boldsymbol{\Sigma}_{ij}| > \epsilon) \leq \sum_{i \leq j} P(|\hat{\boldsymbol{\Sigma}}_{ij} - \boldsymbol{\Sigma}_{ij}| > \epsilon) \leq \frac{p(p-1)}{2} C \exp(-n\epsilon),$$

which converges to zero if $n/log(p(n)) \to \infty$. Consequently, the consistency of $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ directly leads to the result that $P(\hat{\mathcal{D}}_n = \tilde{\mathcal{D}}_n) \to 1$ as $n/log(p(n)) \to \infty$.

Although $q_{(j)}$, $j = 1, \ldots, n$ are unknown in reality, the argument above is enough to prove the consistency of our parametric bootstrap method. This is because our bootstrap approach for the outlier detection focuses on the relationship between the sample and the bootstrap sample rather than a true null distribution. In other words, By repeating generating a bootstrap sample of size n, we estimate $\tilde{\mathcal{D}}_n$, whose consistency to $\hat{\mathcal{D}}_n$ as $n/log(p(n)) \to \infty$ is obtained regardless of what a true distribution is.

3.5 Simulation Study

In this section we carry out some simulations to see the performance of the proposed method for outlier detection in HDLSS data. In Section 3.5.1, we investigate the performance when there is only one outlier. In Section 3.5.2, we study the case when there are more than one outlier, which implies possible masking effect. In that section, we also give an argument that masking effect in HDLSS is rare in general.

3.5.1 Single Outlier Case

In Section 3.2 we introduced three types of distance measures for high-dimensional data: the centroid, rMH, and MDP distances. It is natural to suggest a potential outlier based on the maximum distance since three distances basically measure the level of each data point's remoteness compare to the other observations. However, in some case, there are conflicts between distance measures about which observation has the largest distance; the order statistics for the distance of each data point may be different across the distance measures. We found that these conflicts are influenced by the covariance structure that the data set has. In the following example we compare three types of distances under the three different scenarios: variables are uncorrelated, strongly correlated, or under the mixed condition of these two extreme cases. For simplicity we assume that there is only one outlier in each case.

Let us generate a sample of size 49 from p-variate normal distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, where $p = 500$. We consider three types of covariance structures. In Setting I, we use $\mathbf{\Sigma} = \mathbf{I}$, i.e., the variables are uncorrelated. In Setting II, we set a compound symmetric matrix, where diagonal elements are 1 and off-diagonal elements are ρ (we set $\rho = 0.7$). This case is when the variables are highly correlated. In Setting III, we consider a mixed situation of these two extreme cases in which some variables are highly correlated and some others are less correlated. In fact, such covariance is more realistic other than those of two previous settings. For this setting, we compute a $p \times p$ matrix by $\mathbf{\Gamma \Lambda \Gamma}^{\mathrm{T}}$, where $\mathbf{\Gamma}$ is an orthonormal

matrix and $\mathbf{\Lambda}$ is a diagonal matrix that contains eigenvalues, i.e., $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_p\}$. The eigenvalues are set to $\lambda_j = pj^{-1}/3$, $j = 1, \ldots, p$. The first 50 eigenvalues, $\lambda_1, \ldots, \lambda_{50}$ are displayed in Figure 3.8. The rest of eigenvalues, $\lambda_{51}, \ldots, \lambda_{500}$, gradually decrease toward zero. We convert this covariance matrix to the correlation matrix $\boldsymbol{\rho} = \{\rho_{ij}\}$, and see the correlation coefficient values. In this example, the correlation coefficients vary as follows: $|\rho_{ij}| < 0.1$ (40.12%), $0.1 \leq |\rho_{ij}| < 0.2$ (32.59%), $0.2 \leq |\rho_{ij}| < 0.3$ (18.85%), $0.3 \leq |\rho_{ij}| < 0.4$ (7.09%), $0.4 \leq |\rho_{ij}| < 0.5$ (1.29%), and $|\rho_{ij}| \geq 0.5$ (0.05%). In addition, in order to produce an orthonormal matrix $\mathbf{\Gamma}$, we first generate a random $p \times p$ matrix in which the elements are random numbers from $unif(0, 1)$. Then, the p column vectors are orthonormalised by the Gram-Schmidt process.

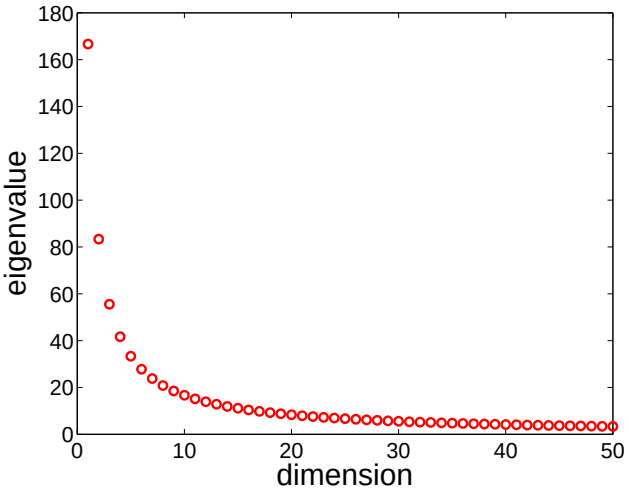


Figure 3.8. Eigenvalues of covariance matrix of Setting III.

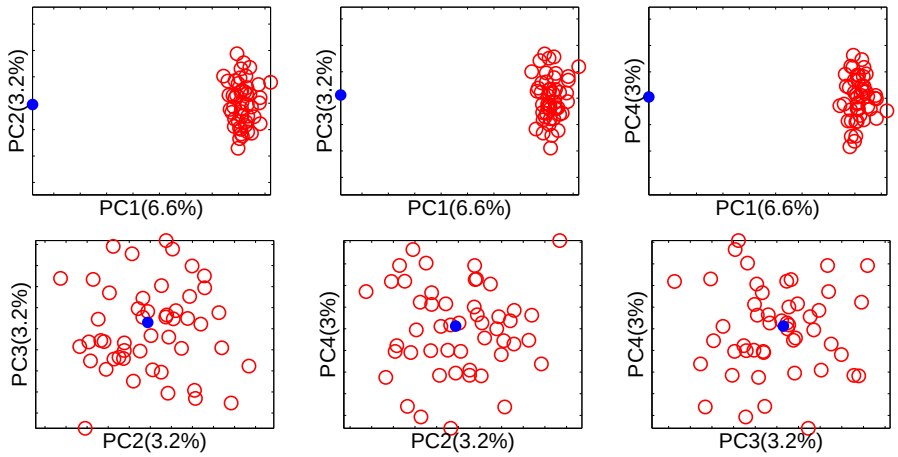
Once a data set without an outlier is generated, we add the 50th data point as the outlier in the data set. An outlying data point is constructed along with a random vector $\mathbf{v}_0$. To decide this random vector, we randomly choose ten eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_{10}$ from the p eigenvectors of $\mathbf{\Sigma}$, and also generate ten random numbers $a_1, \ldots, a_{10}$ from $unif(-1, 1)$, then we construct a linear combination of the vectors with the length of 1 as follows: $\mathbf{v}_0 = (\sum_{i=1}^{10} a_i \mathbf{v}_i)/\| \sum_{i=1}^{10} a_i \mathbf{v}_i \|$. Thus, the outlying data point can be expressed by $outlier = a_0 \mathbf{v}_0 + \mathbf{e}$, where $\mathbf{e}$ is a noise vector from $\mathcal{N}_p(\mathbf{0}, \ .05\,\mathbf{I}_p)$. We set $a_0 = 40$.

With the updated data set, we draw the principal component (PC) scatter plot to see whether this outlier can be seen. Figure 3.9 shows the projections onto the the first four principal components under Settings I to III. The potential outlier is marked by blue solid dot, and the 49 other normal data points are marked by red circle. In (a), PC1 direction clearly shows the outlier separated from the rest of data. In (b), PC2 directions clearly shows the outlier. But in (c), we notice that the first four principal components fail to recognize the outlier. In this case, more principal components are searched for until the outlier appears, and finally it is found in PC7 direction (2.8 %). But, this search may not always work since individual PC does not provide much information as dimension grows. Even though we find the direction that reveals the outliers, it only accounts for 2.8 % of total variance in this example. In general, using PC plots for the purpose of outlier detection may not be efficient in high dimension.

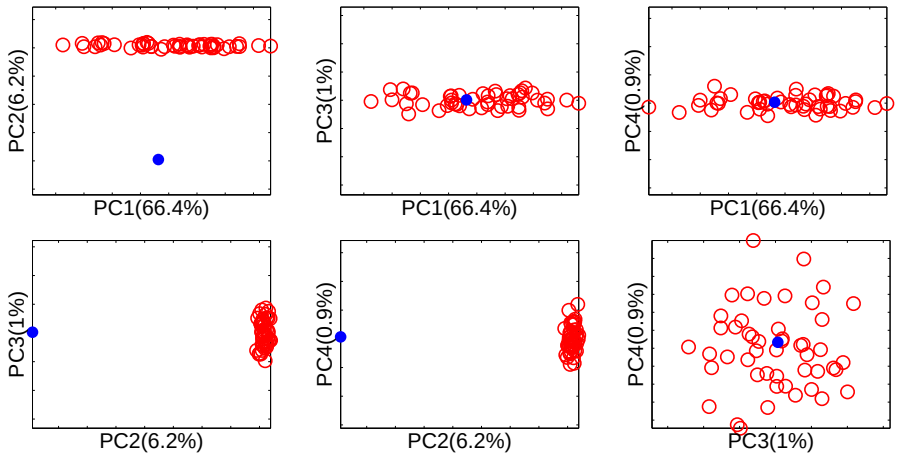
We apply our outlier detection method for this data set. we first calculate the one-vs-others distances of 50 data points using all three types of distances, and draw the quantile-quantile plot as described in Section 3.4.2. Figure 3.10 shows the results. In (a), three distance measures equivalently perform well for identifying outliers. In (b), rMH and MDP distances identify the outlier well but CD does not. In (c), similar to (b), the rMH and MDP distances work well but the CD could not find the outlier. We tried this simulation for different dimensions, such as $p = 100$ and 300, and obtained similar results, which can be found in Figures 3.11-3.13.

3.5.2 Multiple Outliers and Masking Effect

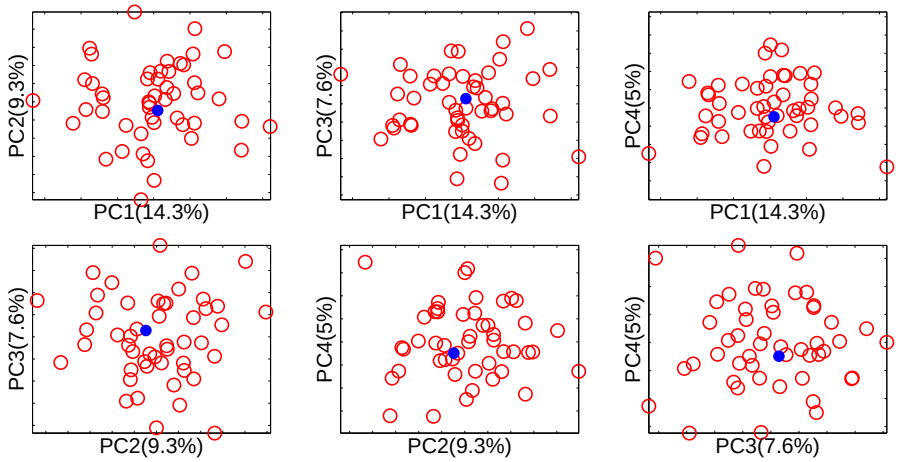
Researchers often face multiple outliers or clusters of outliers in a data set. In this section we examine whether our proposed method is effective under this circumstance. In Section 3.4.2 we propose a sequential algorithm that the potential outliers are tested one by one. As we mentioned, the reason why we follow these steps is that we prevent $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ from being



(a) Setting I

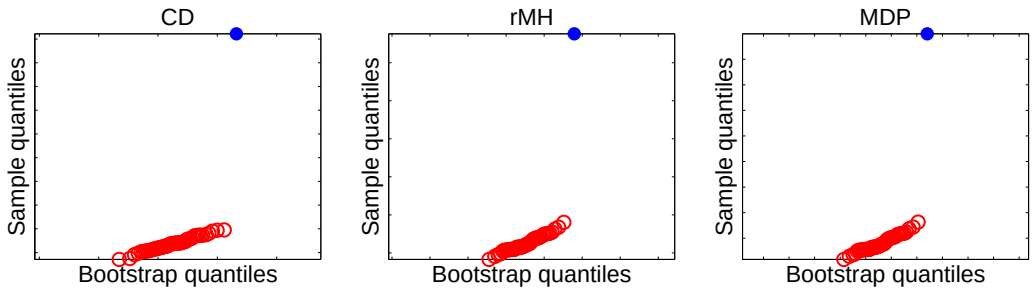


(b) Setting II

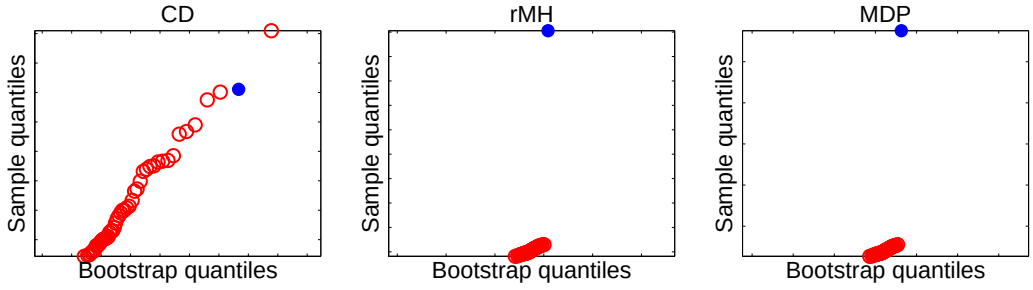


(c) Setting III

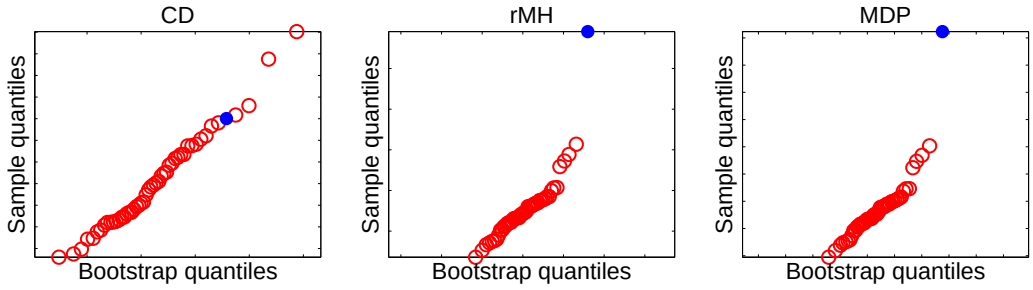
Figure 3.9. PC scatter plot from the data set that has a potential outlier. In (a) and (b), the first and second PC clearly show the outlier, respectively. But in (c), the first four PC fail to recognize the outlier.



(a) Setting I



(b) Setting II



(c) Setting III

Figure 3.10. Performance of three distance measures for outlier detection in HDLSS data. In (a), all three distance measures works well for identifying an outlier. Meanwhile, in (b) and (c), the centroid distance (CD) fails to detect the outlier, but the ridge mahalanobis (rMH) and maximal data piling (MDP) distances identify the outlier regardless of covariance structure.

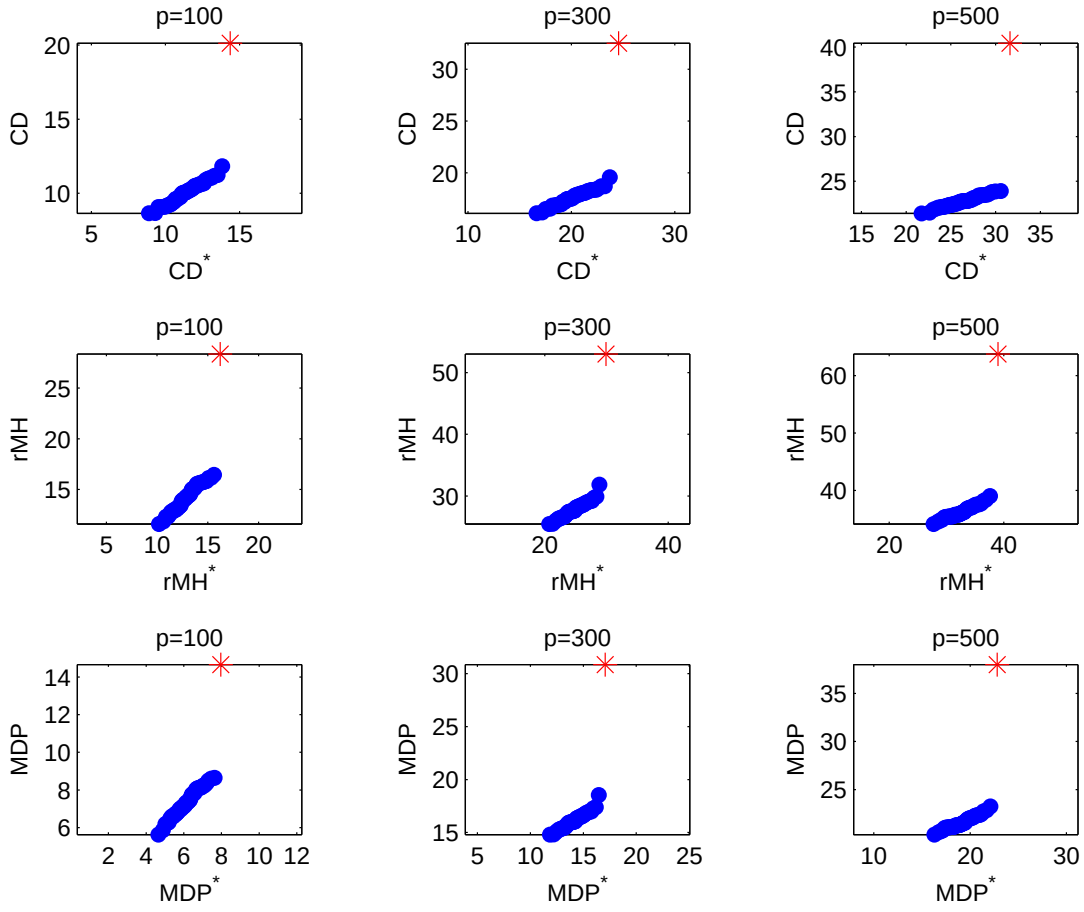


Figure 3.11. (Setting I) Outlier detection results by three distance measures are similar in different dimensions.

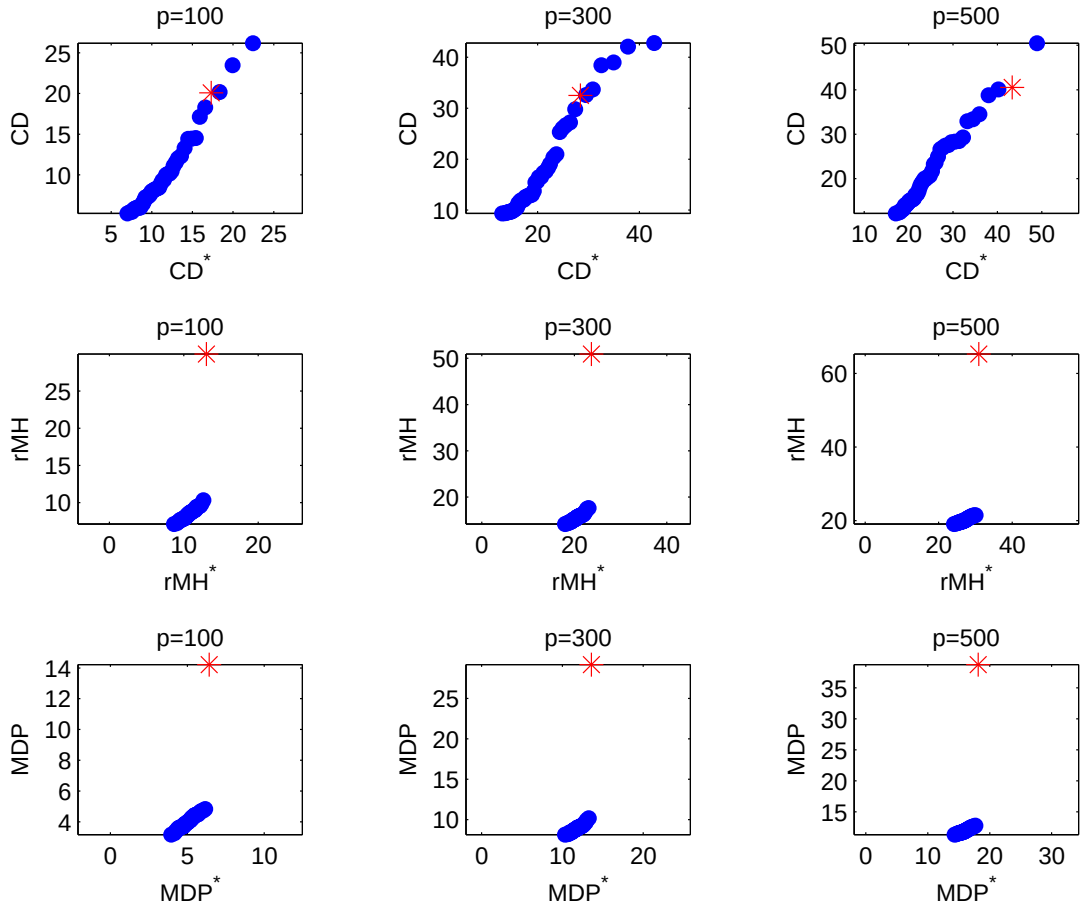


Figure 3.12. (Setting II) Outlier detection results by three distance measures are similar in different dimensions.

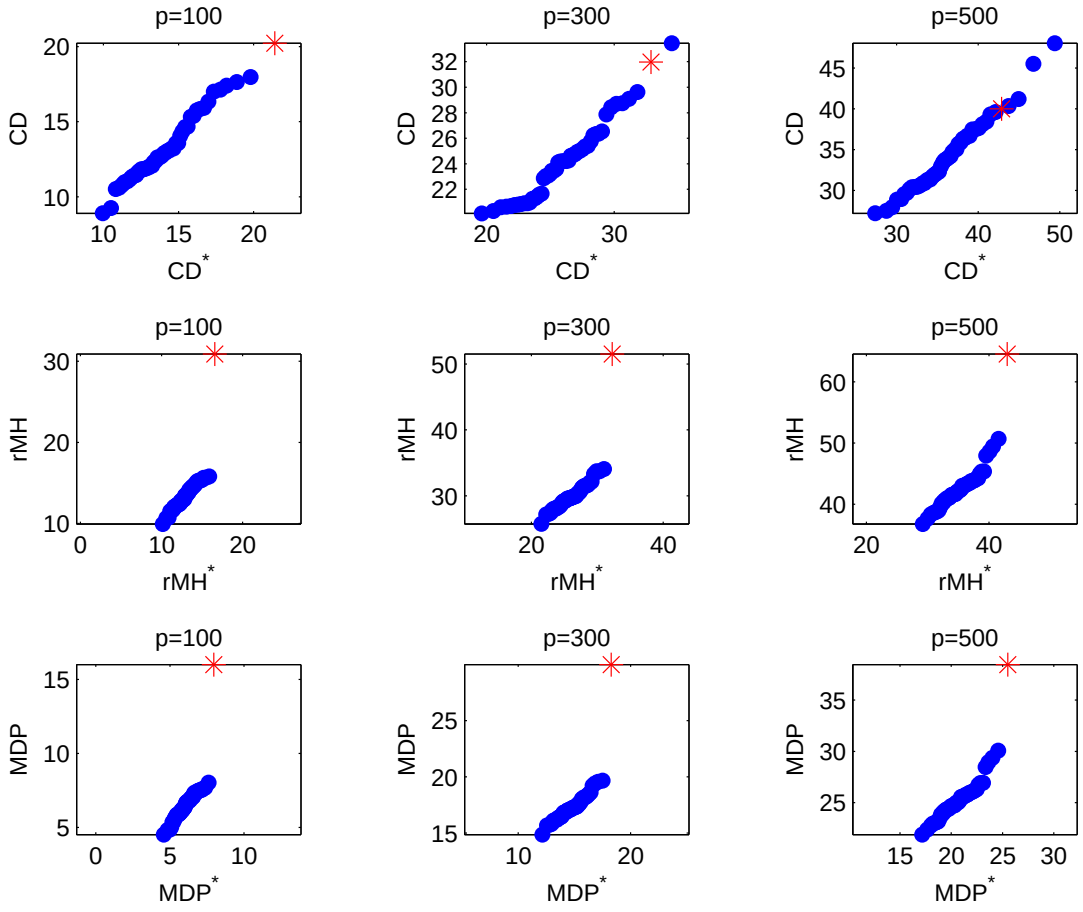


Figure 3.13. (Setting III) Outlier detection results by three distance measures are similar in different dimensions.

contaminated by the outliers, so that the inference from the bootstrap samples, which are drawn from $\mathcal{N}_p(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, is reasonable.

Another benefit of this sequential approach is to prevent the masking effect. Recall that the distance of each observation is defined by the leave-one-out distance, expressed as $\hat{D}_j = d(\mathbf{x}_j, \mathbf{X}_{(-j)})$ in Section 3.2. Suppose there are more than one outlier in a data set. When we calculate the leave-one-out distance from the data, it is possible that the order statistics, $\hat{D}_{(1)} < \ldots < \hat{D}_{(n)}$, are misled by the multiple outliers. In other words, the distance of an outlying data point relative to the other observations can be deflated when some other outliers are included in the data set. If that occurs, we may delete the wrong observation while the true outlier is still kept, resulting in incorrectly estimating $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$. However, this negative scenarios may be not a problem in the proposed outlier detection method thanks to the sequential approach. Even though the true outlier did not appear for the first time through the n ordered distances, it can appear in the next procedure by recalculating the $n-1$ ordered distances, $\hat{D}'_{(1)} < \cdots < \hat{D}'_{(n-1)}$ since removing one observation also removes the possible masking effect carried by that observation.

In the following we run our outlier detection method for two multiple outlier settings. Suppose there is a data set from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $n = 50$ and $p = 500$. For the covariance $\boldsymbol{\Sigma}$, we consider a mixed structure, which is Setting III in Section 3.5.1. In Setting I, we input three outliers that are randomly generated. We replace three observations with the three outlying data points as below

$$\text{outlier1} = d_0\mathbf{v}_{1,0} + \mathbf{e}_1, \tag{3.20}$$

$$\text{outlier2} = d_0\mathbf{v}_{2,0} + \mathbf{e}_2, \tag{3.21}$$

$$\text{outlier3} = d_0\mathbf{v}_{3,0} + \mathbf{e}_3, \tag{3.22}$$

where d_0 is an arbitrary constant that determines the outlyingness (we set $d_0 = 40$), and $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ are independent noise vectors from $\mathcal{N}_p(\mathbf{0},\ .05\,\mathbf{I}_p)$. Also $\mathbf{v}_{1,0}$, $\mathbf{v}_{2,0}$ and $\mathbf{v}_{3,0}$ are

p-dimensional random vectors with the length of 1, expressed by

$$\mathbf{v}_{k,0} \quad = \quad (a_{k,1}\mathbf{v}_{k,1} + \cdots + a_{k,10}\mathbf{v}_{k,10}) \,/ \| a_{k,1}\mathbf{v}_{k,1} + \cdots + a_{k,10}\mathbf{v}_{k,10} \|, \;\; k = 1, 2, 3 \;,$$

where $(\mathbf{v}_{k,1}, \ldots, \mathbf{v}_{k,10})$ are randomly chosen ten vectors from p eigenvectors of $\mathbf{\Sigma}$, which are conducted three times to form three outliers, as denoted by $k = 1, 2, 3$. Also $(a_{k,1}, \ldots, a_{k,10})$ are random numbers from $unif(-1, 1)$.

A data set that includes these three potential outliers is projected onto the principal component directions in Figure 3.14. In this figure, the potential outliers are marked by dot, diamond, and cross, and the 47 normal data points are marked by black circles. The three potential outliers are not shown as the outliers on the first four principal components.

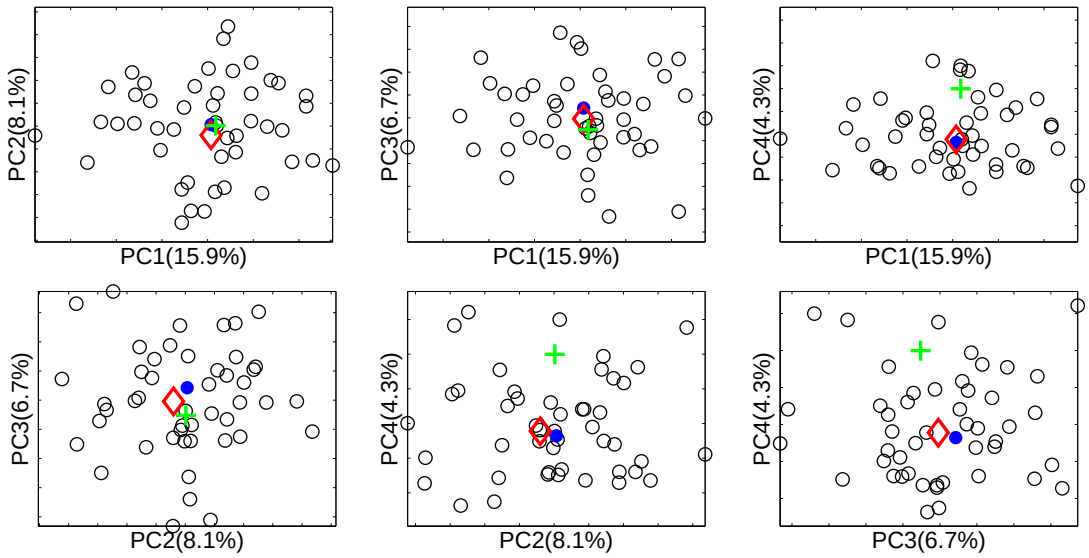


Figure 3.14. PC scatter plot of a data set that have multiple outliers (Setting I). The three potential outliers marked by dot, diamond, and cross are not appeared as the outliers on the first four principal components.

Further, our proposed method for the outlier detection is performed on this Setting I data. We test until $n_0 = 4$. Figure 3.15 displays the results. In each panel, the y-axis displays the sample quantiles which is named by the distances: CD, rMH, and MDP. Also, the x-axis displays the corresponding bootstrap quantiles which is named by CD*, rMH*, and MDP*. In the first row panels, CD does not separate the group outliers from the rest. This result is expected because the simulation is based on a mixed covariance structure, and the CD tends

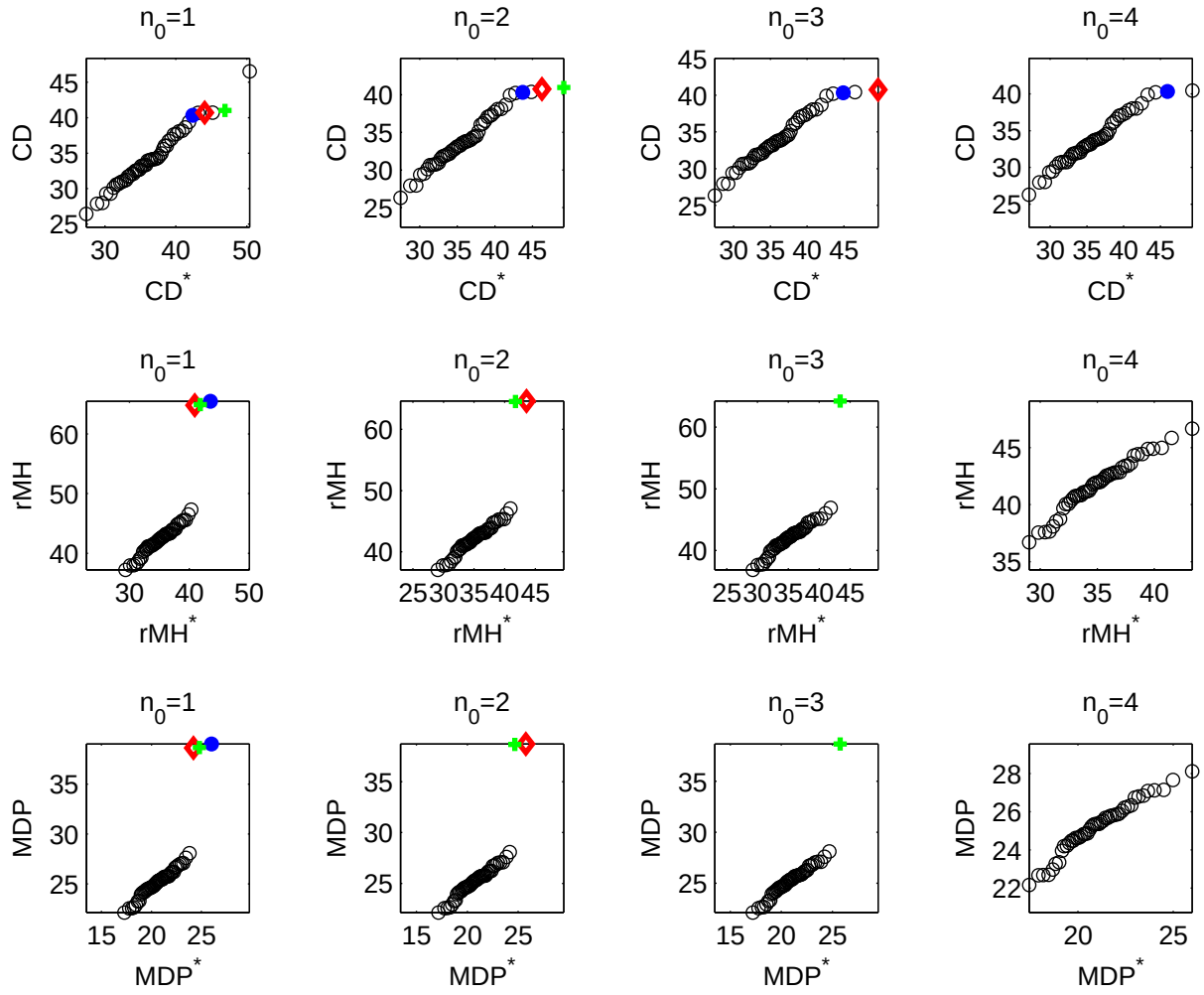


Figure 3.15. Our outlier detection method on multiple outliers (Setting I). In the second and third row panels, the ridge Mahalanobis (rMH) distance and the maximal data piling (MDP) distance clearly separate the group outliers. The sequential procedures lead to the conclusion that there are three outliers in the data set. At $n_0 = 4$, we do not see any evidence of the outliers.

to work for data with little correlation. In the second and third row panels, rMH and MDP separate the group outliers well. When the outlier is removed sequentially, it is clearly seen that there were three outliers in the data set because we do not see any evidence of other outliers at $n_0 = 4$.

In Setting II, we perform a simulation under a challenging situation that some outliers are hard to be detected; we insert three outliers that are rather clustered together. This setting is designed to see whether our method is able to handle the masking phenomenon that might occur by the multiple outliers. In the previous Gaussian data setting, let us construct three outliers following (3.20)-(3.22). But we made these group outliers closer to each other in terms of the pairwise angle. To do this, the random vectors $\mathbf{v}_{1,0}$, $\mathbf{v}_{2,0}$ and $\mathbf{v}_{3,0}$ are constructed as

$$\begin{aligned}
\mathbf{v}_{1,0} &\propto (a_1\mathbf{v}_1 + \cdots + a_6\mathbf{v}_6) + (a_7\mathbf{v}_7 + \cdots + a_{10}\mathbf{v}_{10}),\\
\mathbf{v}_{2,0} &\propto (b_1\mathbf{v}_1 + \cdots + b_6\mathbf{v}_6) + (a_7\mathbf{v}_7 + \cdots + a_{10}\mathbf{v}_{10}),\\
\mathbf{v}_{3,0} &\propto (c_1\mathbf{v}_1 + \cdots + c_6\mathbf{v}_6) + (a_7\mathbf{v}_7 + \cdots + a_{10}\mathbf{v}_{10}),
\end{aligned}$$

where $(\mathbf{v}_1, \ldots, \mathbf{v}_{10})$ are randomly chosen ten vectors from p eigenvectors of $\mathbf{\Sigma}$, and $(a_1, \ldots, a_{10})$, $(b_1, \ldots, b_6)$, and $(c_1, \ldots, c_6)$ are random numbers from $unif(-1, 1)$. The three vectors, $\mathbf{v}_{k,0}$, $k = 1, 2, 3$, are normalized to have the length of 1. Note that all of the three vectors are the linear combinations of $\mathbf{v}_1, \ldots, \mathbf{v}_{10}$ where the coefficients of $\mathbf{v}_1, \ldots, \mathbf{v}_6$ are different, and the other coefficients are the same. Such linear combinations produce closer pairwise angles among the vectors; i.e., $angle(\mathbf{v}_{1,0}, \mathbf{v}_{2,0}) = 63.30°$, $angle(\mathbf{v}_{1,0}, \mathbf{v}_{3,0}) = 59.07°$, and $angle(\mathbf{v}_{2,0}, \mathbf{v}_{3,0}) = 44.42°$.

A data set that includes the three close outliers is projected onto the first four principal components directions as shown in Figure 3.16. It is clearly seen that there are three outliers especially on PC3 direction where the direction accounts for 7.2% of total variance. Note that this plot also shows that the outliers are not seen on some other directions such as PC1, 2, and 4. Individual PC direction is not much informative in high dimension. Luckily, we found that the PC3 represents outliers in this example, but this may be because the outliers

are clustered together. In general, judging outliers based on the first few PC does not always work as previously shown in Figures 3.9 and 3.14.

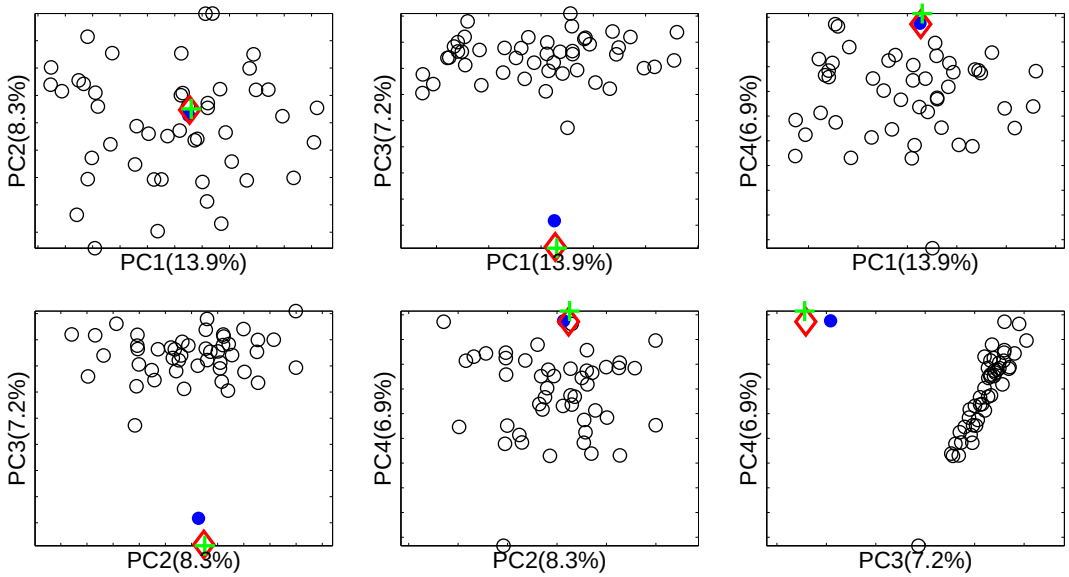


Figure 3.16. PC scatter plot of a data set that have close group outliers (Setting II). It is clearly seen that there are three outliers especially on PC3 direction.

With the Setting II data, we run the proposed outlier detection method. The results are shown in Figure 3.17. In the first row panels, the outliers do not clearly appear as expected since CD is not a good distance measure for non-spherical high dimensional data. In the second and third row panels, rMH and MDP distances show the similar patterns. At $n_0 = 1$, there seems to exist a single outlier, which is a false conclusion. We know there actually exist three outliers in the data. This truth reveals at $n_0 = 3$. Once the diamond is deleted, the cross is clearly seen as the outlier. That says that in the previous step $n_0 = 1$ and 2, there was the masking effect between two data points by the angles of the pair (44.42°).

Usually, when one mentions the masking effect in low dimension ($n > p$), it refers to the problem by using the Mahalanobis distance whose parameters are mean vector and covariance matrix. In other words, the mean and covariance estimations including the outlier(s) themselves can mislead the outlier rejection point (e.g., chi-square distribution criteria), resulting in hiding the true outlier. For this reason, many researchers have attempted to estimate the

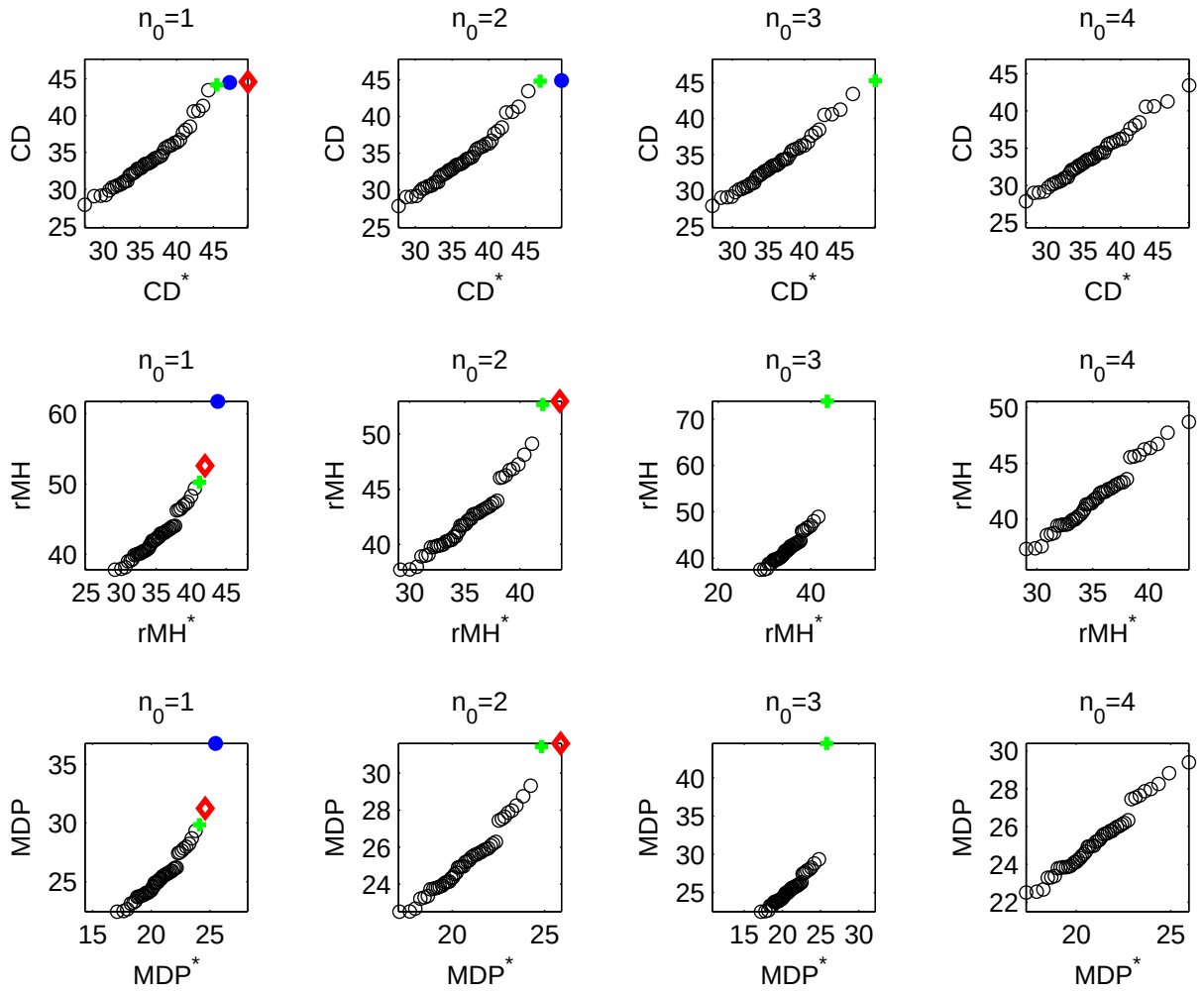


Figure 3.17. Our outlier detection method on multiple outliers (Setting II). In the second and the third row panels, based on $n_0 = 1$ and 2, one may conclude that there exist one outlier. However, at $n_0 = 3$, after the diamond is deleted, the cross is clearly seen as the outlier. Thus we know that there actually exist three outliers in the data. That says that in the previous step $n_0 = 1$ and 2, there was the masking effect between two data points by the angles of the pair.

robust Mahalanobis distance by finding the robust estimators of location and shape parameters (Maronna and Zamar, 2002; Rousseeuw and van Zomeren, 1990). Unfortunately, using the robust estimates do not entirely resolve the masking problem especially when dimension is larger. For example, Rousseeuw and van Zomeren (1990) suggests that their minimum volume ellipsoid estimator (MVE) will be useful for $n/p > 5$ since high dimensionality may cause the collinearity of a few sample points in a multivariate space, which makes detecting outliers harder.

In high dimension ($p > n$), on the other hand, the Mahalanobis distance is no longer in use due to the singularity problem of covariance estimator. Then, one may want to know when the masking phenomenon can occur in high dimension. In Setting II simulation, we addressed the angle between data vectors can be a source of masking effect. The masking effect by the angle is related to the characteristics of MDP distance. According to the MDP distance, the farness of an observation is defined by the orthogonal distance from a data point to the affine subspace that the rest of the data points generate. That means that the masking effect can be arisen when the distance from an outlying data point to the affine subspace is underestimated. Such situation may occur when the angle between the outlying data vector and some other vectors belonging to the affine subspace is much smaller than 90 degrees. In that situation, the perpendicular distance of the outlier to the rest is substantially underestimated, as a result one cannot detect the outlier.

Fortunately, however, such masking phenomenon is highly unlikely when the dimension is large. Hall et al. (2005) addressed some geometric property of high dimensional data vector as $p \to \infty$ for fixed n. Among the facts, we would like to point out that all pairwise angles of the data vectors are approximately perpendicular as dimension increases. Importantly, the fact that the data vectors are almost perpendicular implies that there is extremely less probability of masking effect by the angles between two outliers. Furthermore, the almost perpendicular data vectors are commonly observed in real high-dimensional data. Figure 3.18 shows the histograms of all pairwise angles of the data vectors in real data. The real

data sets are listed in Table 3.1. Each data set denoted by data (i, j) is centered by the mean vector, and then all pairwise angles among the observations are calculated. In this figure. we see that all pairwise angles are piled towards 90 degrees. Therefore, the masking effect is not a problem in the usage of MDP distance as a distance measure for high dimensional data.

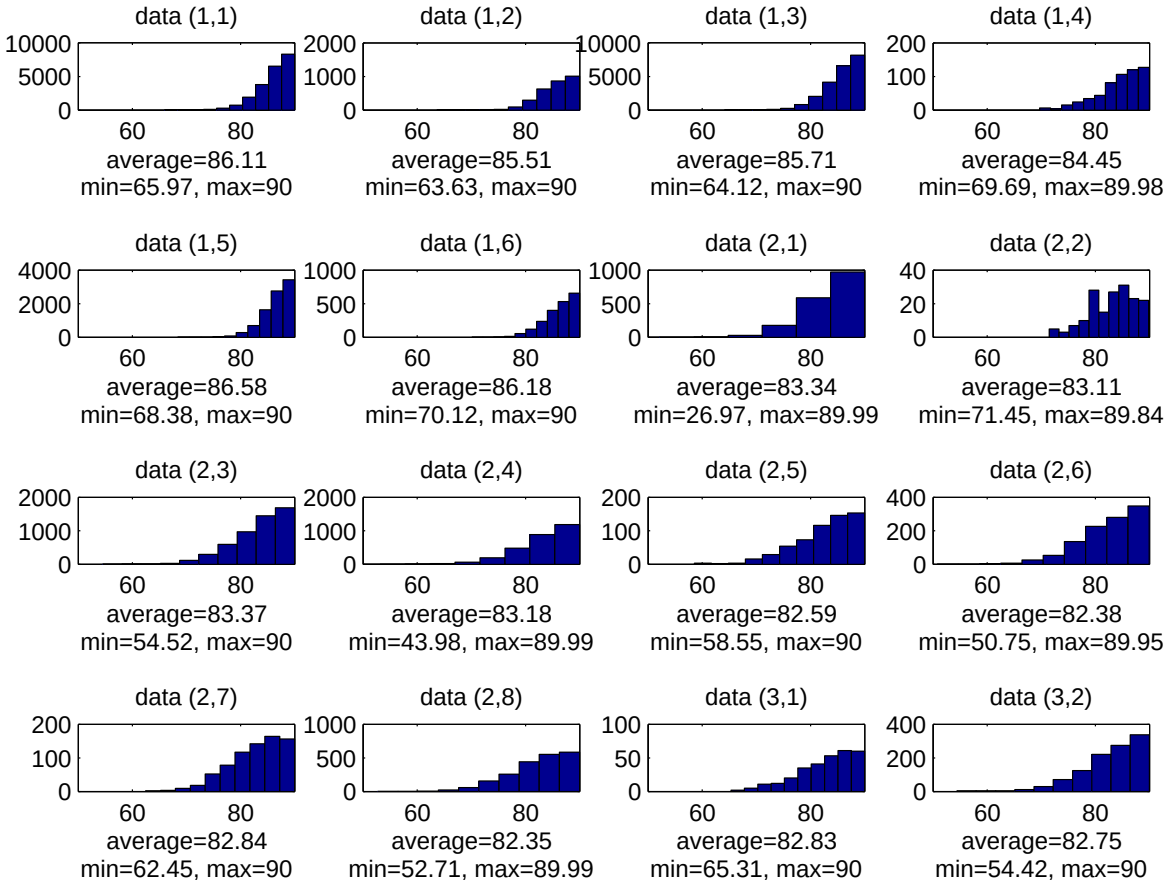


Figure 3.18. Histograms of all pairwise angles among observations in real data sets. The real data sets are listed in Table 3.1. We see that most pairwise angles are piled towards 90 degrees.

3.6 Real Data Example

3.6.1 Data Sets

Our outlier detection method is implemented with real microarray data: breast cancer data (Data 1), lung cancer data (Data 2) and luekemia data (Data 3). Those three data have more than one batch, and each batch is partitioned by two biological classes. For example, "breast cancer data, ER+ group" is denoted by data (1,1), "breast cancer data, ER- group" is denoted by data (1,2) and the like. All data sets denoted by data (i, j) are listed in Table 3.1 as below. More details of Data 1 and 2 are given in Table 2.3 in Chapter 2. The luekemia data (Data 3) is publicly available from Dettling (004b). Note that we only select the top 1000 genes based on the variances since large variance often indicates the presence of the outlier.

Table 3.1
List of data sets

Data id	Data name	Sample size	Batch	Source
(1,1)	Breast cancer ER+	209	Batch1	
(1,2)	Breast cancer ER-	77		
(1,3)	Breast cancer ER+	211	Batch2	Table 2.3
(1,4)	Breast cancer ER-	34		
(1,5)	Breast cancer ER+	134	Batch3	
(1,6)	Breast cancer ER-	64		
(2,1)	Lung cancer dead	60	Batch1	
(2,2)	Lung cancer alive	19		
(2,3)	Lung cancer dead	102	Batch2	
(2,4)	Lung cancer alive	76		Table 2.3
(2,5)	Lung cancer dead	35	Batch3	
(2,6)	Lung cancer alive	47		
(2,7)	Lung cancer dead	39	Batch4	
(2,8)	Lung cancer alive	65		
(3,1)	Luekemia 1	25		Dettling (004b)
(3,2)	Luekemia 0	47		

3.6.2 Outlier Detection Results in Real data

We apply the proposed method on the 16 real data sets above in order to examine whether those data sets include outliers. Among these data sets, some suspicious observations are found in data (1,5), (2,2), (2,5), and (2,6).

In the example of Figure 3.19, the qq plots from the proposed method are drawn for data (1,5). In each panel, the data points are labeled by the observation number, 1 to 134. In the second and third row panels, we can see that there are three outliers: 39, 68 and 113. At $n_0 = 1$, the observation 39 is suggested for the outlier in both rMH and MDP distances. At $n_0 = 2$ and 3, the observations 68 and 113 are suggested for the other outliers in both distances but with a difference sequence. The other results for data (2,2), (2,5), and (2,6) are depicted in Figures 3.21, 3.23, and 3.25, respectively.

Furthermore, these results are compared to the results of the PCout method proposed by Filzmoser et al. (2008). The PCout algorithm finds the outliers based on the principal components which contribute to 99 % of total variance. This approach makes sense since the dimension usually comes down to less than sample size, which makes the analysis of data easier. The process of searching outliers on these PC's involves mainly two parts considering location outliers and scatter outliers. The first part to detect the location outlier utilizes a kurtosis measure proposed by Peña and Prieto (2001). Extremely large or small kurtosis coefficients on some directions indicates the presence of the outliers. Therefore, one weights the directions that clearly separate the outliers based on the kurtosis coefficients. In the second part with regard to scatter outliers, they consider another weight for each observation, utilizing the translated biweight function. See Filzmoser et al. (2008) for the technical details.

The comparisons are shown in Table 3.2. In the comparison of data (1,5), for example, the PCout method finds total 15 observations, which is more than what our method finds (three outliers). Also in the rest of the data sets, the PCout method tends to find more outliers. These results are depicted in Figures 3.22, 3.24 and 3.26, respectively.

Table 3.2

Comparison with PCout

Data id	Our method	PCout
(1,5)	39, 68, 113	39, 68, 113 and others (total 15 observations)
(2,2)	5, 9	3, 5, 9
(2,5)	1, 2	1, 2, 15, 28
(2,6)	13	6, 13, 20, 32, 33, 47

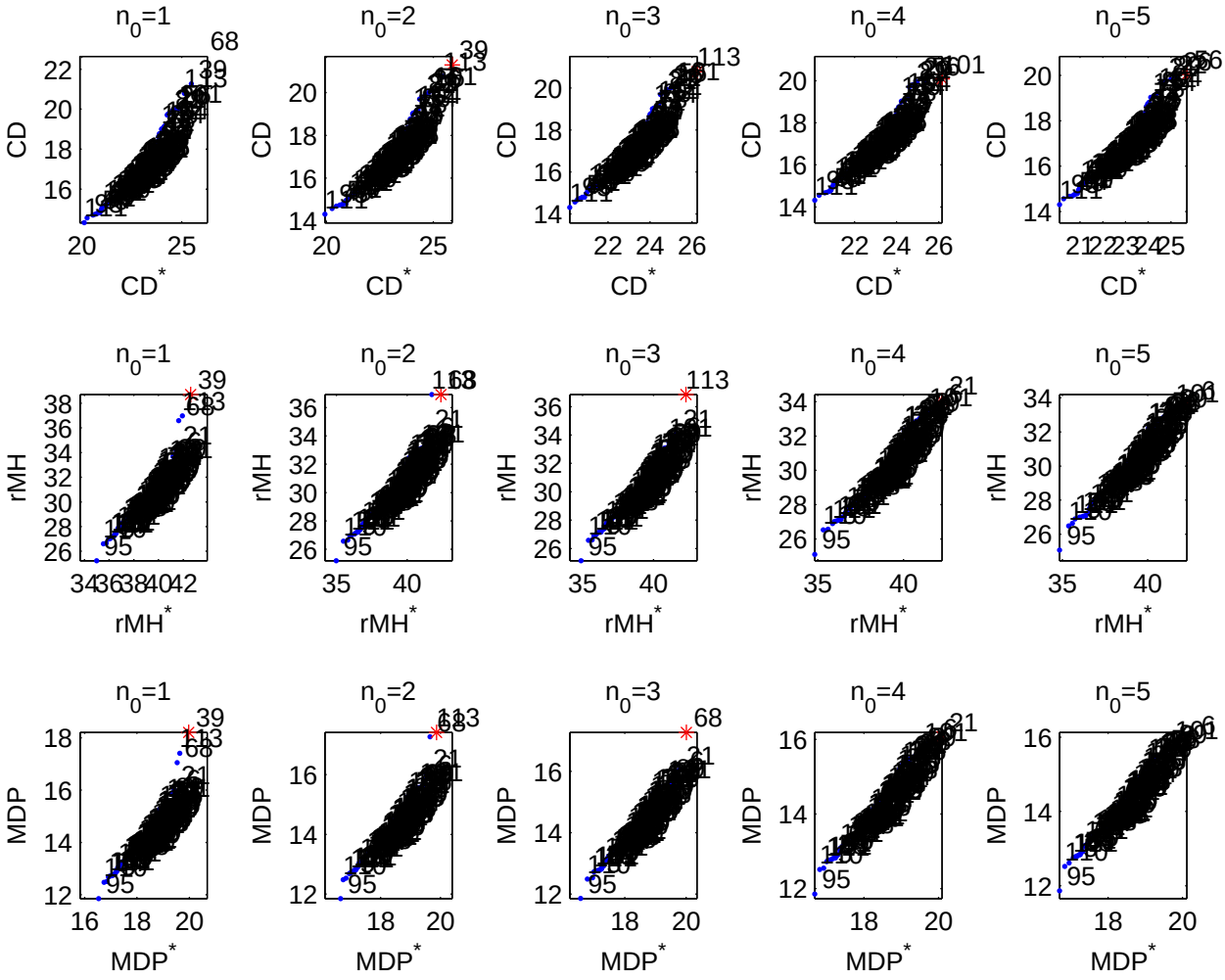


Figure 3.19. Outliers in data (1,5) with our method. Three outliers are found: 39, 68, 113.

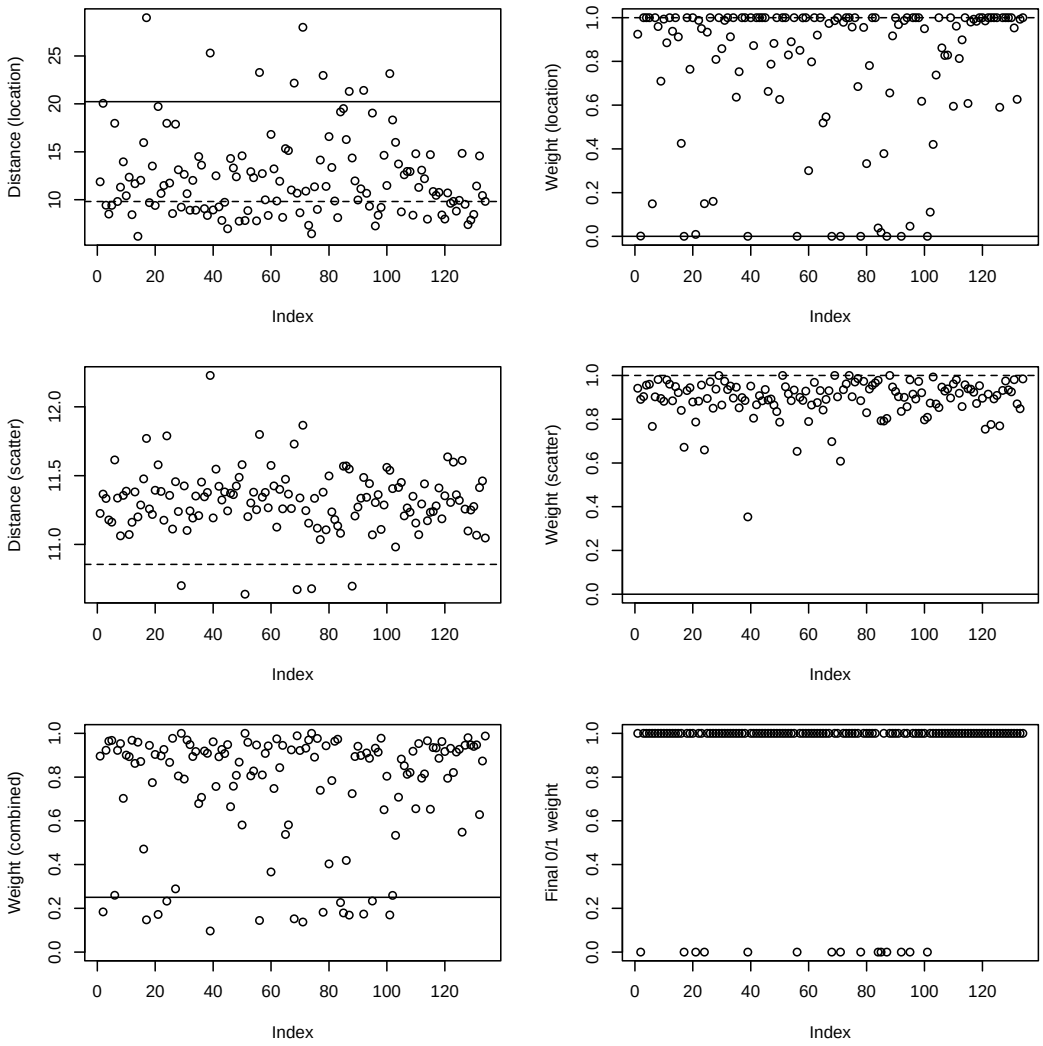


Figure 3.20. Outliers in data (1,5) with PCout. In the right bottom panel, total 15 outliers are found: 39, 68, 113 and others.

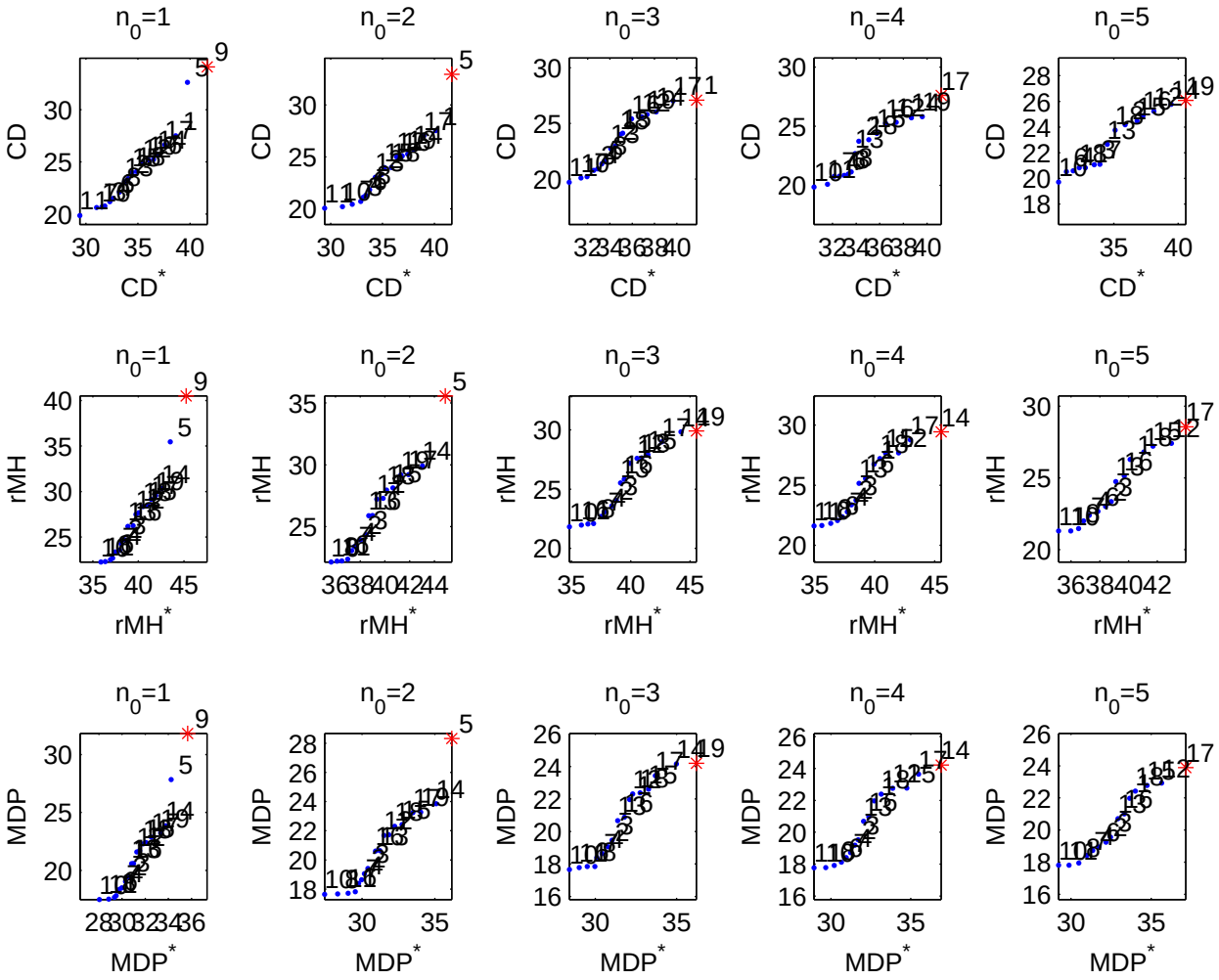


Figure 3.21. Outliers in data (2,2) with our method. Two outliers are found: 5, 9.

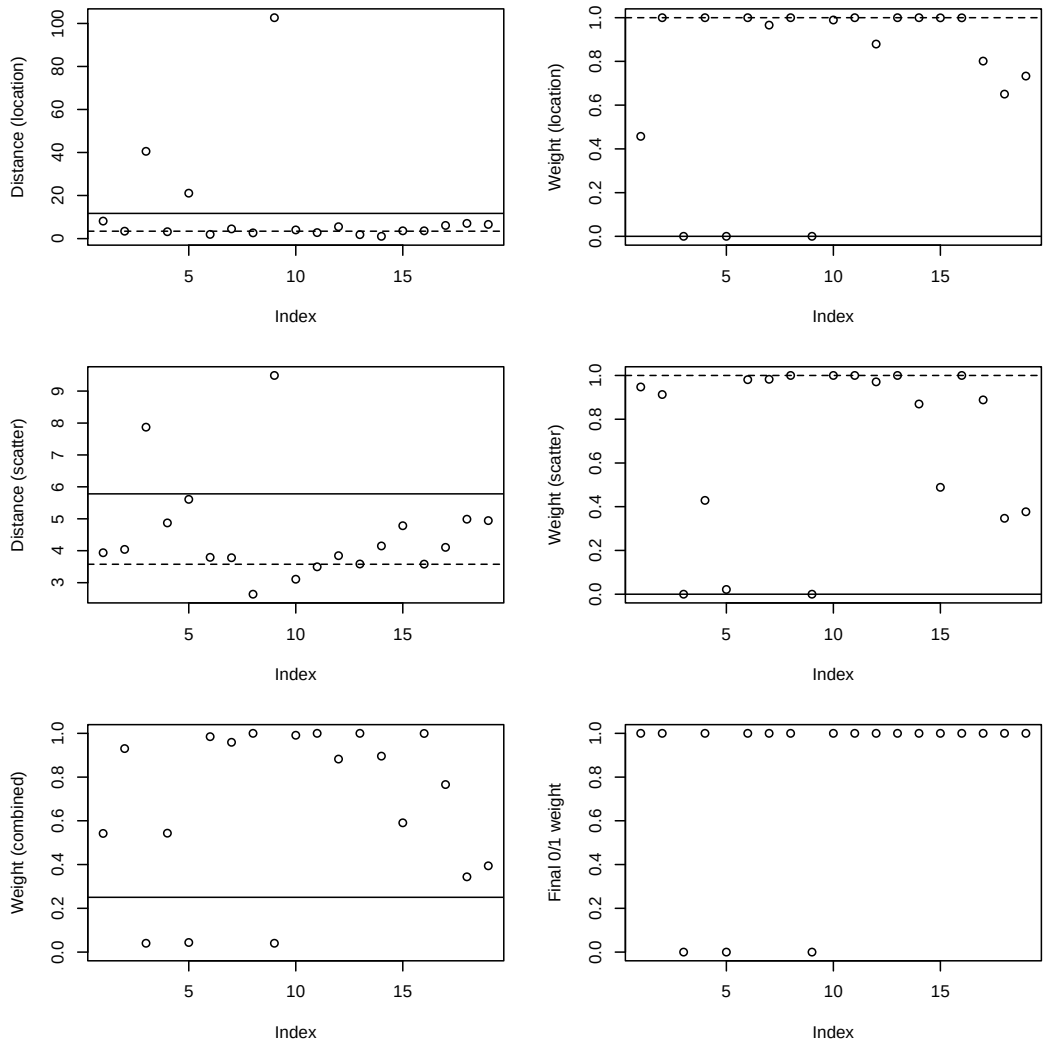


Figure 3.22. Outliers in data (2,2) with PCout. In the right bottom panel, three outliers are found: 3, 5, 9.

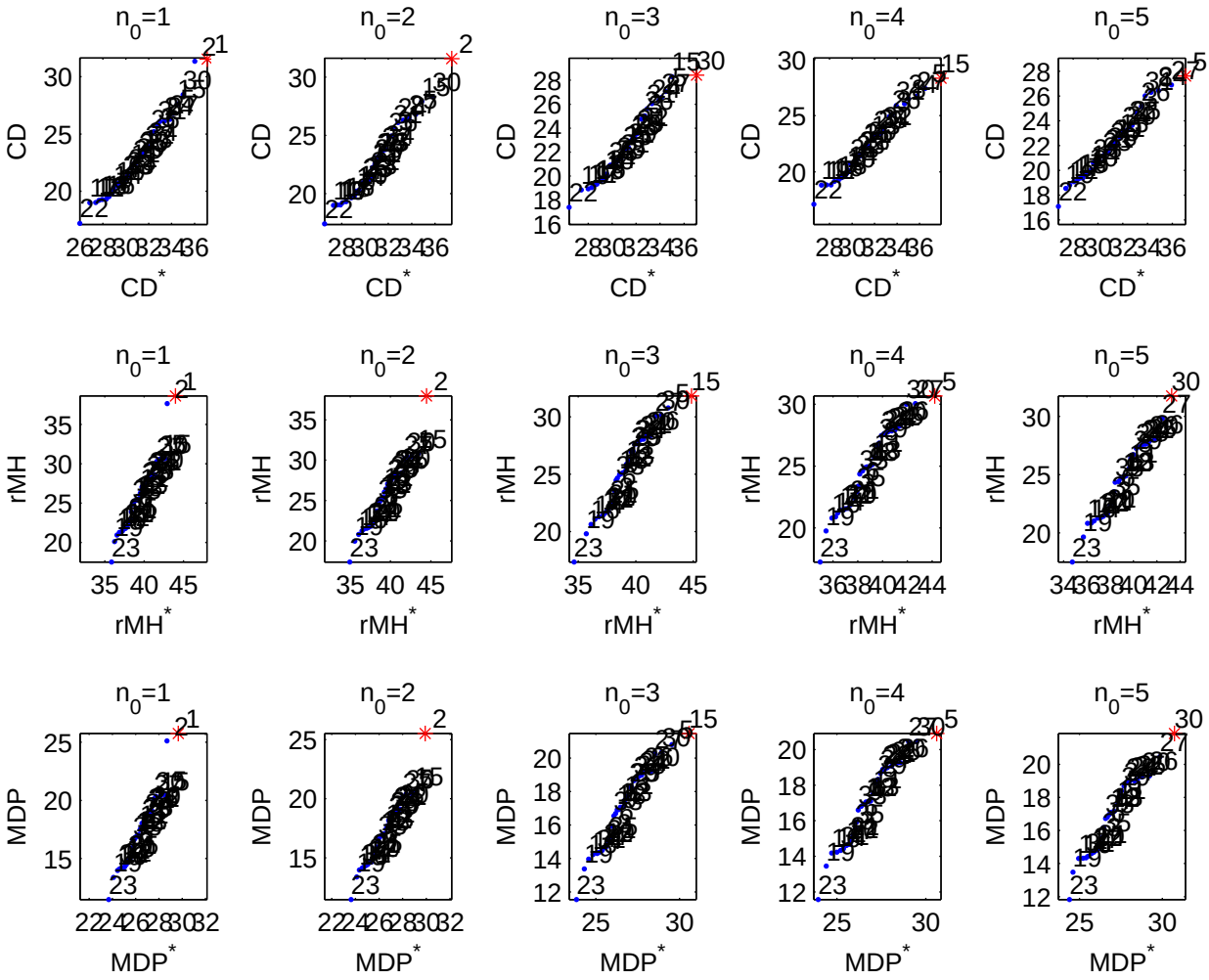


Figure 3.23. Outliers in data (2,5) with our method. Two outliers are found: 1, 2.

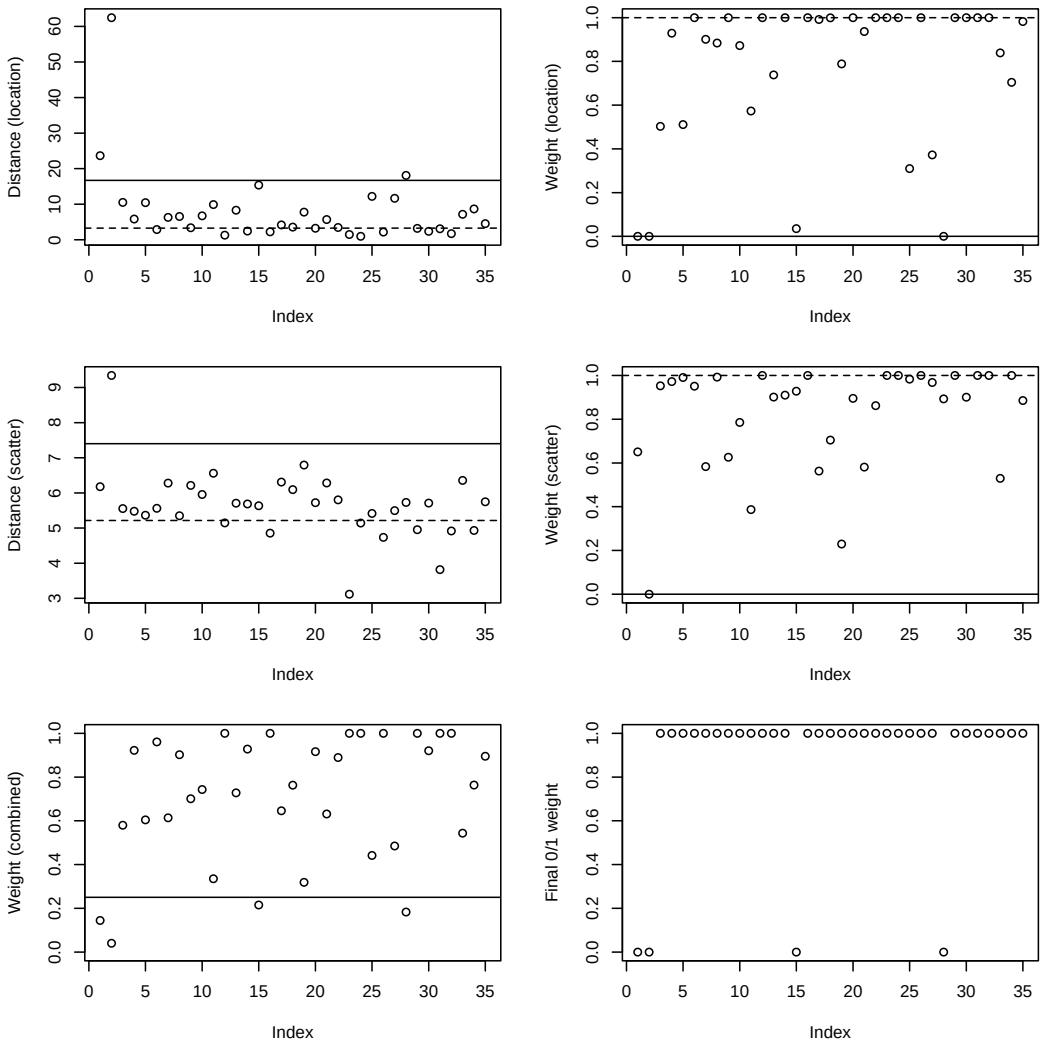


Figure 3.24. Outliers in data (2,5) with PCout. In the right bottom panel, four outliers are found: 1, 2, 15, 28.

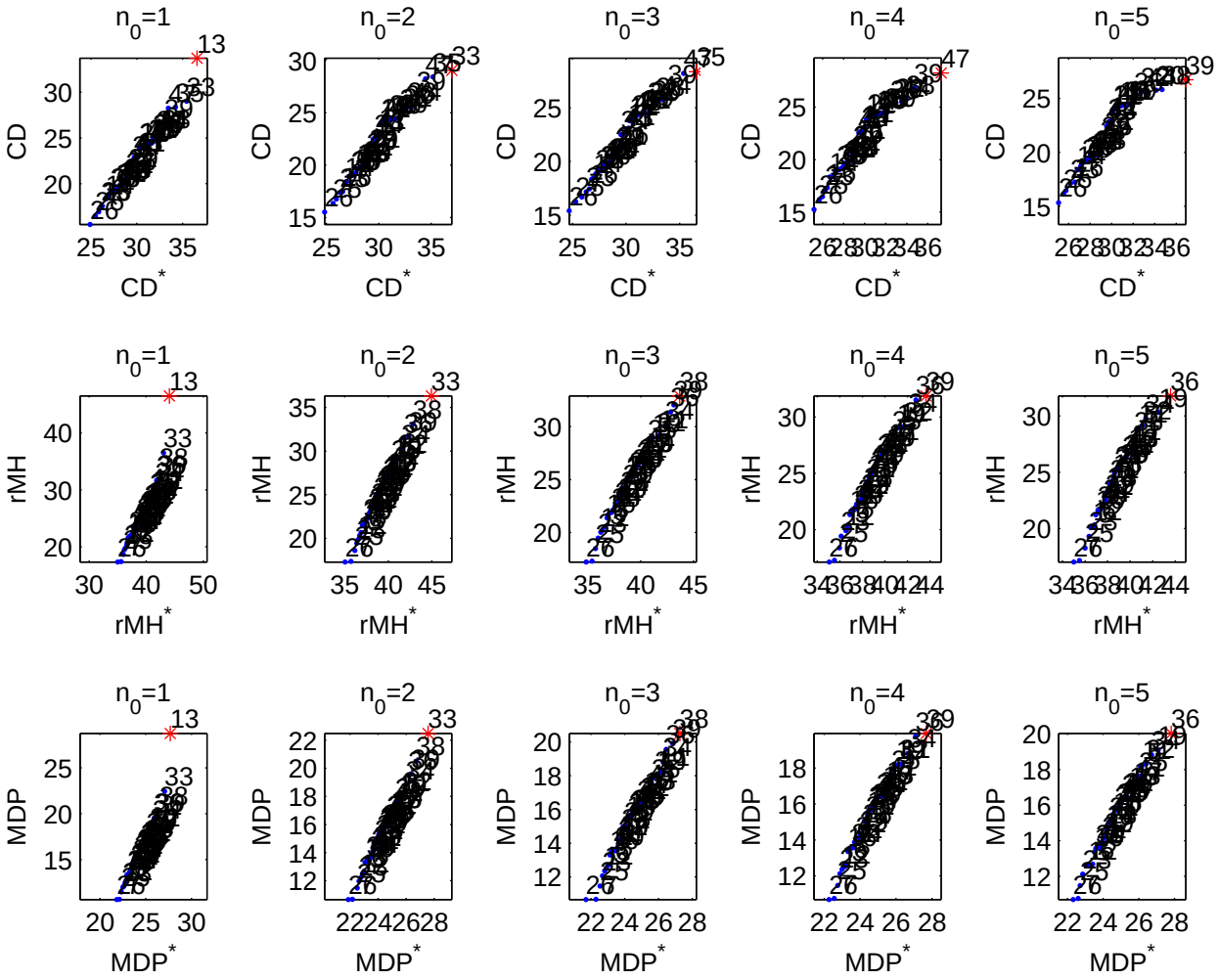


Figure 3.25. Outliers in data (2,6) with our method. one outlier is found: 13.

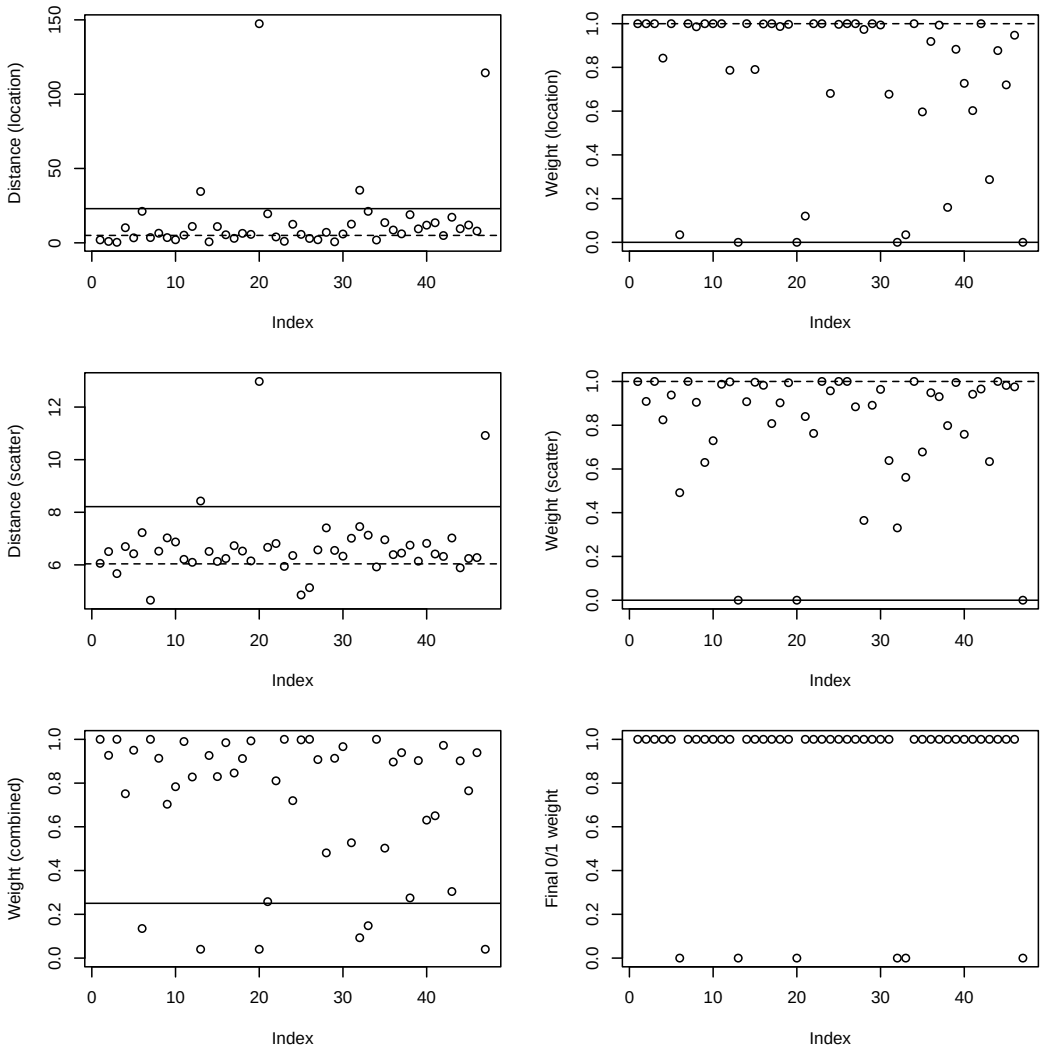


Figure 3.26. Outliers in data (2,6) with PCout. In the right bottom panel, six outliers are found: 6, 13, 20, 32, 33, 47.

3.7 Conclusion

In this chapter, we propose an outlier detection method for HDLSS data, which has not been dealt with in previous works. Specifically, we suggest using three types of distance measures that are useful in the high dimensional space. Also a parametric bootstrap method is applied for introducing a qq plot in which we expect outliers to come out. Through the theoretical properties and a simulation study, the proposed method is shown to be effective in the detection of multiple outliers as well as a single outlier.

Bibliography

Ahn, J., Lee, M., and Lee, J. (2013). Outlier detection in high dimension, low sample size data. *Working paper*.

Ahn, J., Lee, M., and Yoon, Y. (2011). Clustering high dimension, low sample size data using the maximal data piling distance. *Statistica Sinica*, 22(2):443–464.

Ahn, J. and Marron, J. S. (2010). The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259.

Ahn, J., Marron, J. S., Muller, K. E., and Chi, Y.-Y. (2007). High dimension, low sample size geometric representation holds under mild conditions. *Biometrika*, 94:760–766.

Altman, N. (2009). Batches and blocks, samples pools and subsamples in the design and analysis of gene expression studies. In *Batch effects and noise in microarray experiments*, chapter 4, pages 33–50. Wiley.

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

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data*. Wileys and Sons, 3rd edition.

Becker, C. and Gatheer, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, 94(447):947–955.

Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton University Press, Princeon,NJ.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., and Marron, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–114.

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

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Bolstad, B. M., Irizarry, R. A., Astrand, M. A., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformetics*, 19(2):185–193.

Böttinger, E. P., Ju, W., and Zavadil, J. (2002). Applications for microarrays in renal biology and medicine. *Experimental Nehprology*, 10(2):93–101.

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

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011). Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLosOne*, 6(2):e17238.

Cheng, C., Shen, K., Song, C., Luo, J., and Tseng, G. C. (2009). Ratio adjustment and calibration scheme for gene-wise normalization to enhance microarray inter-study prediction. *Bioinformatics*, 25(13):1655–1661.

Claesson, M. J., O'Sullivan, O., Wang, Q., Nikkila, J., Marchesi, J. R., Smidt, H., de Vos, W. M., Ross, R. P., and O'Toole, P. W. (2009). Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS ONE*, 4:e6669. doi:10.1371/journal.pone.0006669.

Davey, R., Savva, G., Dicks, J., and Roberts, I. N. (2007). MPP: a microarray-to-phylogeny pipeline for analysis of gene and marker content datasets. *Bioinformatics*, 23:1023–1025.

Davies, L. (1992). The asymptotics of rousseeuw's minimum volume ellipsoid estimator. *The Annals of Statistics*, 20(4):1828–1843.

Davies, P. L. (1987). Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15(3):1269–1292.

Dettling, M. (2004b). Bagboosting for tumor classification with gene expression data. *Bioinformatics*, 20(18):3583–3593.

Dettling, M. and Buehlmann, P. (2004). Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, 90(1):106–131.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. American Mathematical Society. Math Challengs of the 21st Century.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In Bickel, P. J., Doksum, K., and Hodges, J. L., editors, *A Festschrift for Erich Lehmann*, pages 157–184. Wadsworth, Belmont, CA.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1:107–129.

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences of the USA*, 103:5923–5928.

Fan, J., Fan, Y., and Lv, J. (2006). High dimensional covariance matrix estimation using a factor model. *Technical Report*, Available on the arXiv preprint server:math.ST/0701124.

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

Fan, J., Liao, Y., and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320–3356.

Fan, J. and Lv, J. (2007). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society B*, 70(5):849–911.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, 52:1694–1711.

Friedman, J. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175.

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

Golub, G. H. and Van Loan, C. F. (1996). *Matrix computations*. Johns Hopkins, Baltimore. p.257-258.

Guo, Y., Hastie, T., and Tibshirani, R. (2005). Regularized discriminant analysis and its application in microarrays. *Biostatistics*, 1(1):1–18.

Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension low sample size data. *Journal of the Royal Statistical Society B*, 67(3):427–444.

Hardin, J. and Rocke, D. (2005). The distribution of robust distances. *Journal of Computational Graphical Statistics*, 14:928–946.

Hastie, T. and Tibshirani, R. (2004). Efficient quadratic regularization for expression arrays. *Biostatistics*, 5(3):329–340.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hawkins, D. M. (1980). *Identification of Outliers*. Chapmand and Hall.

Heidecker, B. and Hare, J. M. (2007). The use of transcriptomic biomarkers for personalized medicine. *Heart Failure Reviews*, 12(1):1–11.

Huang, H., Liu, Y., Todd, M. J., and Marron, J. S. (2011). Multiclass distance weighted discrimination with application to batch adjustment. Submitted.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Annals of Statistics*, 37(6B):4104–4130.

Lee, C. H. and Macgregor, P. F. (2004). Using microarrays to predict resistance to chemotherapy in cancer patients. *Pharmacogenomics*, 5(6):611–625.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11:733–739.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161.

Lopuhaä, H. P. (1989). On the relation between s-estimators and m-estimators of multivariate location and covariance. *The Annals of Statistics*, 17(4):1662–1683.

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., Shi, T., Tong, W., Shi, L., Hong, H., Zhao, C., Elloumi, F., Shi, W., Thomas, R., Lin, S., Tillinghast, G., Liu, G., Zhou, Y., Herman, D., Li, Y., Deng, Y., Fang, H., Bushel, P., Woods, M., and Zhang, J. (2010). A comparison of batch effect removal methods for enhancement of prediction performance using MARQ-II microarray gene expression data. *The Pharmacogenomics Journal*, 10:278–291.

Maronna, R., Martin, R., and Yohai, V. (2006). *Robust Statistics: Theory and Methods*. Wiley.

Maronna, R. and Zamar, R. (2002). Robust estimates of location and dispersion for high-dimensional data sets. *Technometrics*, 44(4):307–317.

Maronna, R. A. and Youhai, V. (1995). The behavior of the stahel-donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341.

Marron, J. S., Todd, M., and Ahn, J. (2007). Distance weighted discrimination. *Journal of the American Statistical Association*, 102:1267–1271.

McCall, M. N., Bolstad, B. M., and Irizarry, R. A. (2010). Frozen robust multiarray analysis(fRMA). *Biostatistics*, 11(2):242–253.

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365(9458):488–492.

Montaner, D., Minguez, P., Al-Shahrour, F., and Dopazo, J. (2009). Gene set internal coherence in the context of functional profiling. *BMC Genomics*, 10:doi:10.1186/1471–2164–10–197.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley.

Peña, D. and Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43(3):286–310.

Rousseeuw, P. and van Driessen, K. (1999). A fast algorithms for the minimum covariance determinant estimator. *Technometrics*, 41:212–223.

Rousseeuw, P. J. (1984). Least median of squares resgression. *Journal of the American Statistical Association*, 79:871–881.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In Grossmann, W., Pflug, G., Vincze, I., and Wertz, W., editors, *Mathematical Statistics and Applications*, volume B, pages 283–297. D. Reidel Publishing Company.

Rousseeuw, P. J. (1987). *Robust Regression and Outlier detection*. John Wiley.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639.

Rousseeuw, P. J. and Yohar, V. (1984). Robust regression by means of S-estimators. In Franke, J., Härdle, W., and Martin, R. D., editors, *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statistics No.26, pages 256–272. Springer-Verlag, New York.

Scherer, A. (2009). *Batch effects and noise in microarray experiments: sources and solutions*. Wiley.

Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computational Statistics and Data Analysis*, pages 6535–6542.

Shabalin, A. A., Tjelmeland, H., Fan, C., Perou, C. M., and Nobel, A. B. (2008). Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, 24(9):1154–1160.

Shao, J., Wang, Y., Deng, X., and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39(2):1241–1265.

Shedden, K., Taylor, J. M. G., Enkemann, S. A., Tsao, M. S., Yeatman, T. J., Gerald, W. L., Eschrich, S., Jurisica, I., Venkatraman, S. E., Meyerson, M., Kuick, R., Dobbin, K. K., Lively, T., Jacobson, J. W., Beer, D. G., Giordano, T. J., Misek, D. E., Chang, A. C., Zhu, C. Q., Strumpf, D., Hanash, S., Shepherd, F. A., Ding, K., Seymour, L., Naoki, K., Pennell, N., Weir, B., Verhaak, R., Ladd-Acosta, C., Golub, T., Gruidl, M., Szoke,

J., Zakowski, M., Rusch, V., Kris, M., Viale, A., Motoi, N., Travis, W., and Sharma, A. (2008). Gene-expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine*, 14(8):822–827.

Sims, A. H., Smethurst, G. J., Hey, Y., Okoniewski, M. J., Pepper, S. D., Howell, A., Miller, C. J., and Clarke, R. B. (2008). The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta analysis and prediction of prognosis. *BMC Medical Genomics*, 1(42).

Srivastava, M. S. and Yanagihara, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis*, 101:1319–1329.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirova, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43).

Tan, P. K., Downey, T. J., Spitznagel, E. L., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M., and Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, 31(19):5676–5684.

The Gene Ontology Consortium (2008). The gene ontology project in 2008. *Nucleic Acids Research*, 36(Database issue):D440–D444.

Vachani, A., Nebozhyn, M., Singhal, S., Alila, L., Wakeam, E., Muschel, R., Powell, C. A., Gaffney, P., Singh, B., Brose, M. S., Litzky, L. A., Kucharczuk, J., Kaiser, L. R., Marron, J. S., Showe, M. K., Albelda, S. M., and Showe, L. C. (2007). A 10-gene classifier for distinguishing head and neck squamous cell carcinoma and lung squamous cell carcinoma. *Clinical Cancer Research*, 13(10):2905–2915.

van der Lann, M. J. and Bryan, J. (2001). Gene expression analysis with the parametric bootstrap. *Biostatistics*, 2(4):445–461.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Xu, L., Tan, A. C., Naiman, D. Q., Geman, D., and Winslow, R. L. (2005). Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinfometics*, 21(20):3905–3911.

Yasrebi, H., Sperisen, P., Praz, V., and Bucher, P. (2009). Can survival prediction be improved by merging gene expression data sets? *PLoS ONE*, 4(10):e7431.

Yata, K. and Aoshima, M. (2010). Effective PCA for high-dimension, low-samplesize data with singular value decomposition of cross data matrix. *Journal of Multivariate Analysis*, 101:2060–2077.