New Statistical Tools for Microarray Data and Comparison with Existing Tools

Xuxin Liu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2007

Approved by

Advisor: Dr. J. S. Marron

Reader: Dr. Andrew B. Nobel

Reader: Dr. Yufeng Liu

Reader: Dr. Haipeng Shen

Reader: Dr. Charles M. Perou

ABSTRACT

XUXIN LIU: New Statistical Tools for Microarray Data and Comparison with Existing Tools

(Under the direction of Dr. J. S. Marron)

Microarray technologies have gained tremendous interest from researchers in recent years. The problem we are interested in is how to combine two microarray data, which have systematic batch differences. The reason for the combination is that the combined data set contains more samples which will give improved statistical power. This dissertation covers two topics about microarray batch adjustment. The first topic is about the visualization of paired High Dimension Low Sample Size (HDLSS) data. We propose two interesting directions: the Canonical Parallel and the Canonical Orthogonal Directions (CPD & COD). This pair of directions gives an insightful 2-d parallel view for understanding paired HDLSS data sets. The CPD can be used for adjusting the batch differences. An application to the NCI60 cell lines data shows good performance of this method. The second topic is about the comparison between three commonly used batch adjustment methods: the Support Vector Machine (SVM), the Distance Weighted Discrimination (DWD), and the Prediction Analysis of Microarray (PAM). We show that SVM has some serious problems for the HDLSS data. The DWD method is much more robust than PAM under the Unbalanced Subgroup Model.

The mathematical studies made in this dissertation are in the area of HDLSS asymptotics, in the sense that the sample sizes are fixed and the dimension (the number of genes) goes to infinity. Hall et. al (2004) have studied the geometric structure of the data when the dimension is high. In this dissertation, we study the geometric structure of the data under more complicated models. In the first topic, we give the conditions for the consistency and the strong inconsistency of the CPD under the Linear Shift Model. This model reflects the effects of systematic biases and the random measurement errors. In the second topic, we compare the PAM and the DWD method using the Unbalanced Subgroup Model. Both methods are biased when the dimension goes to infinity. However, DWD is shown to be consistently more robust than PAM. We give the quantitative bias of them.

Keywords: Microarray Batch Adjustment, Principal Component Analysis, Exploratory Data Analysis, High Dimension Low Sample Size Data Analysis, Data Discrimination Methods, Distance Weighted Discrimination, Support Vector Machine, Predication Analysis of Microarray, High Dimension Asymptotics.

ACKNOWLEDGEMENTS

I would like to express my deepest thanks and appreciation to my advisor, Dr. J.S. Marron for his tremendous help during my research and the writing of this dissertation. Dr. Marron lead me to this interesting field of statistical analysis on microarrrays. His sound advice and guidance were invaluable during my research. I would also like to thank my committee members: Dr. Andrew B. Nobel, Dr. Yufeng Liu, Dr. Charles M. Perou and Dr. Haipeng Shen for their suggestions and comments.

This dissertation is dedicated to my mom and dad for their support and encouragement.

CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction and Background

This Chapter is organized as follows: Section 1.1 gives an introduction to microarray data. Section 1.2 discusses the High Dimensional, Low Sample Size (HDLSS) problem. It introduces the multivariate view of microarray data. It also illustrates the principal component direction visualization for HDLSS data. In Section 1.3, the NCI60 cancer cell line data sets are introduced. They will be used for illustration of many different points in the rest of this dissertation. This section also describes the statistical analysis problem of batch adjustment for microarray data sets. Several batch adjustment methods are reviewed and compared. Section 1.4 gives the organization of the rest of the dissertation.

1.1 Microarray Data Introduction

Genes and their products (such as RNA and protein) play an important role in the function of living organisms. The traditional methods of molecular biology generally worked on a “one gene studied in one experiment” basis. The cost was extremely high to get the expressions for thousands of genes, which meant the “whole picture” of gene function was hard to obtain. In recent years, a collection of new technologies called DNA microarrays has attracted tremendous interest among biologists; see Schena *et al.* (1995), Eisen and Brown (1999), and Alter *et al.* (2000). These technologies permit the expression profiling of thousands of genes simultaneously. This highly reduces the costs of collecting gene expression data. Thus researchers can monitor the whole genome and study the interactions among thousands of genes.

Usually a microarray chip contains tens of thousands of spots on a chip of glass or some other material. DNA molecules are immobilized and attached to these spots. There are at least two currently most widely-used formats of DNA microarray technology. One is **single channel microarray**, the other is **two-channel microarray**. An example of single channel microarray is **Oligonucleotide microarray**, i.e. **Affymetrix microarray (Affy)**, developed at Affymetrix, Inc. Affymetrix microarray technology uses synthetic DNA fragments, i.e. oligonucleotides, consisting of around 25 bases. A technique called photolithographical array production is applied to synthesize the oligonucleotides on the chip. An example of two-channel microarray is **cDNA Microarray (cDNA)** , developed at Stanford University. cDNA molecules are usually 0.2 to 5 kb long and are immobilized on the chip using robot spotting (printing).

A microarray experiment consists of three steps: sample preparation and labeling; sample hybridization and washing; and microarray image scanning and processing. We will take the cDNA microarray as a basis for a general discussion of these steps. Other technologies such as the Affymetrix microarray follow similar principals.

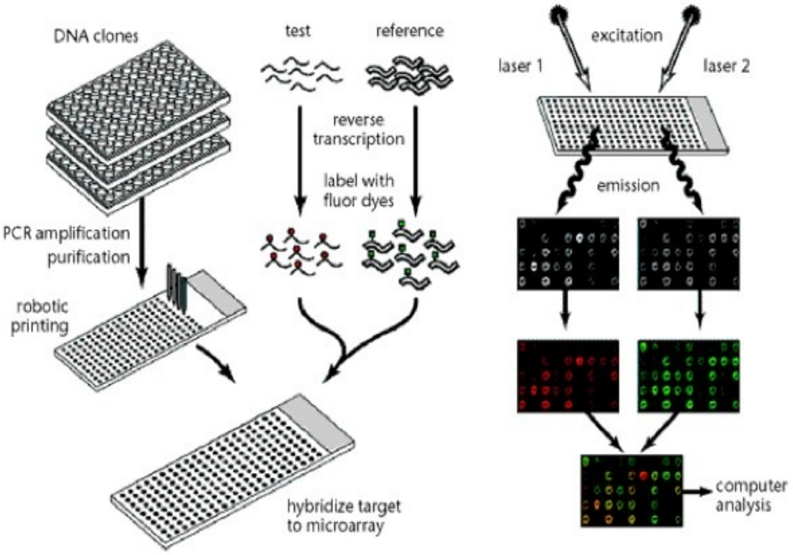


Figure 1.1: Shows the scheme of a cDNA microarray experiment. This figure is taken from Duggan *et al.* (1999).

The general scheme of a cDNA microarray experiment is illustrated in Figure 1.1. For gene expression levels studies, each spot on the chip is representative of a certain gene or a transcript. The total mRNA from the cells in test tissue and in reference tissue is extracted and labeled with two different fluorescent dyes separately, e.g. green dye for the mRNA from the test tissue and red dye for the mRNA from the reference tissue. More precisely, the labeling is done on the nucleotides that are complementary to the isolated mRNA. All the extracted mRNA from both tissues are prepared and hybridized to the immobilized molecules on the spots. The mRNA that did not bind to the immobilized molecules during the hybridization process is washed away. The relative abundance of hybridized molecules on a defined spot can be determined by measuring the fluorescent level of this spot. This is done done by scanning the chip twice with red and green lasers. If the mRNA from the test tissue is abundant, the spot will be green; if the mRNA from the reference tissue is abundant, the spot will be red. If both are equally abundant, the spot will be yellow. If neither are in abundance, the spot will appear black. Thus the relative gene expression level at each spot can be estimated from the fluorescence intensities, i.e. the color for this spot. This method has the advantage of measuring the expression levels for thousands of genes in one experiment.

A microarray experiment produces massive amounts of gene expression data. Figure 1.2 illustrates the organization of one microarray data set. The top row displays the sample (or individual) annotations. The first column on the left shows the gene annotations. The large rectangle displays the gene expression matrix, which is organized in this paper as a $d \times n$ matrix X, where d is the number of genes (rows), and n is the number of the samples (or individuals, i.e. columns). Thus $X_{i,j}$ is the expression value for the ith gene and jth sample (or individual). Sometimes, a microarray data set is organized using the transpose of the above matrix , e.g. each column as a gene and each row as an array (individual); see for example in Irizarry *et al.* (2003).

1.2 High Dimension Low Sample Size data Visualization

There are at least two important view points for the analysis of microarray data $X_{d\times n}$. One is the **gene by gene view**. It treats the gene expression matrix $X_{d\times n}$ as d separate "sets of n numbers". Each set corresponds to the expression values for a single gene. Many analysts choose to study microarray data in this way; see Kuo *et al.* (2002) and Johnson *et al.* (2006). The other view is the **Multivariate view**, which treats the gene expression

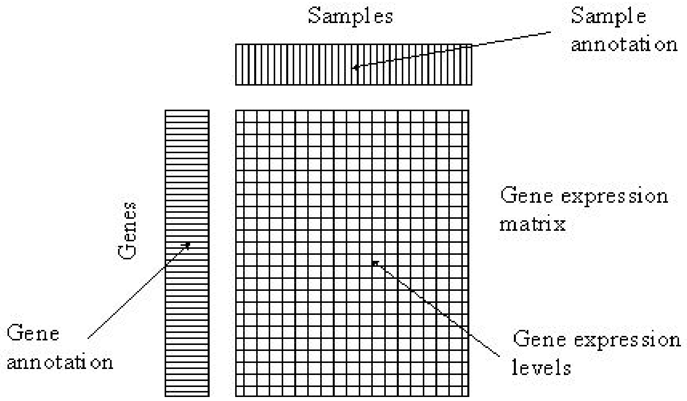


Figure 1.2: The expression matrix for a microarray data set. Each column corresponds to a sample and each row corresponds to a gene. The gene expression values are displayed in a matrix. (This plot is taken from Brazma *et al.* (2004)).

matrix as a set of n $d-$dimensional vectors. The data set contains n data objects. Each data object is a d dimensional vector (the column of the expression matrix), which represents the gene expression values for some specific sample (individual). Since the dimension d is typically much larger than the sample size n, we call this a **High Dimension Low Sample Size (HDLSS)** setting, as studied in Hall *et al.* (2005).

In this section, we will introduce and compare these two viewpoints for HDLSS data. Section 1.2.1 presents the "Gene by Gene" view. Section 1.2.2 introduces the Principal Component Directions view as a the multivariate view method.

1.2.1 Gene by Gene View

The "Gene by gene" view needs to be regarded with healthy skepticism in the analysis of microarray data, because the data are intrinsically multivariate in nature. A toy example in Figure 1.3 is presented to show that "gene by gene view" doesn't provide sufficient insights into the multivariate nature of these data sets.

The toy data set in the Figure 1.3 are for the expression values, measured on 4000 genes (dimensions), and are intended to model an important biological effect with gene expression

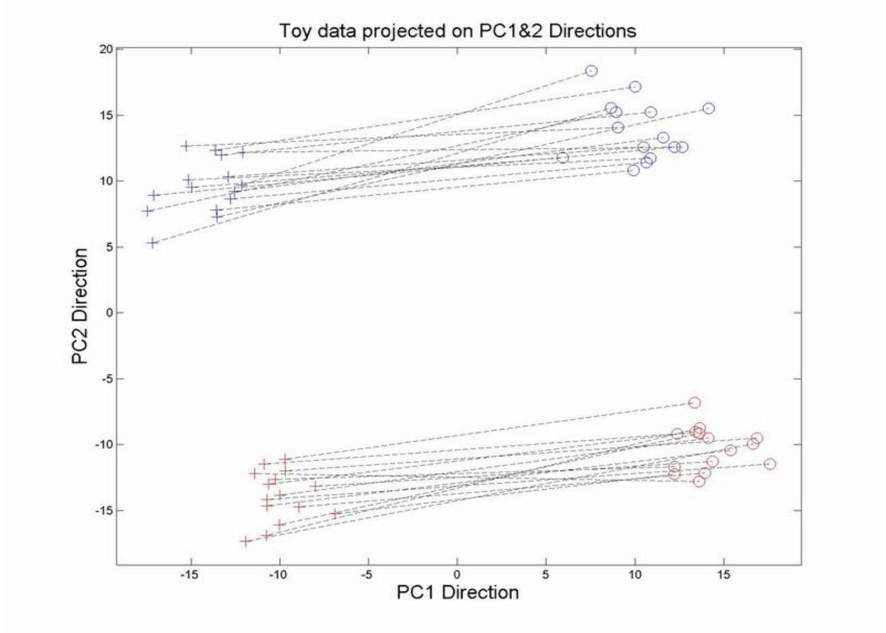


Figure 1.3: Projection view of the toy data, which contain two batches and two biological clusters. Symbols are for the batches and colors are for the biological clusters.

values measured across two batches. There are 30 samples within each batch , split evenly between the two clusters. Hence there are 15 samples in each simulated biological cluster. The entries of each sample are generated from independent Gaussian distributions with standard deviation 1. The means of these entries are taken to be ±0.2, in such a way that there are 4 clusters, where pairs correspond to batches, and within each pair, the clusters simulate an important biological difference. Figure 1.3 shows a two dimensional projection view of the data sets. We will explain more details about the projection directions in the next subsection. In this Figure, each point represents a sample, with expression values for 4000 genes. Two batches are represented by different symbols, and the biological clusters are represented by different colors. Dashed line segments are used to connect associated samples from the two batches. Clearly there is significant batch difference in the data sets (the cloud of crosses are away from the cloud of pluses). Samples from different biological clusters have very different expression values, as shown in Figure 1.3 using colors. However, the very small difference in the means of the entries is an order of magnitude less than the noise level for each gene, so that it is essentially invisible to a gene by gene analysis. This is seen via a gene by gene scatter plot, shown in Figure 1.4 or a gene by gene correlation analysis, as done by Kuo et al. (2002).

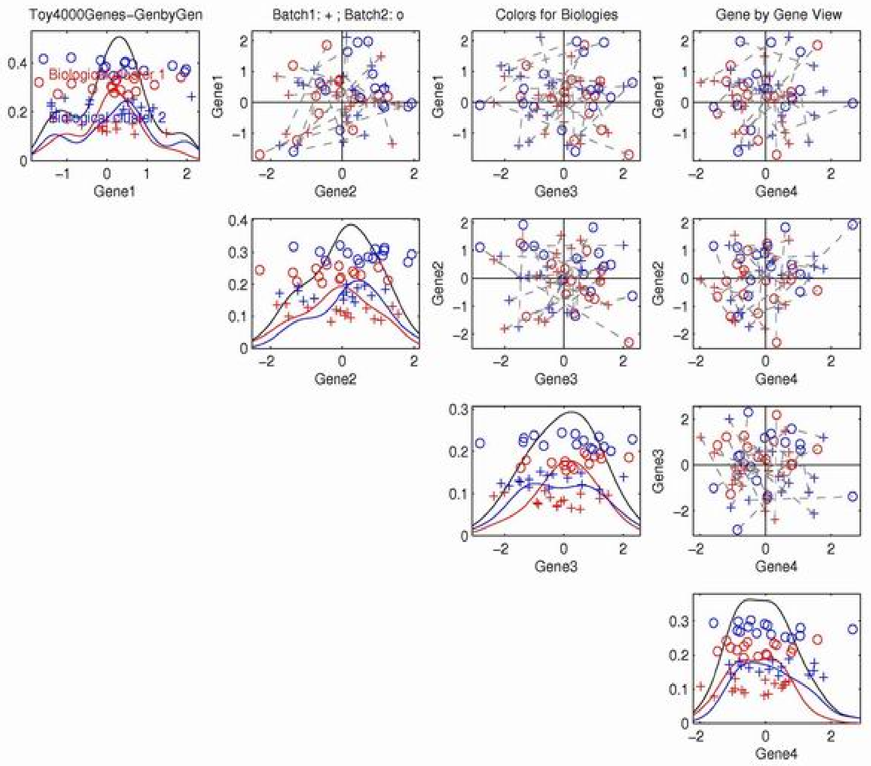


Figure 1.4: Gene by Gene view of the toy data. On-diagonal plots show single gene expression values. Off-diagonal plots are scatter plots of the expression values for two genes. Symbols represent batches, and colors represent biological clusters. The black dashed segments are used to connect the associated samples from the two batches.

Figure 1.4 shows the gene-by-gene view of the simulated data for the first four genes. In these plots, each point represents a sample (i.e. case). Every plot on the diagonal displays the expression values for a single gene. A one-dimensional “jitter plot” (see Tukey and Tukey (1990)) is used with a random vertical coordinate for visual separation of the data points. Also kernel density estimation curves are drawn to provide another view of how the expression values of one single gene are distributed. For example, the subplot in the top row, the first column shows the expression values on the first gene. Three kernel density curves, colored with black, blue and red, are drawn for the all the samples, the blue samples only (biological cluster 1), and red samples only (biological cluster 2) respectively.

In this direction, there is no appropriate separations of batches, or biological clusters. All the off-diagonal plots show the two dimensional scatterplots for the two corresponding genes. For example, the top row, second column subplot shows the projection of the data onto the plane, formed by the first and the second genes. As in Figure 1.3, symbols are used to represent samples from different batches. Colors are used to represent the samples from different biological clusters. Dashed line segments are used to connect the associated samples from the two different batches.

In all the subplots of Figure 1.4, there are no appropriate separations of batches (circles and pluses), or biological types (reds and blues). This is due to the very small difference in the mean values of entries, compared with the noise level for each gene. Thus from the gene by gene view, both batch effect and biology effect are invisible. In the next subsection, we will present the multivariate view of the data, which shows that there are actually some biological and batch effects, which can be seen using an appropriate view.

1.2.2 Multivariate View

The multivariate view treats a microarray data set $X_{d\times n}$ as a cloud of n points in d dimensional space. Due to limitations of the human perceptual system, it is challenging to visually understand the full geometric structure of the data with dimension more than 3. However, we can project the data points onto some carefully chosen directions of interest. There are many interesting directions in HDLSS settings, e.g. you could find a direction (which is not unique) such that the projections of all the samples on this direction are piled up on one single point. Ahn and Marron (2006) developed the maximal data piling direction. In this direction, the projections of some samples are piled up on one single point, the projections of all the other samples are piled up on another single point, and the distance between these two points is maximized. On the webpage for this dissertation (see Liu (2007b)), these types of projections are illustrated for some interesting examples.

Actually, the Gene by Gene view in Figure 1.4 can also be thought of as a multivariate projection view for HDLSS data. The projection directions are the directions of the first four genes (i.e. the first four Euclidean unit vectors). As we have discussed, important batch effects and biological effects are invisible in this gene by gene view, because the difference between batches or biological clusters are very small in a single gene. However, this difference is significant, if all the genes are taken into considerations. E.g. instead of projecting data onto single gene direction, we could project the data onto some linear combinations of the gene directions, such as the principal component directions.

Principal Component Directions View

Principal Component Analysis (PCA) is a classical statistical method, which continues to be widely used for statistical data representation and data compression. For a data set in high dimensional space, PCA finds a set of directions called the **Principal Component directions (PC directions)** such that the first PC direction accounts for as much of the variability in the data as possible, and each succeeding PC direction accounts for as much of the remaining variability as possible. Often, the first several PC directions will express most of the variability in the data. Thus, PC directions are often commonly used to visualize the data. This kind of view for HDLSS data was used by Benito *et al.* (2004), Liu *et al.* (2007) and Marron and Liu (2005).

We use the the toy data in Figure 1.3 to illustrate the idea of the PC projection plot. Note that all the PC directions are the linear combinations of gene directions. Fig 1.5 shows the PC projections of the data on the 1-d directions or 2-d planes formed by the first four PC directions. Every plot on the diagonal has displayed a one-dimensional projection on the PC directions. All the off-diagonal plots show 2-d views of the data projected on the plane formed by the two corresponding PC directions.

The limitation of the gene-by-gene view is made clear in the PCA multivariate scatterplot view of these data in Figure 1.5. Note that the first two principal components (top row, second column) contain the deliberately constructed structure in the data. In particular, the batch effect (indicated by pluses and circles) is clear, shown mostly on the first PC direction. The strong simulated biological effect is shown as two clusters (indicated by the red and blue colors) on the second PC direction.

In the rest of this dissertation, we will focus on the multivariate view of the data.

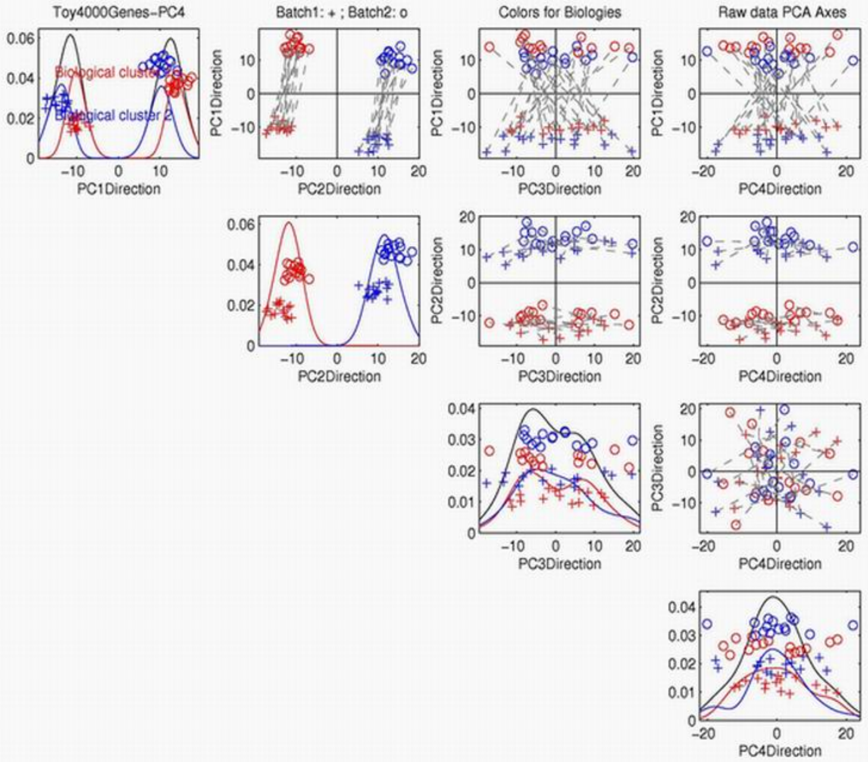


Figure 1.5: Toy data are projected onto the first four PC directions. On-diagonal plots are one-dimensional projection plots. Off-diagonal plots are 2-d projection plots for corresponding PC directions. Symbols represent batches, and colors represents biological clusters. The black dashed segments are used to connect the associated samples from the two batches.

1.3 Microarray Batch Adjustment Methods

1.3.1 NCI60 Cancer Cell Line Data

In 2000, cDNA and Affy microarrays were used to measure the gene expression values among the 60 cell lines from National Cancer Institute's anti-cancer drug screen (NCI60) (Eisen and Brown (1999), and Alter *et al.* (2000)). These cell lines are from different sites of origin, i.e. 7 breast, 5 central nervous system (CNS), 7 colon, 6 leukemia, 8 melanoma, 9 non-small-cell-lung-carcinoma (NSCLC), 6 ovarian, 2 prostate, 9 renal, and 1 unknown. Using cDNA microarrays, 9703 genes were spotted on the chip and the expression values were measured. After excluding those genes with more than two missing data points, the cDNA gene expression data were collected as a 5244 × 60 matrix. Missing data were imputed using K-Nearest Neighbors imputation (KNN, see Troyanskaya *et al.* (2001)). The same list of 60 cancer cell lines were measured with the Affymetrix Microarray Suit 4.0 for 7070 genes. There are some negative values in the Affy, which were set to 1 before taking log2 transformation. We linked genes from cDNA and Affy data sets by mapping their identifiers to Unigene Cluster Identifiers (UCID). Duplicate UCIDs were collapsed by taking the median value within each sample. The paired cDNA and Affy data set were created from the intersection of UCIDs of these two sets. Both the cDNA and the Affy data contain 60 common samples and 2267 common genes. In the rest of the dissertation, We refer the NCI60 data as two such Affy and cDNA data sets with common samples and genes.

Fig 1.6 shows the PC projection view of the NCI60 data, which have a similar format to Figure 1.5. In this figure, The purple circles are the Affy samples, and the green pluses are the cDNA samples. The dashed line segments are used to connect the associated biological samples measured on the two different platforms. Long segments tell us that there are significant differences between the expression values of the associated biological samples measured by cDNA and by Affy. The top row, second column subplot shows the projections of the data on the plane formed by the first and the second PC directions. Note that the differences between cDNA and Affy are mostly along the first PC direction. We also find that the dashed line segments are quite parallel.

1.3.2 Microarray Batch Adjustment

Microarray data contain the expression values for thousands of genes. The measurements tend to be noisy. The noise in the data could be countered by running a large number of arrays, and averaging the results. However, this is currently not practical because array costs are still relatively high. Another approach to reduce the effect of noise is to combine the current data with previously existing data sets, many of which are web available. The combined data set will have larger sample size, which will boost statistical power. However, as noted by Irizarry *et al.* (2003), hurdles to such combinations include biases introduced

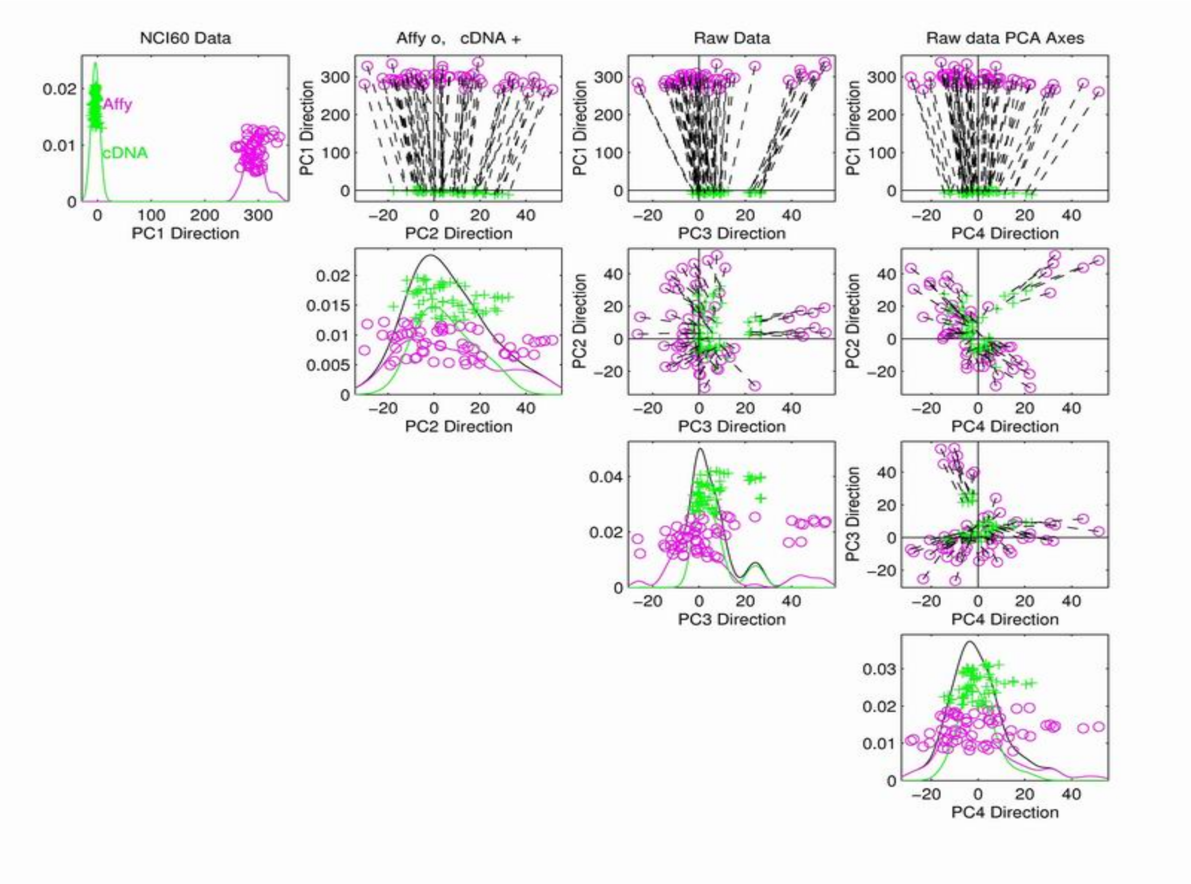


Figure 1.6: NCI60 data are projected onto the first four PC directions. The purple circles are Affy samples. The green pluses are cDNA samples. The black dashed segments are used to connect the same biological samples from the two platforms.

during the sample preparation, manufacture of the arrays, and the processing of the arrays (labelling, hybridization, and scanning, etc). Even more challenging is that the data are especially non-comparable when they are collected using different microarray formats of technologies (e.g. Affymetrix versus cDNA; see Yauk *et al.* (2004)). Systematic biases between data sets are commonplace. As shown in Figure 1.6, the expression values for Affy and cDNA are very different in means and variation of the measurements, even when they are the expression values for the same list of genes and biological samples. We usually use the term "batch adjustment" for the operation of eliminating systematic biases by combining different data sets. In this dissertation, we only consider the combination of two data sets, where the measurements are made for the same list of genes, and may or may not be for the same samples. In this situation, two data sets can be visualized and compared in the same high dimensional gene space.

Some researchers have used the Singular Value Decompositions (SVD/PCA) to correct for systematic biases in the data set of yeast cell cycle experiments (Alter *et al.* (2000)), and to correct for microarray batch bias in a data set containing many soft tissue tumors Nielsen *et al.* (2002)). Recall that the SVD/PCA seeks to find the “directions of greatest variation”. To adjust the batch difference, the variations of the data sets along the SVD direction were totally removed. However, as noticed by Benito et al.(2002), there are some serious problems for this method.

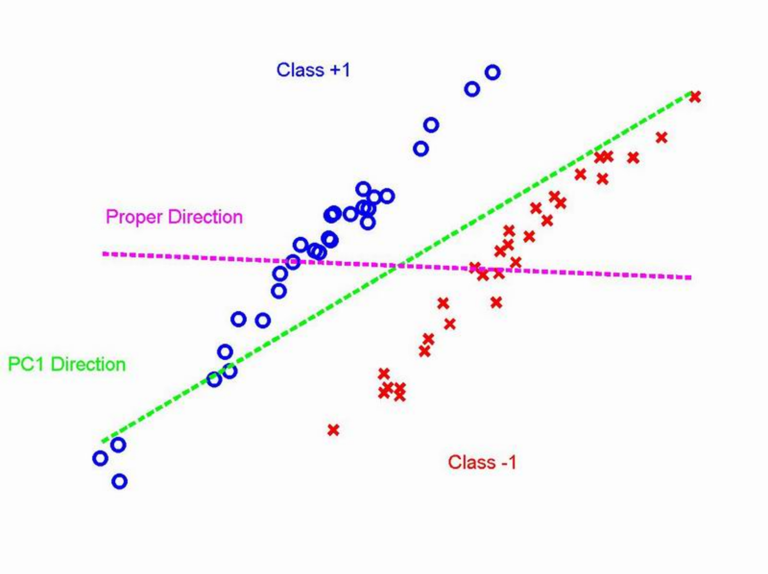


Figure 1.7: Underlying conceptual model shows that the SVD/PCA direction (green dashed line) is not consistent with the batch difference direction (magenta dashed line). Classes are represented by different colors and symbols.

Firstly, it works well only when the direction of the batch difference is consistent with the SVD/PCA direction. This means that the between-group differences are much larger than the within batch variation. Figure 1.7 shows an underlying conceptual model. The observations from two batches are represented by symbols and colors. In this toy data set, the within group variation is much larger than the batch difference. The first SVD/PCA direction (green dashed line) is very different from the actual batch difference direction (magenta dashed line). The adjustment of the data along the first SVD/PCA direction will not eliminate the differences between batches. Notice that SVD/PCA direction doesn’t use the batch memberships of the observations. A natural way to improve the analysis is to make full use of the systematic bias information (i.e. the batch membership of each observation). Instead of choosing the direction to maximize the variation of the full data, i.e. SVD/PCA, we could choose the direction which gives the maximum separation between two batches. In the next subsection, we will introduce and compare several such methods, which separate two batchesas well as possible, in some sense.

In addition to finding a useful direction for systematic batch difference, Benito *et al.* (2004) proposed another improvement over the SVD/PCA adjustment. Alter *et al.* (2000) adjust the batch difference by totally eliminating the variation of each batch along the SVD/PCA direction. This method squashes all the geometric structures in the data along the chosen direction. If there is other important biological variation along this direction, other than systematic batch difference, these important biological differences will disappear when the data are squashed along this direction. This idea is illustrated by a toy data set with two genes, as shown in the left subplot of Figure 1.8. Batches and biological clusters are presented using symbols and colors respectively. The green dashed line shows the first SVD/PCA direction. This direction shows the batch difference, which is in the same direction as the biological differences. If the data are squashed along the SVD/PCA direction, as shown in the top right subplot of Figure 1.8, the batch difference can be successfully removed. However, the differences between biological clusters are removed too, in the sense that the blues and reds are mixed together after adjustment. A possible improvement is to subtract the subpopulation means of the data projected on the given direction. The geometric interpretation of this operation is to shift each cluster along the given direction until they overlap, instead of squashing them along the direction. This preserves any variation in this direction, which is not caused by systematic effects. The bottom right subplot of Figure 1.8 shows the adjusted data after shifting along the SVD/PCA direction. The batch difference has been adjusted in the sense that circles and crosses are mixed well. The important biological structures are preserved in the adjusted data, as shown by the color.

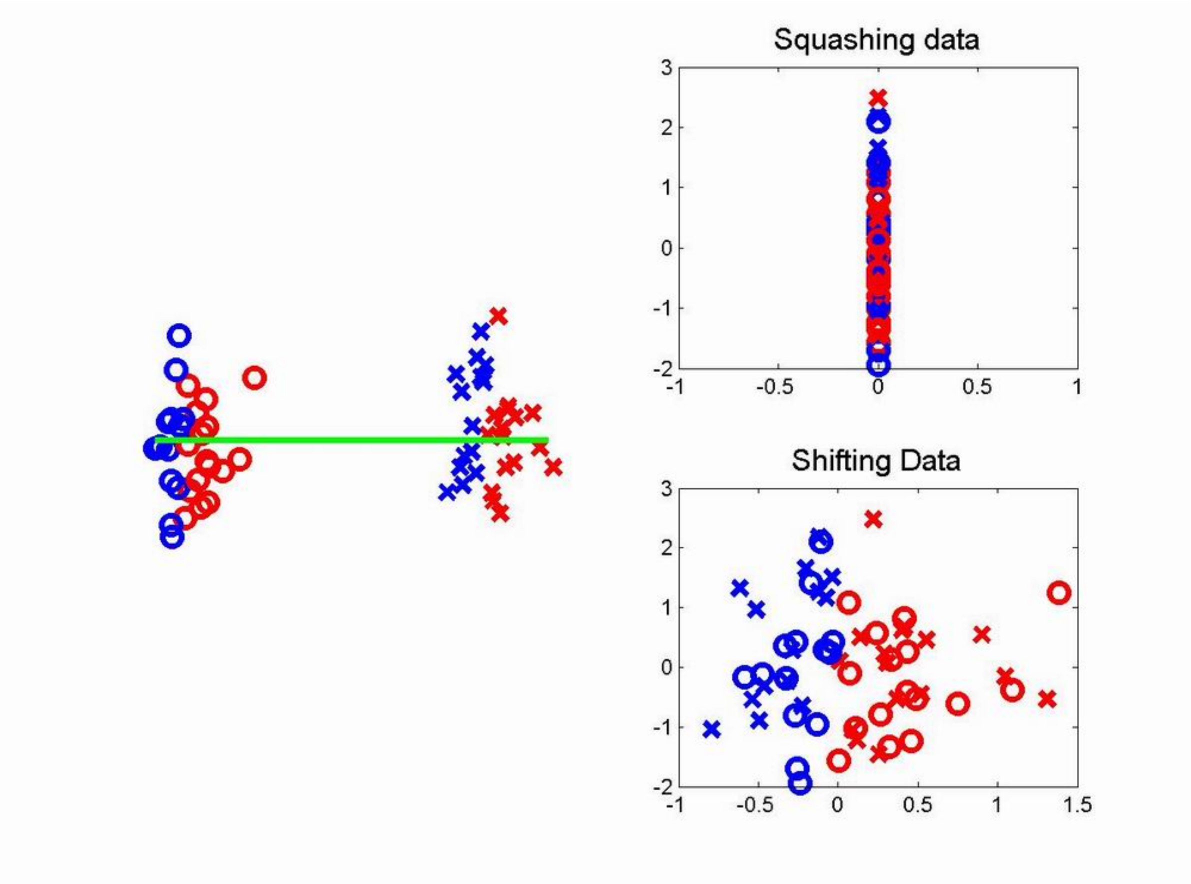


Figure 1.8: Toy data set for comparing the adjustments of squashing and shifting. The left subplot shows the data set before adjustment. Batches are represented by symbols, and the important biological clusters are represented by colors. The green line shows the first SVD/PCA direction. The upper right subplot shows the data after squashing along the direction of the green line. Both batch difference and biological differences are removed. The bottom right subplot shows the data sets after rigidly shifting along the direction of green line. The batch difference is removed and the biological differences are preserved.

1.3.3 Linear Batch Adjustment Methods

In this dissertation, we are mainly interested in linear batch adjustment methods, because they have direct and meaningful geometric interpretations. Using the multivariate view, two HDLSS data sets are treated as two clouds of points in high dimensional space. Linear batch adjustment methods find an appropriate direction and then move the two clouds along this direction until they overlap. The problem of batch adjustment is equivalent to the problem of binary discrimination problem for two data sets. The objective of linear discrimination between two data sets is to find a hyperplane, which separates them as well as possible. The orthogonal direction of the hyperplane gives the maximum separation of two data sets and can be used for adjusting the batch difference. In this Section , several important linear discrimination methods are introduced and compared for the batch adjustment.

Binary Classification (Discrimination) Problem

Here we introduce some mathematical notations for the classification problems, (see Hastie *et al.* (2001)). In the binary classification problem, we use class labels +1 and −1 to represent two different classes. Suppose that we have the training data $\{(\boldsymbol{x_1}, y_1), \cdots (\boldsymbol{x_n}, y_n)\}$. Each $\boldsymbol{x_i} \in \Re^d$ represents the observation vector for the ith sample. Each $y_i = +1$, or -1 represents the class membership for the ith sample. The objective of binary classification is to find a classification rule (classifier) $f(\mathbf{x}) : \Re^d \rightarrow \{-1, 1\}$, which assigns a cluster label ($+1$ or -1) to a given sample $\mathbf{x}$. One goal of $f(x)$ is the consistency with the observed data, i.e. for $(\mathbf{x}, y)$s is in the training data set. A second goal is the prediction of new observations. Sometimes $f(\mathbf{x})$ can be a function from $\Re^d \rightarrow R$. Then the sample is classified to $+1$ if $f(\mathbf{x}) \geqslant 0$, and to -1 if $f(\mathbf{x}) < 0$.

Linear Discrimination Problem

If the classifier $f(\mathbf{x})$ is a linear function of $\mathbf{x}$, we call $f(\mathbf{x})$ is a linear classifier, i.e.

$$f(\mathbf{x}) = \boldsymbol{w}^T\mathbf{x} + b \tag{1.1}$$

where $\boldsymbol{w}$ is a d dimensional vector, and b is the threshold for the classification. The class label $+1$ or -1 is given to the sample $\mathbf{x}$, if $f(\mathbf{x}) \geqslant 0$ or $f(\mathbf{x}) < 0$. Using the multivariate view, each sample $\mathbf{x}$ is a point in d dimensional space. A linear classifier attempts to find a $d-1$ dimensional hyperplane, which separates two the classes $+1, -1$ as well as possible. The vector $\boldsymbol{w}$ defines the orthogonal direction of the separation hyperplane.

The batch adjustment method, corresponding to the above linear classifier is to move two data sets along the normal direction of $\boldsymbol{w}$ to eliminate systematic batch differences. The problem of finding the best batch adjustment direction is equivalent to find the best linear classifier (discrimination hyperplane). In the following, we will take a further look at some basic and widely used discrimination methods. The comparison between them will be further studied in Chapter 3.

Nearest Centroid method

Suppose $X_{d\times n_1}$ and $Y_{d\times n_2}$ are two clusters of d dimensional data. Using the multivariate view, they are treated as two clouds of points in d dimensional gene space. The nearest centroid method uses the within class sample mean as the representative for each cluster. Every sample is classified to the cluster with nearest centroid to this sample. This is a linear discrimination method in the sense that the normal direction w of this method is the normalized direction vector which connects the centroids of the two clusters. Thus

$$\boldsymbol{w} = \frac{\overline{\mathbf{x}} - \overline{\mathbf{y}}}{||\overline{\mathbf{x}} - \overline{\mathbf{y}}||},$$

where $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are the sample mean vectors of the two classes. Tibshirani *et al.* (2002) uses this direction for adjusting the batch difference in their Predicton Analysis of Microarray (PAM) software.

The gene by gene view of the PAM method is that the observations for every gene are subtracted by their within batch mean for this gene. Using multivariate view, this adjustment has a very simple multivariate geometric interpretation. It can be treated as rigidly shifting two clusters such that their centroids are moved to the origin. After adjusting within group mean, the mean value for every gene is zero. To preserve the variation of the mean values of genes, the observations for each gene are added by the mean value of this gene across two batches. The geometric interpretation of this adjustment and the previous within batch mean adjustment is to rigidly shift two clusters along the direction which connects two centroids, until both centroids move to the centroid of two clusters. Instead of moving two clusters to the centroid of two clusters, some researchers choose to fix one cluster and move the other cluster to the first one until their centroids overlap. This method preserves the mean values of genes on the chosen batch.

No matter what kind of centroids adjustment, they are the results of shifting two clusters along the direction which connects two centroids. From now on, we call this direction as **the PAM direction**. In addition to adjust the mean, the PAM software has a step to adjust batch variation differences. However, in this dissertation, we focus on the batch difference adjustment, and hence we won't consider the variation adjustment.

The PAM adjustment has been shown to work very well for many data sets, see Tibshirani *et al.* (2002). It involves easy calculation and has a simple geometric interpretation. However, the PAM method is not robust if there are outliers, which are away from the main population. Johnson *et al.* (2006) proposed the empirical Bayesian methods to improve the robustness. In Chapter 3, Section 3.2, we will study other properties of the PAM direction. In particular, PAM is not asymptotically robust for combining two data sets with unbalanced subgroup sample sizes, when the number of genes goes to infinity.

Note that every observation has some influence on locating the PAM direction. However, it is natural to think that those points which are close to the separating hyperplane are more important than the observations which are away from the separation hyperplane. Another discrimination method, called the Support Vector Machine (SVM), directly addresses this problem.

Support Vector Machine (SVM)

SVM, (see Vapnik (1982), Vapnik (1995), Burges (1998) and Liu (2007a)) is a popular linear discrimination method. It is introduced in two cases: when the data are linear separable, and when they are not. In this dissertation, we will focus on the separable case, because two HDLSS data sets are linear separable with probability one, if the data follow distributions that are absolutely continuous with respect to d dimensional lebesgue measure. Consider a linear classifier $f(\mathbf{x}) = \boldsymbol{w}^T\mathbf{x} + b$, as in Section 1.3.3. A special linear classifier, called SVM classifier, involves an interesting choice of $\boldsymbol{w}$ and b. The SVM first finds two hyperplane margins (over $\boldsymbol{w}$ and b) which are defined by $f(\mathbf{x}) = 1$ or -1, such that there are some observations on the margins and there are no observations between these two margins. The points on the margin are called "support vectors". Usually there are multiple choices over w and b for the margins, when the data are separable in HDLSS settings. The SVM finds w and b such that the distance between two the margins $\frac{2}{||\boldsymbol{w}||^2}$ is maximized. The hyperplane between the two margins: $f(\mathbf{x}) = 0$ is the SVM discrimination hyperplane. Given $\boldsymbol{w}$ and b, the class label $+1$ is given to a new sample $\boldsymbol{x_i}$, if $f(\boldsymbol{x_i}) \geqslant 0$ and the class label -1 is given if $f(bxi) < 0$. The SVM can be interpreted as the solution of the following optimization problem over $\boldsymbol{w}$ and b:

$$\begin{aligned} & minimize \;\; \frac{1}{2}||\boldsymbol{w}||^2 \\ & subject\ to \quad y_i \times f(\boldsymbol{x_i}) \geqslant 1, \quad i = 1, \cdots, n. \end{aligned} \tag{1.2}$$

where y_i represents the class membership of the ith sample $\boldsymbol{x_i}$ in the training data set. The normalized direction vector of $\boldsymbol{w}$ represents the SVM direction. The constrains $y_i * f(\boldsymbol{x_i}) \geqslant 1 \quad i = 1, \cdots, n$ indicate that the $f(\mathbf{x})$ must classify all the samples in the training data set correctly. The SVM classifier gives as accurate predication to the class membership of new samples as possible, in the sense of maximizing the distance between two margins.

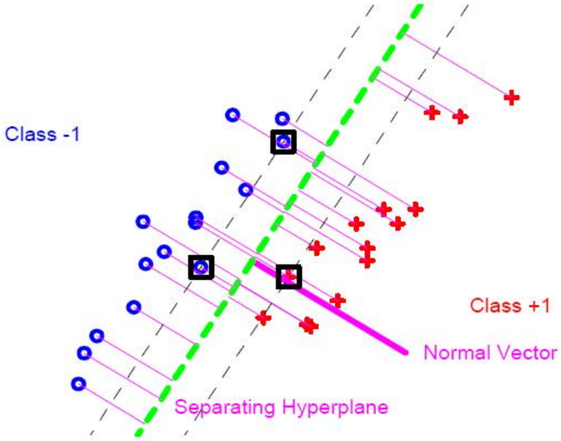


Figure 1.9: SVM hyperplane to separate two classes, represented by crosses and pluses for a two dimensional toy data set. The Purple normal vector is used for batch adjustment.

Figure 1.9 shows the SVM method for classifying a $2-d$ toy data set, with the two classes represented by blue circles and red pluses respectively. The two grey thin dashed lines show the two hyperplanes for the margins ($\{\mathbf{x} : f(\mathbf{x}) = \pm 1\}$), with some support vectors (black boxes) on the margins. The SVM finds two margins (over $\boldsymbol{w}$ and b) such that the distance between them is maximized. The green dashed line between two margins represents the discrimination hyperplane ($\{\mathbf{x} : f(\mathbf{x}) = 0\}$). The observations on the left side of this hyperplane are classified to the class with label -1 (the class of blue circles). The observations on the right side of this hyperplane are assigned to the class with label $+1$ (the class of red pluses). The purple normal vector of the hyperplane is the direction showing the batch difference. It can be used for adjusting the batch difference by rigidly shifting the blue class and the red class along this direction. The SVM method has been shown to be very successful in a variety of classification problems. However, as noticed by Marron and Todd (2002) and Benito *et al.* (2004), the SVM can be improved in HDLSS settings. There are two main drawbacks of the SVM method. Firstly, the SVM suffers from a substantial data piling on the margins, which could lead to biased batch adjustment. Secondly, only those observations on the margins have an influence on locating the SVM hyperplane; the observations which are away from the margins have no influence at all. For example, in Figure 1.9, if you move off-margin blue circles to anywhere on the left side of the above margin, the discrimination hyperplane won't change at all. These two problems of the SVM will be studied more precisely in Chapter 3. Marron *et al.* (2005) have addressed these problems by the development of Distance Weighted Discrimination (DWD) method.

Distance Weighted Discrimination (DWD)

The DWD method, developed by Marron *et al.* (2005) is an improvement upon the Support Vector Machine (see Burges (1998)) in HDLSS contexts, as explained by Benito *et al.* (2004). Suppose two classes are separable, which is very likely for HDLSS data. Again, suppose the separating hyperplane is $f(\mathbf{x}) = \boldsymbol{w}^T\mathbf{x}+b$. Denote the distance from the observation $\boldsymbol{x_i}$ to the hyperplane as r_i (see Figure 1.10). DWD finds the hyperplane that minimizes the sum of the inverse distances. This gives larger influence to those points which are close to the hyperplane relative to the points that are farther away from the hyperplane. For separable classes, the DWD method is the solution of the following optimization problem,

$$\begin{aligned} &\textit{minimize} \quad \sum_{i=1}^{n} \frac{1}{r_i} \\ &\textit{subject to} \quad y_i \times f(\boldsymbol{x_i}) \geqslant 1, \qquad i = 1, \cdots, n. \end{aligned} \tag{1.3}$$

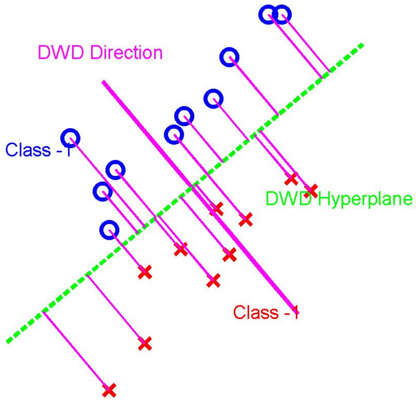


Figure 1.10: DWD hyperplane to separate crosses and pluses. The Purple Normal Vector is the DWD direction.

As shown in Figre 1.10, DWD finds a linear hyperplane (Green) to separate the two clouds of points (blue circles and red pluses) as well as possible, in the sense of minimizing the sum of the inverse distances from the samples to the hyperplane. The normal direction of the hyperplane is called **the DWD direction**. The computing of this hyperplane can be formulated as a Second-Order Cone Programming (SOCP) problem and is solved using the software package SDPT3 (for Matlab), which is web-avaible at Toh *et al.* (2006). The DWD direction has been shown to provide effective bias adjustment for many situations by Benito *et al.* (2004), including effective across-platform adjustment. In Chapter 3, we will demonstrate the robustness of DWD method, compared with PAM, when the dimension d goes to infinity.

1.4 Organization of the Dissertation

This dissertation covers two different aspects of microarray data adjustment, which are organized as two chapters. In each chapter, we will introduce the motivation of the problem, review the literature, and present our work.

Chapter 2 is about HDLSS parallel directions. We propose two interesting directions: the canonical parallel direction and the canonical orthogonal direction. This pair of directions gives an insightful 2-d view for understanding paired HDLSS data sets. The algorithm to produce these two directions is developed in this chapter. Under some mild conditions, these two directions exist and are unique. The canonical parallel direction shows the differences between batches and can be used for adjusting the differences. The mathematical properties of this direction are studied using a *Linear Shifted Model*, for which, we know the theoretical canonical parallel direction between two data sets. We present and prove the asymptotic properties of the empirical canonical parallel direction, as the dimension d increases. We explore the *Consistency* and the *Strong Inconsistency* of the empirical direction under different conditions. Simulated data sets are used to verify the asymptotic results.

Chapter 3 is about the comparison between three linear batch adjustment methods, SVM, DWD, and PAM. First, several examples are presented to illustrate the limitation of the SVM method, especially for HDLSS data. Secondly, DWD and PAM are compared under an *Unbalanced Subgroup Model.* We discover that DWD is more robust than PAM, when the two data sets have unbalanced subgroup sample sizes. We study this problem for two cases. In the first case, when the dimension is fixed and the subgroup sample sizes become more and more unbalanced, DWD is consistently more robust than PAM. In the other case, when the subgroup sample sizes are unbalanced and fixed, as the dimension goes to infinity, DWD is also much more robust than PAM. Thus, the PAM direction has remarkably inferior asymptotic properties, compared to DWD, when the dimension is high.

CHAPTER 2

HDLSS Canonical Parallel Direction

In the HDLSS settings, it's challenging to view the full geometric structure of the data because the dimension of the data d is large. A common approach is to choose some directions and view the projections of the data on the $1-d$ or $2-d$ subspaces determined by these directions. In Chapter 1, Section 1.2, we introduced the gene by gene view and the principal component directions view. In this chapter, Section 2.1, we produce two novel directions, called the **Canonical Parallel Direction** and the **Canonical Orthogonal Direction**. These two directions provide a new and useful $2-d$ subspace to show different aspects of the two data sets. We give the theorems for the existence and uniqueness of these two directions. In Section 2.2, we develop algorithms to generate canonical parallel and canonical orthogonal directions. The algorithms indicate the existences and uniqueness of the two directions. The canonical parallel direction is the one showing batch difference. We use it to adjust the differences between Affy and cDNA in NCI60 data. A visual diagnosis shows good performance of this adjustment. In Section 2.3, we study the asymptotic properties of the empirical canonical parallel direction in a **linear shift model**, for which, the theoretical canonical parallel direction is known. We identify the conditions which assure the convergence of the empirical direction to the theoretical one, and conditions which give strong inconsistency.

2.1 Visualization and Adjustment using the Canonical Parallel Direction

Two microarray data sets $X_{d\times n}$ and $Y_{d\times n}$ are called **paired**, if $x_{i,j}$ and $y_{i,j}$ (the ith row, jth column of the two data sets, ($i = 1, \cdots, d, \quad j = 1, \cdots, n$) are the measurements for the same gene and related biological samples. For paired data sets, the multivariate view treats these two data sets as two clouds of points in d dimensional space. Each cloud contains n points. Since the two data sets are paired, an insightful illustration is to use a line segment to connect the associated points from the two data sets. The vector of the line segment shows the differences of measurements between each pair of associated points.

The top row, second column of Figure 1.6 in Chapter 1 shows the projections of the NCI60 data on the plane formed by the PC1 and PC2 directions. The difference between the two data sets is mostly in the PC1 direction. We notice that all the line segments are almost parallel, but not exactly. Actually we can replace PC1 and PC2 by two other directions such that the projections of the data on the plane formed by these two directions have all of the line segments exactly parallel. The left plot in Figure 2.1 shows one such *parallel projection.*

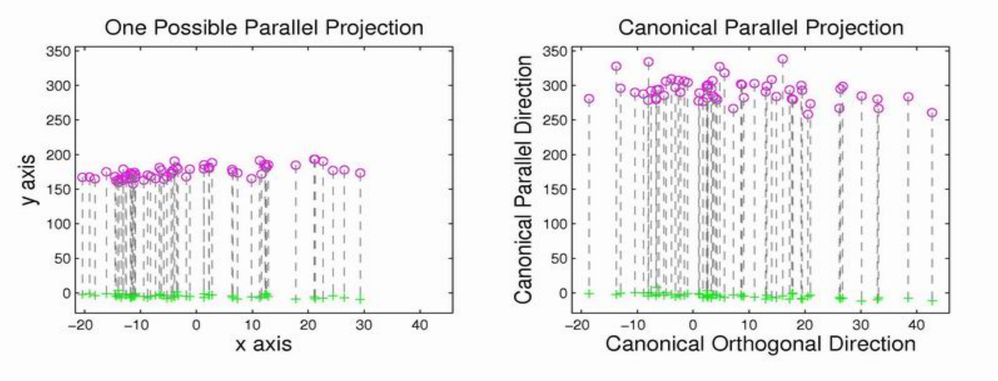


Figure 2.1: Left Plot: NCI60 data are projected on two specific directions which make all the line segments parallel. Symbols and colors are the same as in Figure 1.6. Right Plot: NCI60 data are projected onto the plane formed by the canonical parallel direction and the canonical orthogonal direction.

In the left plot of Figure 2.1, the y-axis is a direction showing the differences between the two data sets. The x-axis is a direction that makes all the line segments parallel. In HDLSS settings, there are many direction vectors of x and y axes, which will also give a parallel projection. A special choice among these, called the **Canonical Parallel Direction** and the **Canonical Orthogonal Direction** is shown in the right plot of Figure 2.1. Among all the possible parallel projection plots, this plot shows the most variability in the data, i.e. on the x-axis, the variation of the projected data is maximized; on the y-axis, the sum of the squared projected lengths of line segments is maximized. This projection plot shows the differences between batches as well as possible, since the y-axis highlights the differences between batches. The definitions of the two canonical directions are given in the following:

Definition 2.1.1. Assume $X_{d \times n}$ and $Y_{d \times n}$ are paired HDLSS matrices $(d > n)$. Associated samples are connected using dashed line segments. The d dimensional direction vector is called the **Canonical Parallel Direction (CPD)**, denoted as $\mathbf{v}_{cpd}$, if the projections of the line segments (i.e. columns of $X - Y$) have the maximum, over all direction vectors in $\Re^d$, sum of squared lengths.

Definition 2.1.2. Assume that $X_{d \times n}$ and $Y_{d \times n}$ are paired HDLSS matrices $(d > n)$. The d dimensional direction vector, which satisfies the following conditions, is called the **Canonical Orthogonal Direction (COD)**, denoted as $\mathbf{v}_{cod}$:

- this direction $\mathbf{v}_{cod}$ is orthogonal to all the directions of line segments (i.e. to all column vectors of $X - Y$);
- the projections of the column vectors of X along $\mathbf{v}_{cod}$ have the maximum, over all direction vectors in $\Re^d$, variation, i.e. the sum of the squared distances from each projection to the center of the projections.

The above two definitions indicate that these two canonical directions can be derived separately and they are orthogonal to each other. The CPD is the direction, which shows the differences between batches as much as possible. The COP is the direction which makes all the projected line segments parallel. The projections of the data onto the plane spanned by these two directions have all the line segments parallel and show as much of the variability in the data as possible among such vectors. Under some mild conditions, these two directions exist and are unique. The following two theorems give the conditions for the existence and uniqueness of these two directions separately.

Theorem 2.1.1. *(Existence and Uniqueness of CPD)*
Suppose $X_{d\times n} = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_n})$ *and* $Y_{d\times n} = (\boldsymbol{y_1}, \cdots, \boldsymbol{y_n})$ *are paired HDLSS matrices* $(n < d)$. *The* $\mathbf{v}_{cpd}$ *between* X *and* Y *exists and is unique (modulo the ± flip of direction) if the first eigenvalue of* $(X-Y)(X-Y)^T$ *is positive and strictly larger than all the rest eigenvalues.*

Proof. This theorem will be proved when the derivation for this direction is given in Section 2.2. In real data analysis, the conditions in this theorem are very likely to be satisfied. From the deviations in Section 2.2, we will show the CPD is the first eigenvector of $(X-Y)(X-Y)^T$. Suppose the eigenvalues of $(X-Y)(X-Y)^T$ are $\lambda_1, \lambda_2, \cdots, \lambda_d$. Because the rank of $(X-Y)(X-Y)^T$ is no larger than n $(n < d)$. Among these eigenvalues, at most n of them are nonnegative. If the first eigenvalue is positive and strictly larger than the others, the first eigenvector exists and is unique (modulo the ± flip of direction). Otherwise, suppose the first two eigenvalues are the same, i.e. $\lambda_1 = \lambda_2 > 0$, then the first two eigenvectors could be any pair of orthogonal basis vectors in an two dimensional plane, and hence the first eigenvector is not unique.

□

Theorem 2.1.2. *(Existence and Uniqueness of COD)*
Suppose $X_{d\times n} = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_n})$ *and* $Y_{d\times n} = (\boldsymbol{y_1}, \cdots, \boldsymbol{y_n})$ *are paired HDLSS matrices* $(n < d)$. *If all the columns of* X *and* Y *are independent and they are from distributions which are absolutely continuous with respect to* d *dimensional Lebesgue measure, the* $\mathbf{v}_{cod}$ *between* X *and* Y *exists and is unique almost surely (modulo the ± flip of direction).*

Proof. We could give weaker conditions for the existence and uniqueness of COD. However, they are very complicated. Note that when the conditions in this theorem are satisfied, the matrices X, Y and $X - Y$ are full rank almost surely. And their eigenvalues are not the same almost surely. The algorithms for the COD will be given in Section 2.2 and it indicates the proof for this theorem. The conditions are very likely to be satisfied in the real data analysis.

$\square$

The NCI60 data projected onto the plane generated by $\mathbf{v}_{cpd}$ and $\mathbf{v}_{cod}$ are shown in the right plot of Figure 2.1. The differences between cDNA and Affy are shown clearly on the CPD. It looks similar to the left plot of Figure 2.1. Both of them show that all line segments are parallel. However, they are not the same. On the x axis, the data points spread from around -20 to 30 on the left plot, and from around -20 to 40 on the right plot. On the y axis, the data points are distributed from 0 to around 150 in the left plot, and from 0 to around 300 on the right plot. Thus the right plot shows much stronger differences between the two data sets and much more variations on the x axis.

As shown in Figure 2.1, the CPD shows the systematic difference between Affy and cDNA. This difference can be eliminated by shifting two data sets along the CPD until the two centers overlaps, as we have done for the other linear adjustment method in Chapter 1, Section 1.3. Affy data have much larger variation than cDNA data. Thus after linear shifting, we standardize each column of the data (each entry is subtracted by the column mean, and divided by the column standard deviation) to adjust the variation difference. The adjusted data are projected onto the first four PC direction of the Raw data, in order to compare with the projection view of the raw data.

In Figure 2.2, line segments become much shorter than those in Figure 1.6, which indicate the systematic batch difference has been successfully removed. In addition, some biological clusters emerge in Figure 2.2 for the data after adjustment. E.g. in the second row, third column subplot, a cluster, colored as read, shows up in the right part of the plot. This cluster has been examined to be a cluster of melanoma cancer cell lines. In the second row, forth column subplot, there is a cluster in the top corner, colored as blue. It has been examined to the cluster of leukemia cell lines. These two clusters can not be seen clearly

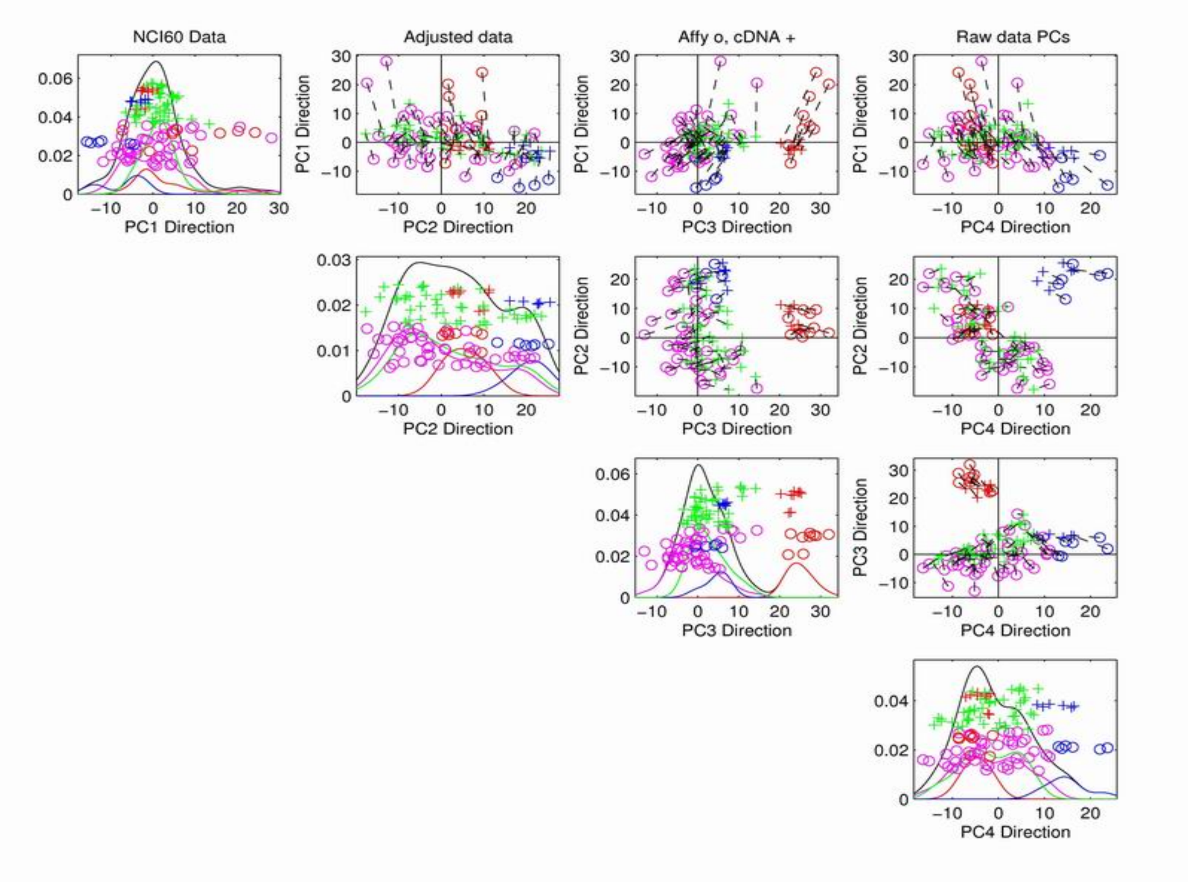


Figure 2.2: The NCI60 data are adjusted using CPD and then are column standardized. The adjusted data are projected onto the first four PC directions of the raw data. Symbols and colors are the same as in Figure 1.6.

in Figure 1.6. Thus by adjusting data along the CPD, we boost statistical power to detect some biological clusters.

In the next section, we will present the algorithms for producing the CPD and the COD for paired HDLSS data sets. The algorithms indicate the proofs for Theorem 2.1.1 and Theorem 2.1.2.

2.2 Canonical Parallel Direction (CPD) and Canonical Orthogonal Direction (COD)

In Section 2.2.1, we review some fundamental results about linear algebra and the Principal Component Analysis (PCA). There are many papers and books about PCA. One recommended reference is the book by Jolliffe (2002). In Section 2.2.2, we give algorithms to produce the CPD and the COP. The algorithms establish the existence and uniqueness of these two directions and can be treated as the proofs of Theorem 2.1.1 and Theorem 2.1.2.

2.2.1 Linear Algebra and PCA Overview

Definition 2.2.1. A matrix M is called **symmetric**, if it equals its transpose.
A matrix M is called a **square matrix**, if it has the same number of rows and columns.

Lemma 2.2.1. *For a real-valued symmetric square matrix $M_{d\times d}$, there exists an* ***eigenvalue decomposition****,*

$$M = VDV^T,$$

such that $D_{d\times d}$ is a diagonal matrix,

$$D = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_d \end{pmatrix},$$

and $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d \geqslant 0$ are called ***eigenvalues****; $V_{d\times d}$ is an orthonormal matrix, which means $V^TV = VV^T = I$. The columns of $V = (\boldsymbol{v_1}, \cdots, \boldsymbol{v_d})$ are called* ***eigenvectors****. Specifically, $\boldsymbol{v_i}$ is called the ith eigenvector.*

Note that if λ_1 is positive and strictly larger than the rest eigenvalues, the first eigenvector $\boldsymbol{v_1}$ exist and is unique. If the columns of X are independent with each other and are from distributions which are absolutely continuous with respect to d dimensional lebesgue measure, then λ_1 is positive and strictly larger than the rest eigenvalues almost surely, which means that the first eigenvector $\boldsymbol{v_1}$ exists and is unique almost surely (modulo the ± flip of direction).

Lemma 2.2.2. *Suppose* $X_{d\times n} = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_n})$ *is a real-valued matrix. If we view the columns of* X *as vectors in the* d *dimensional Euclidean space, the first eigenvector of* XX^T *is the direction such that the projections of all the column vectors on this direction have the maximum sum of squared length.*

This result is very well known; see Jolliffe (2002). The details of the proof are written out here because a very similar idea is used for the computation of canonical directions.

Proof. Assume $X_{d\times n} = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_n})$, where $\boldsymbol{x_i}$ is the *ith* column of X. Given any normalized direction vector $\boldsymbol{\mu} \in \mathcal{R}^d$ (i.e. $\|\boldsymbol{\mu}\| = 1$), the projection of $\boldsymbol{x_i}$ in this direction is denoted as $P_{\boldsymbol{\mu}}(\boldsymbol{x_i})$. Then the sum of squared lengths of the projected column vectors of X are

$$\begin{aligned}
\sum_{i=1}^{n} \|P_{\boldsymbol{\mu}}(\boldsymbol{x_i})\|^2 &= \sum_{i=1}^{n} \|\langle \boldsymbol{x_i}, \boldsymbol{\mu}\rangle \boldsymbol{\mu}\|^2 = \sum_{i=1}^{n} \langle \boldsymbol{x_i}, \boldsymbol{\mu}\rangle^2 \|\boldsymbol{\mu}\| \\
&= \sum_{i=1}^{n} \langle \boldsymbol{x_i}, \boldsymbol{\mu}\rangle^2 = \sum_{i=1}^{n} (\boldsymbol{x_i}^T \boldsymbol{\mu})^2 \\
&= \sum_{i=1}^{n} \boldsymbol{\mu}^T \boldsymbol{x_i} \boldsymbol{x_i}^T \boldsymbol{\mu} \\
&= \boldsymbol{\mu}^T X X^T \boldsymbol{\mu}
\end{aligned}$$

Since XX^T is a real-valued symmetric square matrix, according to Lemma 2.2.1, there is an eigenvalue decomposition, such that

$$XX^T = VDV^T.$$

Thus

$$\sum_{i=1}^{n} \|P_{\boldsymbol{\mu}}(\boldsymbol{x_i})\|^2 = \boldsymbol{\mu}^T X X^T \boldsymbol{\mu} = (\boldsymbol{\mu}^T V) D (\boldsymbol{\mu}^T V)^T.$$

Because $\boldsymbol{\mu}^T V = \boldsymbol{\mu}^T(\boldsymbol{v_1}, \cdots, \boldsymbol{v_d}) = (\langle \boldsymbol{\mu}, \boldsymbol{v_1}\rangle, \cdots, \langle \boldsymbol{\mu}, \boldsymbol{v_d}\rangle)$, and $D = diag(\lambda_1, \cdots, \lambda_d)$, we have

$$\sum_{i=1}^{n} \|P_{\boldsymbol{\mu}}(\boldsymbol{x_i})\|^2 = \sum_{i=1}^{d} \lambda_i \langle \boldsymbol{\mu}, \boldsymbol{v_i} \rangle^2.$$

Since V is an orthonormal matrix, we have$\boldsymbol{\mu} = \sum_{i=1}^{d} \langle \boldsymbol{\mu}, \boldsymbol{v_i} \rangle \boldsymbol{v_i}$. It follows that

$$\begin{aligned}\sum_{i=1}^{d} \langle \boldsymbol{\mu}, \boldsymbol{v_i} \rangle^2 &= \langle \boldsymbol{\mu}, \sum_{i=1}^{d} \langle \boldsymbol{\mu}, \boldsymbol{v_i} \rangle \boldsymbol{v_i} \rangle \\ &= \langle \boldsymbol{\mu}, \boldsymbol{\mu} \rangle \\ &= \|\boldsymbol{\mu}\|^2 = 1.\end{aligned}$$

If the eigenvalues are ordered, e.g. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, $\sum_{i=1}^{n} \|P_{\boldsymbol{\mu}}(\boldsymbol{x_i})\|^2 = \sum_{j=i}^{d} \lambda_i \langle \boldsymbol{\mu}, \boldsymbol{v_i} \rangle^2$ is maximized (over $\boldsymbol{\mu}$) by putting a maximal amount of the energy in the largest direction, i.e. $\boldsymbol{\mu} = \boldsymbol{v_1}$, and

$$max \sum_{i=1}^{n} \|P_{\boldsymbol{\mu}}(\boldsymbol{x_i})\|^2 = \lambda_1$$

The direction which maximizes the sum of squared projected lengths in this direction is the first eigenvector of XX^T. Again, as in Lemma 2.2.1, if λ_1 is positive and strictly larger than the rest eigenvalues, the first eigenvector $\boldsymbol{v_1}$ exist and is unique. If the columns of X are independent with each other and are from distributions which are absolutely continuous with respect to d dimensional lebesgue measure, then λ_1 is positive and strictly larger than the rest eigenvalues almost surely, which means that the first eigenvector $\boldsymbol{v_1}$ exists and is unique almost surely (modulo the ± flip of direction).

□

Lemma 2.2.3. *$X_{d\times n} = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_n})$ can be viewed as n points in the d dimensional Euclidean space. The center of these points is expressed as $\bar{\boldsymbol{x}} \doteq \frac{1}{n}(\boldsymbol{x_i} + \cdots + \boldsymbol{x_n})$. We defined $\bar{X}$ as a matrix with n duplicate columns, $\bar{\boldsymbol{x}}$, which means $\bar{X} = (\bar{\boldsymbol{x}}, \cdots, \bar{\boldsymbol{x}})$. Then, the first eigenvector of $(X - \bar{X})(X - \bar{X})^T$ is the direction such that the projections of these n points on this direction have the maximum variation.*

Lemma 2.2.3 is also well known; see Jolliffe (2002). The proof of this lemma is very similar to the proof of Lemma 2.2.2.

Proof. $X_{d\times n} = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_n})$. Given a direction vector $\boldsymbol{\mu} \in \mathcal{R}^d$. (i.e. $\|\boldsymbol{\mu}\| = 1$). The projection of $\bar{\boldsymbol{x}}$ in this direction is $P_{\boldsymbol{\mu}}(\bar{\boldsymbol{x}})$. The center of all these projections is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} P_{\boldsymbol{\mu}}(\boldsymbol{x_i}) &= \sum_{i=1}^{n}\frac{1}{n}\langle \boldsymbol{x_i}, \boldsymbol{\mu}\rangle\boldsymbol{\mu} \\ &= \langle \bar{\boldsymbol{x}}, \boldsymbol{\mu}\rangle\boldsymbol{\mu} \\ &= P_{\boldsymbol{\mu}}(\bar{\boldsymbol{x}})\end{aligned}$$

The center of the projections is exactly the projection of $\bar{\boldsymbol{x}}$ in this direction, $P_{\boldsymbol{\mu}}(\bar{\boldsymbol{x}})$. Thus the variation of the projections of the data on the direction $\boldsymbol{\mu}$ is

$$\begin{aligned}\sum_{i=1}^{n}\|P_{\boldsymbol{\mu}}(\boldsymbol{x_i}) - P_{\boldsymbol{\mu}}(\bar{\boldsymbol{x}})\|^2 &= \sum_{i=1}^{n}\|\langle \boldsymbol{x_i}, \boldsymbol{\mu}\rangle\boldsymbol{\mu} - \langle \bar{\boldsymbol{x}}, \boldsymbol{\mu}\rangle\boldsymbol{\mu}\|^2 \\ &= \sum_{i=1}^{n}\|\langle(\boldsymbol{x_i} - \bar{\boldsymbol{x}})\boldsymbol{\mu}\rangle\boldsymbol{\mu}\|^2 = \sum_{i=1}^{n}\langle(\boldsymbol{x_i} - \bar{\boldsymbol{x}})\boldsymbol{\mu}\rangle^2\|\boldsymbol{\mu}\| \\ &= \sum_{i=1}^{n}\langle(\boldsymbol{x_i} - \bar{\boldsymbol{x}}), \boldsymbol{\mu}\rangle^2 = \sum_{i=1}^{n}((\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T\boldsymbol{\mu})^2 \\ &= \sum_{i=1}^{n}\boldsymbol{\mu}^T(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T\boldsymbol{\mu} \\ &= \boldsymbol{\mu}^T(X - \bar{X})(X - \bar{X})^T\boldsymbol{\mu}\end{aligned}$$

The rest of the argument is very similar to the proof for Lemma 2.2.2 . We conclude that when μ is the first eigenvector direction of $(X - \bar{X})(X - \bar{X})$, the projections of the data on this direction have the maximum variation. This first eigenvector is also called the first principal component direction of the matrix X. Again, if the columns of X, Y are independent with each other and are from distributions which are absolutely continuous with respect to d dimensional lebesgue measure, the first eigenvector of $(X - \bar{X})(X - \bar{X})^T$ exists and is unique almost surely (modulo the ± flip of direction).

$\square$

2.2.2 Algorithm for the CPD and COD

In this part, the algorithms for the computations of the two canonical directions are developed. We will also discuss the existence and uniqueness of these two directions. The discussion results in the proofs of Theorem 2.1.1 and Theorem 2.1.2.

Again, we assume that $X_{d\times n} = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_n})$ and $Y_{d\times n} = (\boldsymbol{y_1}, \cdots, \boldsymbol{y_n})$ are paired HDLSS data sets, which means that $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ $(i = 1, \cdots, n)$ are the expression vectors for associated samples. E.g. for the NCI60 data, we have X as the expression matrix for the cDNA samples and Y as the expression matrix for the corresponding Affy samples, measured on the same list of genes. The direction vectors of the line segments which connect the same sample from different platforms are the columns of $X - Y$.

Algorithm for the CPD

We intend to find a vector $\mathbf{v}_{cpd}$ which maximizes the sum of squared lengths of the projected line segments in this direction. That is to maximize

$$\sum_{i=1}^{n} \|P_{\mathbf{v}_{cpd}}(\boldsymbol{x_i} - \boldsymbol{y_i})\|^2 = \mathbf{v}_{cpd}^T (X - Y)(X - Y)^T \mathbf{v}_{cpd} \quad (over \ \ \mathbf{v}_{cpd}).$$

According to Lemma 2.2.2, $\mathbf{v}_{cpd}$ is the first eigenvector of $(X - Y)(X - Y)^T$, which can be easily calculated by eigenvalue analysis.

If the first eigenvalue of $(X - Y)(X - Y)^T$ is strictly larger than all the rest eigenvalues, the first eigenvector of $(X - Y)(X - Y)^T$ exists and is unique (modulo the ± flip of direction), which means that the CPD exists and is unique. This proves Theorem 2.1.1.

Algorithm for the Canonical Orthogonal Direction

Before we give the algorithm for COD, we first introduce some definitions and lemmas about the linear algebra.

Definition 2.2.2. A nonzero vector $\boldsymbol{\mu} \in \mathcal{R}^d$ is called a **normalized direction vector**, if $\|\boldsymbol{\mu}\| = 1$.

Definition 2.2.3. We define the following notations:

$\mathcal{H}_X$: the space spanned by the column vectors of X.

$\mathcal{H}_{[X,Y]}$: the space spanned by all the column vectors of X and Y.

$\mathcal{H}_{X-Y}$: the space spanned by the column vectors of $X - Y$.

Definition 2.2.4. ${\mathcal{H}_X}^{\perp}$: the orthogonal complement of the space $\mathcal{H}_X$ in $\mathbb{R}^d$, which means $\mathcal{H}_X \oplus {\mathcal{H}_X}^{\perp} = \mathbb{R}^d$.

$\mathcal{H}_{[X,Y]/X}$ is defined as the orthogonal complement of the space $\mathcal{H}_X$ in the space $\mathcal{H}_{[X,Y]}$, which means $\mathcal{H}_X \oplus \mathcal{H}_{[X,Y]/X} = \mathcal{H}_{[X,Y]}$.

Lemma 2.2.4. *Let $\mathcal{H}$ be any proper subspace of $\mathcal{R}^d$. $\mathcal{H}^{\perp}$ is the orthogonal complement of the space $\mathcal{H}$. For any nonzero vector $\boldsymbol{\mu} \in \mathcal{R}^d$, there exist two normalized vectors $\boldsymbol{\mu_1} \in \mathcal{H}$ and $\boldsymbol{\mu_2} \in \mathcal{H}^{\perp}$, such that $\boldsymbol{\mu}$ has an **orthogonal decomposition:***

$$\boldsymbol{\mu} = \langle \boldsymbol{\mu}, \boldsymbol{\mu_1} \rangle \boldsymbol{\mu_1} + \langle \boldsymbol{\mu}, \boldsymbol{\mu_2} \rangle \boldsymbol{\mu_2}.$$

Note that If $\boldsymbol{\mu} \notin \mathcal{H}$ and $\boldsymbol{\mu} \notin \mathcal{H}^{\perp}$, the two such directions $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are unique (modulo the ± flip of direction). The $\boldsymbol{\mu_1}$ is actually the direction vector of the projection of $\boldsymbol{\mu}$ onto the space $\mathcal{H}$, and the $\boldsymbol{\mu_2}$ is the direction vector of the projection of $\boldsymbol{\mu}$ onto the space $\mathcal{H}^{\perp}$.

Suppose $\mathbf{v}_{cod}$ is the canonical orthogonal direction in Theorem 2.1.2. According to

Lemma 2.2.4, it can be orthogonally decomposed into two directions such that

$$\mathbf{v}_{cod} = \langle \mathbf{v}_{cod}, \boldsymbol{\mu_1} \rangle \boldsymbol{\mu_1} + \langle \mathbf{v}_{cod}, \boldsymbol{\mu_2} \rangle \boldsymbol{\mu_2},$$

where $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$ are normalized vectors and $\boldsymbol{\mu_1} \in \mathcal{H}_{[X,Y]}$, $\boldsymbol{\mu_2} \in \mathcal{H}_{[X,Y]}{}^{\perp}$. Then the projection of any vector $\mathbf{v} \in \mathcal{H}_{[X,Y]}$ on this normalized direction $\mathbf{v}_{cod}$ can be expressed as:

$$\begin{aligned} P_{\mathbf{v}_{cod}}(\mathbf{v}) &= \langle \mathbf{v}, \mathbf{v}_{cod} \rangle \mathbf{v}_{cod} \\ &= \langle \mathbf{v}, (\langle \mathbf{v}_{cod}, \boldsymbol{\mu_1} \rangle \boldsymbol{\mu_1} + \langle \mathbf{v}_{cod}, \boldsymbol{\mu_2} \rangle \boldsymbol{\mu_2}) \rangle \mathbf{v}_{cod} \\ &= \langle \mathbf{v}, \langle \mathbf{v}_{cod}, \boldsymbol{\mu_1} \rangle \boldsymbol{\mu_1} \rangle \mathbf{v}_{cod} + \langle \mathbf{v}, \langle \mathbf{v}_{cod}, \boldsymbol{\mu_2} \rangle \boldsymbol{\mu_2} \rangle \mathbf{v}_{cod} \\ &= \langle \mathbf{v}_{cod}, \boldsymbol{\mu_1} \rangle \langle \mathbf{v}, \boldsymbol{\mu_1} \rangle \mathbf{v}_{cod} + \langle \mathbf{v}_{cod}, \boldsymbol{\mu_2} \rangle \langle \mathbf{v}, \boldsymbol{\mu_2} \rangle \mathbf{v}_{cod}. \end{aligned}$$

Since $\mathbf{v} \in \mathcal{H}_{[X,Y]}$, we have $\langle \mathbf{v}, \boldsymbol{\mu_2} \rangle = 0$ (because $\boldsymbol{\mu_2} \in \mathcal{H}_{[X,Y]}{}^{\perp}$). Thus ,

$$P_{\mathbf{v}_{cod}}(\mathbf{v}) = \langle \mathbf{v}_{cod}, \boldsymbol{\mu_1} \rangle \langle \mathbf{v}, \boldsymbol{\mu_1} \rangle \mathbf{v}_{cod}. \tag{2.1}$$

Recall that Definition 2.1.2 requires that $\mathbf{v}_{cod}$ firstly needs be orthogonal to all the direction vectors of the line segments, which means it is orthogonal to the space $\mathcal{H}_{X-Y}$, thus

$$P_{\mathbf{v}_{cod}}(X - Y) = 0 \Longrightarrow P_{\mathbf{v}_{cod}}(X) = P_{\mathbf{v}_{cod}}(Y).$$

Since X and Y have exactly the same projections on the direction $\mathbf{v}_{cod}$, the second condition in Definition 2.1.2 actually assures that the COD is the one which maximizes the variability of the projections of the data in this direction. The projection of the ith sample of X can be expressed as $P_{\mathbf{v}_{cod}}(\boldsymbol{x_i})$. The center of the samples in X is $\bar{\boldsymbol{x}}$. Thus the variability of the projected data on $\mathbf{v}_{cod}$ is

$$\sum_{i=1}^{n} \| P_{\mathbf{v}_{cod}}(\boldsymbol{x_i} - \bar{\boldsymbol{x}}) \|^2.$$

Since $\boldsymbol{x_i} - \bar{\boldsymbol{x}} \in \mathcal{H}_{[X,Y]}$, we have

$$\begin{aligned}\sum_{i=1}^{n} \|P_{\mathbf{v}_{cod}}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})\|^2 &= \sum_{i=1}^{n} \|\langle \mathbf{v}_{cod}, \boldsymbol{\mu_1}\rangle \langle \boldsymbol{x_i} - \bar{\boldsymbol{x}}, \boldsymbol{\mu_1}\rangle \mathbf{v}_{cod}\|^2 \\ &= \sum_{i=1}^{n} \langle \mathbf{v}_{cod}, \boldsymbol{\mu_1}\rangle^2 \mathbf{v}_{cod}^T (\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T \mathbf{v}_{cod},\end{aligned}$$

where $\boldsymbol{\mu_1} \in \mathcal{H}_{[X,Y]}$.

In order to maximize this variation, we choose $\mathbf{v}_{cod}$ such that $\langle \mathbf{v}_{cod}, \boldsymbol{\mu_1}\rangle = 1$. This means that $\mathbf{v}_{cod} \in \mathcal{H}_{[X,Y]}$, i.e. the maximizing direction is in the subspace generated by the data. Considering $\mathbf{v}_{cod} \perp \mathcal{H}_{X-Y}$ (because $\mathbf{v}_{cod}$ is orthogonal to all the direction vectors of the line segments), we have $\mathbf{v}_{cod} \in \mathcal{H}_{[X,Y]/(X-Y)}$. This also means $\mathbf{v}_{cod} \in \mathcal{H}_{[X-Y,Y]/(X-Y)}$, since $\mathcal{H}_{[X,Y]} = \mathcal{H}_{[X-Y,Y]}$.

Next, we will derive a set of basis vectors for the space $\mathcal{H}_{[X-Y,Y]/(X-Y)}$. Suppose the matrix $[X-Y, Y]$ has an orthogonal-triangular decomposition

$$[X-Y,Y]_{d\times 2n} = Q_{d\times 2n} R_{2n\times 2n},$$

where R is an upper triangular matrix, and Q is a $d\times 2n$ unitary matrix ($Q^T Q = I_{2n\times 2n}$). As we mentioned in Theorem 2.1.2, the columns of X and Y are from continuous distributions, which assumes that $[X-Y,Y]$ is a full rank matrix a.s. and hence both Q and R are full rank matrices a.s. These two matrices exist and are unique if we ignore the direction ± flip in Q and ignore the sign of the corresponding entries in R. We decompose Q as $Q = [Q_1, Q_2]$, where Q_1 is the first n columns, and Q_2 is the last n columns of Q. Because R is a full rank upper triangular matrix, Q_1 forms a basis for the space $\mathcal{H}_{X-Y}$ and Q_2 forms a set basis vectors for the space $\mathcal{H}_{[X-Y,Y]/(X-Y)}$, i.e.

$$\mathcal{H}_{Q_1} = \mathcal{H}_{X-Y},$$

$$\mathcal{H}_{Q_2} = \mathcal{H}_{[X-Y,Y]/(X-Y)}.$$

Since $\mathbf{v}_{cod} \in \mathcal{H}_{[X-Y,Y]/(X-Y)} = \mathcal{H}_{Q_2}$, it can be expressed as a linear combination of the columns of Q_2 , say

$$\mathbf{v}_{cod} = Q_2 C,$$

where C is an $n \times 1$ vector.

The variation to be maximized (over $\boldsymbol{\mu}$, i.e. over C) is :

$$\begin{aligned}
\sum_{i=1}^{n} \|P_{\mathbf{v}_{cod}}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})\|^2 &= \sum_{i=1}^{n} \mathbf{v}_{cod}^T(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T \mathbf{v}_{cod} \\
&= \sum_{i=1}^{n} C^T(Q_2^T(\boldsymbol{x_i} - \bar{\boldsymbol{x}}))((\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T Q_2)C \\
&= C^T(Q_2^T(X - \bar{X}))(Q_2^T(X - \bar{X}))^T C.
\end{aligned}$$

From Lemma 2.2.3, in order to maximize the above variability, we choose C as the first eigenvector of

$$Q_2^T(X - \bar{X})(X - \bar{X})^T Q_2.$$

To produce the canonical orthogonal direction, we first calculate Q_2 by the orthogonal-triangular decomposition of $[X - Y, Y]$, then we get C as the first eigenvector of $Q_2^T(X - \bar{X})(X - \bar{X})^T Q_2$ by the eigenvalue analysis. The canonical orthogonal direction is

$$\mathbf{v}_{cod} = Q_2 C.$$

When the columns of X and Y are independent with each other and are from distributions which are absolutely continuous with respect to d dimensional lebesgue measure, each of X, Y, and $X - Y$ is a full rank matrix a.s. Thus, the orthogonal-triangular decomposition exists and is unique a.s (modulo the ± flip of directions). Also, the first eigenvector of $Q_2^T(X - \bar{X})(X - \bar{X})^T Q_2$ exists and is unique a.s. These establishes the existence and uniqueness of the COD, and can be treated as the proof for Theorem 2.1.2.

Note that $\mathbf{v}_{cpd}$ is the first eigenvector of $(X - Y)(X - Y)^T$, thus $\mathbf{v}_{cpd} \in \mathcal{H}_{X-Y}$. The COD is orthogonal to all the directions of line segments, i.e. $\mathbf{v}_{cod} \in \mathcal{H}_{[X-Y,Y]/(X-Y)}$. Thus we have $\mathbf{v}_{cod} \perp \mathbf{v}_{cpd}$. The definitions of these two directions assure that they are orthogonal to each other, hence we could derive CPD and COD separately.

2.3 Asymptotic results for the CPD

The CPD shows the systematic differences between two paired HDLSS data. It could be used for adjusting the these differences, as we have done for the NCI60 data in Section 2.1. In the previous Section, we have given the algorithms to produce CPD and these algorithms indicate the existence and uniqueness of the CPD, under some mild conditions. In this Section, we will study the asymptotic properties of the CPD using a **linear shift model**, when the sample sizes are fixed and the dimension increases to infinity. In Section 2.3.1, we discussed three types of viewpoints to study asymptotic properties. Section 2.3.2 introduces a linear shift model, which is an underlying conceptual model for studying the batch difference between two HDLSS data sets with Gaussian errors. In Section 2.3.3, we study the asymptotic properties of the CPD for two data sets under the linear shift model. Section 2.3.4 gives the simulation verification for the results in Section 2.3.3.

2.3.1 Three Types of Asymptotic Studies

Using multivariate view, a random matrix $X_{d,n}$ are viewed as n vectors in d dimensional space, or n samples from the distribution of a d dimensional variable. There are at least three types of asymptotic viewpoints to study a random matrix $X_{d,n}$. We call them the n asymptotics, the (d, n) asymptotics and the d asymptotics.

The n asymptotics

The n asymptotics studies the problem when the dimension of the variable d is fixed and the sample size n goes to infinity. This is the traditional mathematical statistical setting, such as the normal approximation to Maximum Likelihood Estimation (MLE); the central limit theorem and so on. However, the concepts that are revealed by this approach are not very relevant to HDLSS data analysis, because the sample size n is small, and even smaller than the dimension d.

The (d,n) asymptotics

The (d,n) asymptotic studies the problem when both d and n increase to infinity. This research falls in the area, called *random matrices*, see Silverstein (1989), Bai *et al.* (1988). The main problems include the distribution of the eigenvalues, the spectral measure of a random symmetric matrix and so on. Fujikoshi (2004) reviewed some (d,n) asymptotic results. Johnstone (2001) studied the distribution of the first eigenvalue of the random matrix, when the dimension d and the sample size n both increase to infinity and the ratio of them goes to 0, a constant, and ∞ respectively.

The d asymptotics

The third type of asymptotics is d asymptotics, which means that the sample size n is fixed and the dimension d goes to infinity. This viewpoint is much more practical than the first two, especially in micorarray data analysis. Hall *et al.* (2005) studied the geometric representation of a random matrix $X_{d,n}$, when the dimension is high. From multivariate view, each column of X_{dn} is a point in d dimensional space. The matrix X_{dn} can be represented as a cloud of n points in the d dimensional space. Hall *et al.* (2005) conclude that when d goes to infinity, under some mild conditions, these points converge to the vertices of a simplex with all the edges of the same length, after scaling by a constant $d^{-\frac{1}{2}}$. They also study and compare the d asymptotic properties of several discrimination methods, such as SVM, PAM and DWD. We will discuss these results in Chapter 3. Ahn *et al.* (2005) establish the same result as in Hall *et al.* (2005) under a milder condition with Gaussian assumptions, which will be discussed in Theorem 2.3.2.

In this dissertation, we will focus on the d asymptotics for HDLSS data. The $d-$asymptotics provide an important viewpoint of HDLSS data. E.g, for a microarray data set, It explains what will happen if the number of measured genes increases. In the next subsection, we will introduce an underlying conceptual model, called the *linear shift model* to study the CPD between two data sets.

2.3.2 Linear Shift Model

Suppose that $\{(X^{(1)}, Y^{(1)}) \cdots, (X^{(d)}, Y^{(d)}), \cdots\}$ is a series of paired HDLSS random matrices, where the dimensions of these paired matrices are $1 \times n, \cdots, d \times n, \cdots$ respectively. For example, the first paired matrices $(X^{(1)}, Y^{(1)})$ are the expression values for 1 genes, and the paired matrices $(X^{(d)}, Y^{(d)})$ are the expression values for d genes, $d = 1, 2, 3, \cdots$. From now on, any variable with superscript (d) indicates that it is specifically for the data with d genes.

Using the multivariate view, each of $X^{(d)}_{d\times n} = (\boldsymbol{x_1}^{(d)}, \cdots, \boldsymbol{x_n}^{(d)})$ and $Y^{(d)}_{d\times n} = (\boldsymbol{y_1}^{(d)}, \cdots, \boldsymbol{y_n}^{(d)})$ is a cloud of n points in the d dimensional space $(d \geqslant n)$. We construct the linear shift model, such that

$$\begin{aligned} \boldsymbol{x_i}^{(d)} &= \boldsymbol{s_i}^{(d)} + \boldsymbol{\epsilon}^{(d)}_{1,i}, && (2.2) \\ \boldsymbol{y_i}^{(d)} &= \boldsymbol{s_i}^{(d)} + \mathbf{v}^{(d)} + \boldsymbol{\epsilon}^{(d)}_{2,i} \qquad (i = 1, 2, \cdots, n). && (2.3) \end{aligned}$$

The $\boldsymbol{s_i}^{(d)}$ represents the vector for the true expression values of d genes in the ith array, and it is unknown. In the batch $X^{(d)}$, the observation vector of the ith array is the sum of $\boldsymbol{s_i}^{(d)}$ (true expression values) and $\boldsymbol{\epsilon}^{(d)}_{1,i}$ (random errors). In the other batch, $Y^{(d)}$, the observations have systematic batch difference $\mathbf{v}^{(d)}$, from the observations in the batch $X^{(d)}$. The systematic difference $\mathbf{v}^{(d)} = (v^{(d)}_1, \cdots, v^{(d)}_d)^T$ is a d dimensional vector. The asymptotic norm of the triangular sequence $\{\mathbf{v}^{(d)}(d = n+1, \cdots)\}$ is of the order cd^{α}, in the sense that

$$\lim_{d\to\infty} \| \frac{1}{cd^{\alpha}} \mathbf{v}^{(d)} \| = 1, \tag{2.4}$$

where c is a constant and α is the parameter which describes how fast the length of the systematic differences increase as the dimension d goes to infinity. For example, $\mathbf{v}^{(d)}_{d\times 1} = (1, \cdots, 1)^T$ has $c = 1, \quad \alpha = \frac{1}{2}$. The difference vectors are the same for any pair of arrays, i.e. $(\boldsymbol{x_i}^{(d)}, \boldsymbol{y_i}^{(d)}), \quad i = 1, 2, \cdots, n$. We define the normalized direction vector of $\mathbf{v}^{(d)}$ as $\mathbf{v}^{(d)}_t$, i.e.

$$\mathbf{v}^{(d)}_t = \frac{1}{cd^{\alpha}} \mathbf{v}^{(d)} \tag{2.5}$$

then the asymptotic norm of $\mathbf{v}_t^{(d)}$ is 1, i.e.

$$\lim_{d\to\infty} \|\mathbf{v}_t^{(d)}\| = 1, \tag{2.6}$$

The errors vectors $\boldsymbol{\epsilon}_{1,i}^{(d)}, \boldsymbol{\epsilon}_{2,i}^{(d)}$ $(i = 1, \cdots, n;\ d = n+1, \cdots)$ are i.i.d random variables and follow the multivariate Gaussian distribution with mean zero and a given sequence covariance matrices $\{\Sigma^{(d)} \quad (d = n+1, \cdots)\}$.

If we define

$$S^{(d)} = (\boldsymbol{s_1}^{(d)}, \cdots, \boldsymbol{s_n}^{(d)})_{d\times n}, \quad V_t^{(d)} = (\mathbf{v}_t^{(d)}, \cdots, \mathbf{v}_t^{(d)})_{d\times n},$$

$$\Upsilon_1^{(d)} = (\boldsymbol{\epsilon}_{1,1}^{(d)}, \cdots, \boldsymbol{\epsilon}_{1,n}^{(d)})_{d\times n}, \quad \Upsilon_2^{(d)} = (\boldsymbol{\epsilon}_{2,1}^{(d)}, \cdots, \boldsymbol{\epsilon}_{2,n}^{(d)})_{d\times n},$$

Equations (2.2) and (2.3) can be expressed as

$$X^{(d)} = S^{(d)} + \Upsilon_1^{(d)}, \tag{2.7}$$

$$Y^{(d)} = S^{(d)} + cd^{\alpha} V_t^{(d)} + \Upsilon_2^{(d)} \tag{2.8}$$

Figure 2.3 shows how the data sets are constructed. Each point in Figure 2.3 represents a d dimensional vector of the expression values for an array. The black dots are the true expression values for the arrays in batch $X^{(d)}$, i.e. $\boldsymbol{s_i}^{(d)}$s in Equations (2.2) and (2.3). Blue dots represent the observations in the batch $X^{(d)}$, each of which deviates from the true expression value $\boldsymbol{s_i}$ by a Gaussian random variable. The dashed line segments are used to connect the associated pairs. These line segments show the systematic difference vector, i.e. $\mathbf{v}_t^{(d)}$, which are exactly the same for all the paired samples. The true expression values in the batch $Y^{(d)}$, shown as black diamonds, have systematic differences with those of the batch $X^{(d)}$. The red diamonds represent the observations in the batch $Y^{(d)}$, which deviate from the true expression values by a Gaussian random variables.

2.3.3 The Consistency and Inconsistency of the empirical CPD

In the linear shift model, shown in Figure 2.3, the observed data are the blue dots and the red diamonds. If there are no measurement errors, i.e. $\boldsymbol{\epsilon}_{1,i}^{(d)} = 0, \quad \boldsymbol{\epsilon}_{2,i}^{(d)} = 0$ $(i = 1, 2, \cdots, n)$,

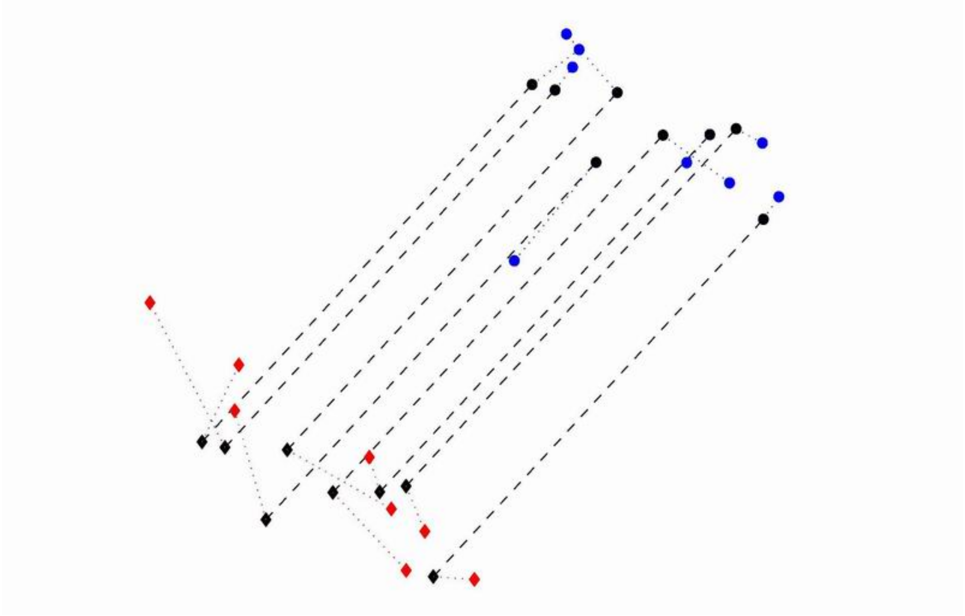


Figure 2.3: The underlying conceptual linear shift model. Blue points represent the observations in the batch $X^{(d)}$. Red diamonds represent the observations in the batch $Y^{(d)}$. Dashed lines show the direction of the systematic batch difference.

all the pair vectors which connect the blue dots and the red diamonds are in the same direction as $\mathbf{v}^{(d)}$, i.e. they are all parallel. Thus the direction vector $\mathbf{v}_t^{(d)}$ represents the theoretical Canonical Parallel Direction (**theoretical CPD**), in the sense that if there are no measurement errors, the batch difference will be totally removed after rigid shifting of the blue and red classes along this theoretical CPD. The data sets we observed $X^{(d)}$ and $Y^{(d)}$, i.e. blue dots and red diamonds, are driven by Gaussian errors. Using the algorithm in Section 2.2, we produce an empirical Canonical Parallel Direction (**empirical CPD**), denoted as $\mathbf{v}_e^{(d)}$. Because of the measurement errors in the data, the empirical CPD is usually different from the theoretical CPD. If we measure more and more genes, i.e. d goes to infinity, what will be the difference between them?

Note that the empirical CPD $\mathbf{v}_e^{(d)}$ is a direction vector, i.e. $\|\mathbf{v}_e^{(d)}\| = 1$ and the theoretical CPD has asymptotic norm 1, i.e. $\lim\limits_{d\to\infty} \|\mathbf{v}_t^{(d)}\| = 1$. We use the Absolute value of the Inner Product (AIP) between the theoretical and the empirical CPD, i.e. $AIP = |(\mathbf{v}_t^{(d)})^T \mathbf{v}_e^{(d)}|$ to evaluate the similarity between them. Thus $AIP \to 1$ in probability (for any given $\epsilon > 0$, $\lim\limits_{d\to\infty} P(|AIP - 1| > \epsilon) = 0$) means that $\mathbf{v}_t^{(d)}$ and $\mathbf{v}_e^{(d)}$ are asymptotically the same in probability (modulo the ± flip of direction), which we called the *consistency* of $\mathbf{v}_e^{(d)}$. The statement $AIP \to 0$ in probability (for any given $\epsilon > 0$, $\lim_{d\to\infty} P(AIP > \epsilon) = 0$) indicates that $\mathbf{v}_t^{(d)}$ and $\mathbf{v}_e^{(d)}$ are asymptotically orthogonal in probability, which is called the *strong inconsistency* of $\mathbf{v}_e^{(d)}$.

The algorithm in Chapter 2, Section 2.2.2 indicates that the CPD between two data sets $X^{(d)}$ and $Y^{(d)}$ is the first eigenvector of the matrix $X^{(d)} - Y^{(d)}$. Thus studying the d asymptotic properties of the CPD is similar to studying the d asymptotic properties of the first eigenvector of a $d \times n$ matrix. The following theorem presents the d asymptotic results for the CPD between $X^{(d)}$ and $Y^{(d)}$ under the linear shift model, when the sequence of covariance matrices $\Sigma^{(d)}$ $(d = n+1, \cdots)$ of the errors is a sequence of identity matrices.

Theorem 2.3.1. *In the linear shift model of Section 2.3.2, if the sequence of covariance matrices of the errors* Σ_d $(d = n+1, \cdots)$ *is a sequence of identity matrices* $I_{d\times d}$ $(d = n+1, \cdots)$, *depending on the assumed value of* α *(note:* $\mathbf{v}_t^{(d)} = \frac{1}{cd^\alpha}\mathbf{v}^{(d)}$*), we have the following conclusions for the empirical CPD* $\mathbf{v}_e^{(d)}$ *and the theoretical CPD* $\mathbf{v}_t^{(d)}$ *between* $X^{(d)}$ *and* $Y^{(d)}$. *As the sample size n is fixed, defining that* $AIP = |(\mathbf{v}_t^{(d)})^T \mathbf{v}_e^{(d)}|$

1: if $\alpha > \frac{1}{2}$, $\mathbf{v}_e^{(d)}$ *is asymptotically the same as* $\mathbf{v}_t^{(d)}$ *in probability, i.e.* $AIP \to 1$ *in prob. as* $d \to \infty$ *(consistency of direction)*

2: if $\alpha < \frac{1}{2}$, $\mathbf{v}_e^{(d)}$ *is asymptotically orthogonal to* $\mathbf{v}_t^{(d)}$ *in probability, i.e.* $AIP \to 0$ *in prob. as* $d \to \infty$ *(strong inconsistency of direction)*

Proof. This theorem will be proved as a special case of Theorem 2.3.3. □

Each of $X^{(d)}$ and $Y^{(d)}$ is a cloud of points in d dimensional space. As studied by Hall *et al.* (2005), the clouds expand to the vertices of a randomly rotated simplex, with all the edges having the same lengths. The speed of expansion as d goes to infinity, is decided by the covariance matrices of the errors. E.g. if the covariance matrices are identity matrices, this speed is $d^{\frac{1}{2}}$. When d goes to infinity, the length of the systematic differences also increase, with the speed of d^α. If the two data clouds expand faster than the systematic difference, then the two clouds will finally overlap. The within group variation will dominate the systematic differences and the empirical CPD between them will be asymptotically orthogonal to the theoretical one. We call this the **strong inconsistency** of the empirical CPD. Note that if two random vectors $\boldsymbol{v_1}^{(d)}$, $\boldsymbol{v_2}^{(d)}$ are independent and are from the d dimensional standard Gaussian distribution, they are asymptotically orthogonal to each other in probability, i.e.

$$\frac{|(\boldsymbol{v_1}^{(d)})^T(\boldsymbol{v_2})^{(d)}|}{\|\boldsymbol{v_1}^{(d)}\|\|\boldsymbol{v_2}^{(d)}\|} \to 0 \ \ in \ \ probability. \tag{2.9}$$

If the length of the systematic difference increases faster than the expansion of the two populations, then the systematic difference will dominate the variation of each cloud and the empirical CPD will converge to the theoretical CPD in probability. This is called the **consistency** of the empirical CPD.

In the real data analysis, the covariance matrices of the errors $\Sigma^{(d)} \ \ (d = n+1, \cdots)$ are not assured to have such simple structures as identity matrices. Hall *et al.* (2005) gave some conditions, under which the asymptotic geometric representation of a matrix $X_{d\times n}$ is the the same as if the covariance matrices are identity matrices. Ahn *et al.* (2005) assume that columns of $X_{d\times n}$ follow the Gaussian distribution and obtain the same conclusion when the eigenvalues of $\Sigma^{(d)}$ are "sufficiently diffuse", which are weaker conditions than those in Hall *et al.* (2005).

Theorem 2.3.2. *For a fixed number n, consider a sequence of random matrices* $\{X^{(1)}, \cdots, X^{(d)}, \cdots\}$, *where* $X^{(d)}$ *is a* $d\times n$ *matrix* $(d = 1, 2, \cdots)$. *The columns of* $X^{(d)}$ *are from the d dimensional normal distribution with mean zero and the covariance matrix* $\Sigma^{(d)}$. *Let* $\lambda_1^{(d)} \geqslant \cdots \geqslant \lambda_d^{(d)}$ *be the ordered triangular array of eigenvalues of the covariance matrices* $S_D^{(d)}$ $(d = 1, 2, \cdots)$, *and let* $S_D^{(d)}$ $(d = 1, 2, \cdots)$ *be the the corresponding uncentered dual sample covariance matrices, i.e.* $\Sigma^{(d)} = (X^{(d)})^T X^{(d)}$. *Suppose the eigenvalues of* $\Sigma^{(d)}$ *are sufficiently diffuse, in the sense that*

$$\frac{\sum_{j=1}^{d}(\lambda_j^{(d)})^2}{(\sum_{j=1}^{d}\lambda_j^{(d)})^2} \longrightarrow 0 \quad as \ d \longrightarrow \infty. \tag{2.10}$$

Then the sample eigenvalues behave as if they are those of the identity covariance in the *sense that* $\frac{S_D^{(d)}}{c^{(d)}} \longrightarrow I_n$ *in probability, as* $d \to \infty$*, where* $c^{(d)} = \sum_{j=1}^{d} \lambda_j^{(d)}$.

The theorem says that if the eigenvalues of the covariance matrices satisfy the condition (2.10), the data become spherical as the dimension d increases. In this situation, all the eigenvalues of the scaled sample uncentered covariance matrices $\frac{S_D^{(d)}}{c^{(d)}} = \frac{(X^{(d)})^T X^{(d)}}{c^{(d)}}$ converge to 1 in probability. The condition (2.10) means that there is no dominant set of eigenvalues. Ahn *et al.* (2005) give some cases where this conditions holds:

- Constant: $\lambda_1^{(d)} = \cdots = \lambda_d^{(d)} = C^{(d)}$, where $C^{(d)}$ can be a constant, or a function of d, because
$$\frac{\sum_{j=1}^{d}(\lambda_j^{(d)})^2}{(\sum_{j=1}^{d}\lambda_j^{(d)})^2} = \frac{d(C^{(d)})^2}{(dC^{(d)})^2} = \frac{1}{d} \longrightarrow 0 \quad as\ d \longrightarrow \infty.$$

- Fixed Blcok, Small α: $\lambda_1^{(d)} = \cdots = \lambda_k^{(d)} = c_1 d^{\alpha}, \lambda_{k+1}^{(d)} = \cdots = \lambda_d^{(d)} = c_2$, where $\alpha < 1, c_1, c_2 > 0$, because
$$\frac{\sum_{j=1}^{d}(\lambda_j^{(d)})^2}{(\sum_{j=1}^{d}\lambda_j^{(d)})^2} = \frac{kc_1^2 d^{2\alpha} + (d-k)c_2^2}{(kc_1 d^{\alpha} + (d-k)c_2)^2} = \frac{O(d \vee d^{2\alpha})}{O(d^2)} \longrightarrow 0 \quad as\ d \longrightarrow \infty.$$

- Polynomial: $\lambda_j^{(d)} = j^{-\beta}, j = 1, \cdots, d, \forall \beta > 0$, because
$$\frac{\sum_{j=1}^{d}(\lambda_j^{(d)})^2}{(\sum_{j=1}^{d}\lambda_j^{(d)})^2} = \frac{\sum_{j=1}^{d} j^{-2\beta}}{(\sum_{j=1}^{d} j^{-\beta})^2} = \frac{O(d^{-2\beta+1})}{O(d^{-2\beta+2})} \longrightarrow 0 \quad as\ d \longrightarrow \infty.$$

Ahn *et al.* (2005) also give some cases where the condition (2.10) doesn't hold:

- Fixed Blcok, Large α: $\lambda_1^{(d)} = \cdots = \lambda_k^{(d)} = c_1 d^{\alpha}, \lambda_{k+1}^{(d)} = \cdots = \lambda_d^{(d)} = c_2$, where $\alpha \geqslant 1, c_1, c_2 > 0$, because
$$\frac{\sum_{j=1}^{d}(\lambda_j^{(d)})^2}{(\sum_{j=1}^{d}\lambda_j^{(d)})^2} = \frac{kc_1^2 d^{2\alpha} + (d-k)c_2^2}{(kc_1 d^{\alpha} + (d-k)c_2)^2} \longrightarrow c_1 \quad as\ d \longrightarrow \infty.$$

- Exponential: $\lambda_j^{(d)} = \gamma^j, j = 1, \cdots, d, \forall 0 < \gamma < 1$, because
$$\frac{\sum_{j=1}^{d}(\lambda_j^{(d)})^2}{(\sum_{j=1}^{d}\lambda_j^{(d)})^2} = \frac{(1-\gamma)^2(1-\gamma^{2d})}{(1-\gamma)^2(1-\gamma^{(d)})^2} \longrightarrow \frac{1-\gamma}{1+\gamma} \quad as\ d \longrightarrow \infty.$$

- Finite Support: $\lambda_j^{(d)} = c_1, j = 1, \cdots, k, \lambda_{k+1}^{(d)} = \cdots = \lambda_d^{(d)} = 0, k < d, c_1^{(d)} > 0$, because

$$\frac{\sum_{j=1}^{d} (\lambda_j^{(d)})^2}{(\sum_{j=1}^{d} \lambda_j^{(d)})^2} = \frac{1}{k}$$

Based on the above results, we are going to study the convergence of the CPD for paired data sets $X^{(d)}, Y^{(d)}$ in the linear shift model with covariance matrices $\Sigma^{(d)}$ $(d = n+1, \cdots)$.

Theorem 2.3.3. *In the linear shift model of Section 2.3.2, assume that the eigenvalues of the covariance matrices* $\Sigma^{(d)}$ $(d = n+1, \cdots)$ *are* $\lambda_1^{(d)} \geqslant \cdots \geqslant \lambda_d^{(d)}$ *and they are sufficiently diffuse as in (2.10). Suppose that* $c^{(d)} = \sum_{i=1}^{d} \lambda_i^{(d)}$ *and* $\lim_{d \to \infty} \frac{log(c^{(d)})}{log(d)} = h$, *where* h *is a constant. Depending on the assumed value of* α, *we have the following conclusions for the empirical CPD* $\mathbf{v}_e^{(d)}$ *and the theoretical CPD* $\mathbf{v}_t^{(d)}$ *between* $X^{(d)}$ *and* $Y^{(d)}$.
As the sample size n *is fixed, again defining that* $AIP = |(\mathbf{v}_t^{(d)})^T \mathbf{v}_e^{(d)}|$,
1: if $\alpha > \frac{h}{2}$, $\mathbf{v}_e^{(d)}$ *is asymptotically the same as* $\mathbf{v}_t^{(d)}$ *in probability, i.e.* $AIP \to 1$ *in prob. as* $d \to \infty$ *(consistency of direction)*
2: if $\alpha < \frac{h}{2}$, $\mathbf{v}_e^{(d)}$ *is asymptotically orthogonal to* $\mathbf{v}_t^{(d)}$ *in probability, i.e.* $AIP \to 0$ *in prob. as* $d \to \infty$ *(strong inconsistency of direction)*

Notice that Theorem 2.3.1 is a special case of Theorem 2.3.3. When $\Sigma^{(d)} = I_d$, all the eigenvalues are 1, which are sufficiently diffuse, in the sense of Equation (2.10). Note that $\sum_{i=1}^{d} \lambda_i^{(d)} = d$ and $h = \lim_{d \to \infty} \frac{log(d)}{log(d)} = 1$. The results in Theorem 2.3.3 indicate the results in Theorem 2.3.1. Hence, we only need to prove Theorem 2.3.3. The proof of Theorem 2.3.3 will be given in Section 2.3.5.

The conclusions in Theorem (2.3.3) provide a way to examine the effect of random errors, when calculating the empirical CPD. As we have discussed before, when d goes to infinity, the two data clouds $X^{(d)}$ and $Y^{(d)}$ are expanding. Although the covariance matrices of their columns are not identity matrices, Theorem 2.3.2 indicates that these two clouds still expand to the vertices of two randomly rotated simplices respectively, normalized the eigenvalues of the covariance matrices are sufficiently diffuse. The asymptotic properties of the empirical CPD depend on the comparison between the speed of cloud expansion, i.e. $d^{h/2}$ and the speed of the increasing systematic difference, i.e. d^{α}. When $\alpha > h/2$, the systematic difference dominates the variation within each group, i.e. the variation of random errors. Hence the empirical CPD converges to the theoretical one. One the other hand, when $\alpha < h/2$, the systematic difference is relatively small, and the two approximating simplices completely overlap. In this situation, the empirical CPD is a random direction vector, thus it is orthogonal to the theoretical CPD, as seen in (2.9).

The above two theorems give the consistency and inconsistency of the empirical CPD when the eigenvalues are sufficiently diffuse. Sometimes, there is one or more eigenvalues, which dominate all the others, see the given examples which follow Theorem 2.3.2. When the condition 2.10 is not satisfied, the consistency of the empirical direction not only depends on the constant α but also the structure of the covariance matrix. The next theorem studies the data sets with a very special covariance matrix, called the **Spike Covariance Matrix**. In this situation, the condition (2.10) is not satisfied.

Theorem 2.3.4. *Two paired data sets $X^{(d)}$ and $Y^{(d)}$ are constructed as in the linear shift model, see Section 2.3.2. Suppose the covariance matrix of the measurement errors has a "spike structure", in the sense that the eigenvalues of Σ_d are $\lambda_{1,d} = d^{\beta}, \lambda_{d,2} = \cdots = \lambda_{d,d} = 1$, where $\beta \geq 1$ (If $\beta < 1$, the condition (2.10) holds; see Theorem 2.3.3). Define the d dimensional vector $\mathbf{v}_s^{(d)} = (1, 0, \cdots 0)^T$ as the* ***Spike Direction****.*

As the sample size n is fixed,

1: if $2\alpha > \beta$, $\mathbf{v}_e^{(d)}$ is asymptotically the same as the empirical CPD $\mathbf{v}_t^{(d)}$ in probability, i.e. $|(\mathbf{v}_t^{(d)})^T \mathbf{v}_e^{(d)}| \to 1$ in prob. as $d \to \infty$ (consistency of direction)

2: if $2\alpha < \beta$, $\mathbf{v}_e^{(d)}$ is asymptotically the same as the spike direction $\mathbf{v}_s^{(d)}$ in probability, i.e. $|(\mathbf{v}_s^{(d)})^T \mathbf{v}_e^{(d)}| \to 1$ in prob. as $d \to \infty$.

When the covariance matrices have the above spike structure, the first eigenvalue dominates all the other eigenvalues. When d goes to infinity, neither cloud of $X^{(d)}$ and $Y^{(d)}$ expands to the vertices of a rotated simplex. The expansion is mainly along the spike direction $\mathbf{v}_s^{(d)}$, with the rate of speed $d^{\beta/2}$. Again, the systematic differences increase with the speed of d^{α}. The asymptotic properties of the empirical CPD depend on the comparison between α and $\beta/2$. Note that in Theorem (2.3.4), when $2\alpha > \beta$, the empirical CPD $\mathbf{v}_e^{(d)}$

converge to the theoretical CPD $\mathbf{v}_t^{(d)}$. However, it may or may not be orthogonal with the spike direction $\mathbf{v}_s^{(d)}$, because the theoretical CPD is not necessarily orthogonal with the spike direction.

2.3.4 Simulation Study

In this subsection, we present several simulation data sets to illustrate the results in Theorem 2.3.1, 2.3.3, and 2.3.4 respectively.

Simulation 1 for Theorem 2.3.1

In the linear shift model, we set $\mathbf{v} = (\frac{1}{\sqrt{d}}, \cdots, \frac{1}{\sqrt{d}})^T$, The random errors $\boldsymbol{\epsilon_i}$s are i.i.d and are from $N(0, I_d)$. We independently generate data sets $(X^{(d)}, Y^{(d)})$ as in (2.7) for $n = 20$, the dimensions $d = 40 + 2^1, 40 + 2^2, \cdots, 40 + 2^{15}$ and $\alpha = -1, 0, 0.25, 0.5, 0.75, 1$. There are totally $15 \times 6 = 90$ pairs of data sets. For each pair of ($X^{(d)}$ and $Y^{(d)}$), the empirical CPD $\mathbf{v}_e^{(d)}$ is calculated using the algorithm introduced in Section 2.2. The theoretical CPD is the normalized direction vector $\mathbf{v} = (\frac{1}{\sqrt{d}}, \cdots, \frac{1}{\sqrt{d}})^T$. We calculate the Absolute Inner Product (AIP) between these two directions. The results are organized in Figure 2.4:

Each plot in Figure 2.4 shows the Absolute Inner Products (AIPs) between the empirical CPD and the theoretical CPD for the paired data sets, simulated with a given α. The AIPs are plotted against the dimension d. The three subplots in the top row illustrate the results for $\alpha = -1, 0, 0.25$ respectively. Since $\alpha < 0.5$, according to Theorem 2.3.1, the AIPs converge to 0 in probability, which are shown by the curves in these three subplots. When $\alpha = 0.25$, the AIPs aren't close to 0 until $d = 40 + 2^{15}$. However, the trend of the convergence is clear. When $\alpha = 0.5$, the AIPs vary between 0.97 to 0.98, as shown in the second row, first column subplot. There is no trend of convergence to 0 or 1. When $\alpha = 0.75$ and 1, the AIPs converge to 1, which is shown the the second row, second column and third column subplots. These two subplots indicate that the empirical CPD converges to the theoretical CPD and hence verify the consistency of the empirical parallel direction.

Note that the scales on the y axes are different in these subplots. In order to com-

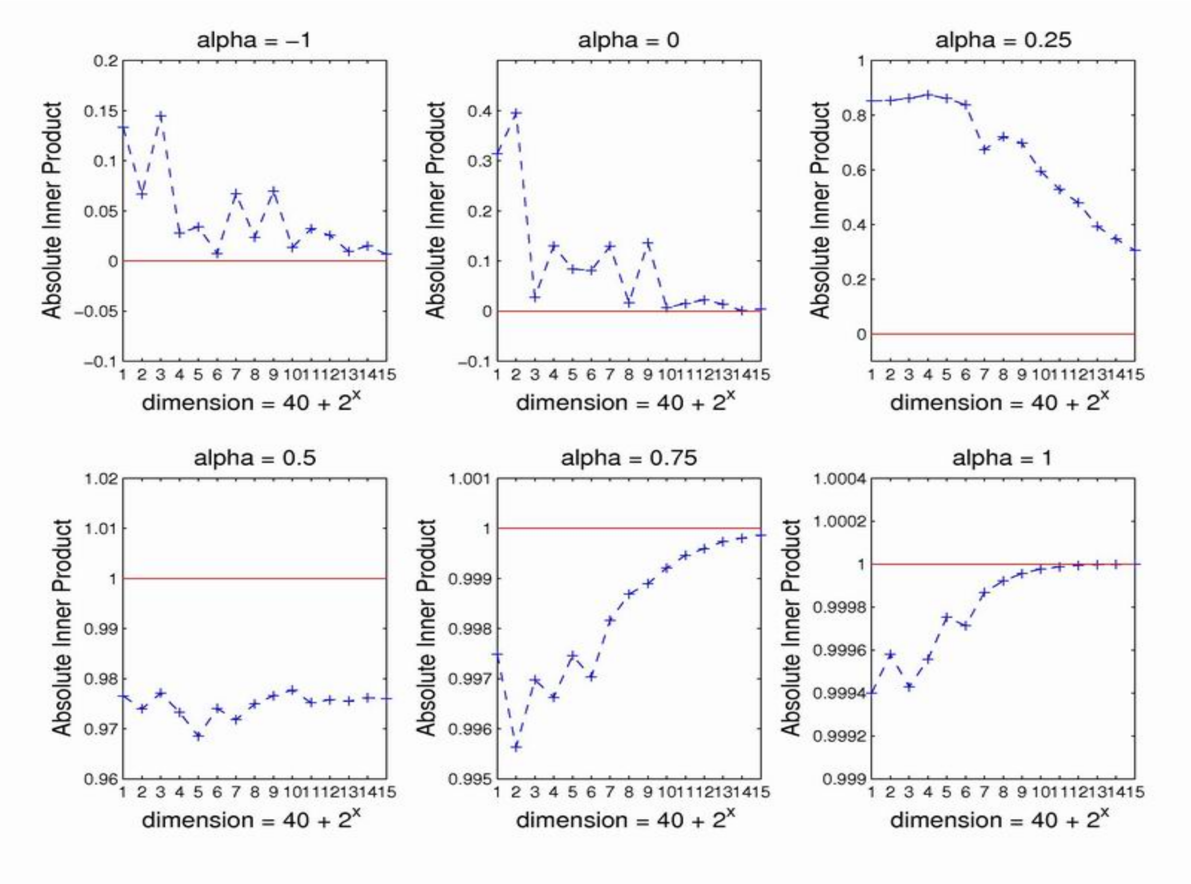


Figure 2.4: Simulation results for Theorem 2.3.1. Each subplot illustrates the results for data sets with a choice of α. The three plots in the first row indicate the strong inconsistency of the empirical CPD, i.e. the AIP converges to 0. The last two plots in the second row illustrate the consistency of the empirical CPD. The second row, first colum subplot is for the data sets with $\alpha = 0.5$, which shows no trend of convergence. These plots are consistent with the conclusions in Theorem 2.3.1.

pare the speeds of the convergence for the data sets with different values of α, in Figure 2.5, we show the same results as in Figure 2.4. All the subplots in Figure 2.5 have the same axes.

When $\alpha < 1/2$, the three subplots in the top row indicate that small α leads to fast convergence to 0. When $\alpha > 1/2$, it's not easy to compare the convergence speed for $\alpha = 0.75$ and $\alpha = 1$, using Figure 2.5. From the last two subplots in Figure 2.4, we can see clearly that larger α leads to faster speed of convergence to 1.

Simulation 2 for Theorem 2.1.2

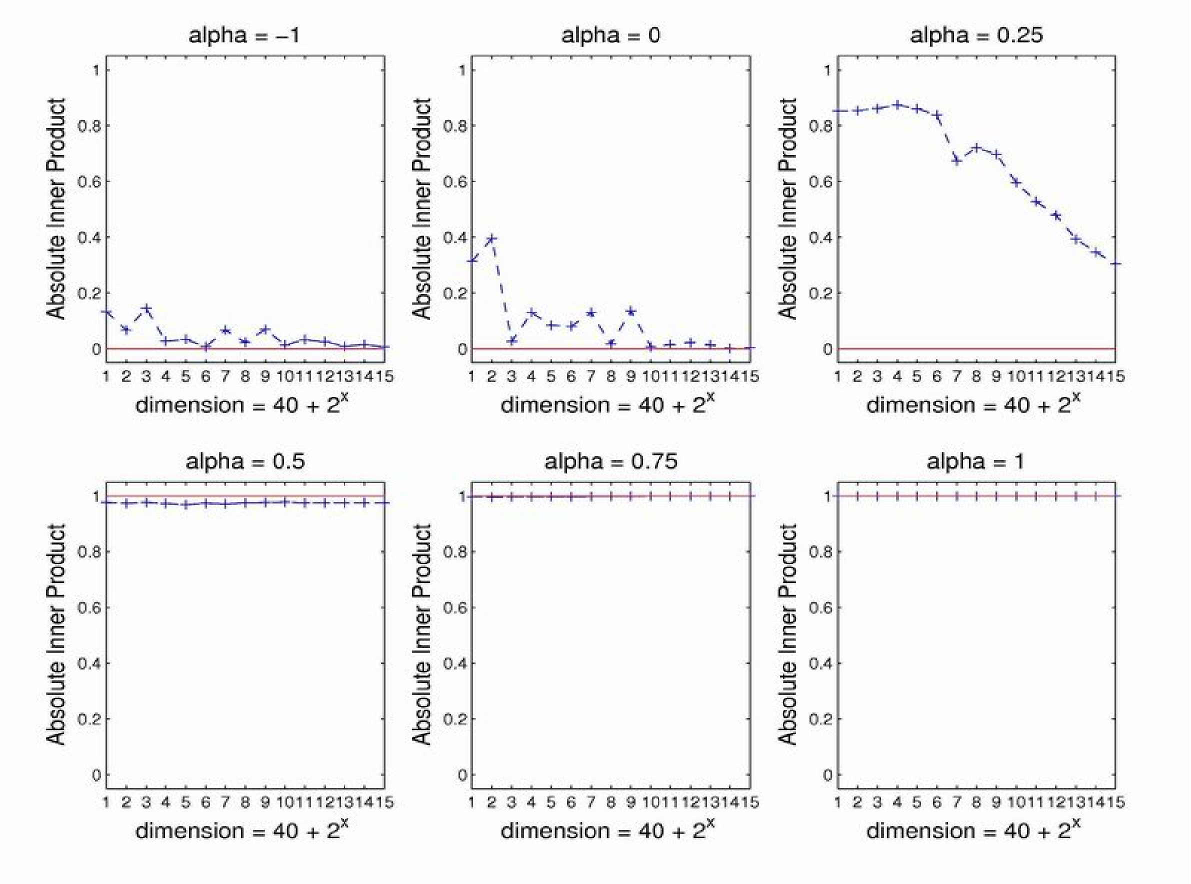


Figure 2.5: Simulation results for Theorem 2.3.1. It shows the same results as in Figure 2.4, using the same axes for each subplot. This plot shows the convergence ($\alpha = 0.75, 1.0$), and strong inconsistency ($\alpha = -1, 0, 0.25$) more clearly.

In Theorem 2.3.3, although the covariance matrices are not identity matrices. The eigenvalues of the covariance matrices are sufficiently diffused, i.e. they satisfy the condition (2.10). We generate paired data sets $(X^{(d)}, Y^{(d)})$ as in (2.7) with $\Sigma_d = diag(d^{\beta}, 1, \cdots, 1)$. normalized $\beta \leqslant 1$, the condition (2.10) is satisfied. In this simulation, we choose $\beta = 0.5$. Similar with Simulation 1, paired data sets are simulated with $n = 20$, $d = 40 + 2^1, 40 + 2^2, \cdots, 40 + 2^{15}$ and $\alpha = -1, 0, 0.25, 0.5, 0.75, 1$. The results are shown in the Figure 2.6.

The conclusions are the same as Simulation 1. When $\alpha = -1, 0, 0.25$, the AIPs converge to 0. When $\alpha = 0.5$, there is no trend of convergence to 0 or 1. When $\alpha = 0.75, 1$, the AIPs converge to 1. The first eigenvalue of the covariance matrices is larger than the rest eigenvalues. However, it doesn't dominate the other eigenvalues, i.e. the eigenvalues are

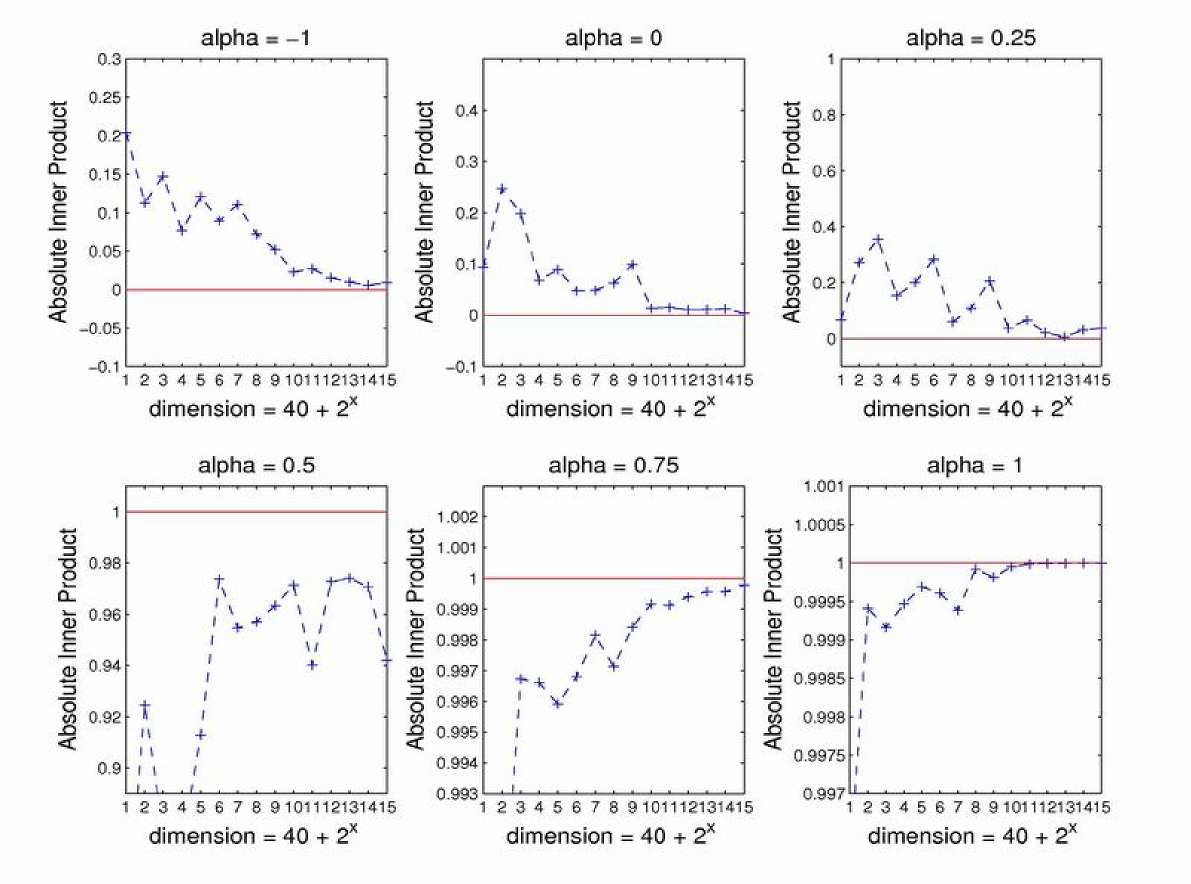


Figure 2.6: Simulation results for Theorem 2.3.3. Each subplot illustrates the results for data sets with a choice of α. The three plots in the first row indicate the strong inconsistency of the empirical CPD, i.e. the AIP converges to 0. The last two plots in the second row illustrate the consistency of the empirical CPD. The second row, first colum subplot is for the data sets with $\alpha = 0.5$, which shows no trend of convergence. These plots are consistent with the conclusions in Theorem 2.3.3.

sufficiently diffuse, as in (2.10). According to Theorem 2.3.2, the sample eigenvalues behave as if they are those of the identity covariance matrices. Thus we obtain similar asymptotic properties as in Theorem 2.3.1.

The value of α also has effect on the speed of convergence. The conclusions are the same as in Simulation 1, i.e. when $\alpha < 1/2$, the three subplots in the top row show that small α leads to fast convergence to 0; when $\alpha > 1/2$, larger α leads to faster speed of convergence to 1.

Simulation 3 for Theorem 2.3.4

In Theorem 2.3.4, the covariance matrices have a "spike structure", i.e. $\Sigma_d = diag(d^\beta, 1, \cdots, 1)$ with $\beta > 1$. Because $\beta > 1$, the condition (2.10) doesn't hold, i.e. the first eigenvalue dominate all the others. We set $\beta = 2$ for the spike covariance matrices in our simulated data. Multiple pairs of data sets are simulated with $n = 20$, $\alpha = 0.5, 1, 1.5$ and the dimensions $d = 40 + 2^1, 40 + 2^2, \cdots, 40 + 2^{15}$. We choose the theoretical CPD as $\mathbf{v} = (\frac{1}{\sqrt{d}}, \cdots, \frac{1}{\sqrt{d}})^T$. The spike direction (the first eigenvector of the covariance matrix) is $\mathbf{v}_s = (1, 0, \cdots, 0)^T$. The AIPs between the empirical CPD and the theoretical CPD are computed and shown in the top row of Figure 2.7. We also compute the AIPs between the empirical CPD and the spike direction, which are shown in the second row of Figure 2.7.

The two subplots in the first column are for the data sets with $\alpha = 0.5$. Because $2\alpha < 2 = \beta$, Theorem 2.3.4 indicate that the empirical CPD will asymptotically converge to the spike direction, which is exactly what we observed in the second row, first column subplot. The plot on the top row, first column shows that this empirical CPD is also asymptotically orthogonal to the theoretical CPD. Note that this is not already true and It depends on your data settings. In our simulated data, the chosen theoretical CPD $\mathbf{v} = (\frac{1}{\sqrt{d}}, \cdots, \frac{1}{\sqrt{d}})^T$ is asymptotically orthogonal with the spike direction $\mathbf{v}_s = (1, 0, \cdots, 0)^T$. The middle two subplots of Figure 2.7 illustrate that there are no trend of convergence when $2\alpha = \beta$. The third columns are for the data sets with $\alpha = 1.5$. Since $2\alpha > 2 = \beta$, the second conclusion in Theorem 2.3.4 implies that the empirical CPD converge to the theoretical CPD, which is shown in the top row, third column subplot. Again, the empirical CPD is not necessary orthogonal to the spike direction. In our data setting, we have them orthogonal to each other, as shown in the top row, third column subplot.

For this simulation, we tried different β values to study the speed of convergence. We found that when $2\alpha < \beta$, larger β leads to a faster speed of convergence to the spike direction; when $2\alpha > \beta$, larger β leads to a slower speed of convergence to the theoretical CPD. The additional plots are presented on the website for this dissertation at Liu (2007b).

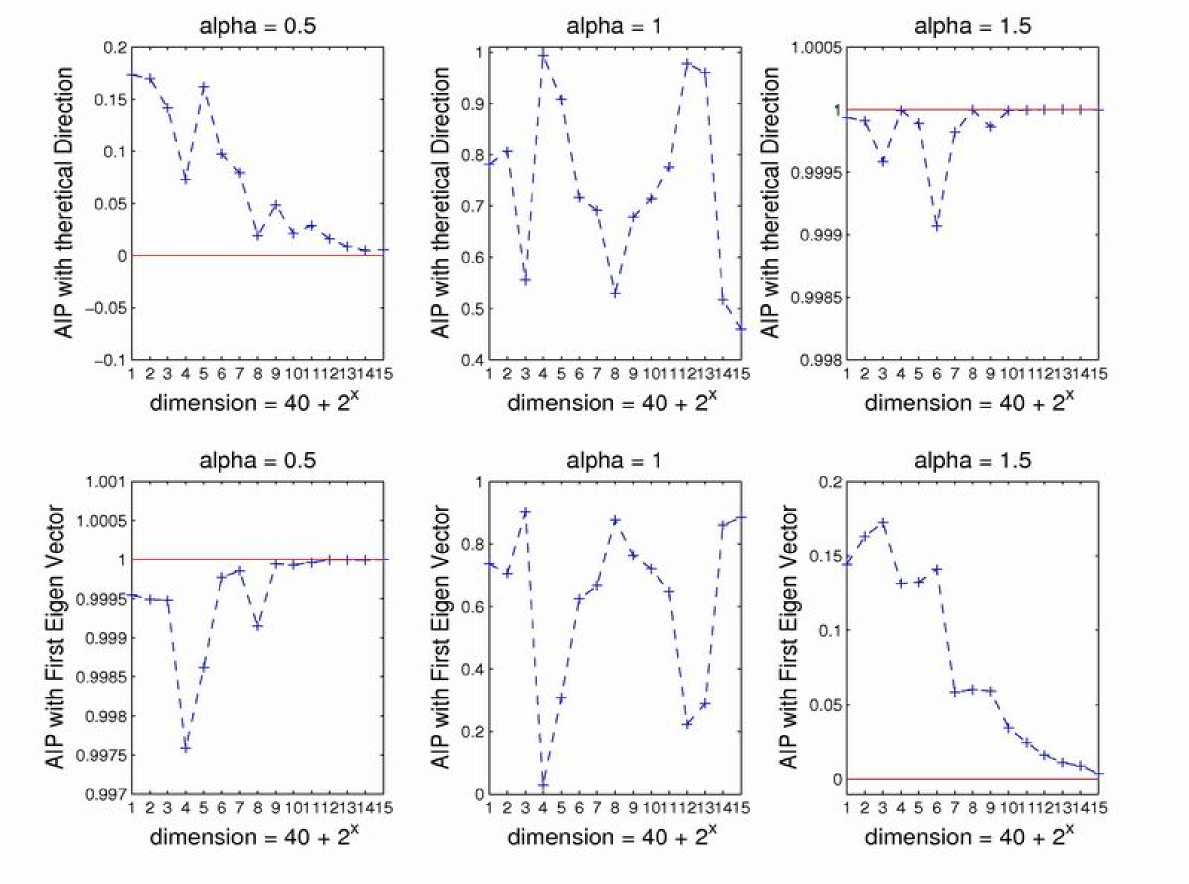


Figure 2.7: Simulation results for Theorem 2.3.4. The two subplots in each column illustrate the results for data sets with a choice of α, i.e. the two subplots in the first column corresponding to the data sets with $\alpha = 0.5$. Three subplots in the first row illustrate the AIPs between the empirical CPD and the theoretical CPD. The three subplots in the seond row show the AIPs between the empirical CPD and the spike direction.

2.3.5 Proofs of the Theorems

In this subsection, we will prove Theorem 2.3.1, 2.3.3 and 2.3.4. As we have discussed before, Theorem 2.3.1 is a special case of 2.3.3, i.e. all the covariance matrices are identity matrices. The proof of Theorem 2.3.3 indicates the proof for Theorem 2.3.1. In the following, we will first prove Theorem 2.3.3.

Proofs for Theorem 2.3.3

Define $Z^{(d)} = X^{(d)} - Y^{(d)}$ as the differences between the paired matrices. According to Equation (2.7) and (2.8) in the linear shift model, we have

$$Z^{(d)} = cd^{\alpha} V_t^{(d)} + \Upsilon_1^{(d)} - \Upsilon_2^{(d)},$$

where $V_t^{(d)} = (\mathbf{v}_t^{(d)}, \cdots, \mathbf{v}_t^{(d)})$, and $\lim\limits_{d \to \infty} \|\mathbf{v}_t^{(d)}\| = 1$. Define $\Upsilon^{(d)} = \Upsilon_1^{(d)} - \Upsilon_2^{(d)}$, then the columns of $\Upsilon^{(d)}$ follow a Gaussian distribution with means 0 and covariance matrix $2\Sigma^{(d)}$, since $\Upsilon_1^{(d)}$ and $\Upsilon^{(d)}$ are independent from Gaussian distribution with mean 0 and covariance matrix $\Sigma^{(d)}$. The batch difference matrix between $X^{(d)}$ and $Y^{(d)}$ can be expressed as

$$Z^{(d)} = cd^{\alpha} V_t^{(d)} + \Upsilon^{(d)}. \tag{2.11}$$

As we have concluded in Section 2.2.2, the empirical CPD is the first eigenvector of $Z^{(d)}(Z^{(d)})^T = (cd^{\alpha} V_t^{(d)} + \Upsilon^{(d)})(cd^{\alpha} V_t^{(d)} + \Upsilon^{(d)})^T$.

Because the proof of Theorem 2.3.3 is quite complicated, we organize them into the following steps:

Step 1: We first show that it is enough to assume that $\Sigma^{(d)}$ is a diagonal matrix. Suppose $\Sigma^{(d)}$ has the following Singular Value Decomposition (SVD)

$$\Sigma^{(d)} = F \Lambda F^{-1},$$

where F is an $d \times d$ orthonormal matrix, i.e. $FF^T = I_d$; Λ is a diagonal matrix of eigenvalues. Multiply both sides of Equation (2.11) by F^{-1}, as follows

$$F^{-1} Z^{(d)} = cd^{\alpha} F^{-1} V_t^{(d)} + F^{-1} \Upsilon^{(d)}. \tag{2.12}$$

Define $Z^{*(d)} = F^{-1} Z^{(d)}$, $V_t^{*(d)} = F^{-1} V_t^{(d)}$ and $\Upsilon^{*(d)} = F^{-1} \Upsilon^{(d)}$, then Equation (2.12) is equivalent to

$$Z^{*(d)} = cd^{\alpha} V_t^{*(d)} + \Upsilon^{*(d)} \tag{2.13}$$

where suppose $V_t^{(*d)} = (\mathbf{v}_t^{(*d)}, \cdots, \mathbf{v}_t^{(*d)})$, then $\mathbf{v}_t^{(*d)} = F^{-1} \mathbf{v}_t^{(d)}$. Now, for the new difference matrix $Z^{*(d)}$, the theoretical CPD is the direction vector of $\mathbf{v}_t^{*(d)}$ and the matrix of random errors is $\Upsilon^{*(d)}$. The covariance matrix for the random errors $F^{-1} \Upsilon^{(d)}$ is 2Λ, which is a diagonal matrix. Because $F^{-1}F = I_d$, $V_t^{*(d)}$ has the same asymptotic length as $V_t^{(d)}$, i.e. $\lim_{d\to\infty} \|\mathbf{v}_t^{*(d)}\| = 1$. Notice that if the first eigenvector of $Z^{(d)}(Z^{(d)})^T$ is $\mathbf{v}_e^{(d)}$, then the first eigenvector of $Z^{*(d)}(Z^{*(d)})^T = F^{-1}Z^{(d)}(Z^{(d)})^T F$ is $\hat{\mathbf{v}}_e{}^{*(d)} = F^{-1}\mathbf{v}_e^{(d)}$. To study the relations between $\mathbf{v}_e^{(d)}$ and $\mathbf{v}_t^{(d)}$, $\hat{\mathbf{v}}_e{}^{*(d)}$ and $\mathbf{v}_t^{*(d)}$, we calculate the inner products, i.e. $|(\mathbf{v}_t^{(d)})^T\mathbf{v}_e^{(d)}|$ and $|(\mathbf{v}_t^{*(d)})^T\hat{\mathbf{v}}^{*(d)}|$, as follows

$$|(\mathbf{v}_t^{*(d)})^T\hat{\mathbf{v}}^{*(d)}| = |(\mathbf{v}_t^{(d)})^T F F^{-1}\hat{\mathbf{v}}^{*(d)}| = |(\mathbf{v}_t^{(d)})^T\mathbf{v}_e^{(d)}| \tag{2.14}$$

Since two inner products are the same, if we could prove the relations of $\mathbf{v}_t^{*(d)}$ and $\hat{\mathbf{v}}^{*(d)}$ for the new data $Z^{*(d)}$, the same results between $\mathbf{v}_t^{(d)}$ and $\mathbf{v}_e^{(d)}$ hold too, by Equation (2.14). Thus it is enough to assume that $\Sigma^{(d)}$ is a diagonal matrix, the same results hold when it is not. From now on, we assume that $\Sigma^{(d)}$ is a diagonal matrix.

Step 2: Asymptotic properties of the uncentered Dual Sample Covariance Matrix. Suppose $X_{d\times n}$ is a HDLSS data. The **uncentered Dual Sample Covariance Matrix** of X is defines as X^TX, which is a $n \times n$ matrix, denotes as S_D. The **uncentered Sample Covariance Matrix** of X is defines as XX^T, which is a $d\times d$ matrix, denoted as S_P. When n is fixed and d goes to infinity, the dimension of S_P is increasing and it's hard to study the asymptotic properties of it's eigenvalues directly. Since S_P and S_D have exactly the same nonnegative eigenvalues, we can study the eigenvalues of S_D to obtain the asymptotic properties of the eigenvalues of S_P.

Define $\mathbf{1_n}$ as the $n \times 1$ vector with all the entries as 1, then

$$V_t^{(d)} = \mathbf{v}_t^{(d)} 1_n^T$$

Suppose $\Upsilon^{(d)} = (\boldsymbol{\epsilon}_1^{(d)}, \cdots, \boldsymbol{\epsilon}_n^{(d)})$, where $\boldsymbol{\epsilon}_1^{(d)}$ follows the Gaussian distribution with mean zero and covariance matrix $2\Sigma^{(d)}$. Hence Equation (2.11) implies

$$Z^{(d)} = cd^{\alpha}\mathbf{v}_t^{(d)}\mathbf{1_n}^T + \Upsilon^{(d)}. \tag{2.15}$$

The uncentered dual sample covariance matrix of $Z^{(d)}$ is

$$\begin{aligned} S_D^{(d)} &= (Z^{(d)})^T Z^{(d)} = (cd^\alpha \mathbf{v}_t^{(d)} \mathbf{1_n}^T + \Upsilon^{(d)})^T (cd^\alpha \mathbf{v}_t^{(d)} \mathbf{1_n}^T + \Upsilon^{(d)}) \\ &= c^2 d^{2\alpha} \mathbf{1_n} (\mathbf{v}_t^{(d)})^T \mathbf{v}_t^{(d)} \mathbf{1_n}^T + cd^\alpha \mathbf{1_n} (\mathbf{v}_t^{(d)})^T \Upsilon^{(d)} + cd^\alpha (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)} \mathbf{1_n}^T + (\Upsilon^{(d)})^T \Upsilon^{(d)} \\ &\equiv A + B_1 + B_2 + C \qquad (2.16) \end{aligned}$$

where

$$A = c^2 d^{2\alpha} \mathbf{1_n} (\mathbf{v}_t^{(d)})^T \mathbf{v}_t^{(d)} \mathbf{1_n}^T$$

$$B_1 = cd^\alpha \mathbf{1_n} (\mathbf{v}_t^{(d)})^T \Upsilon^{(d)}$$

$$B_2 = cd^\alpha (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)} \mathbf{1_n}^T$$

$$C = (\Upsilon^{(d)})^T \Upsilon^{(d)}$$

The uncentered Dual sample covariance matrix is the sum of four terms A, B_1, B_2 and C. Next, we will study the asymptotic properties of them separately.

- **The asymptotic properties of A.**

$$\begin{aligned} A &= c^2 d^{2\alpha} \mathbf{1_n} (\mathbf{v}_t^{(d)})^T \mathbf{v}_t^{(d)} \mathbf{1_n}^T \\ &= (cd^\alpha)^2 \mathbf{1_n} (\mathbf{v}_t^{(d)})^T (\mathbf{v}_t^{(d)}) \mathbf{1_n}^T \\ &= (cd^\alpha)^2 \mathbf{1_n} \|\mathbf{v}_t^{(d)}\|^2 \mathbf{1_n}^T \qquad (2.17) \end{aligned}$$

  Because $\lim_{d\to\infty} \|\mathbf{v}_t^{(d)}\| = 1$, as in Equation (2.4), we obtain

$$\lim_{d\to\infty} \frac{1}{(cd^\alpha)^2} A = \mathbf{1_n} \mathbf{1_n}^T = J_n \qquad (2.18)$$

  where J_n is an $n \times n$ matrix with all the entries as 1.

- **The asymptotic properties of B_1 and B_2.**

  Note that $B_1 = B_2^T$. Thus we only need to focus on B_2. Recall that $\Upsilon^{(d)} = (\boldsymbol{\epsilon}_1^{(d)}, \cdots, \boldsymbol{\epsilon}_n^{(d)})$, where $\boldsymbol{\epsilon}_i^{(d)} = (\epsilon_{i,1}^{(d)}, \cdots, \epsilon_{i,d}^{(d)})^T$, and $\boldsymbol{\epsilon}_i^{(d)}$ follows a d dimensional Gaussian distribution with mean 0 and covariance matrix $2\Sigma^{(d)}$. Also recall that $\mathbf{v}_t^{(d)} =$

$(v_{t,1}^{(d)}, \cdots, v_{t,n}^{(d)})$. Since $B_2 = cd^\alpha (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)} \mathbf{1_n}^T$, the ith row, jth column of B_2 is

$$B_2(i,j) = cd^\alpha (\boldsymbol{\epsilon}_i^{(d)})^T \mathbf{v}_t^{(d)} = cd^\alpha \sum_{k=1}^{d} \epsilon_{i,k}^{(d)} v_{t,k}^{(d)} \quad (i,j = 1, \cdots, n)$$

Next, we will prove that

$$\frac{1}{cd^\alpha (2c^{(d)})^{1/2}} B_2(i,j) \to 0 \ \ in\ probability, \quad as\ d \to \infty.$$

For any given $\epsilon > 0$, using Chebyshev's inequality, we get

$$P(|\frac{1}{cd^\alpha (2c^{(d)})^{1/2}} B_2(i,j)| > \epsilon) \quad \leqslant \quad \frac{E|\frac{1}{cd^\alpha (2c^{(d)})^{1/2}} B_2(i,j)|}{\epsilon} \tag{2.19}$$

For any random variable x with a finite mean, $(E|x|)^2 \leqslant E|x|^2$, because $E|x|^2 - (E|x|)^2 = var(|x|) \geqslant 0$. Thus

$$\begin{aligned} (E|\frac{1}{cd^\alpha (2c^{(d)})^{1/2}} B_2(i,j)|)^2 \quad &\leqslant \quad E|\frac{1}{cd^\alpha (2c^{(d)})^{1/2}} B_2(i,j)|^2 \\ &= \quad \frac{1}{2c^{(d)}} E(\sum_{k=1}^{d} \epsilon_{i,k}^{(d)} v_{t,k}^{(d)})^2 \end{aligned} \tag{2.20}$$

Since $E(\sum_{k=1}^{d} \epsilon_{i,k}^{(d)} v_{t,k}^{(d)}) = 0$, and $2\Sigma^{(d)}$ is a diagonal matrix, i.e. $\epsilon_{i,k_1}^{(d)}$ and $\epsilon_{i,k_2}^{(d)}$ are independent $(k_1 \neq k_2)$, we have

$$\begin{aligned} E(\sum_{k=1}^{d} \epsilon_{i,k}^{(d)} v_{t,k}^{(d)})^2 \quad &= \quad var(\sum_{k=1}^{d} \epsilon_{i,k}^{(d)} v_{t,k}^{(d)}) \\ &= \quad \sum_{k=1}^{d} (v_{t,k}^{(d)})^2 var(\epsilon_{i,k}^{(d)}) \\ &= \quad \sum_{k=1}^{d} (v_{t,k}^{(d)})^2 2\lambda_k^{(d)} \end{aligned} \tag{2.21}$$

Using the CauchySchwarz inequality,

$$(\sum_{k=1}^{d}(v_{t,k}^{(d)})^2 2\lambda_k^{(d)})^2 \quad \leqslant \quad \left[\sum_{k=1}^{d}(v_{t,k}^{(d)})^4\right]\left[\sum_{k=1}^{d}(2\lambda_k^{(d)})^2\right] \tag{2.22}$$

Summarizing results from (2.19) to (2.22), we obtain

$$\begin{aligned} P(|\frac{1}{cd^{\alpha}(2c^{(d)})^{1/2}}B_2(i,j)| > \epsilon) \quad &\leqslant \quad \frac{1}{\epsilon}\frac{1}{(2c^{(d)})^{1/2}}\left[\sum_{k=1}^{d}(v_{t,k}^{(d)})^4\right]^{1/4}\left[\sum_{k=1}^{d}(2\lambda_k^{(d)})^2\right]^{1/4} \\ &= \quad \frac{1}{\epsilon}\left[\frac{\sum_{k=1}^{d}(\lambda_k^{(d)})^2}{(c^{(d)})^2}\right]^{1/4}\left[\sum_{k=1}^{d}(v_{t,k}^{(d)})^4\right]^{1/4} \end{aligned} \tag{2.23}$$

Because $\lim_{d\to\infty}\sum_{k=1}^{d}(v_{t,k}^{(d)})^2 = \lim_{d\to\infty}\|\mathbf{v}_t^{(d)}\| = 1$, it follows that

$$\lim_{d\to\infty}\sum_{k=1}^{d}(v_{t,k}^{(d)})^4 \quad \leqslant \quad \lim_{d\to\infty}\sum_{k=1}^{d}(v_{t,k}^{(d)})^2 = 1. \tag{2.24}$$

Recall that the eigenvalues of $\Sigma^{(d)}$ have been assumed to be sufficiently diffuse (see Equation (2.10)) and also recall that $c^{(d)} = \sum_{i=1}^{d}\lambda_i^{(d)}$. Then,

$$\frac{\sum_{k=1}^{d}(\lambda_k^{(d)})^2}{(c^{(d)})^2} = \frac{\sum_{j=1}^{d}(\lambda_k^{(d)})^2}{(\sum_{k=1}^{d}\lambda_k^{(d)})^2} \longrightarrow 0 \quad as\ d \longrightarrow \infty. \tag{2.25}$$

From (2.23), (2.23) and (2.23), it follows that

$$P(|\frac{1}{cd^{\alpha}(2c^{(d)})^{1/2}}B_2(i,j)| > \epsilon) \quad \to \quad 0 \quad as\ d \to \infty. \tag{2.26}$$

This implies that

$$|\frac{1}{cd^{\alpha}(2c^{(d)})^{1/2}}B_2(i,j)| \to 0 \ in\ probability \quad as\ d \to \infty. \tag{2.27}$$

An assumption of Theorem (2.3.3) is that $c^{(d)}$ has the following asymptotic property,

$$\lim_{d\to\infty} \frac{log(c^{(d)})}{log(d)} = h. \tag{2.28}$$

Combining (2.27) and (2.28), we conclude that for $i, j = 1, \cdots, n$,

$$\begin{aligned} \frac{1}{\sqrt{2}d^{h/2}} \frac{1}{(2c^{(d)})^{1/2}} B_2(i,j) \quad &= \quad (\frac{c^{(d)}}{d^h})^{1/2} \frac{1}{cd^\alpha (2c^{(d)})^{1/2}} B_2(i,j) \\ &\longrightarrow \quad 0 \qquad in\ probability,\ as\ d \to \infty \end{aligned} \tag{2.29}$$

Thus

$$\frac{1}{\sqrt{2}d^{h/2}} \frac{1}{(cd^\alpha)} B_2 \quad \longrightarrow \quad 0 \qquad in\ probability,\ as\ d \to \infty \tag{2.30}$$

The same result holds for B_1.

$$\frac{1}{\sqrt{2}d^{h/2}} \frac{1}{(cd^\alpha)} B_1 \quad \longrightarrow \quad 0 \qquad in\ probability,\ as\ d \to \infty \tag{2.31}$$

- **Asymptotic properties of C.**

  Recall that $C = (\Upsilon^{(d)})^T \Upsilon^{(d)}$. We use $C(i,j)$ to represent the ith row, jth column of C.

  When $i = j, \quad (i, j = 1, 2, \cdots, n)$,

$$C(i,i) = \sum_{k=1}^{d} (\epsilon_{i,k}^{(d)})^2.$$

Recall that $c^{(d)} = \sum_{i=1}^{d} \lambda_i^{(d)}$. Next, we are going to prove that

$$\frac{1}{2c^{(d)}} C(i,i) \longrightarrow 1,\ in\ probability,\ as\ d \to \infty \tag{2.32}$$

Because the random variables $\epsilon_{i,k_1}^{(d)}$ and $\epsilon_{i,k_2}^{(d)}$ are independent ($k_1 \neq k_2$), we have

$$E\left[\frac{1}{2c^{(d)}}\sum_{k=1}^{d}(\epsilon_{i,k}^{(d)})^2\right] = \frac{1}{2c^{(d)}}\sum_{k=1}^{d}E(\epsilon_{i,k}^{(d)})^2 = \frac{\sum_{i=1}^{d}2\lambda_i^{(d)}}{2c^{(d)}} = 1$$

For any given $\epsilon > 0$, according the Chebyshev's inequality,

$$\begin{aligned} P\left[|\frac{1}{2c^{(d)}}\sum_{k=1}^{d}(\epsilon_{i,k}^{(d)})^2 - 1| > \epsilon\right] &\leqslant \frac{var(\frac{1}{2c^{(d)}}\sum_{k=1}^{d}(\epsilon_{i,k}^{(d)})^2)}{\epsilon^2} \\ &= \frac{\sum_{k=1}^{d}var((\epsilon_{i,k}^{(d)})^2)}{(2c^{(d)})^2\epsilon^2} \end{aligned} \quad (2.33)$$

Since $\epsilon_{i,k}^{(d)}$ follows $N(0, \lambda_k^{(d)})$, then $var((\epsilon_{i,k}^{(d)})^2) = (\lambda_k^{(d)})^2$. It follows that

$$\begin{aligned} P\left[|\frac{1}{2c^{(d)}}\sum_{k=1}^{d}(\epsilon_{i,k}^{(d)})^2 - 1| > \epsilon\right] &\leqslant \frac{\sum_{k=1}^{d}(2\lambda_j^{(d)})^2}{(2c^{(d)})^2}\frac{1}{\epsilon^2} \\ &= \frac{\sum_{k=1}^{d}(\lambda_j^{(d)})^2}{(c^{(d)})^2}\frac{1}{\epsilon^2} \end{aligned} \quad (2.34)$$

Again because the eigenvalues of $\Sigma^{(d)}$ are sufficiently diffused, i.e. in Equation (2.10)

$$\frac{\sum_{k=1}^{d}(\lambda_j^{(d)})^2}{(c^{(d)})^2} = \frac{\sum_{j=1}^{d}(\lambda_j^{(d)})^2}{(\sum_{j=1}^{d}\lambda_j^{(d)})^2} \longrightarrow 0 \quad as\ d \longrightarrow \infty.$$

Thus

$$P\left[|\frac{1}{2c^{(d)}}(\sum_{k=1}^{d}(\epsilon_{i,k}^{(d)})^2) - 1| > \epsilon\right] \longrightarrow 0 \quad as\ d \longrightarrow \infty.$$

This means that

$$\frac{1}{2c^{(d)}}(\sum_{k=1}^{d}(\epsilon_{i,k}^{(d)})^2) \longrightarrow 1 \quad in\ probability. \quad (2.35)$$

Similar derivations as (2.32) to (2.35) can be found at Hall *et al.* (2005).

When $i \neq j \quad (i, j = 1, 2, \cdots, n)$,

$$C(i,j) = \sum_{k=1}^{d} \epsilon_{i,k}^{(d)} \epsilon_{j,k}^{(d)}.$$

Note that $E(\frac{1}{2c^{(d)}} \sum_{k=1}^{d} \epsilon_{i,k}^{(d)} \epsilon_{j,k}^{(d)}) = 0$. Using the similar derivations as (2.32)- (2.35), for any $\epsilon > 0$,

$$\begin{aligned} P\left[|\frac{1}{2c^{(d)}} \sum_{k=1}^{d} \epsilon_{i,k}^{(d)} \epsilon_{j,k}^{(d)}| > \epsilon \right] &\leqslant \frac{var(\frac{1}{2c^{(d)}} \sum_{k=1}^{d} \epsilon_{i,k}^{(d)} \epsilon_{j,k}^{(d)})}{\epsilon^2} \\ &= \frac{\sum_{k=1}^{d} var(\frac{1}{2c^{(d)}} \epsilon_{i,k}^{(d)} \epsilon_{j,k}^{(d)})}{\epsilon^2} \\ &= \frac{\sum_{k=1}^{d} (2\lambda_k^{(d)})^2}{(2c^{(d)})^2} \frac{1}{\epsilon^2} \\ &= \frac{\sum_{k=1}^{d} (\lambda_k^{(d)})^2}{(c^{(d)})^2} \frac{1}{\epsilon^2} \end{aligned} \tag{2.36}$$

Again, the eigenvalues of $\Sigma^{(d)}$ have been assumed to be sufficiently diffuse

$$P\left[|\frac{1}{2c^{(d)}} \sum_{k=1}^{d} \epsilon_{i,k}^{(d)} \epsilon_{j,k}^{(d)}| > \epsilon \right] \longrightarrow 0 \ \ as\ d \to \infty.$$

This indicates that

$$\frac{1}{2c^{(d)}} \sum_{k=1}^{d} \epsilon_{i,k}^{(d)} \epsilon_{j,k}^{(d)} \longrightarrow 0 \ \ in\ probability,\ as\ d \to \infty. \tag{2.37}$$

Combining the results in (2.35) and (2.37), we conclude the element-wise convergence for C:

$$\frac{1}{2c^{(d)}} C \longrightarrow I_n \ \ in\ probability,\ as\ d \to \infty. \tag{2.38}$$

Again in Theorem 2.3.3, we assume that $\lim_{d \to \infty} \frac{log(c^{(d)})}{log(d)} \to h$, which means $\lim_{d \to \infty} \frac{c^{(d)}}{d^h} \to$

1. Thus

$$\frac{1}{2d^h}C = \left[\frac{c^{(d)}}{d^h}\right]\left[\frac{1}{2c^{(d)}}C\right] \longrightarrow I_n \ \ \textit{in probability}, \ \textit{as } d \to \infty. \tag{2.39}$$

From the above discussions, we have the following results for the asymptotic properties of A, B_1, B_2 and C :

$$\lim_{d\to\infty} \frac{1}{c^2 d^{2\alpha}} A = J_n;$$

$$\frac{1}{\sqrt{2}d^{h/2}} \frac{1}{(cd^\alpha)} B_1 \longrightarrow 0 \qquad \textit{in probability}, \ \textit{as } d \to \infty;$$

$$\frac{1}{\sqrt{2}d^{h/2}} \frac{1}{(cd^\alpha)} B_2 \longrightarrow 0 \qquad \textit{in probability}, \ \textit{as } d \to \infty;$$

$$\frac{1}{2d^h} C \longrightarrow I_n \ \ \textit{in probability}, \ \textit{as } d \to \infty.$$

Recall from (2.16), the uncentered dual sample covariance matrix

$$S_D^{(d)} = A + B_1 + B_2 + C \tag{2.40}$$

Next, we study the asymptotic properties of $S_D^{(d)}$, with respect to different values of α.

- **The case when** $\alpha > h/2$.

  We multiply both sides of (2.40) by $\frac{1}{c^2 d^{2\alpha}}$, according to the asymptotic properties of A, B_1, B_2, C, we conclude that

$$\begin{aligned} \frac{1}{c^2 d^{2\alpha}} S_D^{(d)} &= \frac{1}{c^2 d^{2\alpha}} A + \frac{1}{c^2 d^{2\alpha}} B_1 + \frac{1}{c^2 d^{2\alpha}} B_2 + \frac{1}{c^2 d^{2\alpha}} C \\ &= \frac{1}{c^2 d^{2\alpha}} A + \frac{1}{cd^{\alpha-h/2}} \left[\frac{1}{d^{h/2}} \frac{1}{cd^\alpha} B_1\right] + \frac{1}{cd^{\alpha-h/2}} \left[\frac{1}{d^{h/2}} \frac{1}{cd^\alpha} B_2\right] + \frac{1}{c^2} \frac{1}{d^{2\alpha-h}} \left[\frac{1}{d^h} C\right] \\ &\longrightarrow J_n + 0 + 0 + 0 = J_n \ \ \textit{in probability}, \ \textit{as } d \to \infty. \end{aligned}$$

Thus,

$$\frac{1}{c^2d^{2\alpha}}S_D^{(d)} \longrightarrow J_n \ \ in\ probability,\ as\ d \to \infty. \tag{2.41}$$

- The case when $\alpha < h/2$

We multiple both sides of (2.40) by $\frac{1}{2d^h}$. Then according to the asymptotic properties of A, B_1, B_2, C, we conclude that

$$\begin{aligned}
\frac{1}{2d^h}S_D^{(d)} &= \frac{1}{2d^h}A + \frac{1}{2d^h}B_1 + \frac{1}{2d^h}B_2 + \frac{1}{2d^h}C \\
&= \frac{1}{2}c^2d^{h-2\alpha}\left[\frac{1}{c^2d^{2\alpha}}A\right] + \frac{1}{\sqrt{2}}cd^{\alpha-h/2}\left[\frac{1}{\sqrt{2}d^{h/2}}\frac{1}{cd^{\alpha}}B_1\right] \\
&\quad + \frac{1}{\sqrt{2}}cd^{\alpha-h/2}\left[\frac{1}{\sqrt{2}d^{h/2}}\frac{1}{cd^{\alpha}}B_2\right] + \left[\frac{1}{2d^h}C\right] \\
&\longrightarrow 0+0+0+I_n = I_n \ \ in\ probability,\ as\ d \to \infty.
\end{aligned}$$

Thus,

$$\frac{1}{2d^h}S_D^{(d)} \longrightarrow I_n \ \ in\ probability,\ as\ d \to \infty. \tag{2.42}$$

- **The case when $\alpha = h/2$.**

We multiple both sides of (2.40) by $\frac{1}{2d^h}$ and get

$$\begin{aligned}
\frac{1}{2d^h}S_D^{(d)} &= \frac{1}{2d^h}A + \frac{1}{2d^h}B_1 + \frac{1}{2d^h}B_2 + \frac{1}{2d^h}C \\
&= \frac{1}{2}c^2d^{h-2\alpha}\left[\frac{1}{c^2d^{2\alpha}}A\right] + \frac{1}{\sqrt{2}}cd^{\alpha-h/2}\left[\frac{1}{\sqrt{2}d^{h/2}}\frac{1}{cd^{\alpha}}B_1\right] \\
&\quad + \frac{1}{\sqrt{2}}cd^{\alpha-h/2}\left[\frac{1}{\sqrt{2}d^{h/2}}\frac{1}{cd^{\alpha}}B_2\right] + \left[\frac{1}{2d^h}C\right] \\
&= \frac{1}{2}c^2\left[\frac{1}{c^2d^{2\alpha}}A\right] + \frac{1}{\sqrt{2}}c\left[\frac{1}{\sqrt{2}d^{h/2}}\frac{1}{cd^{\alpha}}B_1\right] + \frac{1}{\sqrt{2}}c\left[\frac{1}{\sqrt{2}d^{h/2}}\frac{1}{cd^{\alpha}}B_2\right] + \left[\frac{1}{2d^h}C\right] \\
&\longrightarrow \frac{1}{2}c^2J_n + I_n \ \ in\ probability,\ as\ d \to \infty.
\end{aligned}$$

Thus,

$$\frac{1}{2d^h} S_D^{(d)} \longrightarrow \frac{1}{2} c^2 J_n + I_n \quad \textit{in probability}, \ \textit{as} \ d \to \infty. \tag{2.43}$$

Step 3: The first eigenvector of the uncentered sample covariance matrix $S_P^{(d)}$. The uncentered sample covariance matrix is defined as

$$\begin{aligned} S_P^{(d)} &= Z^{(d)} (Z^{(d)})^T \\ &= (cd^\alpha \mathbf{v}_t^{(d)} \mathbf{1_n}^T + \Upsilon^{(d)})(cd^\alpha \mathbf{v}_t^{(d)} \mathbf{1_n}^T + \Upsilon^{(d)})^T \\ &= c^2 d^{2\alpha} \mathbf{v}_t^{(d)} \mathbf{1_n}^T \mathbf{1_n} (\mathbf{v}_t^{(d)})^T + cd^\alpha \mathbf{v}_t^{(d)} \mathbf{1_n}^T (\Upsilon^{(d)})^T + cd^\alpha \Upsilon^{(d)} \mathbf{1_n} (\mathbf{v}_t^{(d)})^T + \Upsilon^{(d)} (\Upsilon^{(d)})^T \\ &= n \times c^2 d^{2\alpha} \mathbf{v}_t^{(d)} (\mathbf{v}_t^{(d)})^T + cd^\alpha \mathbf{v}_t^{(d)} \mathbf{1_n}^T (\Upsilon^{(d)})^T \\ &\quad + cd^\alpha \Upsilon^{(d)} \mathbf{1_n} (\mathbf{v}_t^{(d)})^T + \Upsilon^{(d)} (\Upsilon^{(d)})^T \end{aligned} \tag{2.44}$$

We are interested in the relation between the first eigenvector of $S_P^{(d)}$ and the theoretical CPD $\mathbf{v}_t^{(d)}$. From Equation (2.44), we get

$$\begin{aligned} (\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} &= n \times c^2 d^{2\alpha} (\mathbf{v}_t^{(d)})^T \mathbf{v}_t^{(d)} (\mathbf{v}_t^{(d)})^T \mathbf{v}_t^{(d)} + cd^\alpha (\mathbf{v}_t^{(d)})^T \mathbf{v}_t^{(d)} \mathbf{1_n}^T (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)} \\ &\quad + cd^\alpha (\mathbf{v}_t^{(d)})^T \Upsilon^{(d)} \mathbf{1_n} (\mathbf{v}_t^{(d)})^T \mathbf{v}_t^{(d)} + (\mathbf{v}_t^{(d)})^T \Upsilon^{(d)} (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)} \\ &= n \times c^2 d^{2\alpha} \|\mathbf{v}_t^{(d)}\|^4 + cd^\alpha \|\mathbf{v}_t^{(d)}\|^2 \mathbf{1_n}^T (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)} \\ &\quad + cd^\alpha (\mathbf{v}_t^{(d)})^T \Upsilon^{(d)} \mathbf{1_n} \|\mathbf{v}_t^{(d)}\|^2 + (\mathbf{v}_t^{(d)})^T \Upsilon^{(d)} (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)} \\ &\equiv S_1 + S_2 + S_3 + S_4 \end{aligned} \tag{2.45}$$

where

$$S_1 = n \times c^2 d^{2\alpha} \|\mathbf{v}_t^{(d)}\|^4,$$

$$S_2 = cd^\alpha \|\mathbf{v}_t^{(d)}\|^2 \mathbf{1_n}^T (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)},$$

$$S_3 = cd^\alpha (\mathbf{v}_t^{(d)})^T \Upsilon^{(d)} \mathbf{1_n} \|\mathbf{v}_t^{(d)}\|^2,$$

$$S_4 = (\mathbf{v}_t^{(d)})^T \Upsilon^{(d)} (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)}.$$

Note that the dimensions of S_1, S_2, S_3, S_4 are all 1×1, hence

$$S_3 = S_2^T = S_2$$

From (2.45) we have

$$(\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} \quad = \quad S_1 + 2S_2 + S_4 \tag{2.46}$$

Next, we will study the asymptotic properties of S_1, S_2, S_4 respectively. Because $\mathbf{v}_t^{(d)}$ has the asymptotic norm 1, thus we conclude

$$\lim_{d\to\infty} \frac{1}{c^2 d^{2\alpha}} S_1 = \lim_{d\to\infty} n\|\mathbf{v}_t^{(d)}\|^4 = n. \tag{2.47}$$

Recall that $B_1 = cd^\alpha \mathbf{1_n}(\mathbf{v}_t^{(d)})^T \Upsilon^{(d)})$, as in (2.16). Then,

$$\mathbf{1_n}^T (B_1)^T \mathbf{1_n} = cd^\alpha \mathbf{1_n}^T (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)} \mathbf{1_n}^T \mathbf{1_n} = ncd^\alpha \mathbf{1_n}^T (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)},$$

$$\mathbf{1_n}^T B_1 (B_1)^T \mathbf{1_n} = n^2 c^2 d^{2\alpha} (\mathbf{v}_t^{(d)})^T \Upsilon^{(d)} (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)}.$$

Thus

$$S_2 = \frac{1}{n}\|\mathbf{v}_t^{(d)}\|^2 \mathbf{1_n}^T (B_1)^T \mathbf{1_n} = \frac{1}{n}\|\mathbf{v}_t^{(d)}\|^2 \sum_{i=1}^n \sum_{j=1}^n B(i,j)$$

$$S_4 = \frac{1}{n^2 c^2 d^{2\alpha}} \mathbf{1_n}^T B_1 (B_1)^T \mathbf{1_n} = \frac{1}{n^2 c^2 d^{2\alpha}} \sum_{i=1}^n \left[\sum_{j=1}^n B_1(i,j) \right]^2$$

Recall the asymptotic properties of B_1 from (2.29)

$$\frac{1}{\sqrt{2} d^{h/2}} \frac{1}{(cd^\alpha)} B_1(i,j) \longrightarrow 0 \qquad \textit{in probability, as } d \to \infty.$$

It follows that

$$\frac{1}{\sqrt{2} d^{h/2}} \frac{1}{(cd^\alpha)} |B_1(i,j)| \longrightarrow 0 \qquad \textit{in probability, as } d \to \infty.$$

We have the following asymptotic properties for S_2 and S_4:

$$\begin{aligned}
\frac{1}{d^{h/2}cd^{\alpha}}|S_2| &= \frac{1}{n}\frac{1}{d^{h/2}}\frac{1}{cd^{\alpha}}|\sum_{i=1}^{n}\sum_{j=1}^{n}B_1(i,j)| \\
&\leqslant \frac{\sqrt{2}}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{1}{\sqrt{2}d^{h/2}}\frac{1}{cd^{\alpha}}|B_1(i,j)|\right] \\
&\longrightarrow \frac{\sqrt{2}}{n}\times n\times n\times 0 = 0 \qquad (2.48)
\end{aligned}$$

Thus

$$\frac{1}{d^{h/2}cd^{\alpha}}S_2 \longrightarrow 0 \qquad \textit{in probability, as } d\to\infty. \qquad (2.49)$$

For the term S_4, we have

$$\begin{aligned}
\frac{1}{d^h}S_4 &= \frac{1}{n^2d^hc^2d^{2\alpha}}\sum_{i=1}^{n}\left[\sum_{j=1}^{n}B_1(i,j)\right]^2 \\
&\leqslant \frac{2}{n^2}\sum_{i=1}^{n}\left[\sum_{j=1}^{n}\left|\frac{1}{\sqrt{2}d^{h/2}cd^{\alpha}}B_1(i,j)\right|\right]^2 \\
&\longrightarrow \frac{2}{n^2}\times n\times n^2\times 0 = 0 \qquad (2.50)
\end{aligned}$$

Thus

$$\frac{1}{d^h}S_4 \longrightarrow 0 \qquad \textit{in probability, as } d\to\infty. \qquad (2.51)$$

Combining results from (2.47), (2.49) and (2.51), we have the following asymptotic results for $(\mathbf{v}_t^{(d)})^TS_P^{(d)}\mathbf{v}_t^{(d)}$ with different value of α.

- **The case when $\alpha > h/2$.**

$$\begin{aligned}
(\frac{1}{c^2d^{2\alpha}})(\mathbf{v}_t^{(d)})^TS_P^{(d)}\mathbf{v}_t^{(d)} &= \left[\frac{1}{c^2d^{2\alpha}}S_1\right] + 2cd^{h/2-\alpha}\left[\frac{1}{d^{h/2}cd^{\alpha}}S_2\right] \\
&\quad + d^{h-2\alpha}\frac{1}{c^2}\left[\frac{1}{d^h}S_4\right]
\end{aligned}$$

$$\rightarrow \quad n \tag{2.52}$$

- **The case when** $\alpha = h/2$

$$\begin{aligned}
(\frac{1}{c^2 d^{2\alpha}})(\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} &= \left[\frac{1}{c^2 d^{2\alpha}} S_1\right] + 2cd^{h/2-\alpha}\left[\frac{1}{d^{h/2} c d^{\alpha}} S_2\right] \\
&\quad + d^{h-2\alpha}\frac{1}{c^2}\left[\frac{1}{d^h} S_4\right] \\
&= \left[\frac{1}{c^2 d^{2\alpha}} S_1\right] + 2c\left[\frac{1}{d^{h/2} c d^{\alpha}} S_2\right] + \frac{1}{c^2}\left[\frac{1}{d^h} S_4\right] \\
&\rightarrow n
\end{aligned} \tag{2.53}$$

- **The case when** $\alpha < h/2$

$$\begin{aligned}
(\frac{1}{d^h})(\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} &= c^2 d^{h-2\alpha}\left[\frac{1}{c^2 d^{2\alpha}} S_1\right] + 2cd^{\alpha-h/2}\left[\frac{1}{d^{h/2} c d^{\alpha}} S_2\right] \\
&\quad + \left[\frac{1}{d^h} S_4\right] \\
&\rightarrow 0
\end{aligned} \tag{2.54}$$

The $d \times d$ matrix $S_P^{(d)}$ has the same list of eigenvalues as the $n \times n$ dual sample covariance matrix $S_D^{(d)} = (Z^{(d)})^T Z^{(d)}$. Since the rank of $S_P^{(d)}$ is no more than n ($n < d$), it has no more than n positve eigenvalues. Suppose the first n eigenvalues of $S_P^{(d)}$ and $S_D^{(d)}$ are $\hat{\lambda}_1 \geqslant \cdots \geqslant \hat{\lambda}_n \geqslant 0$. Assume that the symmetric $d \times d$ matrix $S_P^{(d)}$ has the following eigenvalue decomposition:

$$S_P^{(d)} = GLG^t = \hat{\lambda}_1 \hat{\boldsymbol{g}}_{\boldsymbol{1}} \hat{\boldsymbol{g}}_{\boldsymbol{1}}^T + \cdots + \hat{\lambda}_n \hat{\boldsymbol{g}}_{\boldsymbol{n}} \hat{\boldsymbol{g}}_{\boldsymbol{n}}^T \tag{2.55}$$

where $L = diag(\hat{\lambda}_1, \cdots, \hat{\lambda}_n)$ is an $n \times n$ diagonal matrix; the $d \times n$ matrix $G = (\hat{\boldsymbol{g}}_{\boldsymbol{1}}, \cdots, \hat{\boldsymbol{g}}_{\boldsymbol{n}})$ contains the first n $d-$dimensional eigenvectors of $S_P^{(d)}$. As we have studied in Section 2.2, the empirical CPD is the first eigenvector of $S_P^{(d)}$, i.e. $\mathbf{v}_e^{(d)} = \hat{\boldsymbol{g}}_{\boldsymbol{1}}$.

Equation (2.55) implies that

$$(\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} = \hat{\lambda}_1 (\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{1}} \hat{\boldsymbol{g}}_{\boldsymbol{1}}^T \mathbf{v}_t^{(d)} + \cdots + \hat{\lambda}_n (\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{n}} \hat{\boldsymbol{g}}_{\boldsymbol{n}}^T \mathbf{v}_t^{(d)}$$

$$= \quad \hat{\lambda}_1 |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 + \cdots + \hat{\lambda}_n |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{n}}|^2 \tag{2.56}$$

Next, we will study the asymptotic property of the first eigenvector $\hat{\boldsymbol{g}}_{\mathbf{1}}$, depending on the values of α and h.

- **The case when** $\alpha > h/2$.

  We have shown in (2.41) that

$$\frac{1}{c^2 d^{2\alpha}} S_D^{(d)} \longrightarrow J_n \ \ \textit{in probability}, \ \textit{as } d \to \infty.$$

Note that the first eigenvalue of J_n is n, and all the rest of the eigenvalues are 0. Thus

$$\frac{1}{c^2 d^{2\alpha}} \hat{\lambda}_1 \longrightarrow n \ \ \textit{in probability}, \ \textit{as } d \to \infty; \tag{2.57}$$

$$\frac{1}{c^2 d^{2\alpha}} \hat{\lambda}_j \longrightarrow 0 \ \ \textit{in probability}, \ \textit{as } d \to \infty, (j = 2, \cdots, n) \tag{2.58}$$

We multiply both sides of Equation (2.56) by $\frac{1}{c^2 d^{2\alpha}}$ and get

$$\begin{aligned} \frac{1}{c^2 d^{2\alpha}} (\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} &= \frac{1}{c^2 d^{2\alpha}} \hat{\lambda}_1 |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 + \\ &\quad \cdots + \frac{1}{c^2 d^{2\alpha}} \hat{\lambda}_n |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{n}}|^2 \end{aligned} \tag{2.59}$$

Because $\lim\limits_{d\to\infty} |\mathbf{v}_t^{(d)}| = 1$ and $|\hat{\boldsymbol{g}}_{\boldsymbol{i}}| = 1$, it follows that for a sufficiently large d_0, when $d > d_0$, $|(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{i}}| \leqslant |\mathbf{v}_t^{(d)}| |\hat{\boldsymbol{g}}_{\boldsymbol{i}}| \leqslant 2$ $(i = 1, \cdots, n)$. If we let $d \to \infty$ on both sides of (2.59), because of (2.57) and (2.58), we have

$$\begin{aligned} \frac{1}{c^2 d^{2\alpha}} (\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} &= \left[\frac{1}{c^2 d^{2\alpha}} \hat{\lambda}_1\right] |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 + \cdots + \left[\frac{1}{c^2 d^{2\alpha}} \hat{\lambda}_n\right] |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{n}}|^2 \\ &\propto_p n |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 + \cdots + 0 \times |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{n}}|^2 \\ &= n |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 \end{aligned} \tag{2.60}$$

where the symbol $\propto_p$ means that two terms are asymptotically the same in probability, i.e. $A(d) \propto_p B(d)$ means that $\frac{A(d)}{B(d)} \longrightarrow 1$ in probability, as $d \to \infty$.

Recall that in (2.41), we have the conclusion

$$\frac{1}{c^2 d^{2\alpha}} (\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} \quad \longrightarrow \quad n \;\; in\ probability,\ as\ d \to \infty \tag{2.61}$$

Hence,

$$|(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 \longrightarrow 1 \;\; in\ probability,\ as\ d \to \infty. \tag{2.62}$$

Since $\mathbf{v}_e^{(d)} = \hat{\boldsymbol{g}}_{\mathbf{1}}$, the result in (2.62) is equivalent to

$$|(\mathbf{v}_t^{(d)})^T \mathbf{v}_e^{(d)}| \longrightarrow 1 \;\; in\ probability,\ as\ d \to \infty. \tag{2.63}$$

This proves the first conclusion in Theorem 2.3.3.

- **The case when** $\alpha = h/2$.

  We have shown in (2.43) that

$$\frac{1}{2d^h} S_D^{(d)} \longrightarrow \frac{1}{2} c^2 J_n + I_n \;\; in\ probability,\ as\ d \to \infty.$$

Note that the first eigenvalue of $\frac{1}{2}c^2 J_n + I_n$ $(c > 0)$ is $\frac{1}{2}c^2 n + 1$, and all the rest of the eigenvalues are 1. Thus

$$\frac{1}{2d^h} \hat{\lambda}_1 \longrightarrow \frac{1}{2} c^2 n + 1 \;\; in\ probability,\ as\ d \to \infty; \tag{2.64}$$

$$\frac{1}{2d^h} \hat{\lambda}_j \longrightarrow 0 \;\; in\ probability,\ as\ d \to \infty, (j = 2, \cdots, n) \tag{2.65}$$

Similiar with (2.59) and (2.60), we get

$$\begin{aligned} \frac{1}{2d^h}(\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} &= \left[\frac{1}{2d^h}\hat{\lambda}_1\right] |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 + \cdots + \left[\frac{1}{2d^h}\hat{\lambda}_n\right] |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{n}}|^2 \\ &\propto_p (\frac{1}{2}c^2 n + 1)|(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 + \cdots + 1 \times |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{n}}|^2 \qquad (2.66) \end{aligned}$$

Recall that from (2.53), we have

$$(\frac{1}{2d^h})(\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} \longrightarrow \frac{1}{2}nc^2 \qquad (2.67)$$

Thus

$$(\frac{1}{2}c^2 n + 1)|(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 + \cdots + |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{n}}|^2 \rightarrow \frac{1}{2}nc^2 \qquad (2.68)$$

Obtaining the value of $|(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2$ does not appear to be straightforward. However, it follows that

$$|(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}| < 1, \quad or \quad |(\mathbf{v}_t^{(d)})^T \mathbf{v}_e^{(d)}| < 1,$$

which means that the empirical CPD is not asymptotically the same as the theoretical CPD.

- **The case when** $\alpha < h/2$.

  We have shown in (2.42) that

$$\frac{1}{2d^h} S_D^{(d)} \longrightarrow I_n \;\; in\ probability,\ as\ d \rightarrow \infty.$$

Note that all the eigenvalues of I_n are 1. Thus

$$\frac{1}{2d^h}\hat{\lambda}_j \longrightarrow 1 \;\; in\ probability,\ as\ d \rightarrow \infty, (j = 1, \cdots, n) \qquad (2.69)$$

Similiar with (2.59) and (2.60), we get

$$\begin{aligned}\frac{1}{2d^h}(\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} &= \left[\frac{1}{2d^h}\hat{\lambda}_1\right] |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 + \cdots + \left[\frac{1}{2d^h}\hat{\lambda}_n\right] |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{n}}|^2 \\ &\propto_p |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 + \cdots + |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{n}}|^2 \qquad (2.70)\end{aligned}$$

Recall that from (2.54), we have

$$(\frac{1}{2d^h})(\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} \rightarrow 0 \qquad (2.71)$$

Thus it follows that

$$|(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 + \cdots + |(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{n}}|^2 \rightarrow 0 \qquad (2.72)$$

Since all $|(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{i}}|^2$ $(i = 1, \cdots, n)$ are nonnegative, (2.72) indicates that

$$|(\mathbf{v}_t^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{i}}|^2 \longrightarrow 0 \quad \textit{in probability}, \ \textit{as } d \rightarrow \infty \quad (i = 1, \cdots, n). \qquad (2.73)$$

The theoretical CPD is asymptotically orthogonal to all the eigenvectors. In particular, it is orthogonal with the first eigenvector, i.e.

$$|(\mathbf{v}_t^{(d)})^T \mathbf{v}_e^{(d)}| \longrightarrow 0 \ \ \textit{in probability}, \ \textit{as } d \rightarrow \infty. \qquad (2.74)$$

This proves the second conclusion in Theorem 2.3.3.

Now we have finished the proof of Theorem 2.3.3. As we have discussed, Theorem 2.3.1 is a special case of 2.3.3, when all the covariance matrices are identity matrices.

Proof of Theorem 2.3.4

The proof of Theorem 2.3.4 is similar to that of Theorem 2.3.3. Since the covariance matrices are diagonal matrices, we don't need to show Step 1, as in the proof of Theorem 2.3.3. The rest of the proof is organized as two steps.

Step A: Asymptotic properties of the uncentered Dual Sample Covariance Matrix. As we have obtained in (2.16), the uncentered dual sample covariance matrix of $Z^{(d)}$ is

$$S_D^{(d)} = A + B_1 + B_2 + C \tag{2.75}$$

where

$$A = c^2 d^{2\alpha} \mathbf{1_n} (\mathbf{v}_t^{(d)})^T \mathbf{v}_t^{(d)} \mathbf{1_n}^T$$

$$B_1 = c d^{\alpha} \mathbf{1_n} (\mathbf{v}_t^{(d)})^T \Upsilon^{(d)}$$

$$B_2 = c d^{\alpha} (\Upsilon^{(d)})^T \mathbf{v}_t^{(d)} \mathbf{1_n}^T$$

$$C = (\Upsilon^{(d)})^T \Upsilon^{(d)}$$

Next we study the asymptotic properties of A, B_1, B_2 and C for the cases when $2\alpha > \beta$ and $2\alpha < \beta$ respectively.

- **The case when $2\alpha > \beta$.**

  We multiply both sides of (2.75) by $\frac{1}{c^2 d^{2\alpha}}$,

$$\frac{1}{c^2 d^{2\alpha}} S_D^{(d)} = \frac{1}{c^2 d^{2\alpha}} A + \frac{1}{c^2 d^{2\alpha}} B_1 + \frac{1}{c^2 d^{2\alpha}} B_2 + \frac{1}{c^2 d^{2\alpha}} C \tag{2.76}$$

  Next, we study the asymptotic properties of the four terms on the right side of Equation (2.76) respectively.

  - **The asymptotic properties of $\frac{1}{c^2 d^{2\alpha}} A$**

    Using the same derivations as in (2.18), we have

$$\lim_{d \to \infty} \frac{1}{(cd^{\alpha})^2} A = \mathbf{1_n} \mathbf{1_n}^T = J_n \tag{2.77}$$

    where J_n is an $n \times n$ matrix with all the entries as 1.

- **The asymptotic properties of $\frac{1}{c^2 d^{2\alpha}} B_1$ and $\frac{1}{c^2 d^{2\alpha}} B_2$**

Again $B_1 = B_2^T$, so we only focus on B_2. The *ith* row, *jth* column of B_2 is

$$B_2(i,j) = cd^\alpha (\boldsymbol{\epsilon}_i^{(d)})^T \mathbf{v}_t^{(d)} = cd^\alpha \sum_{k=1}^{d} \epsilon_{i,k}^{(d)} v_{t,k}^{(d)} \quad (i,j = 1,\cdots,n)$$

Using similar derivations from (2.19) to (2.22), we have for any given $\epsilon > 0$

$$\begin{aligned} P(|\frac{1}{(cd^\alpha)^2} B_2(i,j)| > \epsilon) &\leqslant \frac{1}{\epsilon}\frac{1}{(cd^\alpha)} \left[\sum_{k=1}^{d} (v_{t,k}^{(d)})^4\right]^{1/4} \left[\sum_{k=1}^{d} (2\lambda_k^{(d)})^2\right]^{1/4} \\ &= \frac{1}{\epsilon} \left[\frac{\sum_{k=1}^{d} 4(\lambda_k^{(d)})^2}{(cd^\alpha)^4}\right]^{1/4} \left[\sum_{k=1}^{d} (v_{t,k}^{(d)})^4\right]^{1/4} \end{aligned} \quad (2.78)$$

Again for the second term on the right side

$$\lim_{d\to\infty} \sum_{k=1}^{d} (v_{t,k}^{(d)})^4 \leqslant \lim_{d\to\infty} \sum_{k=1}^{d} (v_{t,k}^{(d)})^2 = 1. \quad (2.79)$$

Recall that in Theorem 2.3.4, $\Sigma^{(d)} = diag(d^\beta, 1, \cdots, 1)$. Because $\beta > 1$ and $2\alpha > \beta$, we have

$$\frac{\sum_{k=1}^{d} 4(\lambda_k^{(d)})^2}{(cd^\alpha)^4} = 4 \times \frac{d^{2\beta} + (d-1)}{c^4 d^{4\alpha}} \longrightarrow 0 \quad as\ d \longrightarrow \infty. \quad (2.80)$$

Thus

$$P(|\frac{1}{(cd^\alpha)^2} B_2(i,j)| > \epsilon) \longrightarrow 0 \quad as\ d \longrightarrow \infty, \quad (2.81)$$

which means that

$$\frac{1}{(cd^\alpha)^2} B_2(i,j) \longrightarrow 0 \ \ in\ probability, \quad as\ d \to \infty. \quad (2.82)$$

- **The asymptotic properties of $\frac{1}{c^2 d^{2\alpha}} C$**

  Recall that $C = (\Upsilon^{(d)})^T \Upsilon^{(d)}$.

  When $i = j, \quad (i, j = 1, 2, \cdots, n)$,

$$C(i,i) = \sum_{k=1}^{d} (\epsilon_{i,k}^{(d)})^2.$$

  For any given $\epsilon > 0$, according Chebyshev's inequality,

$$\begin{aligned} P\left[|\frac{1}{(cd^\alpha)^2} C(i,i) > \epsilon \right] &= P\left[|\frac{1}{(cd^\alpha)^2} \sum_{k=1}^{d} (\epsilon_{i,k}^{(d)})^2 | > \epsilon \right] \leqslant \frac{E(\frac{1}{(cd^\alpha)^2} \sum_{k=1}^{d} (\epsilon_{i,k}^{(d)})^2)}{\epsilon} \\ &= \frac{\sum_{k=1}^{d} var((\epsilon_{i,k}^{(d)})^2)}{(cd^\alpha)^2 \epsilon} \\ &= 2 \times \frac{d^\beta + (d-1)}{c^2 d^{2\alpha} \epsilon} \\ &\longrightarrow 0 \ \ (\beta > 1 \ \ and \ \ 2\alpha > \beta). \end{aligned} \tag{2.83}$$

  Thus

$$\frac{1}{(cd^\alpha)^2} C(i,i) \quad \longrightarrow \quad 0 \ \ in\ probability, \quad as\ d \to \infty. \tag{2.84}$$

  When $i \neq j, \quad (i, j = 1, 2, \cdots, n)$,

$$C(i,j) = \sum_{k=1}^{d} \epsilon_{i,k}^{(d)} \epsilon_{j,k}^{(d)}.$$

  Using similar derivations with (2.83), we have

$$\frac{1}{(cd^\alpha)^2} C(i,j) \quad \longrightarrow \quad 0 \ \ in\ probability, \quad as\ d \to \infty. \tag{2.85}$$

  Hence

$$\frac{1}{(cd^\alpha)^2} C \quad \longrightarrow \quad 0 \ \ in\ probability, \quad as\ d \to \infty. \tag{2.86}$$

Summarizing the results in (2.77), (2.82) and (2.86), we conclude that when $\beta > 1$

and $2\alpha > \beta$,

$$\frac{1}{c^2 d^{2\alpha}} S_D^{(d)} \quad \longrightarrow \quad J_n \ \ \textit{in probability}, \quad \textit{as } d \to \infty. \tag{2.87}$$

- **The case when $2\alpha < \beta$**

  We multiply both sides of (2.75) by $\frac{1}{d^\beta}$,

  $$\frac{1}{d^\beta} S_D^{(d)} \quad = \quad \frac{1}{d^\beta} A + \frac{1}{d^\beta} B_1 + \frac{1}{d^\beta} B_2 + \frac{1}{d^\beta} C$$

  - **The asymptotic properties of $\frac{1}{d^\beta} A$**

    $$\lim_{d\to\infty} \frac{1}{d^\beta} A = \lim_{d\to\infty} \frac{(cd^\alpha)^2}{d^\beta} \frac{1}{(cd^\alpha)^2} A = 0 \times J_n = 0 \tag{2.88}$$

  - **The asymptotic properties of $\frac{1}{d^\beta} B_1$ and $\frac{1}{d^\beta} B_2$**

    Using similar derivations from (2.19) to (2.22), we have for any given $\epsilon > 0$

    $$\begin{aligned} P\left(\left|\frac{1}{d^\beta} B_2(i,j)\right| > \epsilon\right) \quad &\leqslant \quad \frac{1}{\epsilon} \frac{cd^\alpha}{d^\beta} \left[\sum_{k=1}^{d} (v_{t,k}^{(d)})^4\right]^{1/4} \left[\sum_{k=1}^{d} (2\lambda_k^{(d)})^2\right]^{1/4} \\ &= \quad \frac{c}{\epsilon} \left[\frac{4\sum_{k=1}^{d} (\lambda_k^{(d)})^2}{d^{4\beta-4\alpha}}\right]^{1/4} \left[\sum_{k=1}^{d} (v_{t,k}^{(d)})^4\right]^{1/4} \\ &= \quad \frac{c}{\epsilon} \left[\frac{d^{2\beta} + (d-1)}{d^{4\beta-4\alpha}}\right]^{1/4} \left[\sum_{k=1}^{d} (v_{t,k}^{(d)})^4\right]^{1/4} \\ &\longrightarrow \quad 0 \quad \textit{as } d \longrightarrow \infty.\ (\beta > 1 \textit{ and } 2\alpha < \beta) \end{aligned} \tag{2.89}$$

    Hence,

    $$\frac{1}{d^\beta} B_2(i,j) \longrightarrow 0 \ \ \textit{in probability}, \quad \textit{as } d \to \infty. \tag{2.90}$$

  - **The asymptotic properties of $\frac{1}{d^\beta} C$**

    Ahn *et al.* (2005) have studied the asymptotic properties of C as follows,

    Recall that $C = (\Upsilon^{(d)})^T \Upsilon^{(d)}$. Because the covariance matrix for the errors are

$2\Sigma^{(d)} = 2 \times diag(d^{\beta}, 1, \cdots, 1)$, the matrix C can be expressed as:

$$C = 2d^{\beta} W_1 + 2 \sum_{j=2}^{d} W_2,$$

where the W_j's are i.i.d. from the Wishard distribution $\mathcal{W}_n(1, I_n)$. Let

$$U := W_1$$

$$V := \sum_{j=2}^{d} W_2$$

Note that $U \sim \mathcal{W}_n(1, I_n)$ and $U \sim \mathcal{W}_n(d-1, I_n)$ independently. Then dividing C by d^{β} gives

$$\frac{1}{d^{\beta}} C = 2U + \frac{2}{d^{\beta}} V \tag{2.91}$$

As $d \to \infty$, the matrix V has the element-wise convergence, i.e. $\frac{1}{d-1} V \to I_n$. Thus, when $\beta > 1$, $\frac{2}{d^{\beta}} V \to 0$. It follows that

$$\frac{1}{d^{\beta}} C \longrightarrow 2U. \tag{2.92}$$

Combining results in (2.88), (2.90), and (2.92), we conclude that when $\beta > 1$ and $2\alpha < \beta$,

$$\frac{1}{d^{\beta}} S_D^{(d)} \longrightarrow 2U, \tag{2.93}$$

where $U \sim \mathcal{W}_n(1, I_n)$. Since U can be represented as the outer product of a random vector from $\mathcal{N}(0, I_n)$ it's eigenvalue is its inner product, which is a univariate random variable from χ_n^2.

Step B: The first eigenvector of the uncentered sample covariance matrix $S_P^{(d)}$.

- **The case when $2\alpha > \beta$.**

  Using similar derivations from (2.45) to (2.52), we obtain

$$(\frac{1}{c^2 d^{2\alpha}})(\mathbf{v}_t^{(d)})^T S_P^{(d)} \mathbf{v}_t^{(d)} \quad \longrightarrow \quad n \tag{2.94}$$

  As in (2.87), the matrix $S_D^{(d)}$ has the following limits

$$\frac{1}{c^2 d^{2\alpha}} S_D^{(d)} \quad \longrightarrow \quad J_n \ \ \textit{in probability}, \quad \textit{as } d \to \infty.$$

  Again, using similar derivations from (2.57) to (2.63), we have the following result

$$|(\mathbf{v}_t^{(d)})^T \mathbf{v}_e^{(d)}| \longrightarrow 1 \ \ \textit{in probability}, \ \textit{as } d \to \infty. \tag{2.95}$$

  This proves the first conclusion in Theorem 2.3.4.

- **The case when $2\alpha < \beta$.**

  Recall that the spike direction is $\mathbf{v}_s^{(d)} = (1, 0, \cdots 0)^T$. Then

$$\begin{aligned} (\frac{1}{d^\beta})(\mathbf{v}_s^{(d)})^T S_P^{(d)} \mathbf{v}_s^{(d)} &= (\frac{1}{d^\beta})\frac{nc^2 d^{2\alpha}}{d^\beta} + 2\frac{cd^\alpha}{d^\beta}\sum_{i=1}^{n} \epsilon_{1,i} + \frac{1}{d^\beta}\sum_{i=1}^{n}(\epsilon_{1,i})^2 \\ &\equiv S_1 + S_2 + S_3 \end{aligned} \tag{2.96}$$

  Because $\beta > 2\alpha$, we have $S_1 \longrightarrow 0$. Since $var(\epsilon_{i,1}) = 2d^\beta$ and $\beta > 2\alpha$,

$$S_2 = 2\frac{cd^\alpha}{d^\beta}\sum_{i=1}^{n} \epsilon_{1,i} = 2\sqrt{2}\frac{cd^\alpha}{d^{\beta/2}}\sum_{i=1}^{n} \frac{\epsilon_{1,i}}{\sqrt{2}d^{\beta/2}} \to 0 \times \sum_{i=1}^{n} Z_i \longrightarrow 0$$

  where the Z_i follows the standard Gaussian distribution.

$$S_3 = \frac{1}{d^\beta}\sum_{i=1}^{n}(\epsilon_{1,i})^2 = 2\sum_{i=1}^{n}(\frac{\epsilon_{1,i}}{\sqrt{2}d^{\beta/2}})^2 = 2\sum_{i=1}^{n}(Z_i)^2 \sim 2 \times \chi_n^2$$

  Suppose $S_P^{(d)}$ has the following eigenvalue decomposition:

$$S_P^{(d)} = GLG^t = \hat{\lambda}_1 \hat{\boldsymbol{g}}_{\boldsymbol{1}} \hat{\boldsymbol{g}}_{\boldsymbol{1}}{}^T + \cdots + \hat{\lambda}_n \hat{\boldsymbol{g}}_{\boldsymbol{n}} \hat{\boldsymbol{g}}_{\boldsymbol{n}}{}^T$$

This indicates that

$$\frac{1}{d^\beta}(\mathbf{v}_s^{(d)})^T S_P^{(d)} \mathbf{v}_s^{(d)} = \frac{\hat{\lambda}_1}{d^\beta}|(\mathbf{v}_s^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 + \cdots + \frac{\hat{\lambda}_n}{d^\beta}|(\mathbf{v}_s^{(d)})^T \hat{\boldsymbol{g}}_{\boldsymbol{n}}|^2 \tag{2.97}$$

As we have discussed in (2.93), the eigenvalues have the following asymptotic properties

$$\frac{\hat{\lambda}_1}{d^\beta} \longrightarrow \chi_n^2$$

$$\frac{\hat{\lambda}_j}{d^\beta} \longrightarrow 0 \ (j = 2, \cdots, n).$$

The left side of (2.97) also converges to χ_n^2. Since Equation (2.97) holds, we must have

$$|(\mathbf{v}_s^{(d)})^T \hat{\boldsymbol{g}}_{\mathbf{1}}|^2 \longrightarrow 1 \ \ \textit{in probability}, \quad \textit{as } d \to \infty.$$

This is equivalent to

$$|(\mathbf{v}_s^{(d)})^T \mathbf{v}_e^{(d)}|^2 \longrightarrow 1 \ \ \textit{in probability}, \quad \textit{as } d \to \infty,$$

which proves the second conclusion in Theorem 2.3.4.

CHAPTER 3

Comparison among SVM, DWD, and PAM

This chapter studies and compares three batch adjustment methods that were motivated by data discrimination methods. These methods are SVM, DWD and PAM, which have been introduced in Chapter 1, Section 1.3.3. In Section 3.1, we compare the SVM and the DWD methods. Several toy examples are given to illustrate the limitations of SVM, especially for the HDLSS data sets. In Section 3.2, we study the robustness of DWD and PAM under the **Unbalanced Subgroup Model**. DWD will be shown to be much robust than PAM when the dimension is fixed and the subgroup sample sizes become more and more unbalanced. The mathematical problem of interest is to study the d asymptotic properties of DWD and PAM (see Chapter 2, Section 2.3.1). The conclusions are presented in Theorems 3.2.1 and 3.2.2. Simulation studies are given to verify the results in the two theorems. In Section 3.2.5, we give the proofs of Theorems 3.2.1 and 3.2.2.

3.1 The Comparison between DWD and SVM

In Chapter 1, Section 1.3.3, we have given the definitions for several commonly used linear discrimination methods, including SVM, DWD and PAM. Figure 1.9 shows the SVM hyperplane between the two data sets, represented by blue circles and red pluses. As we have discussed, the SVM normal vector (green dashed line) is only affected by those points on the two margins (dashed thin grey lines). The observations which are not on the margins have no effect at all. For example, in Figure 1.9, if you move those off-margin blue circles to the locations which are further away from the margin, the SVM hyperplane will not change at all. This property of SVM could cause serious problems. Next, we will use two toy example to illustrate the drawbacks of SVM, when we use it as a batch adjustment method.

The first drawback is that SVM could produce bias batch adjustment, as shown in Figure 3.1. This toy data contain two batches, represented by blue circles and red crosses. The purpose of linear batch adjustment is to find a direction and shift the two data sets until they overlap. Note that in the toy data, the support vector, i.e. the points on the margins almost form parallel line. Thus the SVM discrimination hyperplane will also be parallel with the two sets of points on the margins and is located halfway between the two margins. The orthogonal direction of the SVM hyperplane is shown using the magenta dashed line, called the SVM direction. Apparently, shifting the two data sets along the SVM direction will not successfully combine the two data sets and will instead produce biased batch adjustment results. The reason is that the SVM ultimately only considers those points on the margins, and totally ignore the effects of other points. A much better batch adjustment direction is shown using the green line. Shifting the two data sets along this direction will successfully eliminate the batch difference.

The second drawback is the data piling problem, especially for the HDLSS data. This problem was first noticed by Marron *et al.* (2005). Figure 3.2 illustrates this problem using two toy data sets, each of which contains 20 samples in 50 dimensional space. The left plot shows a projection view of the data. The magenta line represents the SVM direction. The projection view of the data along the SVM direction is shown in the right bottom plot. Notice that many observations pile up on the margins, which lead to opposite directions of skewness between the two populations. Shifting the data along the SVM direction will not combine the two data sets successfully. A much better projection direction is shown using the dashed magenta line in the left plot. The projection of the data along this direction is shown in the right top plot. The projections of both data sets have an approximately Gaussian shape (uni-modal and symmetric). Shifting the data along this direction will produce a successful data combination. The problem of data piling becomes more and more severe when the dimension of the data increase. This is due to the fact that SVM only maximizes the margin and totally ignores those points off the margins.

Distance Weighted Discrimination (DWD) was proposed by Marron *et al.* (2005) as an

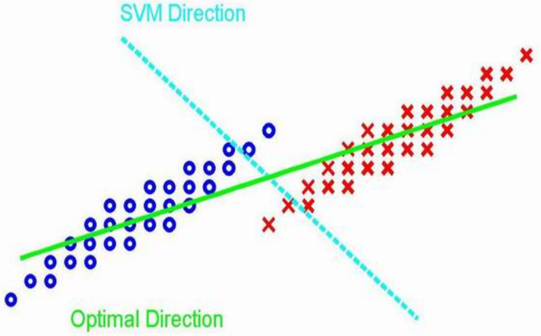


Figure 3.1: Two data sets are represented by blue circles and red crosses respectively. The SVM Direction (magenta dashed line) is orthogonal to the SVM hyperplane, which is determined only by those points on the margin. Combining two data sets along this direction will not produce a good result. A much better batch adjustment direction is shown using the green line.

improvement upon the SVM for the problem of statistical classification (i.e. discrimination), especially in HDLSS problems. In Chapter 1, Section 1.3 gives the definition of the DWD hyperplane between two separable data sets. DWD finds the hyperplane such that the sum of the inverse distances from the samples to the hyperplane is minimized. Thus, instead of only considering the observations on the margin as SVM does, DWD allow the every observation to have some influence. However, those observation close to the hyperplane are much more important than those which are far away from the hyperplane. DWD has been shown to avoid the data piling problem for HDLSS data sets. Benito *et al.* (2004) illustrates this using some toy data sets. Figure 3.3 shows the projection of the toy data along the SVM direction. Although two populations have good separation along this direction, many observations from the two sources are piled up on the margins. The approximate density curves for the two populations are skewed in the opposite direction. These make the shifting the two data sets along the SVM direction unsuccessful to adjust source difference. In

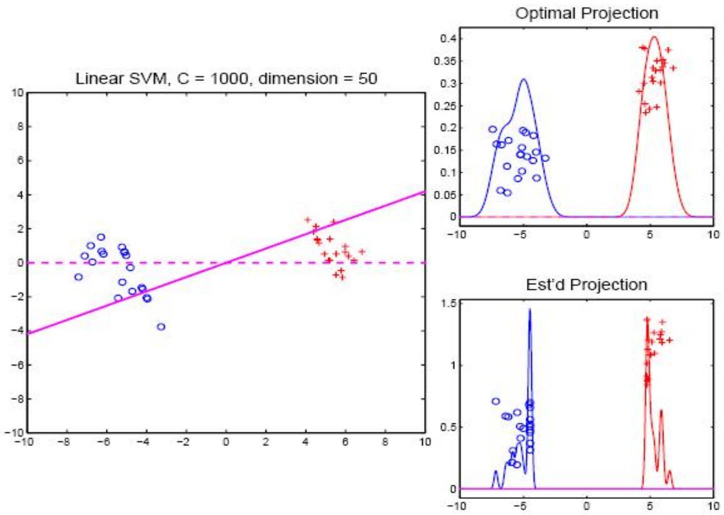


Figure 3.2: This Figure is taken from Marron *et al.* (2005) to illustrate the data piling problem of SVM. The toy data contain two batches, represented by blue circles and red pluses. Each data set contains 20 observations in 50 dimensional space. The left plot shows the projection view of the toy data on the plane formed by the first two principal component directions. The SVM direction is shown using the magenta line. The dashed line shows the optimal direction. The two plots on the right show the projection view of the toy data along the SVM direction and the optimal direction.

this figure, the distance between the two centers of the projected data sets is around 26. Figure 3.4 shows similar projection plot for the toy data in Figure 3.3 using the DWD direction between the two data sets. First of all, the projection of both data sets have smooth Gaussian shape density curve (uni-modal and symmetric). The shifting of the data along the DWD direction will eliminate the source difference and produce successful data combination. Secondly, along the DWD direction the distance between the centers of the two projected population is around 30. Thus the DWD direction provides better separation between the two data sets than the SVM does.

Since SVM has this serious problem of data piling, we will focus on DWD and PAM in the rest of this dissertation.

3.2 The Comparison between PAM and DWD

In Section 3.1, we have studied the good performance of DWD over SVM for adjusting Batch difference. In this Section, we will extend the analysis of the DWD direction, by

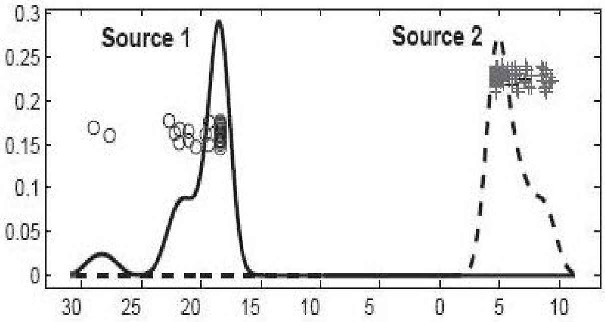


Figure 3.3: This Figure is taken from Benito *et al.* (2004) to illustrate the data piling problem of SVM. Two data sets are projected along the SVM direction. The estimated density curves are fitted for the two populations separately. Similar projection plots can be found on the diagonal of Figure 1.6.

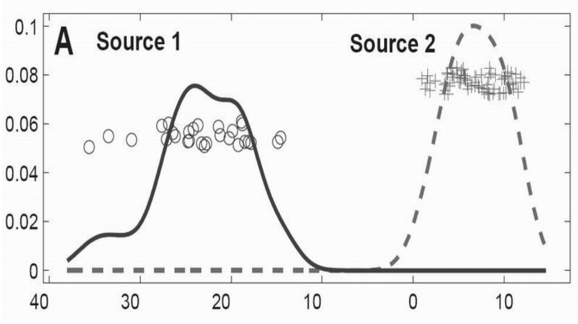


Figure 3.4: This Figure is taken from Benito *et al.* (2004) to illustrate that DWD does not have the data piling problem. The two data sets are projected along the DWD direction. The estimated density curves are fitted for the two populations separately. Similar projection plots can be found on the diagonal of Figure 1.6.

explicitly studying robustness issues due to unbalanced subgroup sample sizes. E.g, both of two microarray data sets contain breast cancer samples and leukaemia samples. But one data set has a much larger proportion of breast cancer samples and smaller proportion of leukaemia samples than the other data set. We say that these two data sets have unbalanced subgroups.

The robustness of DWD due to the unbalanced sugroups effect is compared with another commonly used batch adjustment method: PAM, which has been discussed in Chapter 1 Section 1.3.3. Suppose $X_{d\times n_1}$ and $Y_{d\times n_2}$ are two microarray data sets. With the multivariate view, they are treated as two clouds of points in d dimensional gene space. Using PAM, two clouds are rigidly shifted along the direction, which connects two centroids of the clouds, until these two centroids overlap.

PAM is simple and easily understood. However, PAM doesn't work well when two data sets have unbalanced subgroups. We studied the robustness of DWD and PAM due to the effect of unbalanced subgroup in two ways. In Section 3.2.1, the toy data sets are used to show that DWD is consistently more robust than PAM, when the the dimension of data is fixed and the subgroups sample sizes become more and more unbalanced. In Section 3.2.2 to 3.2.5, we studied the robustness of DWD and PAM directions for the data from the **Unbalanced Subgroup Model**, i.e. the subgroups sample sizes are unbalanced and fixed, and the dimension of the data goes to infinity. The robustness of DWD and PAM are shown in two theorems separately. Section 3.2.4 is about the simulation verifications for these two theorems. The proofs of these two theorems are given in Section 3.2.5.

3.2.1 Robustness of DWD and PAM for Data with Fixed Dimension

This point is explored in the following toy example. In this example, there are 4 clusters in the simulated data, which have 4000 genes. One grouping of the clusters is into two biological subtypes, which could represent treatment or cancer type, represented by color. The other grouping of the data is into systematic effects, which could be protocol, batch or platform effects, represented by different symbols. This design is illustrated in the first column of Figure 3.5, where red and blue are used to illustrate the two biological subgroups, and pluses and circles are used for the systematic effect.

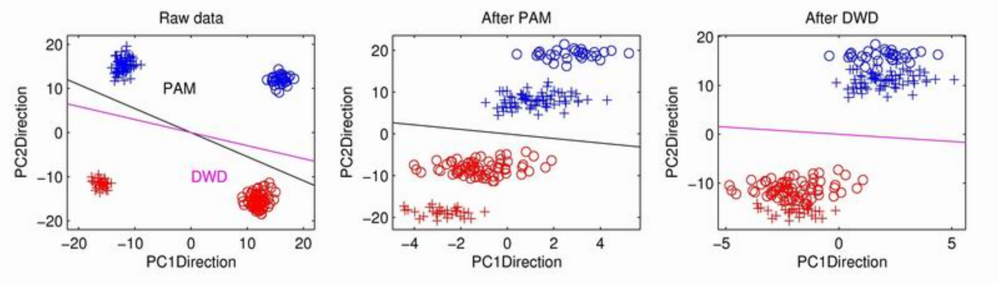
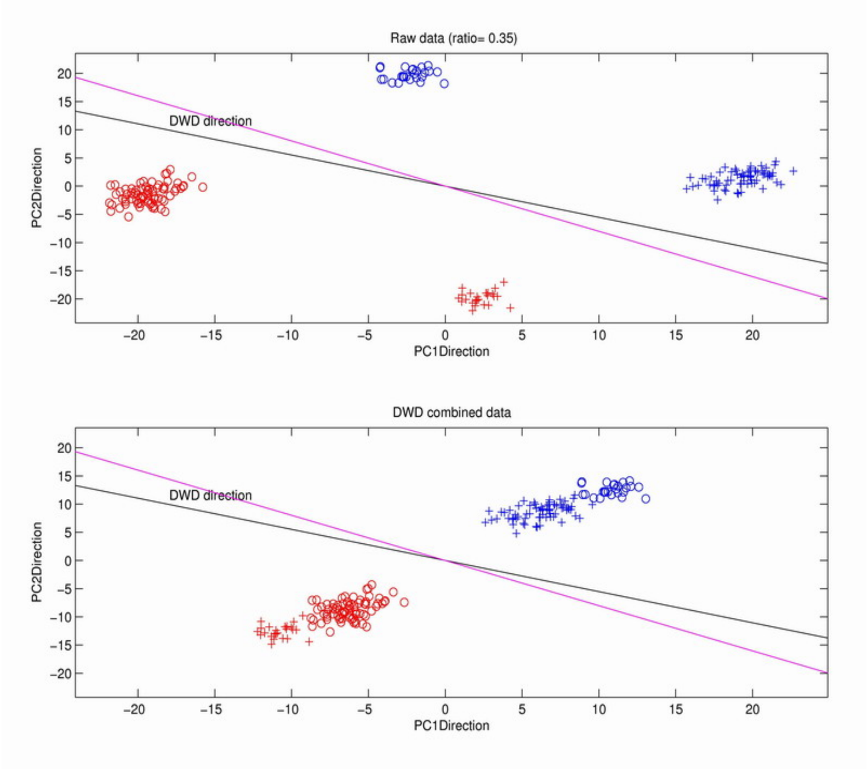


Figure 3.5: Toy example to illustrate the effect of unbalanced subgroup effect. Symbols are for the batches and colors are for the biological effects. The first, second, third columns are the PC projection plots for the Raw, PAM adjusted, and DWD adjusted data, respectively. The purple line is the DWD direction. The black line is the PAM direction.

The goal of DWD and PAM adjustment is to eliminate the systematic effect (i.e. to move the clusters with similar symbols on top of each other), while at same time preserving the biological structure in the data (i.e. to keep the different colors separated). For the data in the left plot of Figure 3.5, an excellent result will have just two clusters, each with a different color, and will have complete overlap of the appropriately colored symbols. This good result can be achieved when the four cluster sizes are balanced. Usually, when the data have unbalanced subgroups, the samples from the sam biological cluster won't totally overlap. In the following, we investigate the performance of PAM and DWD using toy data sets.

The left plot of Figure 3.5 shows the Raw data (data before adjustment). Note that there are relatively fewer blue circles and red pluses, and more blue pluses and red circles. Hence the subgroup sample sizes are unbalanced. In this panel the ratio between the number of blue circles and red circles (similarly between red pluses and blue pluses) is 0.43. The left plot is a projection of the raw data onto the first two Principal Component directions (in 4000 dimensional gene space). This clearly shows the four clusters. The best adjustment is combining the blue clouds together and combining the red clouds together. However, PAM doesn't produce such a good adjustment. The second column plot shows the result of PAM adjustment. The colored clusters have not been brought together. The third column shows the result of DWD adjustment. The colored clusters have now been brought together which indicates that DWD is more robust than PAM due to the unbalanced subgrooups.

To study robustness of DWD over a range of different cluster size ratios, we chose to fix the number of samples for each biological subtype, and to fix the number of samples for each systematic subtype. The number of genes is always 4000 (fixed dimension). Thus unbalance of the subgroup sample sizes is created by removing samples from two of the clusters, and adding the same number to the other two, i.e. removing sample from red pluses to blue pluses and removing samples from blue circles to red circles. DWD is used for adjusting batch difference. The result are shown using a movie ToyMovie-DWDRobust.avi, which is web available from the Liu (2007b). Each frame of the movie shows the projection view of the Raw data and DWD adjusted data with a different subgroup sample size ratio. There the subgroup sample size ratio varies from 1 (perfectly balanced subgroups) to the extremely unbalanced ratio of 0.04. As expected, when the ratio is 1, the adjustment is excellent, with the two colored clusters coming together, and a complete overlapping of the circles and pluses. The overlap is still very good to ratios around 70%. Then overlap lessens down to 35% (chosen to appear in Figure 3.6, as "midway"), after which a space appears between the clusters. While it is an arbitrary choice in a continuum, we feel that the method is quite broken down, when the ratio falls below 20% (i.e. 5 to 1), in the sense that then the gap between the two biggest clusters is actually smaller than the gaps between the clusters of the same colors. Figure 3.6 shows one frame of the movie when the ratio is 35%.


Raw data (ratio= 0.35)

DWD combined data

Figure 3.6: Toy example to illustrate the effect of sub-sample size. Symbols and colors are the same as above. The purple line is the DWD adjustment direction. The black line is the best adjustment direction. Top and bottom panels show the projections of the Raw data and DWD adjusted data onto the plane formed by the first two PC directions.

In summary, we find that DWD is much more robust than PAM. DWD gives very robust performance for data sets where the subgroup sample size ratios are $\frac{2}{3}$ or better. It is still reasonably robust for the ratios down to $\frac{1}{3}$, and still seems to have some benefit for ratios down to $\frac{1}{5}$. When the subgroup sample size ratio is very low, two data sets can't have many samples from the same biological subgroup. In this case, we won't expect any successful adjustment.

In the rest of this Chapter, we will discuss the asymptotic properties of the DWD and the PAM directions as d goes to infinity, when the subgroup sample sizes are unbalanced and fixed. In Section 3.2.2, we propose a statistical model, called **Unbalanced Subgroup Model**, for simulating similar data sets as in Figure 3.6 with four clusters.

3.2.2 Unblanced Subgroup Model

In this section, we use similar notations for the gene expression matrices as in Chapter 2, Section 2.3.2. Suppose that $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{Y}_1$, and $\mathcal{Y}_2$ are four series of HDLSS random matrices, i.e.

$$\mathcal{X}_1 = \{X_1^{(1)}, \cdots, X_1^{(d)}, \cdots\},$$

$$\mathcal{X}_2 = \{X_2^{(1)}, \cdots, X_2^{(d)}, \cdots\},$$

$$\mathcal{Y}_1 = \{Y_1^{(1)}, \cdots, Y_1^{(d)}, \cdots\},$$

$$\mathcal{Y}_2 = \{Y_2^{(1)}, \cdots, X_2^{(d)}, \cdots\}.$$

The variables with superscript (d) indicate that they are specifically for the data with d genes. For example, the four matrices $X_1^{(d)}, X_2^{(d)}, Y_1^{(d)}$ and $Y_2^{(d)}$ are expression matrices for d genes. One grouping of the four series of matrices is into systematic effects, i.e. the batch $\mathcal{X}$, which contains $\mathcal{X}_1$, $\mathcal{X}_2$ and the other batch $\mathcal{Y}$, which contains $\mathcal{Y}_1$ and $\mathcal{Y}_2$. The other grouping of the data is into two biological subgroups, i.e. treatments or cancer types. In this model, we use the subscripts to represent the biological subgroups, i.e. all the samples in $\mathcal{X}_1$ and $\mathcal{Y}_1$ are from the biological subtype 1; all the samples in $\mathcal{X}_2$ and $\mathcal{Y}_2$ are for the biological subtype 2. For mathematical convinces, we study a simplified unbalanced sugroup model, the sample sizes in $\mathcal{X}_1$ and $\mathcal{Y}_2$ are both n, and the sample sizes in $\mathcal{X}_2$ and $\mathcal{Y}_1$ are m $(m < n)$. Thus the total number of samples in each batch is $N = n + m$. The number of samples for each biological subtype is also N. The subgroup sample sizes are unbalanced because there are more biological subtype 1 samples than biological subtype 2 samples in the batch $\mathcal{X}$, and there are less biological subtype 1 samples than the biological subtype 2 samples in the batch $\mathcal{Y}$. The subgroup sample size ratio is defined as $r = \frac{n}{m}$. In our model, we have $r > 1$.

When the number of genes is d, the four expression matrices are $X_1^{(d)}$, $X_2^{(d)}$, $Y_1^{(d)}$ and $Y_2^{(d)}$. From the multivariate view, they are represented by four clusters of points in d dimensional space. We write the four data matrices as

$$X_1^{(d)} = (\mathbf{x}_{1,1}^{(d)}, \cdots, \mathbf{x}_{1,n}^{(d)}),$$

$$X_2^{(d)} = (\mathbf{x}_{2,1}^{(d)}, \cdots, \mathbf{x}_{2,m}^{(d)}),$$

$$Y_1^{(d)} = (\mathbf{y}_{1,1}^{(d)}, \cdots, \mathbf{y}_{1,m}^{(d)}),$$

$$Y_2^{(d)} = (\mathbf{y}_{2,1}^{(d)}, \cdots, \mathbf{y}_{1,n}^{(d)}).$$

In our model, each column vector is generated from the multivariate Gaussian distribution, with the covariance as the identity matrix I_d. The vectors in each cluster have the same mean vector. The following four d dimensional vectors are the mean vectors for the columns in $X_1^{(d)}$, $X_2^{(d)}$, $Y_1^{(d)}$ and $Y_2^{(d)}$ respectively:

$$\mathbf{v}_{x,1}^{(d)} = d^{\alpha-\frac{1}{2}}(-1,\ \ 1, -1,\ \ 1, \cdots)^T,$$

$$\mathbf{v}_{x,2}^{(d)} = d^{\alpha-\frac{1}{2}}(-1, -1, -1, -1, \cdots)^T,$$

$$\mathbf{v}_{y,1}^{(d)} = d^{\alpha-\frac{1}{2}}(\ \ 1,\ \ 1,\ \ 1,\ \ 1, \cdots)^T,$$

$$\mathbf{v}_{y,2}^{(d)} = d^{\alpha-\frac{1}{2}}(\ \ 1, -1,\ \ 1, -1 \cdots)^T.$$

Over the sequence of different numbers of genes, the four mean vectors are represented by four triangular sequences, i.e.

$$\mathcal{V}_{x,1} = \{\mathbf{v}_{x,1}^{(1)}, \cdots, \mathbf{v}_{x,1}^{(d)}, \cdots\},$$

$$\mathcal{V}_{x,2} = \{\mathbf{v}_{x,2}^{(1)}, \cdots, \mathbf{v}_{x,2}^{(d)}, \cdots\},$$

$$\mathcal{V}_{y,1} = \{\mathbf{v}_{y,1}^{(1)}, \cdots, \mathbf{v}_{y,1}^{(d)}, \cdots\},$$

$$\mathcal{V}_{y,2} = \{\mathbf{v}_{y,2}^{(1)}, \cdots, \mathbf{v}_{y,2}^{(d)}, \cdots\}.$$

The asymptotic norms of these four triangular sequences are all d^α in the sense that

$$\lim_{d\to\infty} \frac{1}{d^\alpha}\|\mathbf{v}_{x,1}^{(d)}\| = 1 \tag{3.1}$$

Similar results hold for the other three sequences. Using matrix notation, the four data sets in the unbalanced subgroup model can be expressed as:

$$\begin{aligned} X_1^{(d)} &= \mathbf{v}_{x,1}^{(d)} \times (1_n)^T + \Upsilon_{x,1}^{(d)} \\ X_2^{(d)} &= \mathbf{v}_{x,2}^{(d)} \times (1_m)^T + \Upsilon_{x,2}^{(d)} \\ Y_1^{(d)} &= \mathbf{v}_{y,1}^{(d)} \times (1_m)^T + \Upsilon_{y,1}^{(d)} \\ Y_2^{(d)} &= \mathbf{v}_{y,2}^{(d)} \times (1_n)^T + \Upsilon_{y,2}^{(d)} \end{aligned} \tag{3.2}$$

where 1_n and 1_m represents the n and m dimensional vectors respectively with all entries equal to one; all the $\Upsilon^{(d)}s$ represent measurement errors. Each column of them follows the multivariate gaussian distribution with mean 0 and covariance matrix I_d.

Figure 3.7 illustrates the underlying conceptual structure for the unbalanced subgroup model. The batch effects are represented by symbols, i.e. pluses for the batches $\mathcal{X}$ and circles are for the batch $\mathcal{Y}$. The biological effects are represented by colors, i.e. reds are for the biological subtype 1, blues are for the biological subtype 2. Hence, clockwise from the top row, first column cluster, the four clusters are $X_1^{(d)}$, $Y_1^{(d)}$, $Y_2^{(d)}$ and $X_2^{(d)}$ respectively. The sample sizes are unbalanced in the sense that there are more red pluses than blue pluses and there are less red circles than blue circles.

The unbalanced subgroup model captures an important phenomena in microarray batch adjustment analysis. In the real data analysis, it is very common that two before-adjusted data sets contain unequal proportions of the samples from the same biological subtype. The unbalanced subgroup model studies an extreme case, where the sample size ratio in one batch is r, and it is $\frac{1}{r}$ $(r > 1)$ in the other batch. In next section, we study the effects

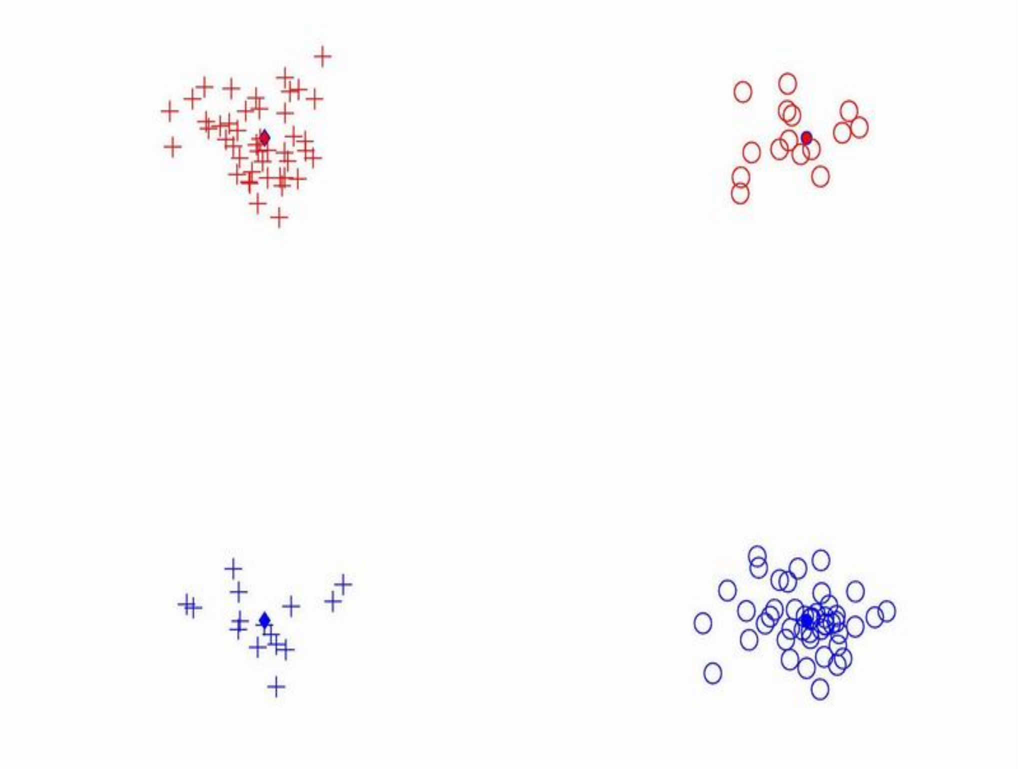


Figure 3.7: Toy example to illustrate the underlying conceptual structure of the unbalanced subgroup model. Symbols are for the batches and the colors are for the biological subgroups.

of unbalanced sample sizes on the batch adjustment, when the dimension tends to infinity. Considering the drawbacks of SVM in Section 3.1, we focus on the comparison between DWD and PAM.

3.2.3 The d Asymptotic Properties of the DWD and PAM directions

Two batches are $[X_1^{(d)}, X_2^{(d)}]$ and $[Y_1^{(d)}, Y_2^{(d)}]$, which are represented by pluses and circles respectively in Figure 3.7. Suppose that we intend to adjust the batch difference between pluses and circles by linearly shifting them along the chosen direction. A successful combination result will have all the blue samples together and all the red samples together. The best combination direction is the direction vector of $\mathbf{v}_{y,1}^{(d)} - \mathbf{v}_{x,1}^{(d)}$ or $\mathbf{v}_{y,2}^{(d)} - \mathbf{v}_{x,2}^{(d)}$, because if there were no measurement noise, the batch differences can be totally removed by shifting the data along this direction. We call the normalized direction vector of $\mathbf{v}_{y,1}^{(d)} - \mathbf{v}_{x,1}^{(d)}$ or $\mathbf{v}_{y,2}^{(d)} - \mathbf{v}_{x,2}^{(d)}$ as **the best combination direction**, denoted as $\mathbf{v}^{(d)}$. Actually, we have

$$\begin{aligned}\mathbf{v}^{(d)} &= \frac{\mathbf{v}^{(d)}_{y,1} - \mathbf{v}^{(d)}_{x,1}}{\|\mathbf{v}^{(d)}_{y,1} - \mathbf{v}^{(d)}_{x,1}\|} \\ &= \frac{\mathbf{v}^{(d)}_{y,2} - \mathbf{v}^{(d)}_{x,2}}{\|\mathbf{v}^{(d)}_{y,2} - \mathbf{v}^{(d)}_{x,2}\|} \\ &= \sqrt{\frac{2}{d}}(1, 0, 1, 0, \cdots)^T. \end{aligned} \tag{3.3}$$

We can also apply the DWD or PAM method, introduced in Chapter 1, Section 1.3.3 to adjust the batch difference. Two combination directions, the DWD direction and the PAM direction are denoted as $\mathbf{v}^{(d)}_{DWD}$ and $\mathbf{v}^{(d)}_{PAM}$ respectively. Figure 3.8 illustrates the best combination direction (black), the PAM direction (megenta) and the DWD direction (green) for adjusting the differences between the two batches in Figure 3.7. It shows that the PAM direction has been driven significantly by the unbalanced sample sizes effect. It tends to the direction which points from the large cluster (red pluses) to the other large cluster (blue circles). The DWD direction has also been driven by the effect of unbalanced sample sizes, however, not as much as the PAM direction. In Section 3.2.1, we have shown that DWD is consistently better than the PAM direction as the dimension d is fixed and the sample sizes becomes more and more unbalanced. From now on, we compare the asymptotic properties of $\mathbf{v}^{(d)}_{DWD}$ and $\mathbf{v}^{(d)}_{PAM}$, when the sample sizes are fixed and unbalanced, and the dimension d goes to infinity.

The Absolute value of Inner Products (AIP) is used to evaluate the similarity between two normed direction vectors. We use AIP because we only care about the acute angle between the two direction vectors (modulo the ± flip of direction). As we have introduced in Section 3.2.2, the asymptotic norms of the mean vectors $\mathbf{v}^{(d)}_{x,1}, \mathbf{v}^{(d)}_{x,1}, \mathbf{v}^{(d)}_{x,1}$ and $\mathbf{v}^{(d)}_{x,1}$ are all d^{α}, thus α represents how fast these four clusters move apart when d goes to infinity. Looking over a range of choices of α, we develop the two following theorems.

Theorem 3.2.1. *(DWD Direction)*

Suppose that the four series of data are generated from the unbalanced subgroup model, as

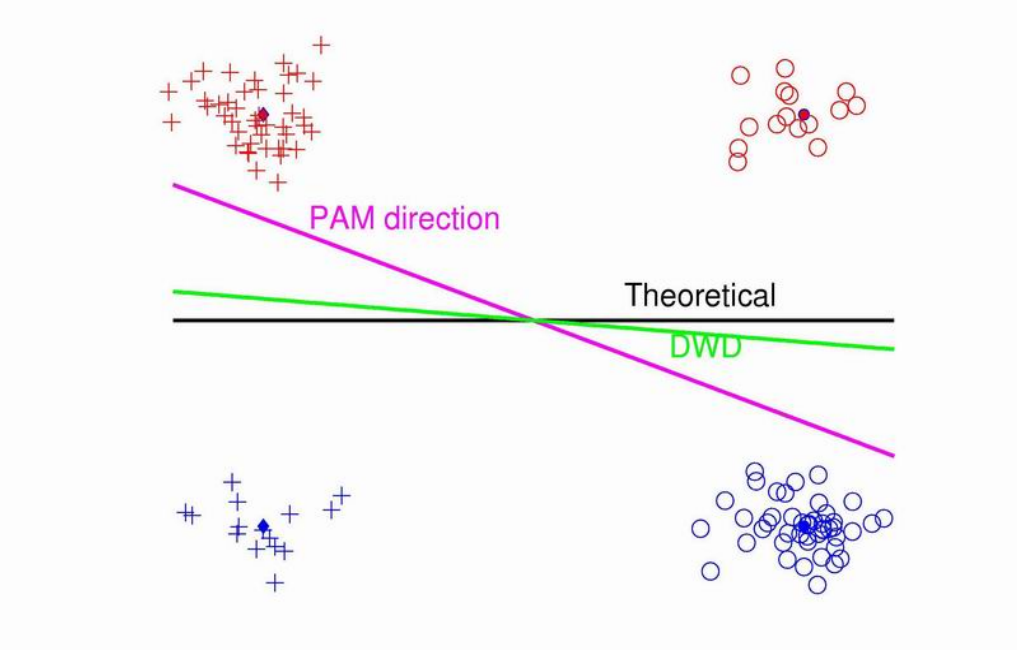


Figure 3.8: This figure illustrates the theoretical, DWD and PAM direction to adjusted the systematic differences between data from two batches. The same data has been shown in Figure 3.7.

in Section 3.2.2. The sample sizes n *and* m *are fixed, and the subgroup sample size ratio* $r = \frac{n}{m}$*. Depending on the value of* α*, we have the following conclusions for the DWD direction* $\mathbf{v}_{DWD}^{(d)}$ *and the theoretical combination direction* $\mathbf{v}^{(d)}$ *between the two batches* $\mathcal{X}$ *and* $\mathcal{Y}$*. Recall that* $AIP = |(\mathbf{v}^{(d)})^T \mathbf{v}_{DWD}^{(d)}|$*,*

1: if $\alpha > \frac{1}{2}$*,* $AIP \longrightarrow \frac{\sqrt[3]{r}+1}{\sqrt{2\sqrt[3]{r^2}+2}}$ *in probability, as* $d \to \infty$*;*

2: if $\alpha < \frac{1}{2}$*,* $\mathbf{v}_{DWD}^{(d)}$ *is asymptotically orthogonal to* $\mathbf{v}^{(d)}$ *in probability, i.e.* $AIP \longrightarrow 0$ *in prob. as* $d \to \infty$*. (strong inconsistency)*

Theorem 3.2.1 presents the asymptotic relations between the DWD direction with the best combining direction $\mathbf{v}^{(d)}$. Similar as we studied in Chapter 2, the samples in each cluster converge to the vertices of a simplex as d goes to infinity. The speed of convergence is determined by the covariance matrix, which is assumed to be the identity matrix I_d. As d increases, the distances between clusters also increase, with the speed of d^{α}. When α is large enough ($> \frac{1}{2}$), the increasing of the distances between batches dominates the variation within each cluster, thus the data act as though there were no errors in the data, under which, the angle between the DWD direction and the best combination direction $\mathbf{v}^{(d)}$ converges to $\theta_{DWD} = cos^{-1}\frac{\sqrt[3]{r}+1}{\sqrt{2\sqrt[3]{r^2}+2}}$. This result is due to the effect of unbalanced sample sizes. Note that when $r = 1$, the angles between them converges to zero. When α is relatively small ($< \frac{1}{2}$), the variations within each cluster dominate the increasing of the distances between batches, so we conclude that the DWD direction is asymptotically orthogonal with the best combination direction. This is called the strong inconsistency of the DWD direction. In Section 3.2.5, we give the details of the proof for Theorem 3.2.1.

Note that the asymptotic properties of the DWD direction are affected by the he unbalanced sample size ratio $r = \frac{n}{m}$. Actually, PAM direction has similar asymptotic properties to the DWD direction. We will we show that the DWD direction is always more robust than the PAM direction.

Theorem 3.2.2. *(PAM Direction)*
Suppose that the four series of data are generated from the unbalanced subgroup model, as in Section 3.2.2. The subgroup sample size ratio is $r = \frac{n}{m}$ *and the total sample size in each batch is* $N = n+m$. *Depending on the value of* α, *we have the following conclusions for the PAM direction* $\mathbf{v}_{PAM}^{(d)}$ *and the theoretical combination direction* $\mathbf{v}^{(d)}$ *between the two batches* $\mathcal{X}$ *and* $\mathcal{Y}$. *As the sample sizes* n *and* m *are fixed, recall that* $AIP = |(\mathbf{v}^{(d)})^T\mathbf{v}_{PAM}^{(d)}|$,
1: if $\alpha > \frac{1}{2}$, $AIP \longrightarrow \frac{r+1}{\sqrt{2r^2+2}}$ *in probability, as* $d \to \infty$;
2: if $\alpha = \frac{1}{2}$, $AIP \longrightarrow \frac{r+1}{\sqrt{2r^2+2+(1/N)(r+1)^2}}$ *in probability, as* $d \to \infty$;
3: if $\alpha < \frac{1}{2}$, $\mathbf{v}_{PAM}^{(d)}$ *is asymptotically orthogonal with* $\mathbf{v}^{(d)}$, *in the sense that* $AIP \longrightarrow 0$ *in probability, as* $d \to \infty$. *(strong inconsistency)*

Theorem 3.2.2 indicates that the PAM direction is always inconsistent with the best combination direction $\mathbf{v}^{(d)}$, as long as the subgroup sample sizes are unbalanced ($r \neq 1$). The asymptotic angle between them can be calculated. Note that $\mathbf{v}_{PAM}$ is strongly affected by unbalanced subgroup sample size ratio r. When $\alpha > \frac{1}{2}$, the angle between the PAM direction and $\mathbf{v}^{(d)}$ converges to a fixed value, $\theta_{PAM} = cos^{-1}(\frac{r+1}{\sqrt{2r^2+2}})$. This angle is not zero as long as $r \neq 1$. In the special case where the subgroup sample sizes are balanced, i.e.

$r = 1$, PAM converges to the best combination direction. When $\alpha = \frac{1}{2}$, the PAM direction never converges to $\mathbf{v}^{(d)}$; even when the subgroup sample sizes are balanced. The asymptotic angle between $\mathbf{v}_{PAM}^{(d)}$ and $\mathbf{v}^{(d)}$ is a little bit different from the angle between them when $\alpha > \frac{1}{2}$. When $\alpha < \frac{1}{2}$, the PAM direction is asymptotically orthogonal to the best combining direction. This is the strong inconsistency of the PAM direction. The details of the proof for Theorem 3.2.2 are given in Section 3.2.5.

Comparison between DWD and PAM

When $\alpha < \frac{1}{2}$, both the DWD and the PAM directions are asymptotically orthogonal with the best combination direction $\mathbf{v}^{(d)}$. When $\alpha > \frac{1}{2}$, Theorem 3.2.1 and 3.2.2 discover a very important difference between PAM and DWD combination. Define that $f(r) = \frac{r+1}{\sqrt{2r^2+2}}$. Some calculations show that $f(r)$ is a decreasing function when $r > 1$. For any $r > 1$, we have $\sqrt[3]{r} < r$, thus

$$\frac{\sqrt[3]{r}+1}{\sqrt{2\sqrt[3]{r^2}+2}} = f(\sqrt[3]{r}) > f(r) = \frac{r+1}{\sqrt{2r^2+2}},\ (r \neq 1).$$

It follows that

$$\theta_{DWD} = cos^{-1}(\frac{\sqrt[3]{r}+1}{\sqrt{2\sqrt[3]{r^2}+2}}) < cos^{-1}(\frac{r+1}{\sqrt{2r^2+2}}) = \theta_{PAM},\ (r \neq 1). \tag{3.4}$$

This indicate that the DWD direction is always more robust than the PAM direction, in the sense that the angle θ_{DWD} is always smaller than the angle θ_{PAM}. Figure 3.9 illustrates these two asymptotic angles, when the sample size ratio changes from 1 to 40. The blue curve shows the angle between the PAM direction and the best combination direction, θ_{PAM}. The red curve shows the angle between the DWD direction and the best combination direction, θ_{DWD}. In this figure, for any given $r > 1$, $\theta_{DWD} < \theta_{PAM}$. Thus this figure is consistent with the conclusion in Inequality (3.4).

Now, we are interested in the quantitative improvement of the DWD direction over the

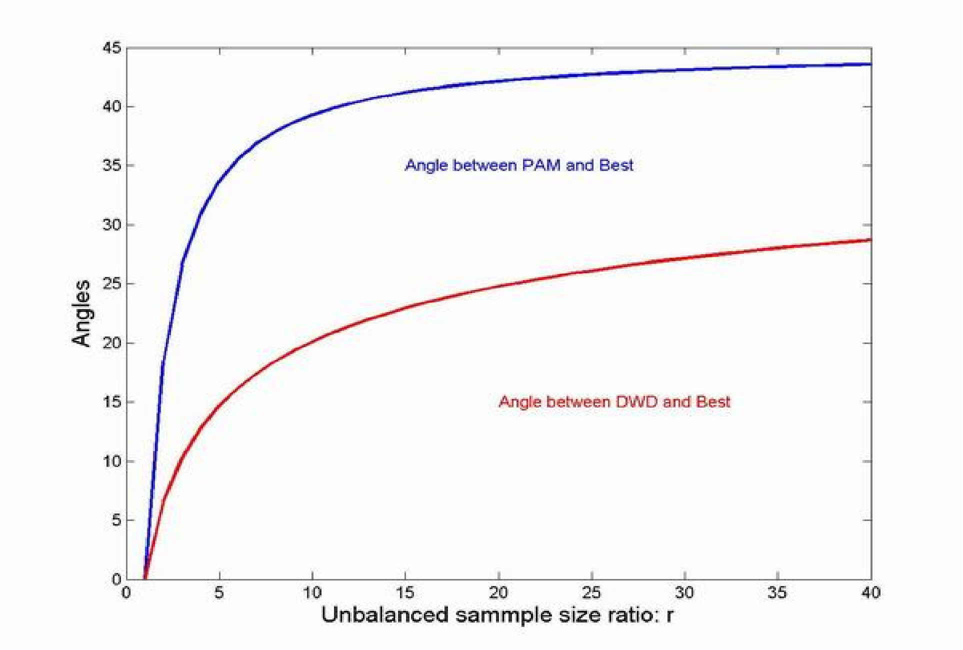


Figure 3.9: Shows the two angles: θ_{PAM} and θ_{DWD} for different choices of r, when $\alpha > \frac{1}{2}$. The DWD direction is consistently better than the PAM direction, when $r > 1$.

PAM direction. The difference between the two asymptotic angles is

$$\theta = cos^{-1}(\frac{r+1}{\sqrt{2r^2+2}}) - cos^{-1}(\frac{\sqrt[3]{r}+1}{\sqrt{2\sqrt[3]{r^2}+2}}) \tag{3.5}$$

To study the change of θ, we plot θs against the subgroup sample size ratios r. The results are shown in Figure 3.10. We chose the sample size ratio r from 1 to 40, and found that the difference between the two angles first increases, then decreases. The difference is maximized at the location, specified by the red dashed line. The exact location can be obtained by taking the derivative with respect to r on Equation (3.5), and solve the equation. This location is $r = 7.21$, at which, $\theta_{DWD} = 17.64$ degrees, $\theta_{PAM} = 37.10$ degrees. The improvement of DWD over PAM is $\theta = 19.47$ degrees. Figure 3.10 indicates that the DWD is much more robust than PAM over a large range of r. Note that the angle between the best combination and the direction which points from the center of the red pluses to the center of blue circles is 45 degress, hence, the improvement of 19.47 degrees at $r = 7.21$ is a very significant improvement.

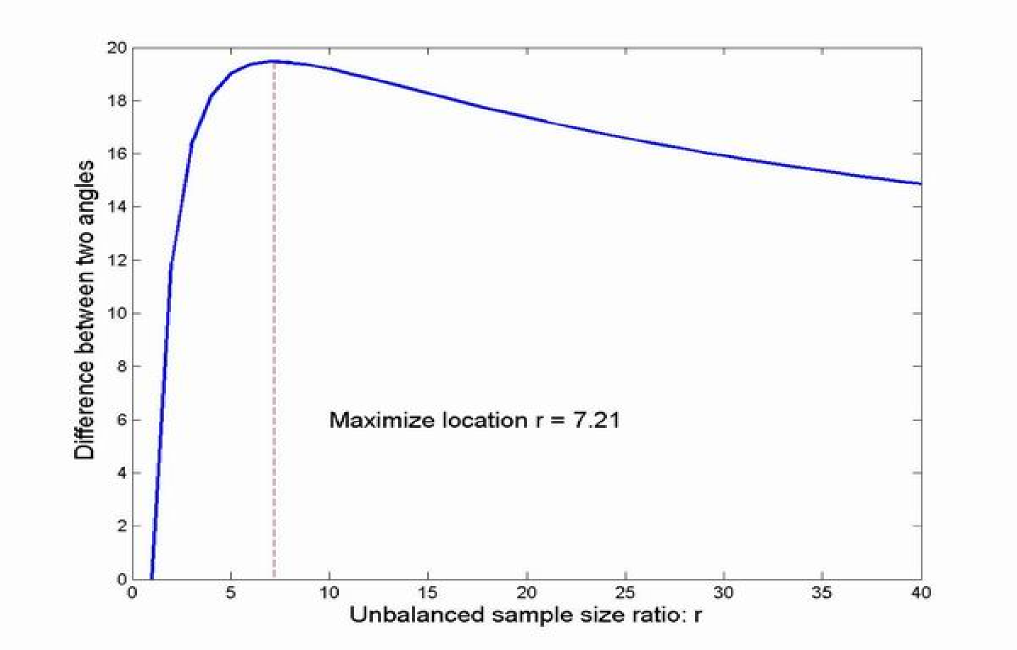


Figure 3.10: This figure shows the difference between the two asymptotic angles: the DWD direction and best combination direction, and the PAM direction and the best combination direction. The red dashed line shows the location, at which the difference is maximized.

3.2.4 Simulation Study

In order to illustrate the conclusions in Theorem 3.2.1 and in 3.2.2, we generate data sets for batch $\mathcal{X}$ and $\mathcal{Y}$ according to the unbalanced subgroup model, with the sample sizes $n = 50$, $m = 10$ and the dimension varying from $2^1, \cdots, 2^{13}$. The subgroup sample size ratio is $r = \frac{n}{m} = 5.0$. The AIPs between $\mathbf{v}_{PAM}$ and the best combining $\mathbf{v}$, $\mathbf{v}_{DWD}$ and $\mathbf{v}$ are calculated and presented in Figure 3.11.

The three plots in the first row of Figure 3.11 illustrate the results in the Theorem 3.2.1. When $\alpha < \frac{1}{2}$, the AIPs converge to 0, as shown in the top row, first column plot. When $\alpha = \frac{1}{2}$, the asymptotic properties of AIPs are unknown. When $\alpha > \frac{1}{2}$, the AIPs converge to $\frac{\sqrt[3]{r}+1}{\sqrt{2\sqrt[3]{r^2}+2}} = 0.9674$, as shown using the red line in the top row, third column plot. The corresponding asymptotic angle between the DWD direction and the best combination direction is 14.68 degrees. The three bottom plots illustrate the asymptotic properties of the AIPs between the PAM direction and the best combining direction for the same data

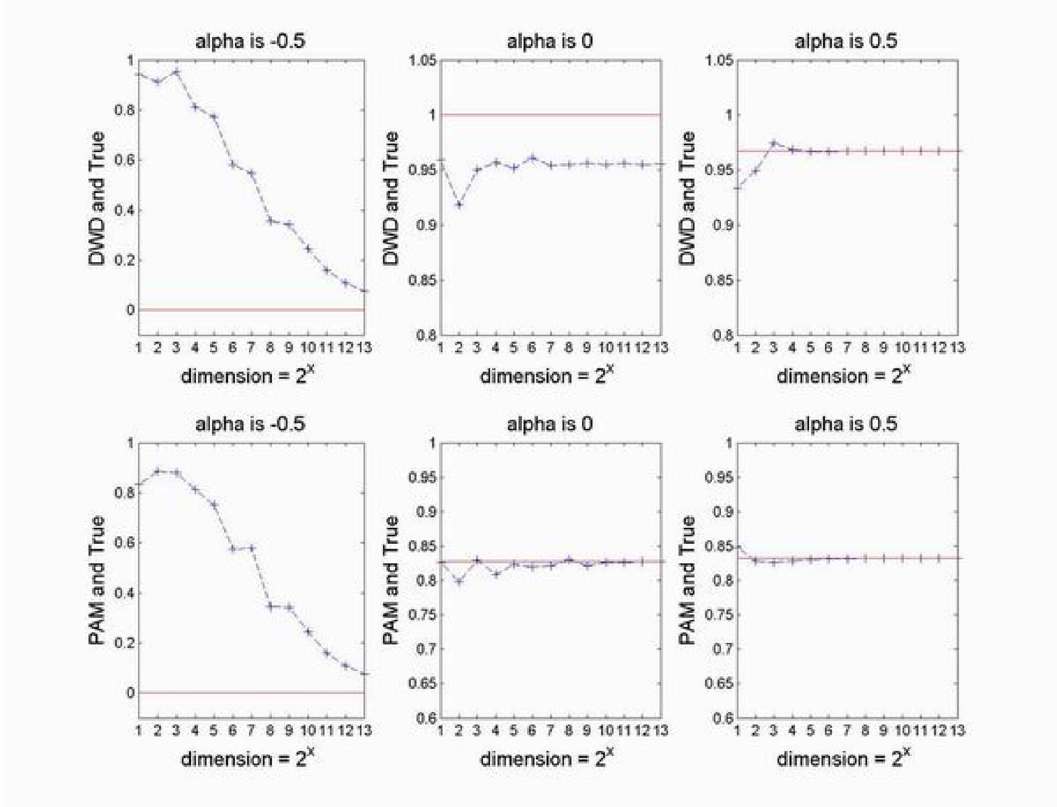


Figure 3.11: Toy example to illustrate the conclusions in Theorem 3.2.1 and 3.2.2. Each column is for a choice of α. The first row is for the AIPs between the DWD direction and the best combination direction. The second row is for the AIPs between the PAM direction and the best combination direction.

sets. As we conclude in Theorem 3.2.2, when $\alpha < \frac{1}{2}$, the AIPs converge to 0; when $\alpha = \frac{1}{2}$, the AIPs converge to a different value $\frac{r+1}{\sqrt{(2r^2+2+(1/N)(r+1)^2)}} = 0.8274$, which is represented by the red line in the bottom row, second column subplot. This corresponds to an angle of 35.56 degrees. when $\alpha > 0.5$, the AIPs converge to $\frac{r+1}{\sqrt{2r^2+2}} = 0.8321$, which is represented by the red line in the bottom row, third column subplot. The corresponding angle is 33.69 degrees, which is large than that of the DWD direction, 14.68 degrees. Three plots on the bottom verify the conclusions in Theorem 3.2.2.

In the unbalanced subgroup model, when the covariance matrix of the noise Σ is not the identity matrix, similar results as Theorem 3.2.1 and 3.2.2 can be obtained by studying the eigenvalues of Σ. The derivations are similar with those in Chapter 2, Section 2.3.3.

3.2.5 Proofs of the Theorems

In the following, we will prove Theorem 3.2.1 and Theorem 3.2.2 separately. Before we give the proofs, we first propose and prove two lemmas. The first lemma is about the DWD direction between two data separable HDLSS data.

Lemma 3.2.1. *The DWD direction* $\mathbf{v}^{(d)}$ *between two separable HDLSS data sets* $X_{d\times n}$ *and* $Y_{d\times m}$ $(d > n)$ *is always in the sample space, i.e.* $\mathbf{v}_{DWD}^{(d)} \in \mathcal{H}_{X,Y}$*.*

Proof. Let $X_{d\times n} = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_n})$ and $Y_{d\times n} = (\boldsymbol{y_1}, \cdots, \boldsymbol{y_n})$ are two separable HDLSS data sets. Using the notation in Chapter 1, Section 2.2.2, the sample space generated by the columns of X and Y is denoted as $\mathcal{H}_{X,Y}$. The orthogonal complementary of $\mathcal{H}_{X,Y}$ is denoted as $\mathcal{H}^{\perp}{}_{X,Y}$. Suppose the normalized direction vector $\mathbf{v}_{DWD}^{(d)}$ is the the DWD direction between X and Y. There exists a nonnegative constant b such that the DWD hyperplane is expressed as between X and Y is

$$\mathcal{H}_{DWD} = \{x : (\mathbf{v}^{(d)})^T x - b = 0, b \geqslant 0, \|\mathbf{v}^{(d)}\| = 1\}.$$

First of all, note that $\mathbf{v}^{(d)}$ is not in the space $\mathcal{H}^{\perp}{}_{X,Y}$. Otherwise, $(\mathbf{v}^{(d)})^T x_i = 0$ $(i = 1, \cdots, n)$, $(\mathbf{v}^{(d)})^T y_j = 0$ $(j = 1, \cdots, m)$, which means that all the observations are on the same side of the hyperplane $\{x : (\mathbf{v}^{(d)})^T x = b\}$. This contradicts the fact of the assumed separability. Next, we will prove that the direction vector $\mathbf{v}^{(d)}$ is in the sample space, i.e. $\mathbf{v}^{(d)} \in \mathcal{H}_{X,Y}$.

Suppose that the DWD direction is not in the sample space $\mathcal{H}_{X,Y}$. According to Lemma 2.2.4, $\mathbf{v}^{(d)}$ has the following orthogonal decomposition

$$\mathbf{v}^{(d)} = c_1\boldsymbol{w}_1 + c_2\boldsymbol{w}_2$$

where $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are two normed direction vectors with $\boldsymbol{w}_1 \in \mathcal{H}_{X,Y}$ and $\boldsymbol{w}_2 \in \mathcal{H}_{X,Y}{}^{\perp}$. Since $\mathbf{v}^{(d)}$ is not in $\mathcal{H}_{X,Y}$ or $\mathcal{H}^{\perp}{}_{X,Y}$, the two constants c_1 and c_2 are positive and $c_1 =$

$(\mathbf{v}^{(d)})^T \boldsymbol{w}_1 < 1$, $c_2 = (\mathbf{v}^{(d)})^T \boldsymbol{w}_2 < 1$. Consider a new classification hyperplane

$$\mathcal{H}_{New} = \{x : \boldsymbol{w}_1^T x - \frac{b}{c_1} = 0\}$$

In the following, we compare the two classification hyperplanes $\mathcal{H}_{DWD}$ and $\mathcal{H}_{New}$. For any column of X, i.e $\boldsymbol{x_i}$,

$$(\mathbf{v}^{(d)})^T \boldsymbol{x_i} - b = (c_1 \boldsymbol{w}_1^T \boldsymbol{x_i} + c_2 \boldsymbol{w}_2^T \boldsymbol{x_i}) - b$$

Since $\boldsymbol{w}_2 \in \mathcal{H}^{\perp}{}_{X,Y}$, we have $\boldsymbol{w}_2^T \boldsymbol{x_i} = 0$. Thus

$$(\mathbf{v}^{(d)})^T \boldsymbol{x_i} - b = c_1 \boldsymbol{w}_1^T \boldsymbol{x_i} - b = c_1(\boldsymbol{w}_1^T \boldsymbol{x_i} - \frac{b}{c_1}) \tag{3.6}$$

It follows that

$$(\mathbf{v}^{(d)})^T \boldsymbol{x_i} - b \leqslant 0 \Longleftrightarrow \boldsymbol{w}_1^T \boldsymbol{x_i} - \frac{b}{c_1} \leqslant 0. \ (i = 1, \cdots, n)$$

Hence $\mathcal{H}_{New}$ gives the same class memberships for the columns of X as $\mathcal{H}_{DWD}$ does. The same result holds for the columns of Y. Thus $\mathcal{H}_{New}$ and $\mathcal{H}_{DWD}$ give the same class memberships for all the columns of X and Y.

From Equation (3.6), we have

$$|(\mathbf{v}^{(d)})^T \boldsymbol{x_i} - b| = c_1 |\boldsymbol{w}_1^T \boldsymbol{x_i} - \frac{b}{c_1}| \tag{3.7}$$

Since $0 < c_1 < 1$, it follows that

$$|(\mathbf{v}^{(d)})^T \boldsymbol{x_i} - b| < |\boldsymbol{w}_1^T \boldsymbol{x_i} - \frac{b}{c_1}| \tag{3.8}$$

Note that the distance from the sample $\boldsymbol{x_i}$ to $\mathcal{H}_{DWD}$ is $|(\mathbf{v}^{(d)})^T \boldsymbol{x_i} - b|$. The distance from $\boldsymbol{x_i}$ to $\mathcal{H}_{New}$ is $|\boldsymbol{w}_1^T \boldsymbol{x_i} - \frac{b}{c_1}|$. Inequality (3.8) indicates that the distance from x_i to $\mathcal{H}_{DWD}$ is always smaller than the distance to $\mathcal{H}_{New}$. This is true for all the columns of X and Y. Recall that DWD solves the following optimization problem:

$$\begin{aligned} &minimize \quad \sum_{i=1}^{n} \frac{1}{r_i} \\ &subject\ to \quad y_i * f(\boldsymbol{x_i}) \geqslant 1, \quad i = 1, \cdots, n. \end{aligned} \tag{3.9}$$

The two hyperplanes $\mathcal{H}_{New}$ and $\mathcal{H}_{DWD}$ give the same cluster membership for any column of X and Y. The sum of inverse distances from the samples to $\mathcal{H}_{New}$ is smaller than the one from the samples to the $\mathcal{H}_{DWD}$. This contradicts the fact that the hyperplane $\mathcal{H}_{DWD}$ is the solution of the optimization problem (3.9). Thus the DWD direction is always in the sample space $\mathcal{H}_{X,Y}$.

□

Suppose $X_1^{(d)}, X_2^{(d)}, Y_1^{(d)}, Y_2^{(d)}$ are the four data sets for d genes in the unbalanced subgroup model (see Section 3.2.2). Define the combined data sets $D^{(d)} = [X_1^{(d)}, X_2^{(d)}, Y_1^{(d)}, Y_2^{(d)}]$. Recall that the sequence of the best combination directions between the batch $\mathcal{X}$ and $\mathcal{Y}$ is $\{\mathbf{v}^{(d)} = \sqrt{\frac{2}{d}}(1, 0, 1, 0, \cdots)^T,\ d = 1, 2, \cdots\}$. We say that the sequence of vectors $\{\boldsymbol{w}^{(d)},\ d = 1, 2, \cdots\}$ is in the sequence of sample space $\{\mathcal{H}_{D^{(d)}},\ d = 1, 2, \cdots\}$, if $\boldsymbol{w}^{(d)} \in \mathcal{H}_{D^{(d)}}$. Lemma 3.2.2 studies the asymptotic relations between $\{\mathbf{v}^{(d)},\ d = 1, 2, \cdots\}$ and any sequence of vectors in the sequence of sample space, when $\alpha < 1/2$.

Lemma 3.2.2. *Suppose $\boldsymbol{w}^{(d)} \in \mathcal{H}_{D^{(d)}}$ is a sequence of nonzero vectors in $\{\mathcal{H}_{D^{(d)}},\ d = 1, 2, \cdots\}$. When $\alpha < 1/2$, this sequence of vectors is asymptotically orthogonal to the best combination vectors $\{\mathbf{v}^{(d)}\ d = 1, 2, \cdots\}$, in the sense that $\frac{(\mathbf{v}^{(d)})^T \boldsymbol{w}^{(d)}}{\|\mathbf{v}^{(d)}\| \|\boldsymbol{w}^{(d)}\|} \longrightarrow 0$ in probability as $d \to \infty$.*

Proof. Any nonzero vector $\boldsymbol{w}^{(d)}$ in the sample space can be expressed as a matrix product $D^{(d)} \times C^{(d)}$, where $C^{(d)}$ is a $2N \times 1$ nonzero constant vector. The cosine of the angle between the two vectors $\boldsymbol{w}^{(d)}$ and $\mathbf{v}^{(d)}$ is

$$\frac{(\mathbf{v}^{(d)})^T \boldsymbol{w}^{(d)}}{\|\mathbf{v}^{(d)}\| \|\boldsymbol{w}^{(d)}\|} = \frac{(\mathbf{v}^{(d)})^T \boldsymbol{w}^{(d)}}{\|\boldsymbol{w}^{(d)}\|}$$

Since $\boldsymbol{w}^{(d)} = D^{(d)} \times C^{(d)}$, it follows that

$$\frac{(\mathbf{v}^{(d)})^T}{\boldsymbol{w}^{(d)}\|\boldsymbol{w}^{(d)}\|} = \frac{(\mathbf{v}^{(d)})^T D^{(d)} \times C^{(d)}}{\|D^{(d)} \times C^{(d)}\|}. \tag{3.10}$$

where $\mathbf{v}^{(d)} = \sqrt{\frac{2}{d}}(1, 0, 1, 0, \cdots)^T$. We first study the square of the denominator,

$$\|D^{(d)} \times C^{(d)}\|^2 = (C^{(d)})^T (D^{(d)})^T D^{(d)} C^{(d)}$$

Since $D^{(d)} = [X_1^{(d)}, X_2^{(d)}, Y_1^{(d)}, Y_2^{(d)}]$, it follows that

$$(D^{(d)})^T D^{(d)} = \begin{pmatrix} (X_1^{(d)})^T X_1^{(d)} & (X_1^{(d)})^T X_2^{(d)} & (X_1^{(d)})^T Y_1^{(d)} & (X_1^{(d)})^T Y_2^{(d)} \\ (X_2^{(d)})^T X_1^{(d)} & (X_2^{(d)})^T X_2^{(d)} & (X_2^{(d)})^T Y_1^{(d)} & (X_2^{(d)})^T Y_2^{(d)} \\ (Y_1^{(d)})^T X_1^{(d)} & (Y_1^{(d)})^T X_2^{(d)} & (Y_1^{(d)})^T Y_1^{(d)} & (Y_1^{(d)})^T Y_2^{(d)} \\ (Y_2^{(d)})^T X_1^{(d)} & (Y_2^{(d)})^T X_2^{(d)} & (Y_2^{(d)})^T Y_1^{(d)} & (Y_2^{(d)})^T Y_2^{(d)} \end{pmatrix}$$

Recall from Equation (3.2), $X_1^{(d)} = \mathbf{v}_{x,1}^{(d)} \times (1_n)^T + \Upsilon_{x,1}^{(d)}$. From the result in Chapter 2, Equation (2.42), we have proven that, when $\alpha < \frac{1}{2}$,

$$\frac{1}{d}(X_1^{(d)})^T X_1^{(d)} \longrightarrow I_n \ \ \textit{in probability}, \ \textit{as } d \to \infty. \tag{3.11}$$

In the same way, we can obtain that

$$\begin{aligned} \frac{1}{d}(X_2^{(d)})^T X_2^{(d)} &\longrightarrow I_m \ \ \textit{in probability}, \ \textit{as } d \to \infty, & (3.12)\\ \frac{1}{d}(Y_1^{(d)})^T Y_1^{(d)} &\longrightarrow I_m \ \ \textit{in probability}, \ \textit{as } d \to \infty, & (3.13)\\ \frac{1}{d}(Y_2^{(d)})^T Y_2^{(d)} &\longrightarrow I_n \ \ \textit{in probability}, \ \textit{as } d \to \infty. & (3.14) \end{aligned}$$

Using the law of large number, when $\alpha < 1/2$, we have the following element-wise convergence,

$$\frac{1}{d}(X_1^{(d)})^T X_2^{(d)} \longrightarrow 0_{n\times m} \ \ \textit{in probability}, \ \textit{as } d \to \infty. \tag{3.15}$$

where $0_{n\times m}$ is the $n \times m$ matrix with all entries equal to 0. In a similar way, we can show that all the off-diagonal matrices in the expression of $(D^{(d)})^T D^{(d)}$ are zeros. Hence

$$\frac{1}{d}(D^{(d)})^T D^{(d)} \longrightarrow I_{2N} \ \ \textit{in probability}, \ \textit{as} \ d \to \infty.$$

It follows that

$$\frac{1}{d}(C^{(d)})^T (D^{(d)})^T D^{(d)} C^{(d)} \longrightarrow (C^{(d)})^T C^{(d)} \ \ \textit{in probability}, \ \textit{as} \ d \to \infty. \tag{3.16}$$

Define the standardized form of $C^{(d)}$ to be $\hat{C} = C^{(d)}/\sqrt{(C^{(d)})^T C^{(d)}}$. Now we study the numerator in Equation (3.10). It follows that

$$\frac{(\mathbf{v}^{(d)})^T D^{(d)} \times C^{(d)}}{\sqrt{(C^{(d)})^T C^{(d)}}} = (\mathbf{v}^{(d)})^T D^{(d)} \times \hat{C}$$

Again, recall that the data can be expressed as in Equation (3.2). Define $\Upsilon^{(d)} = [\Upsilon^{(d)}_{x,1}, \Upsilon^{(d)}_{x,2}, \Upsilon^{(d)}_{y,1}, \Upsilon^{(d)}_{y,2}]$ and write $\hat{C} = (\hat{C}_1; \hat{C}_2; \hat{C}_3; \hat{C}_4)$, where the four vectors have dimensions $n \times 1$, $m \times 1$, $m \times 1$, $n \times 1$ respectively. It follows that

$$\begin{aligned} d^{-1/2}(\mathbf{v}^{(d)})^T D^{(d)} \times \hat{C} \ &= \ d^{-1/2}(\mathbf{v}^{(d)})^T (\mathbf{v}^{(d)}_{x,1} \times (1_n)^T)\hat{C}_1 \\ &\quad + d^{-1/2}(\mathbf{v}^{(d)})^T (\mathbf{v}^{(d)}_{x,2} \times (1_m)^T)\hat{C}_2 \\ &\quad + d^{-1/2}(\mathbf{v}^{(d)})^T (\mathbf{v}^{(d)}_{y,1} \times (1_m)^T)\hat{C}_3 \\ &\quad + d^{-1/2}(\mathbf{v}^{(d)})^T (\mathbf{v}^{(d)}_{y,1} \times (1_n)^T)\hat{C}_4 \\ &\quad + d^{-1/2}(\mathbf{v}^{(d)})^T \Upsilon^{(d)} \hat{C} \end{aligned} \tag{3.17}$$

When $\alpha < 1/2$,

$$\begin{aligned} d^{-1/2}(\mathbf{v}^{(d)})^T (\mathbf{v}^{(d)}_{x,1} \times (1_n)^T)\hat{C}_1 \ &= \ 2\sqrt{2} d^{\alpha - 1/2} \sum_{i=1}^{n} \hat{C}_1(i) \\ &\longrightarrow \ 0 \end{aligned} \tag{3.18}$$

The same results hold for the other three terms in Equation (3.17). The last term in

Equation (3.17),

$$d^{-1/2}(\mathbf{v}^{(d)})^T\Upsilon^{(d)}\hat{C}_{2N} = \sqrt{2}d^{-1}\sum_{k=1}^{d/2}(\sum_{j=1}^{2N}\epsilon_{2k+1,j}\hat{C}(j)) \tag{3.19}$$

Because $\sum_{j=1}^{2N}\epsilon_{2k+1,j}\hat{C}(j)$ follows the standard Gaussian distribution, according to the law of large number, we have

$$d^{-1/2}(\mathbf{v}^{(d)})^T\Upsilon^{(d)}\hat{C}_{2N} \longrightarrow 0 \ \ \textit{in probability}, \ \textit{as } d \to \infty. \tag{3.20}$$

Summaizing Equations (3.17) to (3.20), we have

$$d^{-1/2}(\mathbf{v}^{(d)})^T D^{(d)} \times \hat{C}_{2N} \longrightarrow 0 \ \ \textit{in probability}, \ \textit{as } d \to \infty. \tag{3.21}$$

From Equations (3.10), (3.16) and (3.21), we finally get

$$\frac{(\mathbf{v}^{(d)})^T\boldsymbol{w}^{(d)}}{\|\boldsymbol{w}^{(d)}\|} \longrightarrow 0 \ \ \textit{in probability}, \ \textit{as } d \to \infty.$$

Hence, we have proven Lemma 3.2.2

□

Now, we prove Theorems 3.2.1 and 3.2.2 separately.

Proof of Theorem 3.2.1

The proof is organized as two parts, each of which proves one conclusion in Theorem 3.2.1.

- **The case when $\alpha > \frac{1}{2}$.**

  Consider the ith columns of $X_1^{(d)}$, $\mathbf{x}_{1,i}^{(d)} = \mathbf{v}_{x,1}^{(d)} + \epsilon_{x,1,i}^{(d)}$ $(i = 1, \cdots, n)$, when $d \to \infty$, after scaling by $d^{-\alpha}$, we have

$$d^{-2\alpha}(\mathbf{x}_{1,i}^{(d)})^T\mathbf{x}_{1,i}^{(d)} = d^{-2\alpha}((\mathbf{v}_{x,1}^{(d)})^T\mathbf{v}_{x,1}^{(d)} + 2d^{-2\alpha}(\mathbf{v}_{x,1}^{(d)})^T\epsilon_{x,1,i}^{(d)} + d^{-2\alpha}(\epsilon_{x,1,i}^{(d)})^T\epsilon_{x,1,i}^{(d)} \tag{3.22}$$

When $\alpha > 1/2$, the first term in Equation (3.22) is

$$\begin{aligned} d^{-2\alpha}(\mathbf{v}_{x,1}^{(d)})^T \mathbf{v}_{x,1}^{(d)} &= (\frac{1}{d^\alpha}\|\mathbf{v}_{x,1}^{(d)}\|)^2 \\ &\longrightarrow 1 \end{aligned} \tag{3.23}$$

According to the law of large number, the second term in Equation (3.22) converges to zero, and the third term converges to 0 in probability. Hence,

$$d^{-2\alpha}(\mathbf{x}_{1,i}^{(d)})^T \mathbf{x}_{1,i}^{(d)} \longrightarrow 1.$$

This means that after scaling by a constant $d^{-\alpha}$, the distance from each point in $X_1^{(d)}$ to the origin is 1. Similar results hold for all the samples in batches $\mathcal{X}$ and $\mathcal{Y}$. Again, using the law of large number, we can show that the distance from each sample to it's cluster mean vector satisfies

$$d^{-\alpha}\|\mathbf{x}_{1,i}^{(d)}) - \mathbf{v}_{x,1}^{(d)}\| \longrightarrow 0.$$

It is straightforward to see that any column vector from $X_1^{(d)}$ is asymptotically orthogonal to the one from $Y_1^{(d)}$, in the sense that

$$d^{-2\alpha}(\mathbf{x}_{1,i}^{(d)})^T \mathbf{y}_{1,j}^{(d)} \longrightarrow 0.$$

These results indicate that when $\alpha > 1/2$, the mean vectors dominate the measurement noise in the data. The asymptotic geometric structure of the data is the same as that of the data without any measurement noise. Figure 3.12 shows the asymptotic geometric structure of the data. When $\alpha > 1/2$ and $d \to \infty$, after scaling by a constant $d^{-\alpha}$, all the column vectors in $X_1^{(d)}, X_2^{(d)}, Y_1^{(d)}, Y_2^{(d)}$ converge to their cluster mean vectors V_{x1}, V_{x2}, V_{y1} and V_{y1} respectively as in the Figure, where

$$V_{x1} = d^{-\frac{1}{2}}(-1, \ \ 1, -1, \ \ 1, \cdots)^T,$$

$$V_{x2} = d^{-\frac{1}{2}}(-1,-1,-1,-1,\cdots)^T,$$

$$V_{y1} = d^{-\frac{1}{2}}(\ \ 1,\ \ 1,\ \ 1,\ \ 1,\cdots)^T,$$

$$V_{y2} = d^{-\frac{1}{2}}(\ \ 1,-1,\ \ 1,-1\cdots)^T.$$

The within cluster variation converges to 0. Note that there are n samples located at V_{x1}, V_{y2} and m samples located at V_{x2}, V_{y1}.

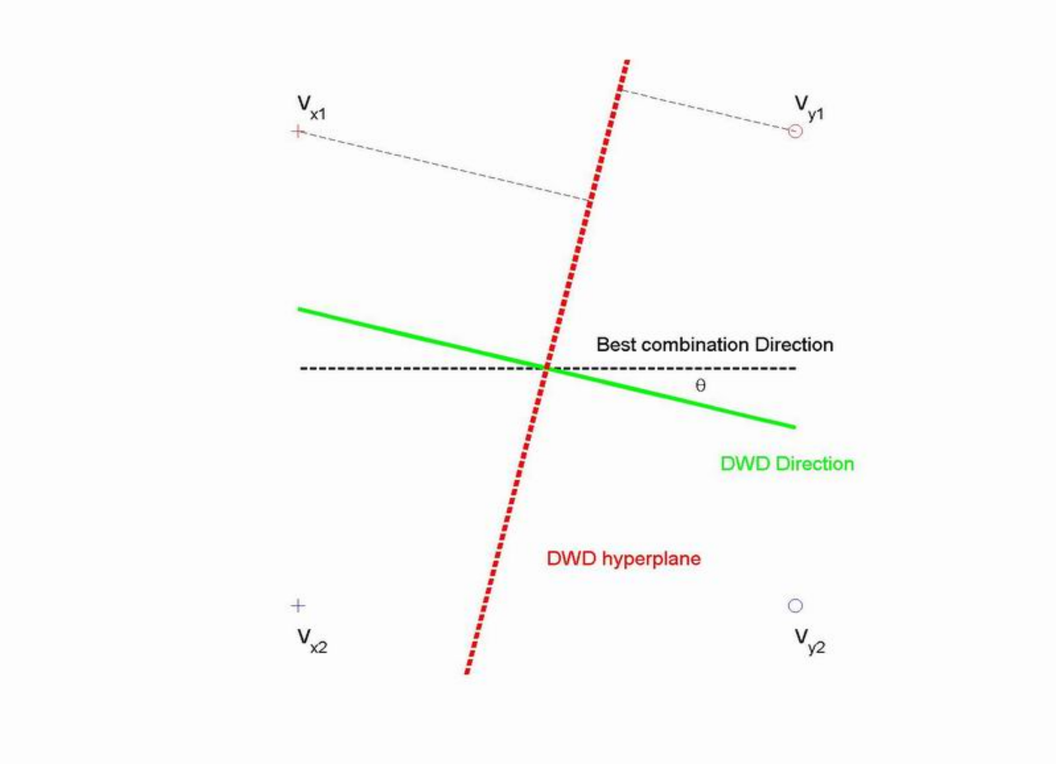


Figure 3.12: Shows the asymptotic geometric structure of the data in the unbalanced subgroup model, when $\alpha > 1/2$ and $d \to \infty$.

In Figure 3.12, the black dashed line shows the best combination direction. Shifting the data along this direction will combine all the samples from the same biological subtype together. Suppose that the DWD direction, shown as the green line, has an angle of θ to the best combination direction. The DWD hyperplane is shown using a red dashed line in the figure. The distance from V_{x1} to the hyperplane can be calculated as $sin(\frac{\pi}{4} + \theta)$. The distance from V_{x2} to the hyperplane can be calculated as $sin(\frac{\pi}{4} - \theta)$. According to the definition of the DWD hyperplane, it finds θ to minimize

$$D(\theta) = \frac{2n}{sin(\frac{\pi}{4}+\theta)} + \frac{2m}{sin(\frac{\pi}{4}-\theta)}$$

Note that $r = \frac{n}{m}$. Solving the equation $\frac{\partial D(\theta)}{\partial \theta} = 0$, we have

$$\hat{\theta} = cos^{-1}(\frac{\sqrt[3]{r}+1}{\sqrt{2\sqrt[3]{r^2}+2}}).$$

Hence, the AIP between the DWD direction and the best combination direction is

$$cos(\hat{\theta}) = \frac{\sqrt[3]{r}+1}{\sqrt{2\sqrt[3]{r^2}+2}}).$$

This proves the first conclusion in Theorem 3.2.1.

- **The case when $\alpha < \frac{1}{2}$.**

  In the unbalanced subgroup model, the four HDLSS data sets are grouped into two batches, which are separable. According to Lemma 3.2.1, the DWD direction $\mathbf{v}_{DWD}^{(d)}$ is in the sample space of the data with d genes. When $\alpha < 1/2$, according the Lemma 3.2.2, the DWD direction is asymptotically orthogonal to the best combination direction $\mathbf{v}^{(d)}$. This proves the second conclusion in Theorem 3.2.1.

Comments on the asymptotics of SVM

As we have shown in Chapter 3, Section 3.1, SVM has a serious data piling problem for HDLSS data. Figure 3.3 shows the projection view of the data on the SVM direction. The data piling problem is quite serious, although the dimension $d = 50$ is not significantly larger than the sample size 20. The batch adjustment using the SVM direction will lead to unsuccessful combination because of the data piling. Thus in Section 3.1, we only compare the asymptotic properties of DWD and PAM.

The asymptotic properties of the SVM direction can be studied similarly as for the DWD direction. In the linear shift model, when $\alpha < 1/2$, because the SVM direction is in the sample spanned space, as is the DWD direction, according to Lemma 3.2.2, the SVM direction is asymptotically orthogonal to the theoretical best combination direction. When $\alpha > 1/2$, as shown in the proofs of Theorem 3.2.1, the four subgroups have a simple asymptotic geometric structure, shown in Figure 3.12. In this figure, all the samples are on the margins, hence the SVM direction is asymptotically the same as the best combination direction. Thus, the SVM direction is more robust than the DWD direction and the PAM direction, when the dimension goes to infinity. However, this does not indicate that we should use the SVM direction for the batch adjustment in real data analysis. Because when the dimension is not very high (around 50), the data piling problem has negative influence on the batch adjustment; when the dimension goes to infinity, the data intrinsically have "the extreme piling" geometric structure as in Figure 3.12, thus the data piling problem will not have a negative influence on the batch adjustment any more.

Proof of Theorem 3.2.2

The proofs are organized into three parts.

- **The case when $\alpha > 1/2$.** As we have concluded in the proofs of Theorem 3.2.1, Figure 3.12 shows the asymptotic geometric structure of the four clusters, when $\alpha > 1/2$. Since the PAM direction is the one which connects the two centers of batches, it follows that

$$\mathbf{v}_{PAM}^{(d)} = \frac{(nV_{x1} + mV_{x2}) - (mV_{y1} + nV_{y2})}{\|(nV_{x1} + mV_{x2}) - (mV_{y1} + nV_{y2})\|} = \frac{1}{\sqrt{d(r^2+1)}}(r+1, r-1, r+1, r-1, \cdots)^T$$

Thus the AIP between the PAM direction and the best combination direction $\mathbf{v}^{(d)} = \sqrt{\frac{2}{d}}(1, 0, 1, 0, \cdots)^T$ is

$$\sqrt{\frac{2}{d}}(1, 0, 1, 0, \cdots) \times \frac{1}{\sqrt{d(r^2+1)}}(r+1, r-1, r+1, r-1, \cdots)^T = \frac{r+1}{\sqrt{2r^2+2}}.$$

This is the first conclusion in Theorem 3.2.2.

- **The case when $\alpha = 1/2$.** When the number of genes is d, recall the matrix expressions of the data in Equation (3.2),

$$\begin{aligned}
X_1^{(d)} &= \mathbf{v}_{x,1}^{(d)} \times (1_n)^T + \Upsilon_{x,1}^{(d)} \\
X_2^{(d)} &= \mathbf{v}_{x,2}^{(d)} \times (1_m)^T + \Upsilon_{x,2}^{(d)} \\
Y_1^{(d)} &= \mathbf{v}_{y,1}^{(d)} \times (1_m)^T + \Upsilon_{y,1}^{(d)} \\
Y_2^{(d)} &= \mathbf{v}_{y,2}^{(d)} \times (1_n)^T + \Upsilon_{y,2}^{(d)}
\end{aligned}$$

When $\alpha = 1/2$, we have

$$\mathbf{v}_{x,1}^{(d)} = (-1, \;\; 1, -1, \;\; 1, \cdots)^T,$$

$$\mathbf{v}_{x,2}^{(d)} = (-1, -1, -1, -1, \cdots)^T,$$

$$\mathbf{v}_{y,1}^{(d)} = (\;\; 1, \;\; 1, \;\; 1, \;\; 1, \cdots)^T,$$

$$\mathbf{v}_{y,2}^{(d)} = (\;\; 1, -1, \;\; 1, -1 \cdots)^T.$$

The center of the batch $\mathcal{X}$ is

$$\begin{aligned}
L_1 &= \frac{1}{m+n}(X_1^{(d)} \times 1_n + X_2^{(d)} \times 1_m) \\
&= \frac{1}{m+n}((n\mathbf{v}_{x,1}^{(d)} + \Upsilon_{x,1}^{(d)} \times 1_n) + (m\mathbf{v}_{x,2}^{(d)} + \Upsilon_{x,2}^{(d)} \times 1_m))
\end{aligned}$$

In the same way, we get the center of the batch $\mathcal{Y}$ is

$$\begin{aligned}
L_2 &= \frac{1}{m+n}(Y_1^{(d)} \times 1_m + Y_2^{(d)} \times 1_n) \\
&= \frac{1}{m+n}((m\mathbf{v}_{y,1}^{(d)} + \Upsilon_{y,1}^{(d)} \times 1_m) + (n\mathbf{v}_{y,2}^{(d)} + \Upsilon_{y,2}^{(d)} \times 1_n))
\end{aligned}$$

The PAM direction is the direction vector which connects two centers, thus

$$\mathbf{v}_{PAM}^{(d)} = \frac{L_2 - L_1}{\|L_2 - L_1\|} \tag{3.24}$$

We first look at the numerator

$$\begin{aligned} L_2 - L_1 &= \frac{1}{m+n}(n(\mathbf{v}_{y,2}^{(d)} - \mathbf{v}_{x,1}^{(d)}) + m(\mathbf{v}_{y,1}^{(d)} - \mathbf{v}_{x,2}^{(d)})) \\ &+ \frac{1}{m+n}(\Upsilon_{y,2}^{(d)} \times 1_n + \Upsilon_{y,1}^{(d)} \times 1_m - \Upsilon_{x,1}^{(d)} \times 1_n - \Upsilon_{x,2}^{(d)}) \\ &\equiv A + B, \end{aligned}$$

where

$$\begin{aligned} A &= \frac{1}{m+n}(n(\mathbf{v}_{y,2}^{(d)} - \mathbf{v}_{x,1}^{(d)}) + m(\mathbf{v}_{y,1}^{(d)} - \mathbf{v}_{x,2}^{(d)})), \\ &= \frac{2}{m+n}(n\mathbf{v}_{y,2}^{(d)} + m\mathbf{v}_{y,1}^{(d)}) \\ B &= \frac{1}{m+n}(\Upsilon_{y,2}^{(d)} \times 1_n + \Upsilon_{y,1}^{(d)} \times 1_m - \Upsilon_{x,1}^{(d)} \times 1_n - \Upsilon_{x,2}^{(d)} 1_m) \end{aligned}$$

The inner product between the PAM direction and the best combination direction is

$$\begin{aligned} (\mathbf{v}^{(d)})^T \mathbf{v}_{PAM}^{(d)} &= \frac{(\mathbf{v}^{(d)})^T (L_2 - L1)}{\|L_2 - L_1\|} \\ &= \frac{(\mathbf{v}^{(d)})^T A + (\mathbf{v}^{(d)})^T B}{\|A + B\|} \end{aligned} \tag{3.25}$$

Next we study the asymptotic properties of the left side term in Equation (3.25). Firstly,

$$\begin{aligned} d^{-1/2} (\mathbf{v}^{(d)})^T A &= d^{-1/2} (\mathbf{v}^{(d)})^T \times \frac{2}{m+n}(n\mathbf{v}_{y,2}^{(d)} + m\mathbf{v}_{y,1}^{(d)}) \\ &= \sqrt{2}. \end{aligned} \tag{3.26}$$

Secondly,

$$\begin{aligned} d^{-1/2}(\mathbf{v}^{(d)})^T B &= d^{-1/2}(\mathbf{v}^{(d)})^T \times \Upsilon_{y,2}^{(d)} \times 1_n + d^{-1/2}(\mathbf{v}^{(d)})^T \times \Upsilon_{y,1}^{(d)} \times 1_m \\ &\quad -d^{-1/2}(\mathbf{v}^{(d)})^T \times \Upsilon_{x,1}^{(d)} \times 1_n - d^{-1/2}(\mathbf{v}^{(d)})^T \times \Upsilon_{x,2}^{(d)} \end{aligned} \tag{3.27}$$

Using the law of large number, we show that $d^{-1/2}(\mathbf{v}^{(d)})^T \times \Upsilon_{y,2}^{(d)} \times 1_n \longrightarrow 0$ as $d \to \infty$. The other three terms in the right side of Equation (3.27) also have the same properties. Thus,

$$d^{-1/2}(\mathbf{v}^{(d)})^T B \longrightarrow 0. \tag{3.28}$$

The square of the denominator in Equation (3.25) is

$$\|A + B\|^2 = A^T A + B^T B + 2A^T B. \tag{3.29}$$

It follows that

$$\begin{aligned} d^{-1} A^T A &= d^{-1}\left(\frac{2}{m+n}\right)^2 (n\mathbf{v}_{y,2}^{(d)} + m\mathbf{v}_{y,1}^{(d)})^T (n\mathbf{v}_{y,2}^{(d)} + m\mathbf{v}_{y,1}^{(d)}) \\ &= \frac{4(n^2+m^2)}{(m+n)^2}. \end{aligned} \tag{3.30}$$

$$\begin{aligned} d^{-1} A^T B &= d^{-1} \frac{2}{m+n} (n\mathbf{v}_{y,2}^{(d)} + m\mathbf{v}_{y,1}^{(d)})^T \\ &\quad \times \frac{1}{m+n} (\Upsilon_{y,2}^{(d)} \times 1_n + \Upsilon_{y,1}^{(d)} \times 1_m - \Upsilon_{x,1}^{(d)} \times 1_n - \Upsilon_{x,2}^{(d)}) \\ &\longrightarrow 0 \end{aligned} \tag{3.31}$$

The asymptotics in Equation (3.31) follow because $d^{-1}(\mathbf{v}_{y,2}^{(d)})^T \Upsilon_{y,2}^{(d)} \times 1_n \longrightarrow 0$, as $d \to \infty$, using the law of large number. All other interaction terms in Equation (3.31) have the same properties. Thus,

$$d^{-1} B^T B = d^{-1}\left(\frac{1}{m+n}\right)^2 (\Upsilon_{y,2}^{(d)} \times 1_n + \Upsilon_{y,1}^{(d)} \times 1_m - \Upsilon_{x,1}^{(d)} \times 1_n - \Upsilon_{x,2}^{(d)})^T$$

$$\times(\Upsilon_{y,2}^{(d)} \times 1_n + \Upsilon_{y,1}^{(d)} \times 1_m - \Upsilon_{x,1}^{(d)} \times 1_n - \Upsilon_{x,2}^{(d)}) \tag{3.32}$$

According to the law of large number,

$$d^{-1}((\Upsilon_{y,2}^{(d)} \times 1_n)^T(\Upsilon_{y,2}^{(d)} \times 1_n) \quad \longrightarrow \quad n, \ as \ d \to \infty. \tag{3.33}$$

Because the columns of $\Upsilon_{y,2}^{(d)}$ and the columns of $\Upsilon_{y,1}^{(d)}$ are independent, again, according to the law of large number, we have

$$d^{-1}((\Upsilon_{y,2}^{(d)} \times 1_n)^T(\Upsilon_{y,1}^{(d)} \times 1_m) \quad \longrightarrow \quad 0, \ as \ d \to \infty. \tag{3.34}$$

Similar results hold for the other terms on the right side of Equation (3.32). Because of Equation (3.32) and results in (3.33) and (3.34), we have

$$d^{-1}B^TB \quad \longrightarrow \quad \frac{2}{m+n} \ as \ d \to \infty \tag{3.35}$$

From the results in (3.29), (3.30), (3.31) and (3.35), we obtain

$$d^{-1}\|A+B\|^2 \quad \longrightarrow \quad \frac{4(n^2+m^2)}{(m+n)^2} + \frac{2}{m+n} = \frac{4(r^2+1)}{(r+1)^2} + \frac{2}{N}. \tag{3.36}$$

Combining the results in (3.25), (3.26), (3.28), and (3.36), it follows that

$$(\mathbf{v}^{(d)})^T\mathbf{v}_{PAM}^{(d)} \longrightarrow \frac{\sqrt{2}}{\sqrt{(\frac{4(r^2+1)}{(r+1)^2} + \frac{2}{N}}} = \frac{r+1}{\sqrt{2r^2+2+\frac{1}{N}(r+1)^2}} \tag{3.37}$$

Hence, we have proven the second conclusion in Theorem 3.2.2

- **The case when** $\alpha < 1/2$ When the number of genes is d, recall that $D^{(d)} = [X_1^{(d)}, X_2^{(d)}, Y_1^{(d)}, Y_2^{(d)}]$. The PAM direction is the direction vector which connects the centers of two batches. Hence

$$\mathbf{v}_{PAM}^{(d)} = \frac{D^{(d)}C}{\|D^{(d)}C\|},$$

where $C = (\underbrace{1\cdots,1}_{N},\underbrace{-1\cdots,-1}_{N})$.
According to Lemma 3.2.2, the PAM direction is asymptotically orthogonal to the best combination direction. Hence we have proven the third conclusion in Theorm 3.2.2

BIBLIOGRAPHY

Ahn J. and Marron J. (2006). The direction of maximal data piling in high dimensional space. *submitted* .

Ahn J., Marron J., Muller K.E. and Chi Y.Y. (2005). The high dimension, low sample size geometric representation holds under mild conditions. *Submitted* .

Alter O., Brown P. and Botstein D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natn. Acad. Sci. USA* .

Bai Z., Silverstein J.W. and Yin Y. (1988). A note on the largest eignevalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis* .

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

Brazma A., Parkinson H., Schlitt T. and Shojatalab M. (2004). A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays URL www.ebi.ac.uk/microarray/biology-intro.html .

Burges C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining Knowl. Disc* .

Duggan D., Bittner D., Chen Y., Meltzer P. and Trent J. (1999). Expression profiling using cdna microarrays. *Nat. Genet.* **21 (1 Suppl)**, 10–14.

Eisen M. and Brown P. (1999). Dna arrays for analysis of gene expression. *Meth. Enzym.* pp. 179–205.

Fujikoshi Y. (2004). Multivariate analysis for the case when the dimension is large copared to the sample size. *Journal of the Korean Statistical Society* .

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

Hastie T., Tibshirani R. and Friedman J. (2001). The elements of statistical learning: Data mining, inference, and prediction .

Irizarry R., Hobbs B., Collin F., Beazer-Barclay Y., Antonellis K., Scherf U. and Speed T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* pp. 249–264.

Johnson W.E., Rabinovic A. and Li C. (2006). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* .

Johnstone I. (2001). On the distribtuion of the largest eigenvalue in principal component analysis. *Annals of Statistics* .

Jolliffe I. (2002). Principal component analysis. *Springer, second Edition* .

Kuo W.P., Jenssen T.K., Butte A.J., Ohno-Machado L. and Kohane I.S. (2002). Analysis of matched mrna measurements from two different microarray technologies. *Bioinformatics* .

Liu X. (2007a). Additional graphics dwd unbalanced su-sample robust URL http://genome.med.unc.edu:8080/caBIG/DWDsubSample/DWDsubSample.htm .

Liu X. (2007b). Additional materials for the dissertation URL www.unc.edu/liux/phd .

Liu X., Parker J., Fan C., MP C. and Marron J. (2007). Visualization of cross platform microarray normalization. *In preparation* .

Marron J. and Liu X. (2005). Website for dwd adjustment URL http://genome.med.unc.edu:8080/caBIG/DWDWebPage/DWDindex.htm .

Marron J. and Todd M. (2002). Distance weighted discrimination. *In revise* .

Marron J., Todd M. and Ahn J. (2005). Distance weighted discrimination. *In revise* .

Nielsen T., West R., Alter O. and Knowling M.e.a. (2002). Molecular characterization of soft tissue tumours: a gene expression study. *Lancet* .

Schena M., Shalon D., Davis R. and Brown P. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science* pp. 467–70.

Silverstein J.W. (1989). On the weak limit of the largest eigenvalues of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis* .

Tibshirani R., Hastie T., Narasimhan B. and Chu G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6567–6572.

Toh K.C., Tutuncu R.H. and Todd M.J. (2006). Sdpt3 URL www.math.nus.edu.sg/mattohkc/sdpt3.html .

Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D. and Altman R. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics* .

Tukey J. and Tukey P. (1990). Strips displaying empirical distributions: I. textured dot strips. *Bellcore Technical Memorandum* .

Vapnik V. (1982). Estimation of dependences based on empirical data. *Berlin:Springer* .

Vapnik V. (1995). The nature of statistical learning theory. *New York:Springer* .

Yauk C.L., Lynn B.M., Williams A. and Douglas G.R. (2004). Comprehensive comparision of six microarray technologies. *Nucleic Acids Research* .