2014

Some recent advances in multivariate statistics: modality inference and statistical monitoring of clinical trials with multiple co-primary endpoints

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

SOME RECENT ADVANCES IN MULTIVARIATE STATISTICS: MODALITY INFERENCE AND STATISTICAL MONITORING OF CLINICAL TRIALS WITH MULTIPLE CO-PRIMARY ENDPOINTS

by

YANSONG CHENG

M.A., Boston University, 2013
B.Sc., Tianjin University of Finance and Economics, 2008

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2014

Approved by

First Reader

Surajit Ray, PhD
Adjunct Assistant Professor of Statistics

Second Reader

Ashis Gangopadhyay, PhD
Associate Professor of Statistics

Third Reader

Mark Chang, PhD
Adjunct Professor of Biostatistics

Acknowledgments

First, I would like to thank my advisor, Dr. Surajit Ray, for his insightful guidance and patience in the past years of my graduate study. I enjoyed this painful but great journey.

I appreciate all other committee members, Dr. Ashis Gangopadhyay, Dr. Mark Chang, Dr. Joe Massaro and Dr. Serkalem Demissie for their efforts of reviewing my dissertation and for those valuable comments and suggestions.

Many thanks to my classmates and friends at Boston University. They made my life at BU more enjoyable.

I am grateful to my parents, Min and Baozhen, for their consistent love, care and encouragement through all the years of my education. They always tried their best to create the best environment for me.

Finally, I want to express my special thanks to my wife, Chen, for her endless support and for those many weekends that we spent at the library together. Without her understanding and support, this dissertation would be impossible. I also want to dedicate this work to our first coming baby, who gave me the strength to finish this dissertation.

SOME RECENT ADVANCES IN MULTIVARIATE STATISTICS: MODALITY INFERENCE AND STATISTICAL MONITORING OF CLINICAL TRIALS WITH MULTIPLE CO-PRIMARY ENDPOINTS

(Order No.)

YANSONG CHENG

Boston University, Graduate School of Arts and Sciences, 2014

Major Professor: Surajit Ray, Adjunct Assistant Professor of Statistics

ABSTRACT

This dissertation focuses on two topics in multivariate statistics. The first part develops an inference procedure and fast computation tool for the *modal clustering* method proposed by Li et al. (2007). The modal clustering, based on the kernel density estimate, clusters data using their associations within a single mode, with the final number of clusters equaling the number of modes, otherwise known as the *modality* of the distribution of the data. This method provides a flexible tool for clustering data of low to moderate dimensions with arbitrary distributional shapes. In contrast to Li and colleagues, we expand their method by proposing a procedure that determines the number of clusters in the data. A test statistic and its asymptotic distribution are derived to assess the significance of each mode within the data. The inference procedure is tested on both simulated and real data sets. In addition, an R computing package is developed (*Modalclust*) that implements the modal clustering procedure using parallel processing which dramatically increases computing speed over the previously available method. This package is available on the Comprehensive R Archive Network (CRAN).

The second part of this dissertation develops methods of statistical monitoring of clinical trials with multiple co-primary endpoints, where success is defined as meeting both endpoints simultaneously. In practice, a group sequential design method is used to stop trials early for promising

efficacy, and conditional power is used for futility stopping rules. In this dissertation we show that stopping boundaries for the group sequential design with multiple co-primary endpoints should be the same as those for studies with single endpoints. Lan and Wittes (1988) proposed the B-value tool to calculate the conditional power of single endpoint trials and we extend this tool to calculate the conditional power for studies with multiple co-primary endpoints. We consider the cases of two-arm studies with co-primary normal and binary endpoints and provide several examples of implementation with simulated trials. A fixed-weight sample size re-estimation approach based on conditional power is introduced. Finally, we discuss the possibility of blinded interim analyses for multiple endpoints using the modality inference method introduced in the first part.

Contents

List of Tables

List of Figures

List of Abbreviations

AMISE	asymptotic mean integrated squared error
CP	conditional power
CRAN	comprehensive R archive network
EM ...	expectation maximization
GSD ..	group sequential design
HMAC	hierarchical mode association clustering
IUT ...	intersection-union test
KDE ..	kernel density estimate
LSE ...	least square estimate
MAC .	mode association clustering
MEM .	modal expectation maximization
MISE .	mean integrated squared error
MSE ..	mean squared error
REM ..	ridgeline expectation maximization
SSR ...	sample size re-estimation

Chapter 1

Introduction

This dissertation focuses on two important topics in multivariate statistics. The first part develops an inference procedure and fast computational tool for the *modal clustering* method proposed by Li et al. (2007). The modal clustering, based on the Kernel Density Estimate (KDE), clusters the data using their associations within a single mode. Therefore, it is also referred to as Mode Association Clustering (MAC). The final number of clusters is equal to the number of modes, known as the *Modality* of the distribution of the data. In detail, Li et al. (2007) proposed the Modal Expectation Maximization (MEM) algorithm, which makes each data point converge to the local maximum (mode) of the KDE. The points that converge to the same mode form a single cluster. The details of the MEM will be reviewed in Section 2.2. MAC provides a flexible nonparametric clustering algorithm since there is no assumption on the distribution of the data. It works well for the data with an arbitrary distributional shape. However, due to the curse of the dimensionality of the KDE, the current version of MAC is limited to the low to moderate dimensions.

Ray and Lindsay (2005) introduced the *ridgeline* concept of a mixture of the multivariate normal distributions. It is useful to understand the geometric feature of the probability density of a mixture distribution, especially for the mixture of two normal components. For the mixture of two homogeneous normal components, the density is bimodal if the Mahalanobis distance between the two mean vectors is greater than 4 (More details are in Section 8.1). In particular, for the data with two dimensions, the density of the data can be considered as a mountain. The ridgeline connects the two modes along the ridge of the mountain and passes through the saddle point, which is the point with the lowest density, between the two modes. Li et al. (2007) provides the Ridgeline Expectation

Maximization (REM) algorithm based on the KDE to carry out the ridgeline between the two modes which are identified by the MEM.

In this dissertation, in contrast to Li et al. (2007), we expand their method by proposing an inferential framework that determines the number of clusters in the data clustered by MAC. In order to do this, we assess the significance of each pair of modes within the data. We propose our inferential framework based on the fact that, if the selected pair of modes are significant, the valley between them will be deep. In other words, the lower density of the two modes should be significantly higher than the density of the saddle point. Based on this, a test statistic is proposed to test the difference of the densities of the mode $\mathbf{x_m}$ and the saddle point $\mathbf{x_s}$. Consequently, the asymptotic distribution of the test statistic is derived based on the asymptotic properties of the KDE. In order to use the asymptotic normality of the KDE, the choice of the bandwidth parameter is important and is carefully chosen. The bandwidth selection involves two steps. The first step involves detecting the modes by MAC. The second step deals with the inference. The inference procedure is applied on some simulated and real data sets.

Based on the fact that the larger bandwidth parameter produces a smoother KDE, i.e., fewer modes/clusters, the MAC algorithm is naturally extended to its hierarchical form (HMAC) by using a series of ascending bandwidth parameters. However, the MAC approach is computationally expensive when the number of objects n is large. It requires that we apply the MEM on each data point to find the local maximum of the density. Note that for HMAC, from the second level onwards, we only need to apply the MAC for the modes of the previous level, and hence the computational cost does not increase at the rate of n. We propose a "divide and conquer" algorithm of clustering by randomly partitioning the data into m partitions and performing the modal clustering on each of those partitions. Then we pool the modes obtained from each of these partitions together to form the set of modes G and apply the HMAC onward. If the user has access to multiple computing cores on the same machine or several processors of a shared memory computing cluster, the divide and conquer algorithm can be seamlessly parallelized. In this dissertation we propose an algorithm to

parallelize the HMAC (PHMAC) and provide comparisons of the performance of the parallel and non-parallel computing approach. The R package *Modalclust* is created to implement the parallel computing algorithm and is available on the Comprehensive R Archive Network (CRAN).

The second part of the dissertation develops the statistical monitoring methods of clinical trials with multiple co-primary endpoints, where success is defined as meeting both endpoints simultaneously. In contrast, there is another type of multiple endpoints named *alternative* primary endpoints, where success is defined as meeting at least one of the multiple endpoints. There is a lot of research literature on this topic. The main issue relating the alternative primary endpoints is the overall Type I error rate will be inflated, since there is more of chance to make Type I error. Meyerson et al. (2007) discusses some general issues of the clinical trials with multiple co-primary endpoints. The Intersection-Union Test (IUT) is the standard hypothesis test for such a problem. The main issue is the power of the IUT is lower than the power of each single endpoint. Furthermore, it is more difficult to reach the significance when the number of the primary endpoints increases. A larger sample size is needed in order to have the desired power of the study. There is few research that has been done to discuss the sample size calculation for the study with multiple co-primary endpoints. Moreover, none have been developed to monitor the trial with co-primary endpoints.

In this dissertation, some statistical methods are developed for monitoring the clinical trials with multiple co-primary endpoints. Current practice involves using a Group Sequential Design (GSD) method to stop trials early for promising efficacy, and using the Conditional Power (CP) for futility stopping rules. The stopping boundaries of the group sequential design for the clinical trials with multiple co-primary endpoints are shown to be the same as the ones for single endpoints. Lan and Wittes (1988) proposed the B-value tool to calculate the CP of a single endpoint and we extend this tool to multiple dimensions to calculate the CP for the studies with multiple co-primary endpoints. To introduce the concepts, we start from the simplest case, in which the study is one-arm (one sample) with two normal co-primary endpoints that have a known covariance structure between the endpoints. Moreover, we extend the methods to be applicable to some common cases of two-arm

studies with co-primary normal and binary endpoints, and then we provide several examples of implementation of our approach with simulated trials. For the co-primary normal endpoints, we consider using the multivariate regression model to adjust for the other covariates. The CP provides a basis for the Sample Size Re-estimation (SSR). A fixed-weight SSR method is introduced. We show that the fixed-weight test statistic will not inflate the Type I error rate. It is worth pointing out that the method introduced in this dissertation is not limited to the area of clinical trials, but also could be applicable to the studies that require reaching the significance on more than one response simultaneously and need a long time period in which to use the interim analysis to monitor the on-going study. At the end, we discuss the possibility of the blinded interim analysis with alternative multiple endpoints using the modality inference introduced in the first part.

This dissertation is organized as follows: Chapter 2 to Chapter 4 focus on the first part of the research: the modality inference and computing of the modal clustering. Chapter 2 reviews the relevant research of the modality inference. It first surveys several existing methods of the modality inference. Next, it reviews the MEM and REM algorithms proposed by Li et al. (2007), which are used to detect the modes of the data and the ridgeline between the two modes. The clustering algorithm MAC based on MEM is also reviewed. Consequently, this chapter reviews several properties of the KDE, since the MAC algorithm and the modality inference procedure are based on the KDE. Chapter 3 develops a new modality inference procedure. In this chapter, we propose the test statistic and its asymptotic distribution. The choice of the bandwidth parameter is discussed. The inference method is tested on several simulated data and two real data sets, including flow cytometry data and Swiss banknotes data. Chapter 4 introduces the parallel computing of the HMAC.

Chapter 5 to Chapter 7 focuses on the second part of the thesis: the statistical monitoring of clinical trials with multiple co-primary endpoints. Chapter 5 reviews some relevant research, the GSD of single endpoint, B-value tool introduced by Lan and Wittes (1988), and the problem of the multiple co-primary endpoints. Chapter 6 discusses the GSD procedure with multiple co-primary endpoints. In this chapter, we prove that the stopping boundaries for the study with multiple co-

primary endpoints are the same as the ones for the studies with single endpoints. Chapter 7 extends the B-value tool to multiple dimensions so that the CP can be calculated for the study with co-primary endpoints. Part III tries to link the two parts of the research. Chapter 8 discusses the idea of using the multivariate modality inference procedure to monitor the trial with alternative multiple primary normal endpoints. Finally, Chapter 9 concludes the dissertation by pointing out several potential directions of the future research.

The major achievements in the dissertation are listed as follows:

- In Chapter 3, a multivariate modality inference procedure is developed. The test statistic, which is based on the contrast between the density heights of the mode and saddle point, is introduced. Consequently, the asymptotic distribution of the test statistic is derived. The inference procedure is applied on several simulated and real data sets.

- Chapter 4 introduces the parallel computing method of the HMAC, which can be used to cluster large data sets.

- Chapter 6 discusses the group sequential test procedure of multiple co-primary endpoints. In this chapter, we prove that the stopping boundaries of the multiple co-primary endpoints are exactly the same as the ones for single endpoint.

- In Chapter 7, the B-value tool is extended to multi-dimensions, named Multivariate B-value tool, to calculate the CP of the clinical trials with multiple co-primary endpoints. The fixed-weight SSR method based on the CP is introduced.

Part I

Multivariate Modality Inference and Parallel Computing of Modal Clustering

Chapter 2

The Problem of Modality Inference

Mode is defined as the local maximum of a probability density. *Modality*, which is the number of the modes, is an important feature of any probability distribution. The natural evolution of multimodality occurs when a distribution is composed by several sub-populations. In practice, it is important to learn how many sub-populations the data has. In general, there are three different, but related, research areas that addresses this issue: (1) the inference on the number of components in the finite mixture model; (2) estimating the number of clusters and/or merging the clusters of a clustering output; and (3) the modality inference. Each of these three approaches addresses the question of how many components the data has from its own angle. For the inference on the number of components in the mixture distribution, the hypothesis is usually $H_0 : K = k$ versus $H_a : K = k+1$ where K is the parameter of the number of components. The likelihood ratio test (LRT) is often used to assess the significance of the hypothesis. However, in general, the distribution of the LRT is very complicated. More details can be found in e.g. McLachlan and Peel (2004, Chap. 6). Depending on the clustering method, the inference procedures on the number of clusters could be different. Tibshirani et al. (2001) proposed the *GAP* statistic, which contains information regarding the distance between the data points within each cluster. In particular, the authors applied the method on the K-means clustering (Lloyd, 1982). Fraley and Raftery (2002) proposed the EM-clustering method, a model based clustering method. The authors selected the ideal number of clusters by using the Bayesian Information Criterion (BIC), which is a model selection tool. However, as it is well established, BIC does not follow the regularity conditions and is inappropriate to use in the problem of determining the number of components.

The modality inference, which is used to assess the number of modes of the data, is often a robust *nonparametric* approach. In practice, if we collect a sample of data, but do not know the underlying distribution, (in statistics, this is called the *nonparametric* setting), frequently we would want to learn the features of the probability distribution by using the collected data. Existing literature addresses the problem of the modality of a univariate data, but generalization to multivariate data is sparse. In this dissertation, we focus on developing modality inference in a multivariate setting.

This chapter reviews the relevant research, which will be used to develop the modality inference method in the next chapter. It is organized as follows: Section 2.1 provides a brief review of several modality inference procedures. Ray and Lindsay (2005) and Li et al. (2007) provided comprehensive tools to learn the features of the probability density of data based on the Kernel Density Estimates (KDE), including detecting the modes and the ridgeline between the two modes. It has been shown in Ray and Lindsay (2005) that the ridgeline contains all the modes and saddle points between the modes of the probability density. Section 2.2 provides a review of these tools. Section 2.3 reviews some properties of the KDE, which is the basis of the inference procedure introduced in Chapter 3, including the choice of the bandwidth parameters, the asymptotic properties, and the curse of dimensionality.

2.1 Relevant Research

There is lot of existing literature that addresses the problem of the modality of univariate probability distribution. These methods can be classified as the test of unimodality, bimodality or multimodality. Alternatively, these methods can be grouped as a global or local test. The global test considers the modality of the entire distribution. In contrast, the local test focuses on the specific region of the density that contains the particular investigated mode instead of considering the entire distribution.

In the case of the global test, Silverman (1981) proposed the most commonly used *critical bandwidth parameter*, the smallest value of the bandwidth parameter h for which the kernel density

estimate with Gaussian kernel

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{N} \phi \left(\frac{x - X_i}{h} \right) \tag{2.1}$$

is k-modal, where $\phi(.)$ is the probability distribution function (pdf) of the standard normal distribution. The h_{crit} is defined as

$$h_{crit} = \inf\{h; \hat{f}(., h) \text{ has at most } k \text{ modes}\}. \tag{2.2}$$

Silverman (1981) showed that by using the pdf of the standard normal distibution as the kernel function, the number of the modes of the density estimate (2.1) is monotone decreasing as h increases. It has more than k modes if and only if $h < h_{crit}$.

To assess the significance, Silverman (1981) suggested to use h_{crit} as the bandwidth parameter, denoted as h_0, and to use the nonparametric bootstrap method proposed by Efron (1979) to sample the reference data, consequently to get the distribution of the test statistic under the null hypothesis. Denoting the bootstrapped sample as X_B, one calculates the reference data as:

$$y_i = (1 + h_0^2/\sigma^2)^{-\frac{1}{2}}(X_{B,i} + h_0 \epsilon_i),$$

where σ^2 is the sample variance of X and ϵ_i are the added noise following standard normal distribution. The scale term $1 + h_0^2/\sigma^2$ ensures that the variance of the reference data y_i is σ^2, which is the same as the original sample variance. To calculate the p-value, one counts the number of times that the kernel density estimate of Y has more than k modes using $h = h_0$, out of the total amount of the bootstrap sampling time.

Among the local test, the one proposed by Minnotte (1997) is widely used. Denoting the mode

as u_{2k} and the saddle point u_{2k-1}, the test statistic is defined as:

$$M_i = \int_{u_{i-1}}^{u_{i+1}} \{\hat{f}(x) - \max(\hat{f}(u_{i-1}), \hat{f}(u_{i+1}))\}_+ dx, \tag{2.3}$$

where i is the ith investigated mode and $\hat{f}(x)$ is the kernel density estimate in (2.1). It can be thought of as the probability mass of the mode above the higher of the two saddle points or anti-modes around it. The advantage of this statistic is that it does not only consider the heights of the mode and saddle point, but also the distance between them. The reference distribution is generated by forcing the distribution flat (uniform) between the point u_{i-1} and u_{i+1} and keep the rest of the distribution the same. It then simulates the reference data from the reference distribution and calculates the test statistic. The distribution of the test statistic under the null hypothesis is obtained by repeating this simulation step and thus the p-value is calculated by the usual bootstrap approach.

Burman and Polonik (2009) proposed a mode hunting tool together with a further test of the significance for the existence of these modes, by using the k-nearest neighbor (KNN) density estimate for multivariate data. The authors proposed an iterative nearest neighbor method for selecting the initial modal candidates and then thinning out this list of modal candidates by eliminating the modes that fail the local parametric test. One needs to choose two values k_1 and k_2 for implementing these two steps. In general, it is required that $k_1 > k_2$. Specifically, the method of finding the initial modal candidates is executed by repeating the following two steps: First, find the initial modal candidate $W_1 = \text{argmin}_{X_j, j\in\{1,\cdots,n\}} \hat{d}_n(X_j)$ where $\hat{d}_n(x)$ is the distance between a point x and its k_1 nearest neighbor. Then, eliminate the k_2 nearest neighbor points around W_1. The remaining data forms a new data set, D. The modal candidate for D can be found, treating D as the new data set. Repeat these two steps until no modal candidates can be found. To shorten the list of modal candidates, for each modal candidate W_j, consider the k_2 nearest neighbor points around it. Under specified assumptions, the local data follows multivariate normal distribution and the mean of these k_2 nearest neighbor points should be the same as W_j. Hotelling's T^2 test is then carried out. In the paper, the significance level of the Hotelling's Test is 0.01. By eliminating those modal candidates

which are significant for the local test and a set of modal candidates, M, is formed. The details of the choice of k_1 and k_2 were discussed in this paper. For the formal pairwise test of existence of the modes in M, Burman and Polonik (2009) proposed the following statistic

$$SB(\alpha) := -logf(\mathbf{x}_\alpha) + \min\{logf(\mathbf{x}_{m1}), logf(\mathbf{x}_{m2})\}, \tag{2.4}$$

where $\mathbf{x}_{m1}$ and $\mathbf{x}_{m2}$ are the two candidate modes and $\mathbf{x}_\alpha = (1-\alpha)\mathbf{x}_{m1} + \alpha\mathbf{x}_{m2}, \alpha \in [0,1]$ is the point on the segment between $\mathbf{x}_{m1}$ and $\mathbf{x}_{m2}$. Note that the test statistic is the logarithm of the ratio of the heights of a point between the two modes, $\mathbf{x}_{m1}$ and $\mathbf{x}_{m2}$, and mode with lower estimated density. The test of $SB < 0$ leads to the conclusion of whether $\mathbf{x}_{m1}$ and $\mathbf{x}_{m2}$ are the two distinct modes. Moreover, using the KNN to estimate $SB(\alpha)$, it is found that the asymptotic distribution of the test statistic $\hat{SB}(\alpha)$ follows the normal distribution. The null hypothesis is rejected if and only if $\hat{SB}(\alpha) \geq \sqrt{\frac{2}{k_1}}\Phi^{-1}(0.95)$.

2.2 Mode Hunting Tool

Li et al. (2007) proposed a set of comprehensive tools to explore the geometric feature of the density estimate of the data. In this section, we review the basic quantities of the modality inference, which are the mode, the saddle point and the ridgeline. We will also discuss the algorithms to determine these quantities under the KDE.

2.2.1 Mode

Mode is defined as the local maximum of a probability density. Traditional techniques of finding local maxima, such as *hill climbing*, work well for univariate data. However, multivariate hill climbing is computationally expensive, thereby limiting its use in high dimensions. Li et al. (2007) proposed an algorithm that solves a local maximum of a KDE by ascending iterations starting from the data points. Since the algorithm is very similar to the Expectation Maximization(EM) algorithm (Dempster et al., 1977), it is named as the Modal Expectation Maximization (MEM). The finite mixture

model can be expressed as

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x), \tag{2.5}$$

with $\sum_{k=1}^{K} \pi_k = 1$ and $f_k(x)$ are the mixing components. Given any initial value $x^{(0)}$, the MEM solves the local maxima of the mixture density by alternating the following two steps until it meets some user defined stopping criterion.

Step 1: Let $p_i = \dfrac{\pi_i f_i(x^{(r)})}{f(x^{(r)})}$, $i = 1, ..., n$.

Step 2: Update $x^{(r+1)} = \underset{x}{\operatorname{argmax}} \sum_{i=1}^{n} p_i \log f_i(x)$.

Details of convergence of the MEM approach can be found in Li et al. (2007). The above iterative steps provide a computationally simpler approach than the grid search method for hill climbing from any starting point $x \in \mathbb{R}^D$, by exploiting the properties of density functions. Given a multivariate kernel K, let the density of the data be given by $f(x|\Sigma) = \sum_{i=1}^{n} \frac{1}{n} K(x - x_i|\Sigma)$, where Σ is the matrix of smoothing parameters. Moreover, in the special case of Gaussian kernels, i.e., $K(x - x_i|\Sigma) = \phi(x \mid x_i, \Sigma)$, where $\phi(\cdot)$ is the pdf of a Gaussian distribution, the update of $x^{(r+1)}$ is simply

$$x^{(r+1)} = \sum_{i=1}^{n} p_i x_i \ .$$

This allows us to avoid the numerical optimization of Step 2. Due to this reason, the normal kernel function is used throughout the methods introduced in this thesis. However, in general, one can also use other kernel functions.

The MEM algorithm can be naturally used to define clusters. If we start the algorithm from each data point, we can cluster the data that converges to the same mode as one group. Li et al. (2007) denotes this algorithm the *Mode Association Clustering* (MAC). If we choose a sequence of bandwidth parameters h, then we can get the hierarchical MAC (HMAC). More details about the computing will be discussed in Chapter 4.

2.2.2 Saddle Point and Ridgeline

Ray and Lindsay (2005) provided the explicit formula for the ridgeline between the two means of the mixture of two multivariate normal distributions. The mixture density of two d-dimensional multivariate normal distributions is:

$$f(\mathbf{x}) = \pi\phi(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\pi)\phi(\mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), x \in \Re^d \tag{2.6}$$

where the $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors and $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are the two covariance matrices of the two mixed multivariate normal components respectively. The ridgeline of the distribution in (2.6) from one mean to another is given by:

$$x(\alpha)^* = [\bar{\alpha}\boldsymbol{\Sigma}_1^{-1} + \alpha\boldsymbol{\Sigma}_2^{-1}]^{-1}[\bar{\alpha}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \alpha\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2], \tag{2.7}$$

where $\alpha \in [0, 1]$ and $\bar{\alpha} = 1 - \alpha$. Ray and Lindsay (2005) showed that all the critical points of the d-dimensional distribution, including the modes and saddle points, are the points on the ridgeline. With different choices of the parameters, the probability density can be unimodal, bimodal or in some special cases, trimodal. See some examples in Ray and Lindsay (2005). In particular, if $\Sigma_1 = \Sigma_2 = \Sigma$ and $\pi_1 = \pi_2 = 0.5$, then the mixture density is bimodal if and only if the Mahalanobis Distance between $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ is greater than 4. Otherwise, it will be unimodal. More details are provided in Section 8.1 The ridgeline provides a useful tool to discover the modality of a mixture of multivariate normal mixtures.

Li et al. (2007) also provided the algorithm to find the ridgeline of the KDE between the two modes identified by MEM, named the Ridgeline EM (REM). Here we provide a brief description of the REM:

Let the density of the two clusters represented by the two modes of interest be f_1 and f_2. We

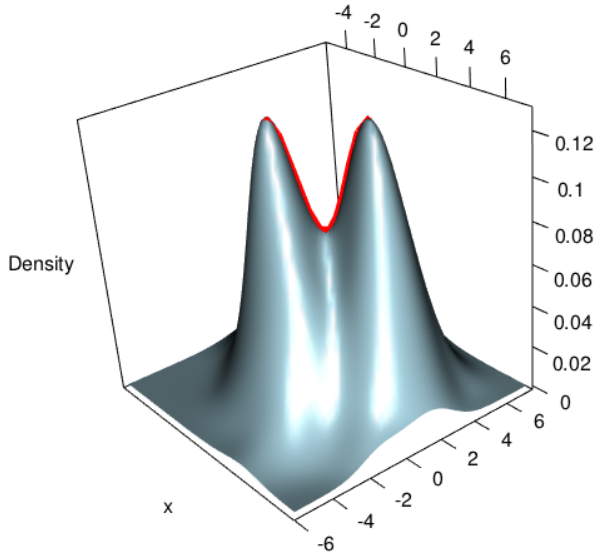

Figure 2.1: Ridgeline example

can consider both f_1 and f_2 as the mixtures of L parametric distributions:

$$f_i(x) = \sum_{k=1}^{L} \pi_{i,k} h_{i,k}(x), i = 1, 2$$

Starting from an initial value $x^{(0)}$, the REM updates x by iterating the following two steps:

Step 1: Compute:

$$p_{i,k} = \pi_{i,k} h_{i,k}(x^{(r)}) / \sum_{j=1}^{L} \pi_{i,j} h_{i,j}(x^{(r)}) \text{ with } k = 1, 2, \cdots L, i = 1, 2$$

Step 2: Update $x^{(r+1)}$:

$$x^{(r+1)} = argmax(1-\alpha) \sum_{k=1}^{L} \pi_{1,k} \log h_{1,k}(x) + \alpha \sum_{k=1}^{L} \pi_{2,k} \log h_{2,k}(x)$$

In the special case where $h_{i,k}(x) = \phi(x|\mu_{i,k}, \Sigma)$, the multivariate normal distribution, the second step becomes $x^{(r+1)} = (1-\alpha) \sum_{k=1}^{L} \pi_{1,k}\mu_{1,k} + \alpha \sum_{k=1}^{L} \pi_{2,k}\mu_{2,k}$ Figure 2.1 illustrates one example of the ridgeline. The point on the ridgeline with the lowest density is the detected saddle point. The REM and MEM introduced in this section provide useful tools to detect the mode and saddle point of the KDE, which provides the basis of the inferential framework introduced in the following section.

2.3 Kernel Density Estimate

In this section, we review some basic properties of the multivariate KDE, which provides the foundation of the modality inference introduced in the next chapter. The multivariate KDE is the most commonly used non-parametric estimate of the probability density of a multivariate random variable. Suppose the d-dimensional vectors $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_n})$ are *i.i.d* samples from the population with some unknown probability density f. The $\mathbf{X_i} = (X_{i1}, X_{i2}, \cdots, X_{id}), i = 1, 2, \cdots, n$. The multivariate KDE is:

$$\hat{f}(\mathbf{x}) = \frac{1}{n\|\mathbf{H}\|} \sum_{i=1}^{n} K(\mathbf{H}^{-1}(\mathbf{x} - \mathbf{X_i})), \tag{2.8}$$

where $K(\cdot)$, a real-valued multivariate kernel function, has the properties of $\int K(\mathbf{z})d\mathbf{z} = 1$, $\int \mathbf{z}K(\mathbf{z})d\mathbf{z} = 0$ and $\int \mathbf{z}\mathbf{z}^{\mathbf{Z}}K(\mathbf{z})d\mathbf{z} = \mu_2(K)\mathbf{I_p}$. Usually $K(\cdot)$ is chosen as the standard multivariate normal density function. $\mathbf{H}$ is the $d \times d$ non-singular positive definite *bandwidth matrix* and $\|\mathbf{H}\|$ is the determinant of $\mathbf{H}$.

The d-dimensional H contains $d(d+1)/2$ number of parameters. In practice, it is difficult to choose the values of H due to such a large number of parameters over d-dimensional space. It is more practical to keep the number of bandwidth parameters as small as possible, but retaining enough to provide good estimates. One approach to reducing the number of bandwidth parameters is to use the simplest model that contains only *one* bandwidth parameter:

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{\mathbf{x} - \mathbf{X_i}}{h}) \tag{2.9}$$

However, if the data have different scales on different dimensions, the KDE of (2.9) will lead to a poor estimation. To avoid the scaling problem, the *sphering transformation* can be applied to the data.

2.3.1 Sphering Transformation

Sphering transformation, also known as *Whitening transformation*, is a linear transformation that makes the data have the identity covariance matrix (Fukunaga, 1990). To carry out the transformation, one computes the spectral decomposition of the sample covariance matrix of $\mathbf{X}$, $\hat{\mathbf{\Sigma}} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathbf{T}}$. Let $\mathbf{Y} = \mathbf{\Lambda}^{-1/2}\mathbf{P}^{\mathbf{T}}\mathbf{X}$, then $Cov(\mathbf{Y}) = \mathbf{I}$. Using the operation $\mathbf{X} = \mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{Y}$, one can transform the data back to the original scale. In this thesis, we will use this transformation and the kernel density estimator 2.9. Figure 2.2 illustrates two examples of original data and sphering transformed data. The original data is shown on the left panel while the transformed data is shown on the right panel. The plots show that after the transformation, both the scale and the "direction" of the data has changed, while the clustering or grouping information of the data is still preserved.

2.3.2 Bandwidth Selection

It is well known that the larger bandwidth parameter will oversmooth the estimate of the density, while the smaller one will under smooth the density estimate. There is lot of literature to describe the choice of the bandwidth parameters based on different criteria. However, there is no unique best choice. In general, the "optimal" choice of the bandwidth is to minimize the Asymptotic Mean Integrated Squared Error (AMISE), which will be defined in the following sequence of definitions. First we consider the expectation of $\hat{f}(\mathbf{x})$ defined in (2.9), which can be calculated as :

$$\begin{aligned} E\left[\hat{f}(\mathbf{x})\right] &= E\left[\frac{1}{nh^d}\sum_{i=1}^{n} K(\frac{\mathbf{X_i}-\mathbf{x}}{h})\right] \\ &= E\left[\frac{1}{h^d}K(\frac{\mathbf{X_i}-\mathbf{x}}{h})\right] \\ &= \int \frac{1}{h^d}K\left(\frac{\mathbf{z}-\mathbf{x}}{h}\right) f(\mathbf{z})d\mathbf{z} \\ &= \int K(\mathbf{u})f(\mathbf{x}-h\mathbf{u})d\mathbf{u} \text{ (using } \mathbf{u} = \tfrac{\mathbf{z}-\mathbf{x}}{h}\text{).} \end{aligned}$$

Using Taylor series expansion, we can write:

$$f(\mathbf{x}-h\mathbf{u}) = f(\mathbf{x}) - \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}'}(h\mathbf{u}) + \frac{1}{2}(h\mathbf{u})'\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}'}(h\mathbf{u}) + o(h^2).$$

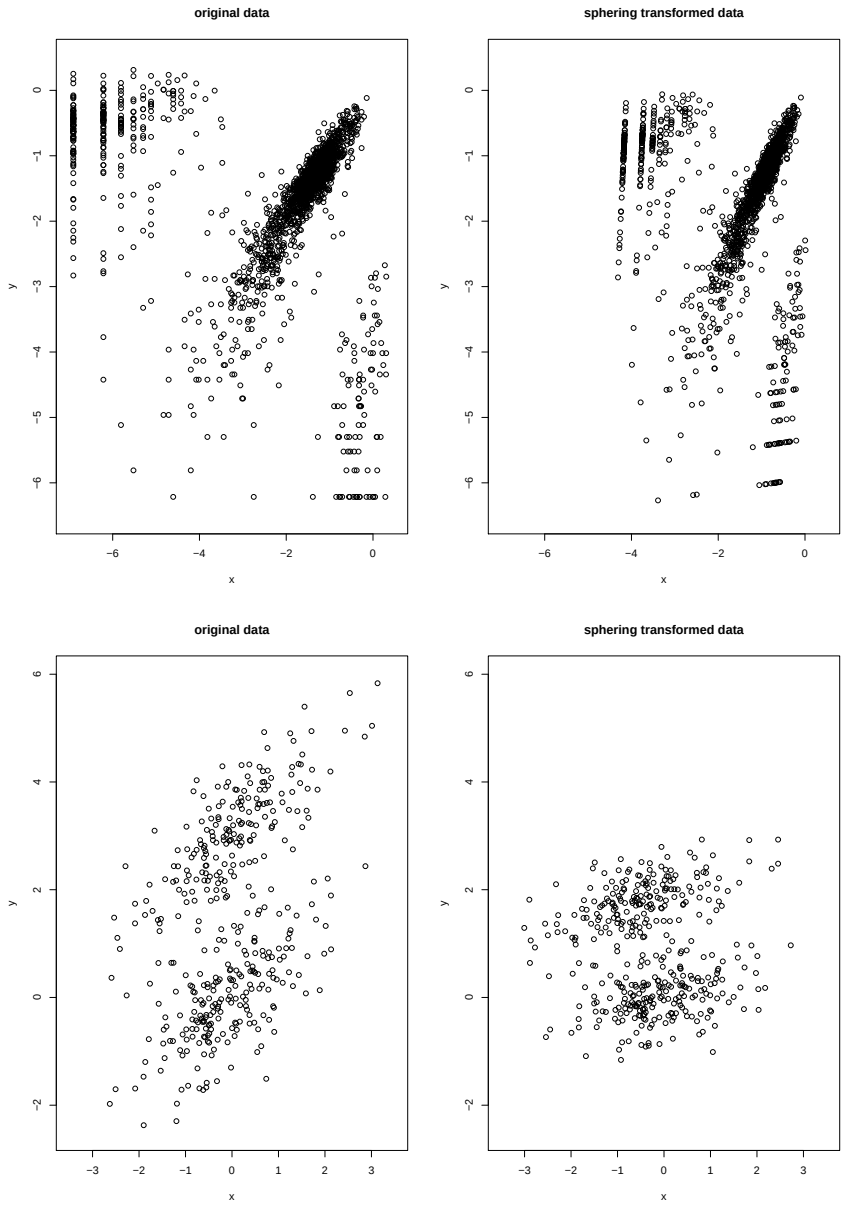

Figure 2.2: Two examples of original data and sphering transformed data

then, the bias term of $\hat{f}(\mathbf{x})$ can be expressed as:

$$\begin{aligned} E\left[\hat{f}(\mathbf{x})\right] - f(\mathbf{x}) &= \frac{h^2}{2} tr\left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} \int \mathbf{u}\mathbf{u}' K(\mathbf{u}) d\mathbf{u}\right) \\ &= \frac{h^2}{2} \mu_2(K) tr\left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'}\right). \end{aligned} \tag{2.10}$$

Similarly, the variance of $\hat{f}(\mathbf{x})$ can be calculated as:

$$\begin{aligned} Var\left[\hat{f}(\mathbf{x})\right] &= Var\left[\frac{1}{nh^d}\sum_{i=1}^{n} K(\frac{\mathbf{X_i} - \mathbf{x}}{h})\right] \\ &= \frac{1}{n} Var\left[\frac{1}{h^d} K(\frac{\mathbf{X_i} - \mathbf{x}}{h})\right] \\ &= \frac{1}{n} E\left[\frac{1}{h^d} K(\frac{\mathbf{X_i} - \mathbf{x}}{h})\right]^2 - \frac{1}{n}\left(E\left[\frac{1}{h^d} K(\frac{\mathbf{X_i} - \mathbf{x}}{h})\right]\right)^2. \end{aligned}$$

Simplifying the above equation, we can get:

$$Var\left[\hat{f}(\mathbf{x})\right] = \frac{f(\mathbf{x})}{nh^d} \int [K(\mathbf{u})]^2 d\mathbf{u} + o(\frac{1}{nh^d}). \tag{2.11}$$

Now we can calculate the Mean Squared Error (MSE) defined as:

$$MSE = Bias^2 + Variance.$$

Using (2.10) and (2.11), the MSE of $\hat{f}(\mathbf{x})$ is:

$$MSE\left[\hat{f}(\mathbf{x})\right] = \frac{h^4}{4}\mu_2^2(K) tr\left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'}\right)^2 + \frac{f(\mathbf{x})}{nh^d} \int [K(\mathbf{u})]^2 d\mathbf{u} + o(h^4) + o(\frac{1}{nh^d}). \tag{2.12}$$

The Mean Integrated Squared Error (MISE) of the KDE is defined as $MISE\left[\hat{f}(\mathbf{x})\right] = \int MSE\left[\hat{f}(\mathbf{x})\right] d\mathbf{x}$. It can be calculated as,

$$MISE\left[\hat{f}(\mathbf{x})\right] = \frac{h^4}{4}\mu_2^2(K) \int tr\left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'}\right)^2 d\mathbf{x} + \frac{1}{nh^d} \int [K(\mathbf{u})]^2 d\mathbf{u} + o(h^4) + o(\frac{1}{nh^d}). \tag{2.13}$$

As $h \to 0$, the asymptotic MISE (AMISE) is given by:

$$AMISE(\hat{f}(\mathbf{x})) = \frac{h^4}{4}\mu_2^2(K) \int tr\left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'}\right)^2 d\mathbf{x} + \frac{\int [K(\mathbf{u})]^2 d\mathbf{u}}{nh^d}. \tag{2.14}$$

To minimize AMISE, we solve the following equation:

$$\frac{\partial}{\partial h} AMISE = h^3 \mu_2^2(K) \int tr\left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'}\right)^2 d\mathbf{x} - \frac{d \int [K(\mathbf{u})]^2 d\mathbf{u}}{nh^{(d+1)}} = 0.$$

Thus the AMISE is minimized as h is:

$$h_{opt} = \left[\frac{d \int [K(\mathbf{u})]^2 d\mathbf{u}}{\mu_2^2(K) \int tr\left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'}\right)^2 d\mathbf{x}}\right]^{\frac{1}{d+4}} n^{-\frac{1}{d+4}}. \tag{2.15}$$

and the minimized AMISE is:

$$AMISE(h_{opt}) = O(n_d^{-\frac{1}{d+4}}). \tag{2.16}$$

Multivariate kernel density estimation was first discussed by Scott (1992), Wand and Jones (1993, 1994) and Sain et al. (1994). Scott (1992) first used the MISE criterion and applied the "normal reference rule". The optimal bandwidth can be approximated by

$$h_j = \sigma_j \left\{\frac{4}{(d+2)n}\right\}^{\frac{1}{d+4}}, \tag{2.17}$$

where σ_j is the standard deviation of the jth dimension and can be estimated by the corresponding sample standard deviation. Sain et al. (1994) employed the least square cross validation and biased cross validation for estimating the optimal bandwidth by minimizing the Integrated Squared Error (ISE) and the asymptotic MISE (AMISE) respectively. Wand and Jones (1994) proposed an extension of the Plug-in method for the univariate data to multivariate data. The main idea of their Plug-in algorithm was to minimize the AMISE. This method is mostly restricted to estimating diagonal bandwidth matrices. "KernSmooth" is the R package that implements this approach (Wand,

2006).

However, Duong and Hazelton (2003) argued that the Plug-in method developed by Wand and Jones (1994) failed to produce a finite bandwidth matrix and suggested an alternative Plug-in method that required less computation of the pilot bandwidths. They also created an R package "ks" with functions that can compute the Plug-in bandwidth matrices for data with up to 6 dimensions. Zhang et al. (2006) proposed a Bayesian approach for multivariate bandwidth selection. It assumes the bandwidth parameter has some prior distribution and uses MCMC to carry out the posterior distribution of the bandwidth parameter. Although the idea is clear and straightforward, in practice, this requires much more computational resources and prior knowledge.

2.3.3 Asymptotic Distribution of KDE

In this section, we provide the details of the asymptotic properties of KDE defined in (2.9).

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X_i}}{h}\right)$$

Recall in Section 2.3, the bias and variance of $\hat{f}(\mathbf{x})$ are given by (2.10) and (2.11). As $h \to 0$, we have

$$E\left[\hat{f}(\mathbf{x})\right] \to f(\mathbf{x}),$$

showing that $\hat{f}(\mathbf{x})$ is asymptotically unbiased. We apply the normal reference rule on the sphering transformed data to choose the bandwidth parameter h as (2.17),

$$h_{normal} = \left\{\frac{4}{(d+2)n}\right\}^{\frac{1}{d+4}}.$$

Note that the term $\sqrt{nh^d}(E[\hat{f}(\mathbf{x}) - f(\mathbf{x})]) = O(\sqrt{nh^{d+4}})$. If we choose h with the optimal convergent rate, which is $h_n = c\left(\frac{1}{n}\right)^{\frac{1}{d+4}}$ defined as (2.15), for example, using normal reference rule, then

$$\sqrt{nh^d}(E[\hat{f}(\mathbf{x}) - f(\mathbf{x})]) \to O(1).$$

The kernel density estimator has the following asymptotic normality property of:

Property 2.1 *If* $h \to 0$ *and* $nh^d \to \infty$ *as* $n \to \infty$ *then*

$$(nh^d)^{\frac{1}{2}}(\hat{f}_{n,h}(\mathbf{x}) - f(\mathbf{x}) - \frac{h^2}{2}\mu_2(K)tr(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}'})) \xrightarrow{d} N(0, f(\mathbf{x})(\int \mathbf{K}^2(\mathbf{u})d\mathbf{u})).$$

The proof of the above property is based on the Liapunov Central Limit Theorem. Li and Racine (2011) provide the details of the proof. From the Property 2.1, it can be seen that by choosing the h at optimal convergent rate, the KDE $\hat{f}_{n,h}(\mathbf{x})$ asymptotically is a biased estimator of $f(\mathbf{x})$. It underestimates the local maxima since the term $tr(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}'})$ in bias is negative at maxima and it overestimates at local minima due to the same reason.

If we choose an h that converges *faster* than the optimal rate, i.e,

$$\begin{aligned} h^* &= o\left(\frac{c}{n}\right)^{\frac{1}{d+4}} \\ &= \left(\frac{c}{n}\right)^{\frac{\gamma}{d+4}} \text{ where } \gamma > 1 \end{aligned}$$

the bias term can be made negligible. The asymptotic distribution then becomes:

Property 2.2 *If* $h \to 0$, $nh^d \to \infty$ *and* $nh^{d+4} \to 0$ *as* $n \to \infty$, *then*

$$(nh^d)^{\frac{1}{2}}(\hat{f}_{n,h}(\mathbf{x}) - f(\mathbf{x})) \xrightarrow{d} N(0, f(\mathbf{x})(\int \mathbf{K}^2(\mathbf{u})d\mathbf{u})).$$

To satisfy the conditions $nh^d \to \infty$ and $nh^{d+4} \to 0$, we should choose h^* as

$$h^* = \left(\frac{c}{n}\right)^{\frac{\gamma}{d+4}} \text{ where } 1 < \gamma < 1 + \frac{4}{d}$$

If the h converges *slower* than the optimal rate, the bias term cannot converge as $n \to \infty$.

Note that the variance term in the asymptotic distribution contains the unknown parameter $f(\mathbf{x})$. Simply by applying the delta method and using the transformation function $g(x) = \sqrt{x}$, the variance

is stabilized and becomes invariant with $f(\mathbf{x})$. The asymptotic distribution in Property 2.1 becomes

$$(nh^d)^{\frac{1}{2}}(\sqrt{\hat{f}_{n,h}(\mathbf{x})} - \sqrt{f(\mathbf{x})}) \xrightarrow{d} N\left(0, \frac{(\int \mathbf{K}^2(\mathbf{u})d\mathbf{u})}{4}\right).$$

Further, it is easy to verify the following property:

Property 2.3 *If $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$, for $\mathbf{x_1} \neq \mathbf{x_2}$, $\hat{f}(\mathbf{x_1})$ and $\hat{f}(\mathbf{x_2})$ are uncorrelated.*

Proof: We consider the covariance between $\hat{f}(\mathbf{x_1})$ and $\hat{f}(\mathbf{x_2})$.

$$\begin{aligned}
& Cov(\hat{f}(\mathbf{x_1}), \hat{f}(\mathbf{x_2})) \\
= \; & Cov\left(\frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{x_1} - \mathbf{X_i}}{h}\right), \frac{1}{nh^d}\sum_{j=1}^{n} K\left(\frac{\mathbf{x_2} - \mathbf{X_j}}{h}\right)\right) \\
= \; & \frac{1}{nh^{2d}} Cov\left(K\left(\frac{\mathbf{x_1} - \mathbf{X}}{h}\right), K\left(\frac{\mathbf{x_2} - \mathbf{X}}{h}\right)\right) \\
= \; & \frac{1}{nh^{2d}}\left[E\left(K\left(\frac{\mathbf{x_1} - \mathbf{X}}{h}\right)K\left(\frac{\mathbf{x_2} - \mathbf{X}}{h}\right)\right) - EK\left(\frac{\mathbf{x_1} - \mathbf{X}}{h}\right)EK\left(\frac{\mathbf{x_2} - \mathbf{X}}{h}\right)\right].
\end{aligned}$$

We want to show that $Cov(\hat{f}(\mathbf{x_1}), \hat{f}(\mathbf{x_2})) \to 0$ if $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$. Consider the term

$$\begin{aligned}
& \frac{1}{nh^{2d}} E\left(K\left(\frac{\mathbf{x_1} - \mathbf{X}}{h}\right)K\left(\frac{\mathbf{x_2} - \mathbf{X}}{h}\right)\right) \\
= \; & \frac{1}{nh^{2d}}\int K\left(\frac{\mathbf{z} - \mathbf{x_1}}{h}\right)K\left(\frac{\mathbf{z} - \mathbf{x_2}}{h}\right) f(\mathbf{z})d\mathbf{z} \\
= \; & \frac{1}{nh^{d}}\int K(\mathbf{u})K(\mathbf{u} + \frac{\mathbf{x_1} - \mathbf{x_2}}{h}) f(\mathbf{x_1} + \mathbf{u}h)d\mathbf{u} \\
= \; & \frac{f(\mathbf{x_1})}{nh^{d}}\int K(\mathbf{u})K(\mathbf{u} + \frac{\mathbf{x_1} - \mathbf{x_2}}{h})d\mathbf{u} \quad \text{since } h \to 0 \\
= \; & O\left(\frac{1}{nh^d}\right) \to 0 \text{ as } nh^d \to \infty.
\end{aligned}$$

For the second last step, since $K(\mathbf{u})K(\mathbf{u} + \frac{\mathbf{x_1} - \mathbf{x_2}}{h})$ is a convolution and hence another density, therefore its integration is bounded by 1. Now we consider the term

$$\frac{1}{nh^{2d}} EK\left(\frac{\mathbf{x_1} - \mathbf{X}}{h}\right)EK\left(\frac{\mathbf{x_2} - \mathbf{X}}{h}\right)$$

$$
\begin{aligned}
&= \frac{1}{nh^{2d}} \int K\left(\frac{\mathbf{z}-\mathbf{x_1}}{h}\right) f(\mathbf{z}) d\mathbf{z} \int K(\mathbf{v}) f(\mathbf{x_2}+\mathbf{v}h) d\mathbf{v} \\
&= \frac{1}{n} \int K(\mathbf{u}) f(\mathbf{x_1}+\mathbf{u}h) d\mathbf{u} \int K(\mathbf{v}) f(\mathbf{x_2}+\mathbf{v}h) d\mathbf{v} \quad \text{since } h \to 0 \\
&= \frac{1}{n} f(\mathbf{x_1}) f(\mathbf{x_2}) \int K(\mathbf{u}) d\mathbf{u} \int K(\mathbf{v}) d\mathbf{v} \\
&= O\left(\frac{1}{n}\right) \to 0 \text{ as } n \to 0.
\end{aligned}
$$

Thus, it follows

$$\lim_{h \to 0} Cov(\hat{f}(\mathbf{x_1}), \hat{f}(\mathbf{x_2})) = 0.$$

Therefore, we proved that $\hat{f}(\mathbf{x_1})$ and $\hat{f}(\mathbf{x_2})$ are asymptotically uncorrelated. Since under the same condition of Property 2.3, $\hat{f}(\mathbf{x_1})$ and $\hat{f}(\mathbf{x_2})$ asymptotically follow normal distributions, we can claim they are asymptotically independent.

2.3.4 Curse of Dimensionality

The phrase *curse of dimensionality* was first introduced in Bellman (1961). As the dimension of the data d increases, it is more difficult to estimate the probability density function. The curse of dimensionality is applicable in many areas of scientific research. In kernel density estimate, if we use (2.9) as the KDE, the minimized AMISE is given by:

$$AMISE(h_{opt}) = O(n_d^{-\frac{1}{d+4}})$$

One needs to increase the sample size several folds to get the same rate of convergence of AMISE in higher dimensions. Suppose n_1 is the sample size of 1-dimensional kernel density estimate and n_d is the sample size of d-dimensional case. Let

$$O(n_1^{-\frac{4}{5}}) = O(n_d^{-\frac{4}{d+4}}) \text{ or } n_d = O(n_1^{\frac{d+4}{5}})$$

Table 2.1: Sample size needed to have the same rate of convergence of AMISE as 1 dimension

d	2	3	5	8	10
n_d	252	631	3982	63096	398108

If we set $n_d = n_1^{\frac{d+4}{5}}$ and let $n_1 = 100$, the sample size needed for different dimensions to have the same rate of convergence of AMISE of 1 dimension are shown in Table 2.3.4. Since MEM, REM and the method which will be introduced in the next chapter are based on the KDE, because of the curse of dimensionality, these methods are limited to low to moderate dimensions. In a "data-rich" environment, the MEM and REM can be applied to high-dimensional data.

Chapter 3

Multivariate Modality Inference

In this chapter, we develop the modality inferential framework. It is a local inference procedure that tests a specific pair of modes, $\mathbf{x_{m_1}}$ and $\mathbf{x_{m_2}}$, of the data. The hypothesis can now be written as

$$\begin{aligned} H_0 &: \mathbf{x_{m_1}} \text{ and } \mathbf{x_{m_2}} \text{ are unimodal} \\ H_a &: \mathbf{x_{m_1}} \text{ and } \mathbf{x_{m_2}} \text{ are bimodal.} \end{aligned} \tag{3.1}$$

This chapter is organized as follows: In Section 3.1, we propose the test statistic along with its asymptotic distribution to assess the significance of the hypothesis in (3.1). Section 3.2 discusses the choice of the bandwidth parameter for the inference procedure. The inference procedure introduced in this chapter combined with the mode hunting tool reviewed in Section 2.2 provides a comprehensive mode hunting and inference procedure. This procedure is summarized in Section 3.3. Furthermore, based on the modality inference, we can make the decision of whether or not to merge the two clusters of the modal clustering algorithm. Section 3.4 applies the mode hunting and inference procedure to the real flow cytometry data and swiss banknotes data as well as some simulated data sets. In addition, Section 3.5 discusses another possible test statistic of the inference that considers the ratio of the density of $\hat{\mathbf{x}}_\mathbf{m}$ and $\hat{\mathbf{x}}_\mathbf{s}$, which is the point on the ridgeline between $\hat{\mathbf{x}}_{\mathbf{m_1}}$ and $\hat{\mathbf{x}}_{\mathbf{m_2}}$ with minimum density..

3.1 Test Statistic and Its Asymptotic Distribution

We denote the one of $\mathbf{x_{m_1}}$ and $\mathbf{x_{m_2}}$ with lower density by $\mathbf{x_m}$. To test the hypothesis defined in (3.1), a natural choice is to compare the density of $\mathbf{x_m}$ against the density of the saddle point $\mathbf{x_s}$,

which is the point on the ridgeline between $\mathbf{x_{m_1}}$ and $\mathbf{x_{m_2}}$ with minimum density. We use Ridgeline EM (REM), which was reviewed in Section 2.2 to determine the saddle point $\hat{\mathbf{x}}_\mathbf{s}$. To identify the interested pair of modes, in practice, when several modes are identified by MEM, it starts with the one with the lowest density and its neighbor mode. Or, one can select the particular pair of modes based on the context of the study. After identifying $\mathbf{x_m}$ and $\mathbf{x_s}$, the hypothesis in (3.1) can be restructured as:

$$\begin{aligned} H_0 &: f(\mathbf{x_m}) = f(\mathbf{x_s}); \\ H_a &: f(\mathbf{x_m}) > f(\mathbf{x_s}). \end{aligned} \tag{3.2}$$

We use $\hat{f}(\hat{\mathbf{x}}_\mathbf{m})$ and $\hat{f}(\hat{\mathbf{x}}_\mathbf{s})$ to make the inference, where $\hat{f}(\cdot)$ is the kernel density estimate of $f(\cdot)$ and $\hat{\mathbf{x}}_\mathbf{m}$ and $\hat{\mathbf{x}}_\mathbf{s}$ are the estimated mode and saddle point of the KDE detected by the Modal EM (MEM) algorithm, which was reviewed in Section 2.2. We believe $\hat{\mathbf{x}}_\mathbf{m}$ is a good estimation of the modal region and is close to the true population mode, and same for the $\hat{\mathbf{x}}_\mathbf{s}$.

Theorem 3.1 *Using Gaussian kernel function, under* H_0 *of* (3.2)*, if* $h \to 0$*,* $nh^d \to \infty$ *and* $nh^{d+4} \to 0$ *as* $n \to \infty$*, then*

$$\sqrt{\hat{f}(\hat{\mathbf{x}}_\mathbf{m})} - \sqrt{\hat{f}(\hat{\mathbf{x}}_\mathbf{s})} \xrightarrow{d} N\left(0, \frac{1}{2nh^d}(\frac{1}{2\sqrt{\pi}})^d\right). \tag{3.3}$$

Proof: Using Property 2.2, Property 2.3 and density in 2.18, we can show that:

$$\sqrt{\hat{f}(\hat{\mathbf{x}}_\mathbf{m})} - \sqrt{\hat{f}(\hat{\mathbf{x}}_\mathbf{s})} \xrightarrow{d} N\left(0, \frac{\int \mathbf{K}^2(\mathbf{u})d\mathbf{u}}{2nh^d}\right). \tag{3.4}$$

$\hat{\mathbf{x}}_\mathbf{m}$ and $\hat{\mathbf{x}}_\mathbf{s}$ are correlated. However, based on Property 2.3, as long as $\hat{\mathbf{x}}_\mathbf{m} \neq \hat{\mathbf{x}}_\mathbf{s}$, $\hat{f}(\hat{\mathbf{x}}_\mathbf{m})$ and $\hat{f}(\hat{\mathbf{x}}_\mathbf{s})$ are asymptotically independent.

Next, we simplify the term $\int \mathbf{K}^2(\mathbf{u})d\mathbf{u}$. For univariate standard normal kernel function: $K(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, we have $\int K^2(x)dx = \frac{1}{2\sqrt{\pi}}$. Therefore, for d-dimensional multivariate nor-

mal kernel function: $K(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} e^{\{-\frac{1}{2}\sum_{j=1}^{d} x_j^2\}}$, we have $\int \mathbf{K}^2(\mathbf{x})d\mathbf{x} = (\frac{1}{2\sqrt{\pi}})^d$. Thus we prove the theorem. This is the test statistic of Hypothesis (3.2) and its asymptotic distribution.

3.2 Choice of the Bandwidth Parameter

In order to use the asymptotic distribution (3.3), the conditions of Property 2.2 must be satisfied. To satisfy the conditions $nh^d \to \infty$ and $nh^{d+4} \to 0$, the bandwidth parameter h^* should be chosen as

$$h^* = \left(\frac{c}{n}\right)^{\frac{\gamma}{d+4}} \text{ where } 1 < \gamma < 1 + \frac{4}{d}$$

However, the range of γ, $1 < \gamma < 1 + \frac{4}{d}$, is still wide if the dimension of the data is not high, e.g., $1 < \gamma < 3$ if $d = 2$. Theoretically, the bias of the KDE can be negligible if γ is within this interval. However, in practice, the selection of γ affects the variance-bias trade off, which affects the inference dramatically. We demonstrate the phenomenon using *logctA20* data set. The description of the data set can be found in R package *Modalclust*, which will be described in the next chapter. *logctA20* is a two-dimensional data with 2166 observations. The scatter plot of the data is shown in Figure 3.1. Using the normal reference rule, the bandwidth parameter used for the MEM is:

$$h_{nrr} = \left\{\frac{4}{(2+2)\times 2166}\right\}^{\frac{1}{2+4}} = 0.278$$

Using the MAC to cluster the data, the output shows that there are four major clusters. Figure 3.2 shows the clustering output as well as the modes, saddle points and ridgeline between the modes. The next step is to test if the four modes are significant. We consider three tests for the three adjacent pairs of modes: the test of Mode 4 against Mode 3, Mode 3 against Mode 1 and Mode 2 against Mode 1. As mentioned at the beginning of this section, in order to use the asymptotic distribution (3.3), we should choose the bandwidth parameter h so that it converges *faster* than the optimal rate. Thus, we should choose

$$h^* = \left(\frac{c}{n}\right)^{\frac{\gamma}{d+4}} \text{ where } 1 < \gamma < 1 + \frac{4}{d}$$

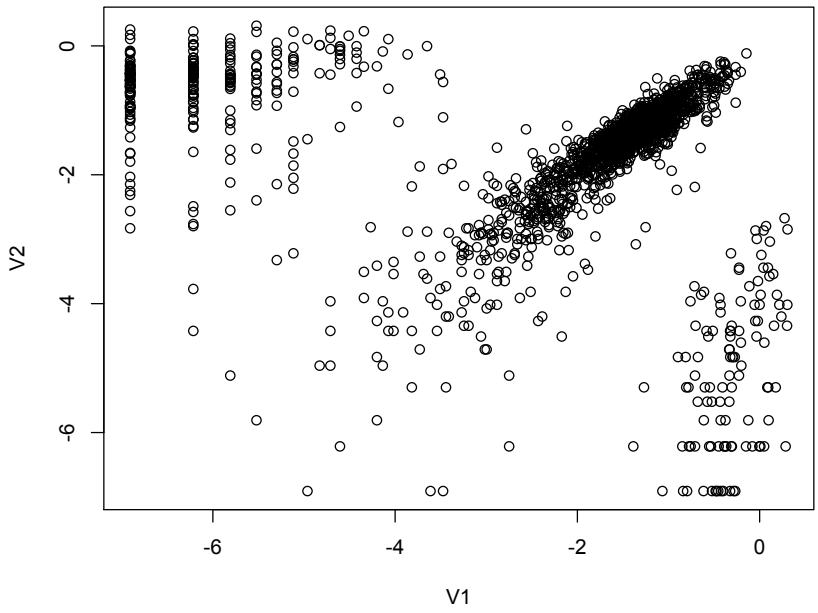

Figure 3.1: Scatter plot of *logctA20* data

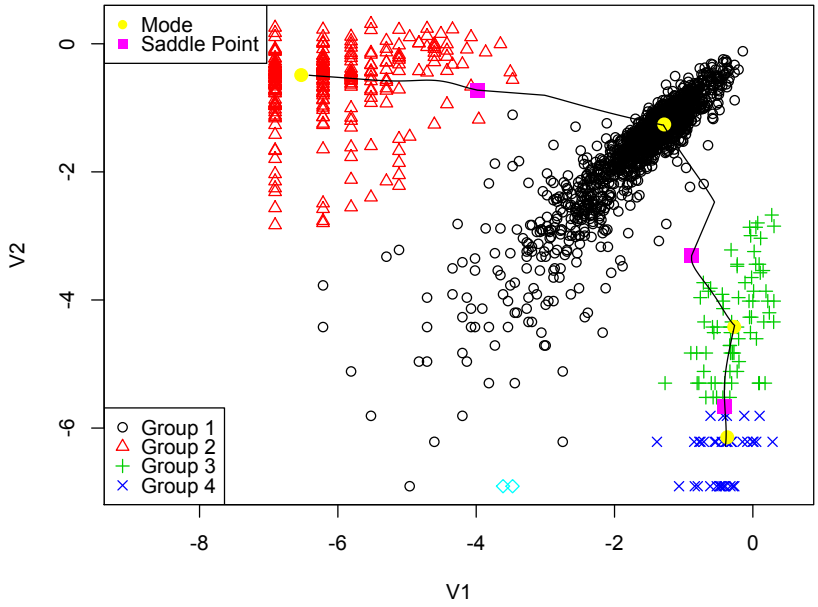

Figure 3.2: Mode, saddle point and ridgeline of *logctA20* data

Figure 3.3 provides the plot of the p-value of the modality test against the choice of the γ. It clearly demonstrates that the value of the γ affects the conclusion of the inference. For Mode 2, the lower values of γ lead to rejecting the null hypothesis, whereas the higher values of γ lead to not rejecting the null hypothesis, In practice, if we choose a small value of γ, the bias term still exists, even though asymptotically it will converge to 0. If we choose a large value of γ, the variance is relatively large and could mislead the conclusion. We suggest to use a small value of γ, which will lead to a large value of h^*.

Remark: Recall in Section 2.3, we reviewed that the bias term in Property 2.1 is $b = \frac{h^2}{2}\mu_2(K)tr(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}'}))$. Note that $b < 0$ at $\hat{\mathbf{x}}_\mathbf{m}$ and $b > 0$ at $\hat{\mathbf{x}}_\mathbf{s}$. Thus, under H_0 of hypothesis in (3.2), the expectation of the test statistic is negative if the bias exists. Therefore, it makes the test conservative.

From the analysis in Figure 3.3, we conclude that Mode 4 is not significant, while Mode 2 and Mode 3 are. Group 4 can be merged with Group 3. The final plot after merging Group 4 with Group 3 is given in Figure 3.4. Note that the resulting modes are all significant.

When we have several mode candidates and when we want to inference on the entire distribution to see how many significant modes, there is a multiplicity issue. One can refer Dmitrienko et al. (2010) for some method to control the overall Type I error rate. We focus on the local significance and do not provide a overall significance of the final result.

3.3 The Procedure of the Mode Hunting and Inference

The inference procedure proposed in the previous section, along with the MEM and REM reviewed in Section 2.2, provides a comprehensive tool for mode hunting and follow-up inference of a data set. In this section, we summarize the procedure.

Step 1: Sphering transformed the data;

Step 2: Use KDE to estimate the density of the data with bandwidth parameter chosen by some

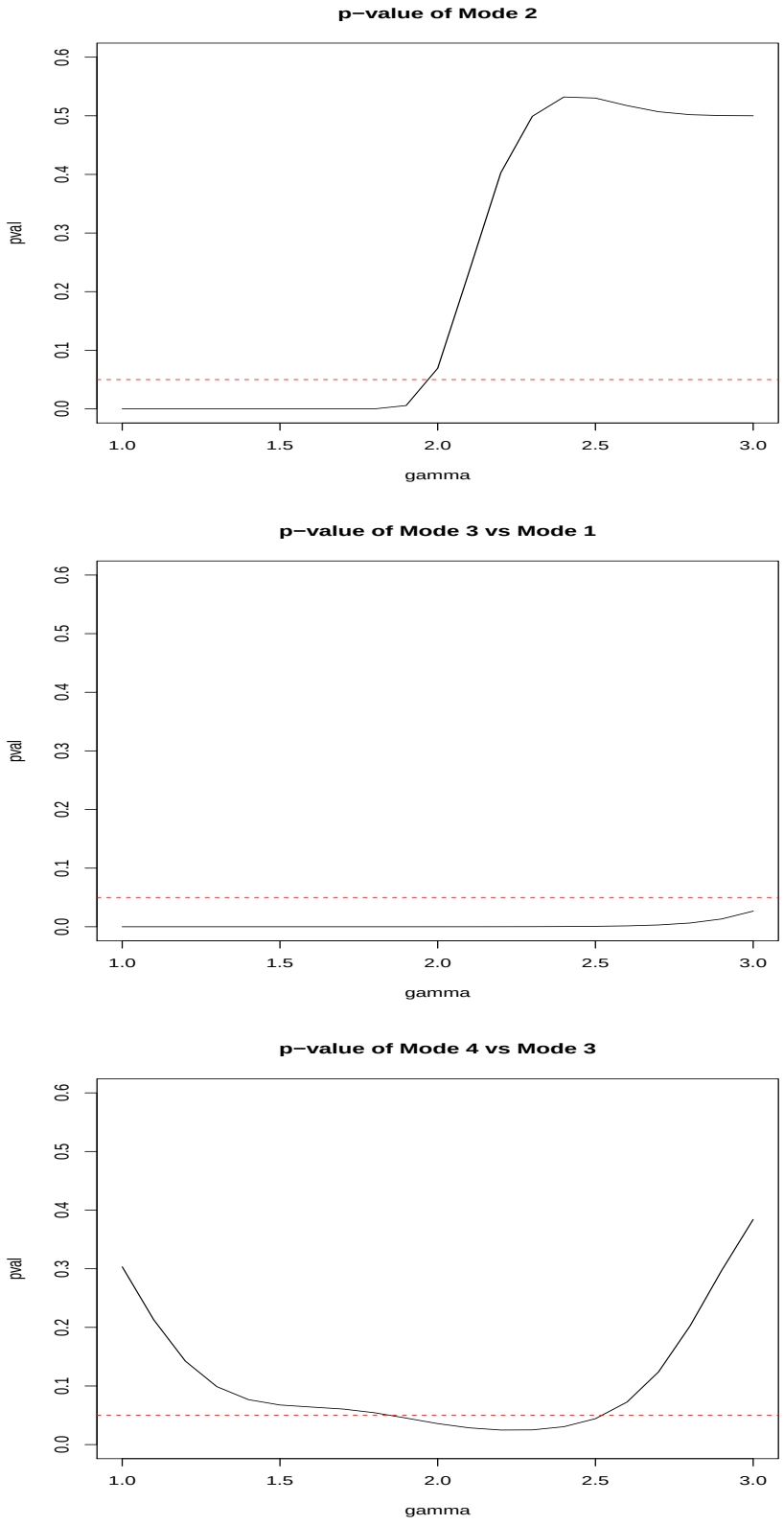

Figure 3.3: p-value of modality inference against γ

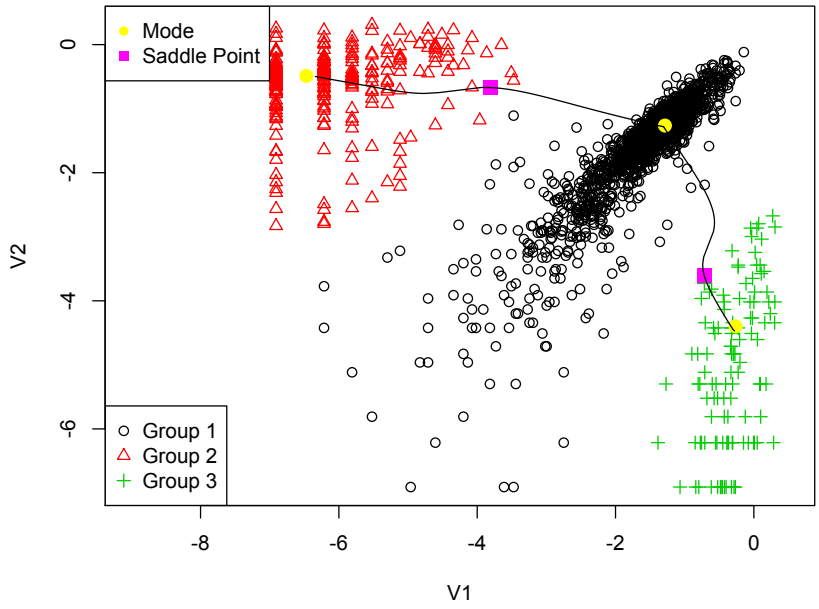

Figure 3.4: Mode, saddle point and ridgeline of the example data after merging

standard method, e.g., the normal reference rule, etc;

Step 3: Identify the modes of KDE using MEM. After determining the pair of modes $\hat{\mathbf{x}}_{\mathbf{m_1}}$ and $\hat{\mathbf{x}}_{\mathbf{m_2}}$, identify the corresponding saddle point $\hat{\mathbf{x}}_{\mathbf{s}}$ by REM;

Step 4: Use $h = (\frac{c}{n})^{\frac{\gamma}{d+4}}$ where $1 < \gamma < 1 + \frac{4}{d}$ and $c = \frac{4}{d+2}$ to calculate $\hat{f}(\mathbf{x_m})$ and $\hat{f}(\mathbf{x_s})$. In particular,we suggest to choose $\gamma = 1.1$ when d is small.

Step 5: Make the inference of the Hypothesis (3.2) based on the asymptotic distribution (3.3).

3.4 Application

This section provides the application of the modality inferential framework on some real and simulated data sets. We start by providing a description of the data sets and follow up by providing the conclusion of the inferential framework.

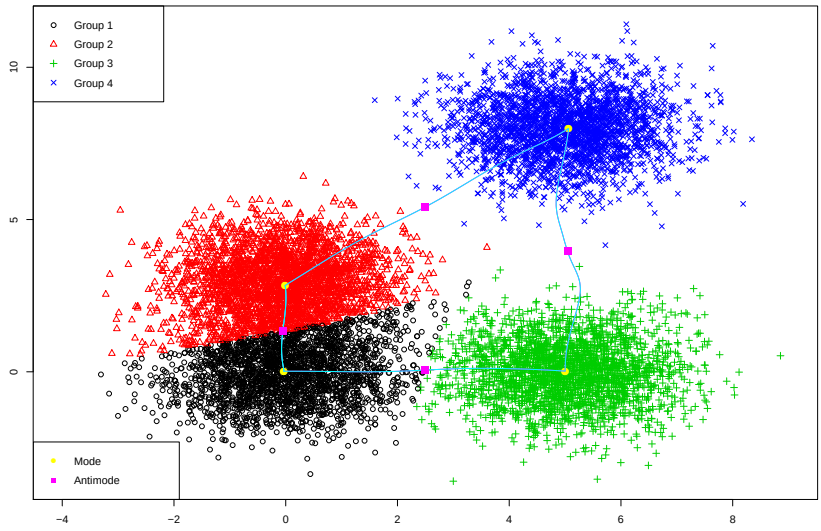

Figure 3.5: The first layer of *four discs* data

3.4.1 Four Discs

The *Four Disks* data is a simulated data. It contains 10000 observations and the data is a mixture of four bivariate normal distributions. The mean vectors are $\boldsymbol{\mu_1} = (0, 0)'$, $\boldsymbol{\mu_2} = (0, 3)'$, $\boldsymbol{\mu_3} = (5, 0)'$, $\boldsymbol{\mu_4} = (5, 8)'$. The data contains multiple layers of the clusters. There are three main clusters with one of them having two sub-clusters. By the simulation design, the Group 1 and 2 shown in Figure 3.5 are two distinct groups. The p-value of Mode 1 compared with Mode 2 is 0.0194. However, Group 1 and 2 are relatively close compared to the other groups. After merging these two groups together, the resulting clusters and the ridgeline between each pair of modes are shown in Figure 3.6. The multiple layers of clusters are common in real life application. The decision of how many clusters the data has is often related to the application area and research question.

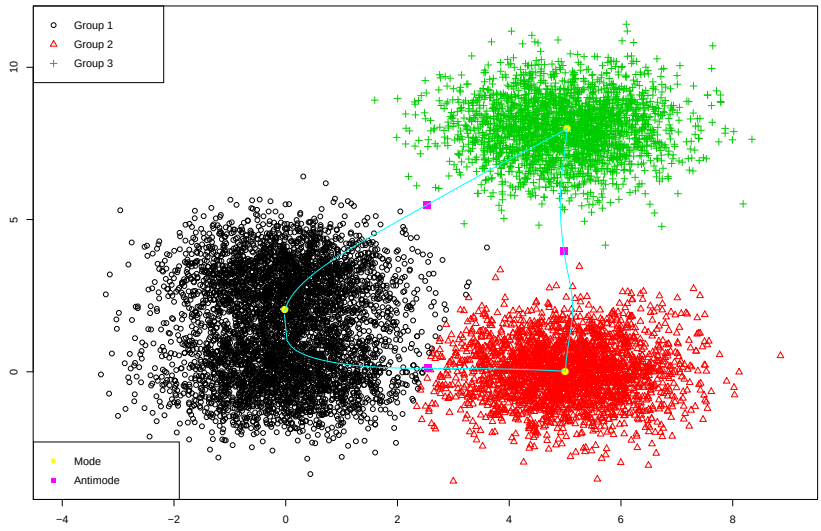

Figure 3.6: The second layer of *four discs* data

Table 3.1: Cluster size of *two half discs* data

Group	1	2	3	4	5	6
Size	82	2	314	2	314	86

3.4.2 3-Dimensional *Two Half Discs*

The 3-dimensional *two half discs* is another simulated data set with 800 samples. It is formed by two half discs with equal size, i.e. 400 samples for each disc. Using $n = 800$ and $d = 3$ for

$$h_{nrr} = \left\{ \frac{4}{(d+2)n} \right\}^{\frac{1}{d+4}},$$

we get $h_{nrr} = 0.373$. The clustering output using the MAC with $h = 0.373$ is shown in Figure 3.7. There are 4 major clusters and the number of samples of each cluster is given in Table 3.1 The inference between some major clusters is carried out and the resulting p-values are given in Table 3.2. It is straightforward to conclude that Group 1 is not significantly distinct from Group 3, and Group 6 is not distinct from Group 5. Group 3 is significantly separated from Group 5 and 6. Group 5 is significantly separated from Group 1 and 3. Thus, we get the conclusion that there are two main groups in this data set.

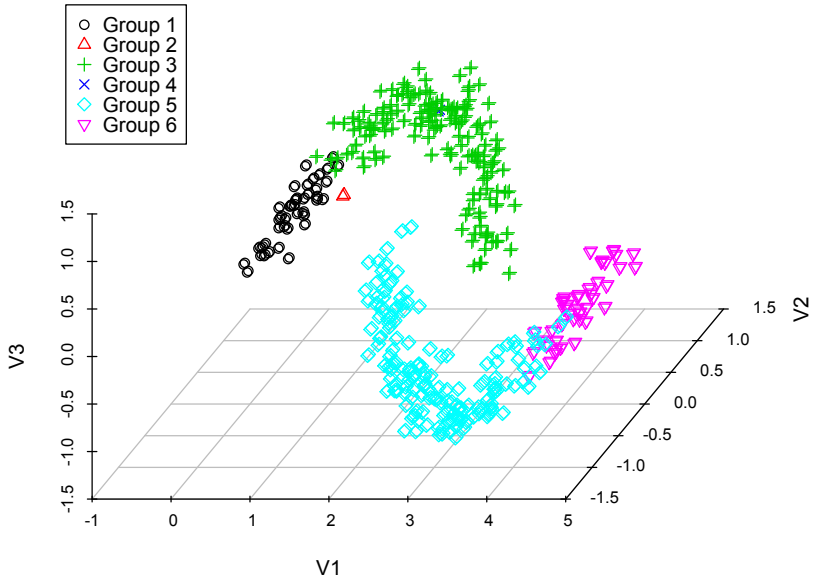

Figure 3.7: 3-D *two half discs* data

Table 3.2: p-value of modality inference on *two half discs* data

Pair of Mode	p-value
1 vs 3	0.0876
1 vs 5	7.805e-5
3 vs 5	4.970e-13
3 vs 6	2.716e-4
5 vs 6	0.0611

3.4.3 Flow Cytometry Data

Flow cytometry is a technology that simultaneously measures and then analyzes multiple physical characteristics of single cell, as they flow in a fluid stream through a beam of light. Flow cytometry is one of the most commonly used platforms in clinical and research labs worldwide. It is used to identify and characterize types and functions of cell populations e.g, dead or live cells, in a sample by measuring the expression of specific proteins on the surface and within each cell.

Flow cytometry data consists of per cell measurements (or events) in the form of scattering of light and fluorescence intensity from the fluorophore-conjugated markers. In a typical flow data analysis workflow, a human analyst visually inspects 2-dimensional scatter plots of a sample, where the dimensions could be scatters, marker intensities, or a combination of these, and it demarcates (or gates) specific populations of interest such as live cells, lymphocytes, etc. Often, gates are drawn around visually discernible congregations of events. For instance, for live gating, the dead cells or debris could be discerned by their small cell size and granularity, which appear as a distribution of points with low forward- and side-scatter values. Forward-Scatter light (FSC) and Side-Scatter light (SSC) reflects two features of the cells and forms a two-dimensional scatter plot. FSC is proportional to cell-surface area or size. SSC is proportional to cell granularity or internal complexity. The manual approach to gating is, however, labor-intensive and subjective, and gating results can vary considerably from one analyst to another. Ray and Pyne (2012) have used the MAC to gate flow cytometry data. However, the inference is distinctly missing. Figure 3.8 is one example of flow cytometry data. The data contains 4905 cells. In this plot, the dead cells have a relatively smaller size compared with the live cells, which shows that the dead cells have a smaller value of FSC and SSC. In the scatter plot, the dead cells are at the bottom left corner. For this data set, using the inference procedure introduced in Section 3.3, we applied the MAC on the data with $h = (1/4905)^{1/6} = 0.243$ and got the two major clusters. It is suspected that the cluster on the bottom left represents the dead cells, while the rest are the live cells. The p-value of the mode existence inference is $p < 0.0001$. Thus, the procedure can automatically identify the dead cell population distinctly from the live cells.

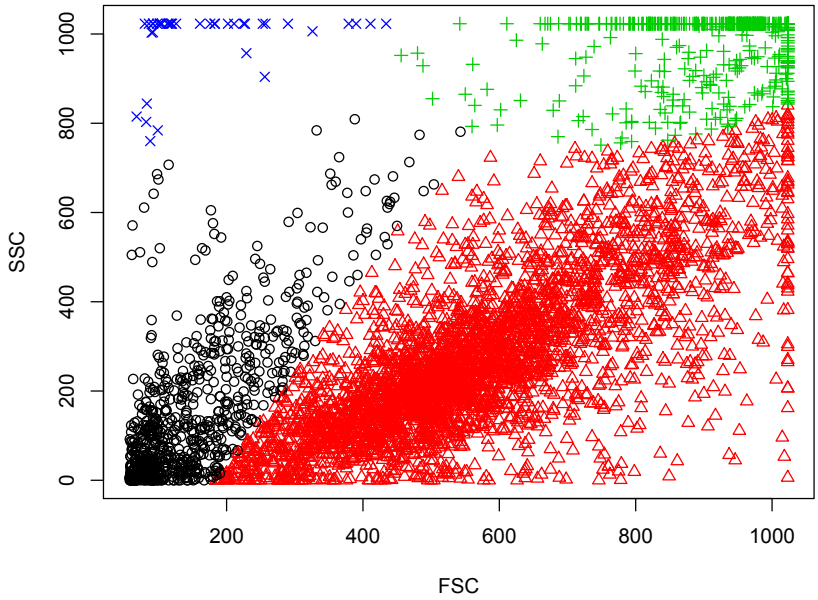

Figure 3.8: One example of flow cytometry data clustered by MAC

3.4.4 Swiss Banknotes

The data set contains 6 measures of 200 Swiss banknotes, where 100 are real and 100 are counterfeit. The 6 measures are:

X_1: Length of the bank note,

X_2: Height of the bank note, measured on the left,

X_3: Height of the bank note, measured on the right,

X_4: Distance of inner frame to the lower border,

X_5: Distance of inner frame to the upper border,

X_6: Length of the diagonal of the inner image.

All measurements are in millimeters. The original banknote image and the measurements are shown in Figure 3.9. In this data set, we know the truth of whether the banknote is real or forged. More information about the data set can be found in Flury and Riedwyl (1988). We use the *spectral degrees of freedom* concept, which was proposed by Lindsay et al. (2008), and supply $h = 1.022$

Figure 3.9: 6 Measurements of *Swiss banknote* data

Table 3.3: MAC output of emphSwiss banknotes data

	real	counterfeit
Group 1	97	4
Group 2	1	0
Group 3	1	0
Group 4	1	96

for the MAC. The MAC output shows there are two major clusters and can capture the two groups well. The output is shown in Table 3.3. Group 1 and Group 4 are the two major clusters. Using $h = 0.517$, the p-value of the corresponding modality inference is 0.001774, which indicates the two clusters are clearly separated.

3.5 Ratio Statistic

This section introduces an alternative test statistic using the ratio of the densities of the mode and the saddle point, which is defined as:

$$r(\mathbf{x}) = \frac{\hat{f}(\hat{\mathbf{x}}_\mathbf{s})}{\hat{f}(\hat{\mathbf{x}}_\mathbf{m})}.$$

The motivation behind this approach comes from a possible drawback of the test statistic defined in Section 3.1. The method introduced in Section 3.1 considered the difference of density heights

between the investigated mode and corresponding saddle point. This sometimes might be misleading. For example, in Figure 3.10, the points represented by ■ are the modes of interests and the points represented by ⧫ are the corresponding saddle points in the two densities. In the left picture, the mode has the density 0.1 and the density of the saddle point is 0.05. These modes and saddle points in the right picture have densities 0.15 and 0.1 respectively. Intuitively, the mode in the left picture is more significant compared to the mode in the right one, even though the difference between the mode and saddle point in these two examples are the same. The test statistic $r(\mathbf{x})$ is more appropriate in this case.

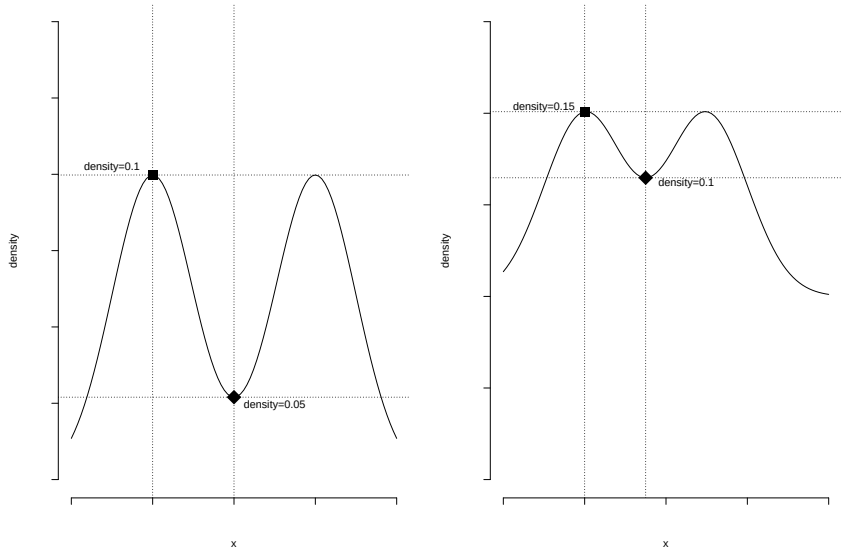

Figure 3.10: The pair of mode and saddle point have same difference but different ratio

In this situation, using the ratio of the two densities is more appropriate than using the difference. However, it is difficult to find the exact or asymptotic distribution of the test statistic $r(\mathbf{x})$ or any transformations. We note that $r(\mathbf{x})$ is very similar to the test statistic $\hat{SB}(\alpha)$ proposed by Burman and Polonik (2009), which has been reviewed in Section 2.1. But, there are some key differences. The ratio we proposed is the ratio of the densities of the saddle point on the ridgeline and the mode with lower density. In Burman and Polonik (2009), the authors used the points on the segment line

between the two modes. We believe it is more appropriate to use the points on the ridgeline rather than the points on the segment line. The other obvious difference is that our method is based on the KDE. Burman and Polonik (2009) considered K-nearest neighbor (KNN) as the estimation of the density and derived the asymptotic distribution of the proposed test statistic. We propose a bootstrap approach and use uniform distribution as the reference distribution and use the empirical null distribution for developing the inferential framework.

The choice of the reference distribution is important but not straightforward. There are many distributions that are unimodal. The uniform distribution has been used as the reference distribution by many researchers. Tibshirani et al. (2001) argued that among the class of the unimodal distribution, the uniform distribution is the most likely to produceR spurious clusters. In this paper, the authors provided two ways of generating the uniform reference data. One is to simulate each dimension uniformly over the the range of the corresponding dimension. The other approach is to generate the reference data from a uniform distribution over a box aligned with the principal components of the data. As we introduced before, the data will be sphering transformed first, and then analyzed; therefore, these two approaches will be the same.

The range of the uniform distribution will affect the null distribution. When calculating the KDE, it mainly relies on the points surrounding the mode and saddle point. The reference data becomes more sparse with the fixed sample size if a wider range of the uniform distribution is used. Then, the number of points surrounding the mode and saddle point are reduced. This can be considered as reducing the sample size. In our research, we simulated the reference data from the range of $(-2, 2)$ in each dimension. The reason for this is that after sphering transformation, the variance of each dimension becomes 1. Under the standard normal distribution, 95% of the data lies in the range of $(-2, 2)$.

It is not computationally expensive to simulate the null distribution. Since the reference data is uniformly distributed, the ratio of the estimated density height between the saddle point and mode

Table 3.4: Critical value of ratio statistic for $d = 2$

sample size	1%	5%	10%
200	0.575	0.679	0.739
400	0.634	0.723	0.781
600	0.659	0.744	0.796
800	0.687	0.761	0.808
1000	0.697	0.777	0.821
1500	0.726	0.802	0.843
2000	0.748	0.814	0.853
3000	0.769	0.836	0.868
5000	0.801	0.858	0.888
10000	0.845	0.886	0.910

is the same as the ratio of any two points not closed to the boundary of the data. Note that the bandwidth parameter for the calculation is $h = h_{opt}$. Table 3.4 gives the critical values of ratio statistic under different sample sizes at significant levels 1%, 5% and 10% for $d = 2$. There are many controversial issues regarding the method introduced in this section. The method to simulate the null distribution introduced in this section works for the hypothesis of comparing unimodal against bimodal distributions only. Even for this hypothesis, choosing the uniform distribution as the reference distribution might not be appropriate. The range of the uniform reference data also affects the null distribution, and further affects the inference results. Some more rigorous research work needs to be done in this direction.

Chapter 4

Parallel Computing of Hierarchical Mode Association Clustering

In Section 2.2, we reviewed the Modal EM (MEM) algorithm. One clustering method follows the MEM naturally, which was also introduced in Li et al. (2007). If we start this algorithm from each data point, we can cluster the data that converges to the same mode into one group. It is named the *Mode Association Clustering* (MAC). Based on the fact that a larger bandwidth parameter h produces a smoother KDE, we can get the hierarchical MAC (HMAC) if we choose a sequence of ascending values of h.

This chapter introduces the methodology of parallel computing of the HMAC (PHMAC). It is organized as follows: Section 4.1 reviews the details of the HMAC algorithm. Section 4.2 introduces the method of PHMAC. We provide the comparisons to show that the parallel computing can dramatically increase the computing speed. The R package *Modalclust* is created to implement the developed algorithm. Section 4.3 describes the usage of *Modalclust* package.

4.1 HMAC

This section reviews the details of the HMAC algorithm. Based on the fact that the larger value of bandwidth parameter h leads to the smoother KDE defined in (2.9), i.e., fewer modes/clusters, more points tend to climb to the same mode by MEM, which was reviewed in Section 2.2. This suggests a natural approach of hierarchy (or "nesting") of the MAC clustering. Given a range of bandwidths $h_1 < h_2 < \cdots < h_L$, the clustering can be performed in an aggregated manner. G_l is defined as

the collection of all the distinct modes obtained by MAC using the σ_l. First, we cluster the data by MAC using h_1, consequently form the collection G_1. For any $l > 1$, we use MAC to cluster the elements in G_{l-1} with $h = h_l$. The modes identified at this level form the collection G_l. We repeat this procedure across all l's. This procedure preserves the hierarchy of clusters, and thus it is named the Hierarchical Mode Association Clustering (HMAC). We summarize the HMAC procedure as follows:

Step 1: Sphering transform the data $\mathbf{X}$ to form a new data set $\mathbf{Y}$.

Step 2: Start with the data $G_0 = \{y_1, ..., y_n\}$ and set level $l = 0$ and initialize the mode association of the ith data point as $\mathcal{P}_0(y_i) = i$.

Step 3: $l \leftarrow l + 1$.

Step 4: Form the KDE in (2.9) using $h = h_l$.

Step 5: Cluster the elements in G_{l-1} by MAC using the KDE in (2.9) with $h = h_l$. Let the set of distinct modes obtained be G_l.

Step 6: If $\mathcal{P}_{l-1}(y_i) = k$ and the kth element in G_{l-1} is clustered to the k'th mode in G_l, then $\mathcal{P}_l(y_i) = k'$. In other words, the cluster of y_i at level l is determined by its cluster representative in G_{l-1}.

Step 7: Stop if $l = L$, otherwise go back to Step 2.

Step 8: Transform $\mathbf{Y}$ back to $\mathbf{X}$.

4.2 Parallel of HMAC

In this section we develop the method of parallel computing of HMAC (PHMAC) and its application together with some comparisons of performance of the parallel and non-parallel approach. The MAC approach is computationally expensive when the number of objects n becomes large. It requires that we use the MEM for each data point to find its local mamximum of the density. Note that for the HMAC, the steps for the level $l = 2$ onwards only need to start the MEM from the modes

of the previous level G_{l-1}, and hence the computational cost does not increase at the rate of n. Fortunately the MAC approach provides a natural framework for a "divide and conquer" clustering algorithm. One can simply divide the data into m partitions, perform modal clustering on each of those partitions, and pool the modes obtained from each of these partitions to form a collection G and apply the HMAC onward. If the user has access to several computing cores of the same machine or several processors of a shared memory computing cluster, the "divide and conquer" algorithm can be seamlessly parallelized. The PHMAC procedure is summarized as follows:

Step 1: Sphering transform the data $\mathbf{X}$ to form a new data set $\mathbf{Y}$.

Step 2: Let $G_0 = \{y_1, ..., y_n\}$. Divide the data (n objects) into m partitions G_o^j randomly, $j = 1, 2, \ldots, m$.

Step 3: Perform HMAC on each of these subsets at the lowest resolution, i.e., using h_1 and get the modes G_1^j, $j = 1, 2, \ldots, m$.

Step 4: Pool the modes from each subset of data to form $G_1 = \bigcup_{j=1}^{m} G_1^j$

Step 5: Perform HMAC starting from Step 2 and $l = 1$ and obtain the final hierarchical clustering.

Step 6: Transform $\mathbf{Y}$ back to $\mathbf{X}$.

Figure 4.1 shows one PHMAC example on the graph. In this figure, (a) shows the simulated data with four clusters along with the contour plot, where the color indicates the final clustering using PHMAC; (b) shows the four random partitions of the unlabeled data along with the modes (red asterisks) at each partition; (c) shows the mode obtained from the four partitions; (d) shows the final modes (green triangles) starting from the modes of the partitioned data. A demonstration of different steps of parallel clustering with four random partitions is given in Figure 4.1. The original data set is partitioned into 4 random subsets, and initial modal clustering is performed within the partitions. In the next step, the modes of each of these partitions are merged to form the overall modal clusters in Figure 4.1(c).

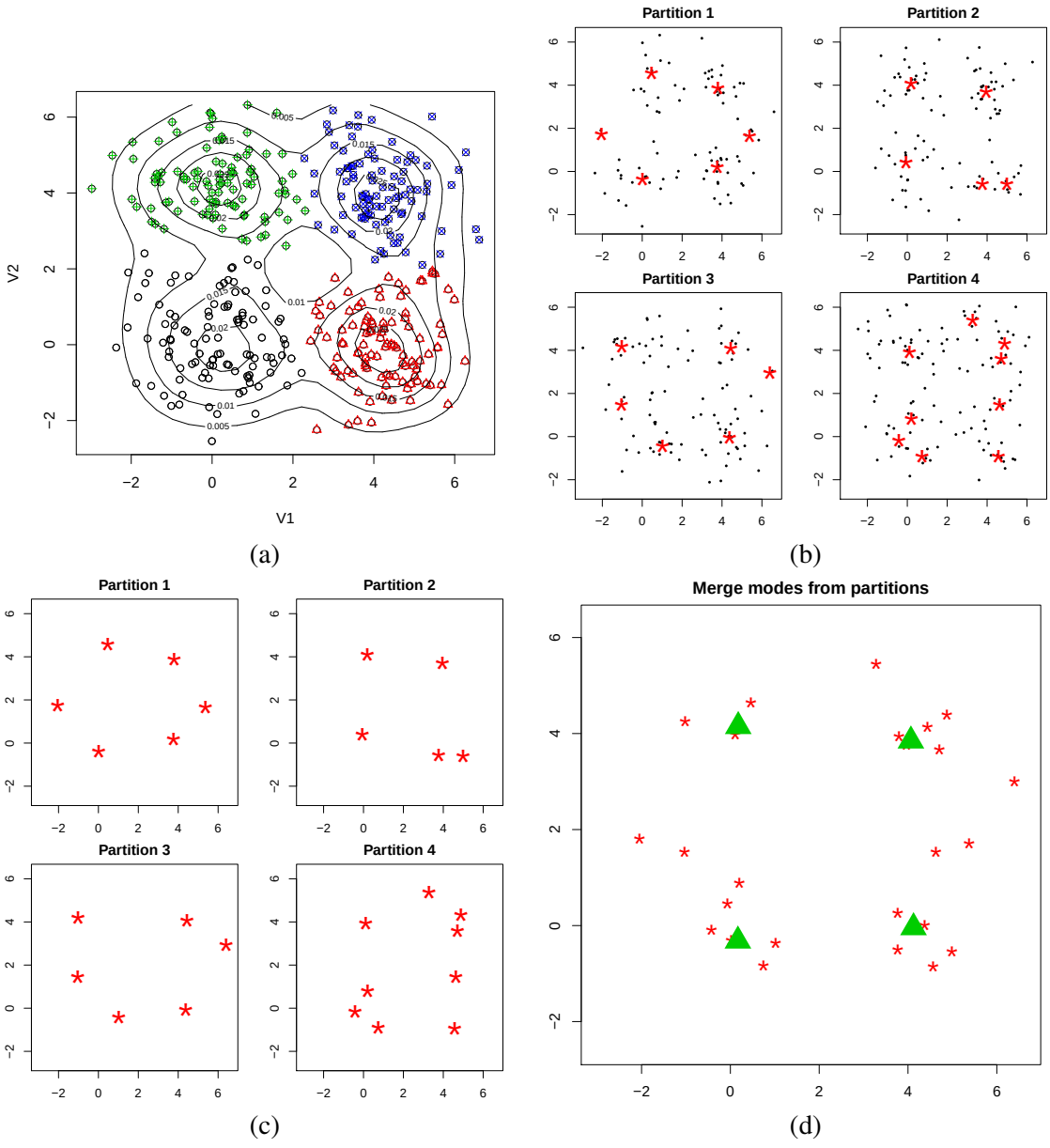

Figure 4.1: Steps in parallel HMAC procedure for a simulated data set

Modes have a natural hierarchy and it is computationally easy to merge modes from different partitions. In practice, we need to decide the best choice of the partition and how many partitions to use. In this section, we provide some guidelines regarding the choices, without exploring their quality in details. In the absence of any other knowledge, one should randomly partition the data. Other choices include partitioning data based on certain coordinates which form a natural clustering, and then taking products of a few of those coordinates to build the overall partition. This strategy might increase the computational speed by restricting the modes within a relatively homogeneous set of observations. Another choice might be to sample the data and build partitions based on the modes of the sampled data.

The PHMAC we proposed uses parallel computing at the first level of HMAC and then use non-parallel computing from the second level onwards. Therefore, the number of partitions to minimize the computational time is a complex function of the number of available processors, the number of observations and the bandwidth parameter of the KDE. If one uses too many partitions, one might speed up the first step, but would have the risk of ending up with too many modes for the next level, where the hill climbing is done from the collection of modes from each partition with respect to the overall density. In contrast, for a large n, if one chooses too few partitions or no partitions, this would lead to a huge computational cost at the first step. Moverover, the choice of the smoothing parameter will also determine how many modes one needs to start from at the merged level.

We compare the computing speed of parallel versus serial clustering using $1, 2, 4, 8$ and 12 multi-core processors. Tests were performed on a 64 bits 4 Quad Core AMD 8384 (2.7 Ghz each core), with 16 GB RAM running Linux Centos 5 and R version 2.11.0 From Table 1, it is clear to observe that parallel computing significantly increases the computing speed. Because the KDE in (2.9) is a sum of kernels centered at every data point, the amount of computation needed to identify the mode associated with a single point grows linearly with n. The computational complexity of clustering all the data by MAC is thus quadratic in n. Suppose we have p processors, then the computing complexity for the MAC is n^2 and for parallel computing of MAC is thus $(n/p)^2$. However,

as discussed before, we can see that the computational speed is not a monotone decreasing function of the number of processors. Theoretically, it is true that more processors can reduce the computing complexity at the initial step. However, in practice, if the data set is not sufficiently large, using more processors may not save time, as it may produce a large number of modes for the next level of HMAC. When the $n = 10,000$ or $n = 50,000$, including more processors provides a dramatic decrease in computing time, whereas for $n = 2,000$, there is no clear decrease in time elapsed when using 4 or 8 processors instead of the maximum 12 processors. For $n = 50,000$, the decrease in computing time from 1 processor to using 12 processors is more than 40 fold (see Figure 4.2), but even if the user is able to use just two processors, the computing time is reduced to $1/3$ of how long a single processor would take. Even for $n = 20,000$, the advantage of using 12 processors is almost 30 fold, whereas for $n = 2,000$, the advantage is only 8 folds. In fact, the lowest time is actually clocked by 8 processors for $n = 2,000$, but using all 12 processors does not increase the time significantly. These comparisons show the potential for parallelizing the modal clustering algorithm and its inherent use for clustering high throughput data.

The R package *Modalclust* was created to implement the HMAC and PHMAC. There are also some plotting tools that give the user a comprehensive visual and understanding of the clustering result. Sources, binaries and documentation of *Modalclust* are available for download from the Comprehensive *R* Archive Network `http://cran.r-project.org/` under the GNU Public License.

Table 4.1: Comparison of computing time (elapsed time in seconds) using different number of processors

	Number of processors				
Data dimensions	1	2	4	8	12
n=2,000, d=2	56.58	17.01	7.84	6.91	8.02
n=2,000, d=20	323.16	128.13	112.42	190.11	250.22
n=2,000, d=40	730.18	560.16	687.79	764.29	753.36
n=10,000, d=2	3849.83	871.33	276.88	145.61	131.22
n=10,000, d=20	8410.96	1694.82	585.33	536.32	459.88
n=50,000, d=2	210295.29	71152.82	23383.61	11959.24	4875.64

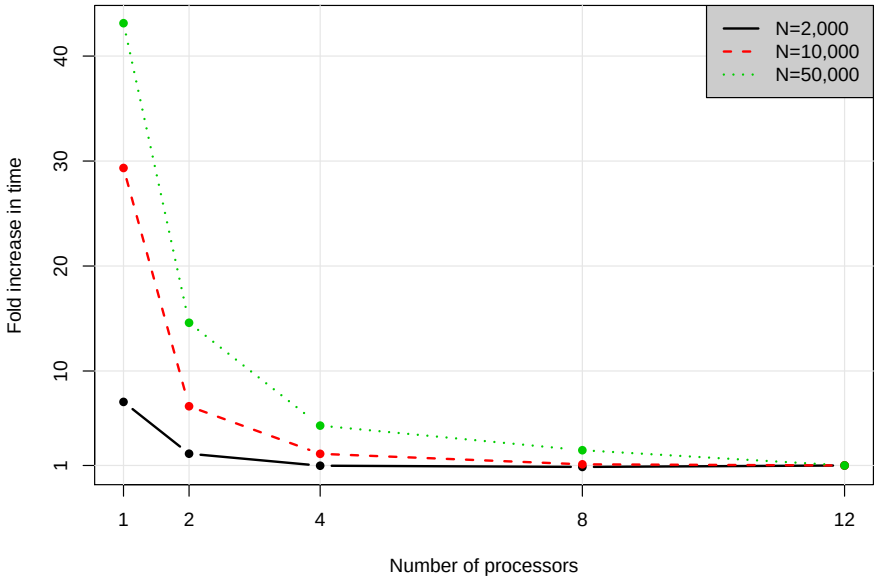

Figure 4.2: Comparison of fold increase in time for clustering two dimensional data of different sample sizes with respect to using 12 processors.

4.3 R Package *Modalclust*

In this section, we demonstrate the usage of the functions and plotting tools that are available in the *Modalclust* package.

4.3.1 Modal Clustering

First, we provide an example of performing modal clustering to extract the subpopulations in the *logcta20* data. The description of the dataset is given in the package. The scatter plot, along with its smooth density, is provided in Figure 4.3. First, we use the following command to download and install the package:

R> install.packages("Modalclust")

R> library("Modalclust")

Using the following command, we can get the standard (serial) HMAC and parallel HMAC using two processors for logctA20 data.

R> logcta20.hmac $<-$ phmac(logcta20,npart=1,parallel=FALSE)

R> logcta20p2.hmac $<-$ phmac(logcta20,npart=2,parallel=TRUE)

Both implementation results are given in Figure 4.4, which clearly identifies the three distinct subpopulations. Other model-based clustering methods, such as EM-clustering or K-means, could not capture the subpopulation structure, as the individual subpopulation is not a normal density. Distance based clustering method e.g., hierarchical clustering, with a range of linkage functions performed even worse.

By default, the function selects an interesting range of smoothing parameters with ten σ^2 values, and the final clustering only shows the results from the levels which produced merging from the previous level. For example, for the logcta20, the smoothing parameters chosen automatically are

R> logcta20.hmac$sigma

[1] 0.26 0.29 0.31 0.34 0.38 0.43 0.49 0.58 0.72 0.94,

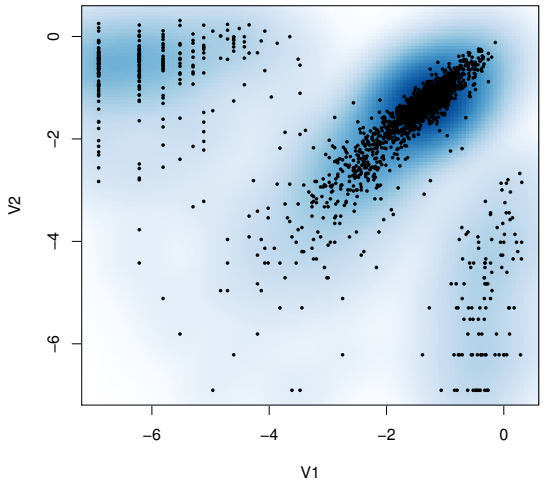

Figure 4.3: Smoothing scatter plot of *logctA20* data.

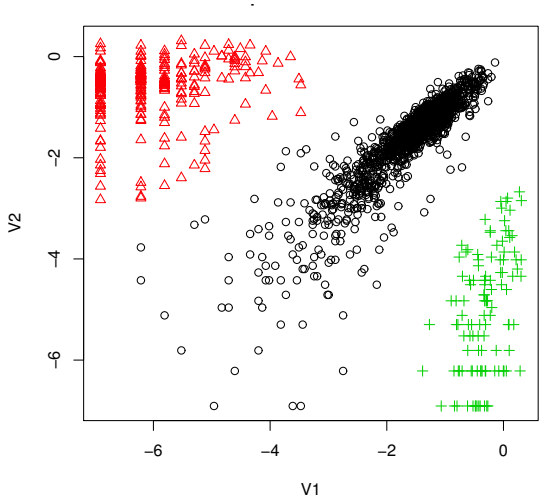

Figure 4.4: HMAC output of *logctA20* data.

which are chosen using the *spectral degrees of freedom* criterion introduced in Lindsay et al. (2008). Though we started with 10 different smoothing levels, the final clustering shows only 6 different levels along with a decreasing number of hierarchical cluster.

R> logcta20.hmac$level

[1] 1 2 3 3 3 4 4 4 5 6

R> logcta20.hmac$n.cluster

[1] 11 7 5 5 5 3 3 3 2 1

The user can also provide smoothing levels using the option *sigmaselect* in *phmac*. There is also the option of starting the algorithm from user defined modes instead of the original data points. This option becomes handy if the user wishes to merge clusters obtained from other clustering methods, e.g., EM-clustering or K-means.

4.3.2 Some Examples of Plotting

There are several plotting functions in *Modalclust*, which can be used to visualize the output from the function *phmac*. The plotting functions are defined on object class *hmac*, which is the default

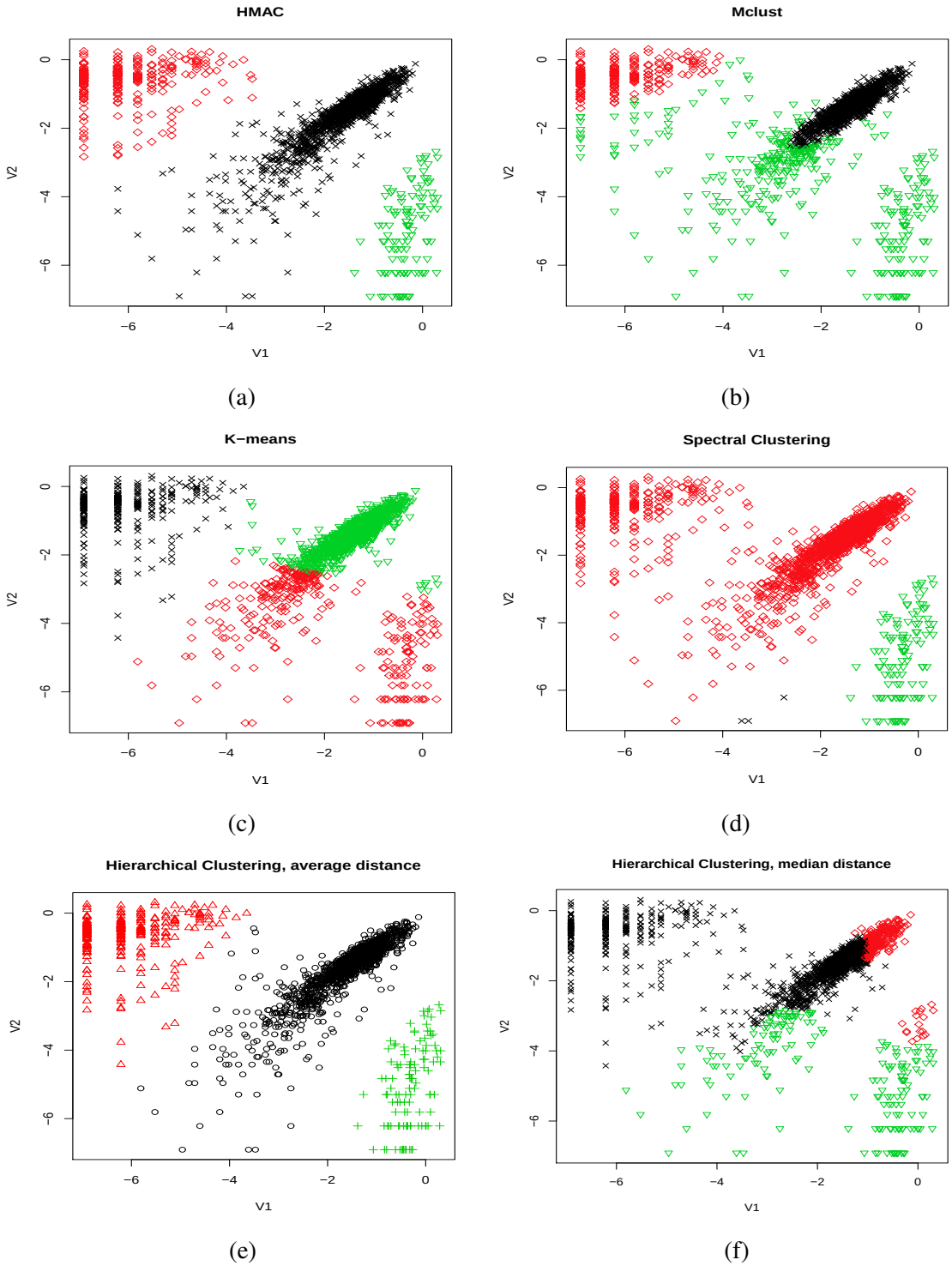

Figure 4.5: *logctA20* data clustering results by different methods

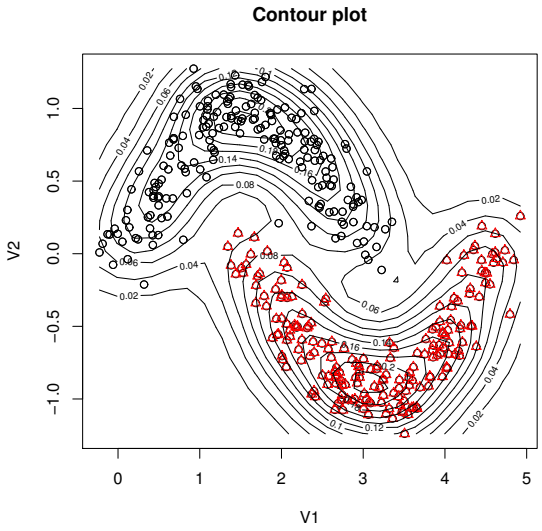

Figure 4.6: The scatter plot of *disc2d* data along with its probability contours.

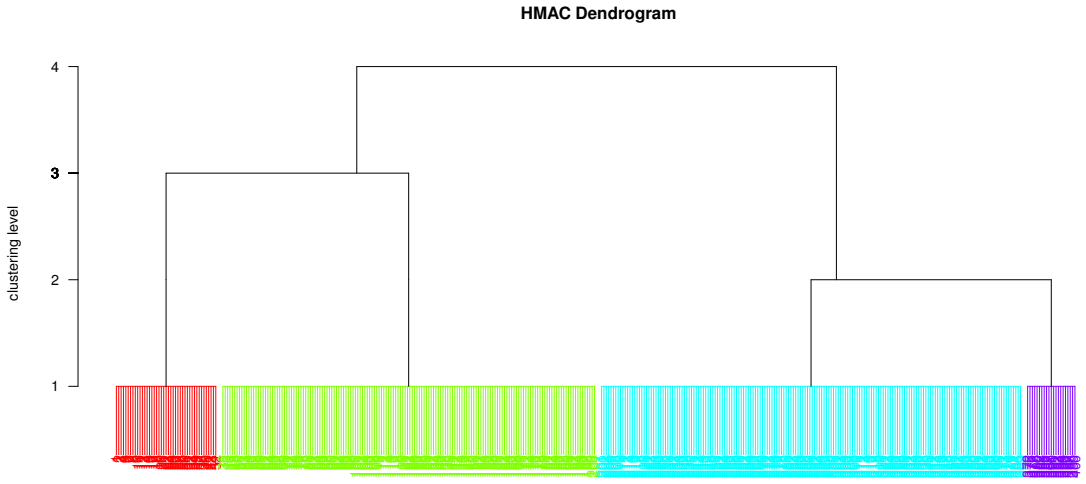

Figure 4.7: Hierarchical tree (Dendrogram) of *disc2d* data showing the clustering at four levels of smoothing.

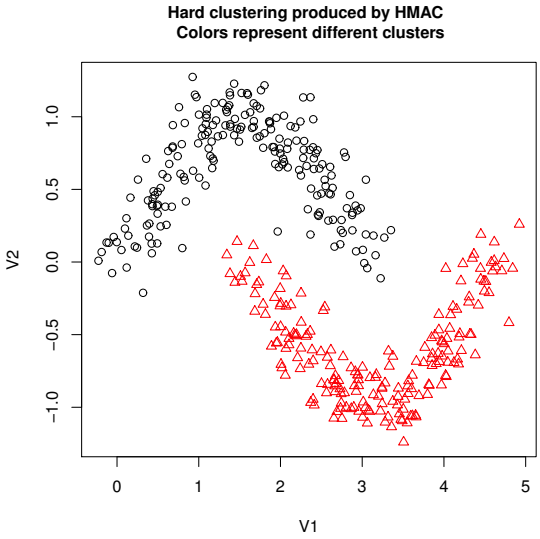

Figure 4.8: Hard clustering for *disc2d* data at level 3.

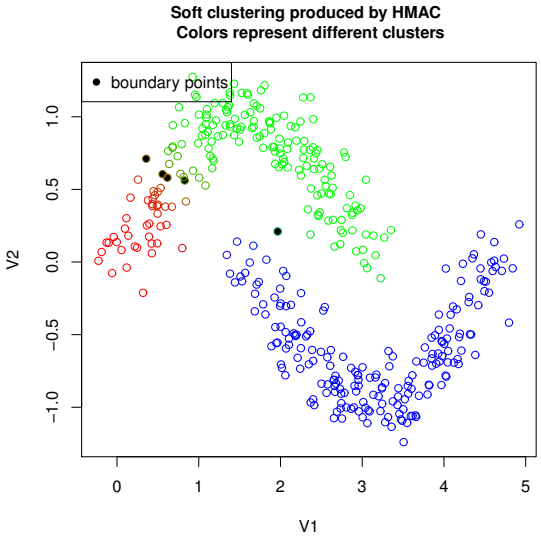

Figure 4.9: Soft clustering for *disc2d* data at level 2.

class of a *phmac* output. These plot functions will be illustrated through a data set named *disc2d*, which has 400 observations displaying the shape of two half discs. The scatter plot of disc2d along with its contour plot are given in Figure 4.6.

First, we introduce the standard *plot* function for an object of class "hmac". This unique and informative plot shows the hierarchical tree obtained from modal clustering. It can be obtained by

R> data("disc2d.hmac")

R> plot(disc2d.hmac)

The dendrogram obtained from the *disc2d* data is given in Figure 4.7. The y-axis gives the different levels, and the tree displays the merging at different levels. There are several options available for drawing the tree, including starting the tree from a specific level, drawing the tree only up to a desired number of clusters, and comparing the clustering results with user defined clusters. There are some other plotting functions that are designed mainly for visualizing clustering results for two dimensional data, although one can provide multivariate extensions of the functions by considering all possible pairwise dimensions. One can obtain the hard clustering of the data for each level using the command

R> hard.hmac(disc2d.hmac)

Alternatively, the user can specify the hierarchical level or the number of desired clusters, and obtain the corresponding cluster membership (hard clustering) of the data. For example, the plot in Figure 4.8 can be obtained by either of the following two commands:

R> hard.hmac(disc2d.hmac,n.cluster=2)

R> hard.hmac(disc2d.hmac,level=3)

Another function, which allows the user to visualize the soft clustering of the data, is based on the posterior probabilities of each observation belonging to the clusters at a specified level. For example, the plot in Figure 4.9 can be obtained using

R > soft.hmac(disc2d.hmac,n.cluster=3)

The plot enables us to visualize the probabilistic clustering of the three cluster model. A user can specify a probability threshold for assigning observations which clearly belong to a cluster or lie in the "boundary" of more than one cluster. Points having posterior probability below the user specified *boundlevel* (default value 0.4) are assigned as boundary points and colored in gray. In Figure 4.9, we have five boundary points among the 400 original observations. Additionally, at any specified level or cluster size, the *plot=FALSE* option in *hard.hmac* returns the cluster membership. Similarly, *plot=FALSE* option in *soft.hmac* returns a list that contains the posterior probability of each observation and boundary points.

R> disc2d.2clust $<-$ hard.hmac(disc2d.hmac,n.cluster=2,plot=FALSE)

R> disc2d.2clust.soft $<-$ soft.hmac(disc2d.hmac,n.cluster=2,plot=FALSE)

Finally, we demonstrate another very useful function for choosing a cluster dynamically from a two dimensional plot. The function *choose.cluster* allows the user to click on any part of a two dimensional plot, and dynamically select the cluster that point belongs to. One can start the display by invoking the command:

R> choose.cluster(disc2d.hmac,n.cluster=2),

which will open up a graphical window with the scatter plot as displayed in the left panel of Figure 4.10. After the user clicks a point anywhere near the upper disc, the points in the cluster con-

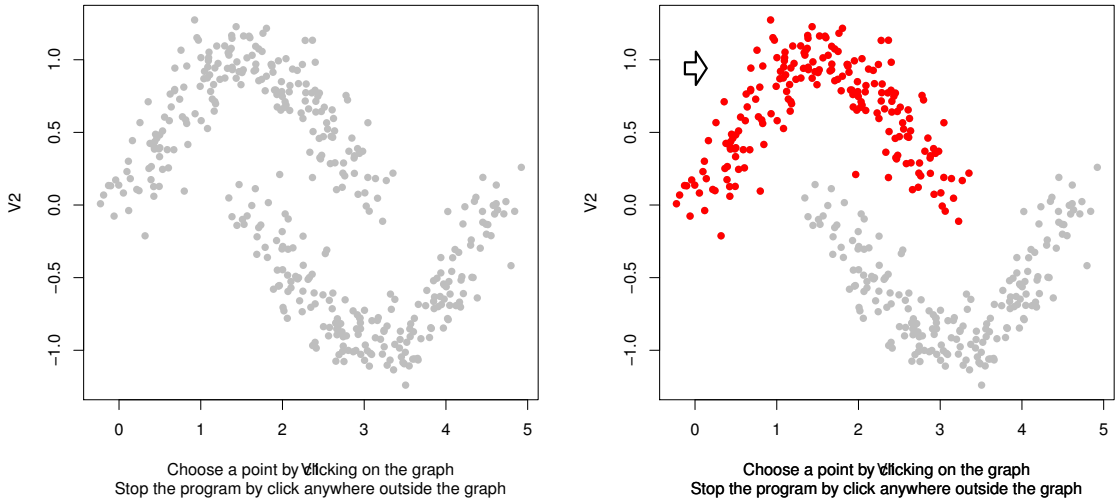

Figure 4.10: Graphical display for choosing the cluster using the function *choose.cluster* for the *logcta20* data at level 3 with 2 clusters. The left panel displays the plot before the click and the right panel highlights the points after the user pointer clicks at the arrow head (⇨).

sisting of upper upper disc will light up as in the right panel of Figure 4.10. If the user clicks any existing data point, all other points associating to the same cluster will light up. One can stop the program by clicking anywhere outside the plot area. This is an useful function and can be used in merging clusters based on "expert opinion" or to design semi-supervised clustering.

Part II

Statistical Monitoring of Clinical Trials with Co-Primary Endpoints

Chapter 5

Review of Relevant Knowledge

Starting from this chapter, we will introduce the second part of the thesis from this chapter: the statistical monitoring of clinical trials with multiple co-primary endpoints. It includes the Group Sequential Design (GSD) to consider stoping the trial early on if the study shows promising efficacy, and to use the Conditional Power (CP) to stop the trial early due to the futility. This chapter reviews some existing statistical monitoring methods that are applied in clinical trial research. Section 5.1 focuses on statistical monitoring tools for clinical trials with one endpoint. Subsection 5.1.1 introduces the basic concepts and methods of GSD and reviews several well-known methods for determining the stopping boundaries of the GSD. Subsection 5.1.2 reviews the B-value tool introduced by Lan and Wittes (1988). The B-value tool is defined as a transformed Z-value of the test statistic at interim analysis. One uses Brownian motion to describe the distribution of the interim B-value. B-value is a convenient tool to calculate the conditional power, which is defined as the probability of rejecting the null hypothesis conditional on the information observed at interim analysis. Section 5.2 introduces the clinical trials with multiple co-primary endpoints. The research in this part of the dissertation focuses on developing GSD methodologies for clinical trials with multiple co-primary endpoints, and to extend the B-value tool to multi-dimensions to calculate the conditional power of such a study.

5.1 Statistical Monitoring with One Primary Endpoint

5.1.1 Group Sequential Design (GSD)

Group Sequential Design (GSD) methods provide a flexible tool to design and monitor the clinical trials. The primary reason of using the GSD is to reduce the sample size. GSD allows the investigators to look at and monitor the process of the study at interim time interval. It allows the study to be stopped early if the study shows significant efficacy or futility at an early stage of the study. It can save resources such as finance and time. Ethically, it reduces the number of patients needed for the study. From the administrative aspect, it allows the investigator to monitor the trial on the right track.

In GSD, the term *information fraction*, denoted as t is defined as the amount of information observed at the interim analysis divided by the amount of total information of the study. If the endpoint is normally distributed or if it is a dichotomized response, the information fraction can be simplified as $t = n/N$, where n is the number of samples observed at the interim analysis and N is the maximum sample size needed. More formally,

$$t = \frac{var(\sum_{i=1}^{n} X_i)}{var(\sum_{i=1}^{N} X_i)}.$$

It also represents the time of interim analysis. For example, $t = 0$ and $t = 1$ represent the beginning and the end of the study respectively.

Let H_0 be the null hypothesis and H_a be the one-sided alternative hypothesis. The K-stage designed study is to collect the sample sequentially by K groups with $K - 1$ interim analyses and one final analysis. In the K-stage designed study, we denote $Z(t_k)$ as the test statistic after the kth group information has been collected. Let z_k be the critical value for the test at the kth stage, also known as the *stopping boundary*. The value of z_k depends on the number of the stages of the test K, the nominal significance level α and the time of the test k. The standard group sequential test

procedure is:

1. At any interim stage k, i.e., for $k = 1, 2, \cdots, K-1$

$$\begin{aligned} &\text{if } Z(t_k) > z_k \qquad \text{stop, reject } H_0; \\ &\text{otherwise} \qquad \text{continue the study.} \end{aligned}$$

2. At the final stage, i.e., $k = K$,

$$\begin{aligned} &\text{if } Z(t_K) > z_K \qquad \text{stop, reject } H_0; \\ &\text{otherwise} \qquad \text{stop, fail to reject } H_0. \end{aligned}$$

It is well known that the overall Type I error rate can be inflated by multiple tests at different time. There are many choices of stopping boundaries. Pocock (1977) proposed a constant nominal significance level. Thus, the stopping boundaries of all the interim analyses are the same. As an alternative to the Pocock's constant nominal significance level, O'Brien and Fleming (1979) proposed an approach to let the significance level increase as the study continues. Therefore, it becomes more difficult to reject H_0 at the early stage of the study. The specific stopping boundary calculation is $z_k = z_K\sqrt{K/k}$. In the group sequential test, most tests are based on asymptotic normality of the test statistic. Thus we can write down the joint distribution of $Z(t_1), Z(t_2), \cdots, Z(t_K)$. The test statistic has the following distribution:

1. $Z(t_1), Z(t_2), \cdots, Z(t_K)$ have a multivariate normal distribution;
2. $E(Z(t_k)) = 0$ and $Var(Z(t_k)) = 1$ for $k = 1, 2, \cdots, K$;
3. $Cov(Z(t_i), Z(t_j)) = \sqrt{t_i/t_j}$ for $t_i \leq t_j$.

The maximum sample size needed in GSD is determined not only by the design parameters, which are usually the effect size of the treatment, but it also depends on the choice of the stopping boundary. Different stopping boundaries will require different sample sizes. More details on sample size

Table 5.1: Stopping boundaries of four methods at one-sided $\alpha = 0.025$

# Looks (K)	t	Pocock	Pocock-Like	O-F	O-F-Like
K=2	$t = 1/2$	2.178	2.157	2.796	2.963
	$t = 1$	2.178	2.201	1.977	1.969
K=3	$t = 1/3$	2.289	2.279	3.471	3.710
	$t = 2/3$	2.289	2.295	2.454	2.511
	$t = 1$	2.289	2.296	2.004	1.995

determination for various stopping rules are provided in Jennison and Turnbull (1999).

Both the Pocock and O'Brien-Fleming tests are required to specify the number and the time of the interim analyses prior to the beginning of the study. The *error spending* approach can help us avoid this issue. It allows the investigators to perform the interim analysis at any time point. It also allows for no restriction on the number of the interim analyses. Lan and DeMets (1983) introduced the two commonly used approaches:

$$(1) \quad \alpha(t) = 2 - 2\Phi(z_{\alpha/2}/\sqrt{t}), \qquad \text{O'Brien-Fleming Like;}$$

$$(2) \quad \alpha(t) = \alpha \ln(1 + (e-1)^t), \qquad \text{Pocock Like,}$$

where Φ is the cumulative distribution function of the standard normal distribution. Using error spending function to determine the stopping boundaries makes the study more flexible. One can determine the timing and the number of the interim analyses adaptively after the study starts. Such studies can be classified as *adaptive design*. Table 5.1.1 provides the stopping boundaries of the four classical methods at a one-sided significance level of 0.025 for $K = 2$ and 3.

5.1.2 B-value Tool

Lan and Wittes (1988) proposed the B-value tool to calculate the conditional power based on the information obtained from the interim study. Here we provide a brief overview of the B-value tool. Consider the following one sample univariate location test problem. Let $X_1, X_2, \cdots, X_N \sim N(\mu, 1)$. We are interested in the $H_0 : \mu = 0$ versus $H_a : \mu > 0$. The test statistic is $Z_N =$

$\sum_{i=1}^{N} X_i/\sqrt{N}$ with rejection region $Z_N \geq c$, where c is the critical value of the test statistic. In this case, the information fraction t can be simplified as n/N. Assume we observe the data group sequentially. After observing n values, the test statistic $Z_n = \sum_{i=1}^{n} X_i/\sqrt{n}$ is the Z-test value of interim analysis at time t. The B-value, which is the transformed Z-value, is defined as

$$B(t) = Z_n\sqrt{t},$$

with expected value $E[B(t)] = \Theta t$, where the drift parameter is $\Theta = \sqrt{N}\mu$. The expectation of B-value is a linear function of t with slope $\Theta = \sqrt{N}\mu$. It changes linearly with information fraction t. Before transformation, we have $E[Z_n] = \sqrt{n}\mu = \sqrt{Nt}\mu$, which is in a quadratic form of t. The B-value makes the prediction of the parameter μ easier in the linear form than it does in the quadratic form. One can easily find the projection of the parameter from the interim analysis to the end of the study. This is the main advantage of using the B-value.

The test statistic of interest at the end of the study is $Z_N = B(1)$. $B(1)$ can be decomposed into two parts at the interim analysis: the observed fixed part $B(t)$ and unobserved random part $B(1) - B(t)$. The decomposition has the following properties.

$$\begin{aligned} &Z_N = B(t) + (B(1) - B(t)); \\ &B(t) \sim N(t\Theta, t); \\ &B(1) - B(t) \sim N(\Theta(1-t), 1-t); \\ &B(t) \text{ and } B(1) - B(t) \text{ are independent.} \end{aligned} \tag{5.1}$$

When data are observed at time t, $B(t)$ is no longer random. The remainder $B(1) - B(t)$ is still random. Therefore, the distribution of the test statistic $B(1)$ conditional on the interim observed $B(t)$ is

$$B(1)|B(t) \sim N(B(t) + \Theta(1-t), 1-t). \tag{5.2}$$

The conditional power becomes the upper percentile of the critical value c of the distribution

$N(B(t)+\Theta(1-t),1-t)$. We can use $\hat{\Theta}=B(t)/t$ to estimate Θ and thus conditional distribution becomes $B(1)|B(t)\sim N(B(t)+\hat{\Theta}(1-t),1-t)$. Therefore, the formula to calculate conditional power under the true alternative hypothesis is:

$$\begin{aligned} CP(\Theta) &= P[B(1)>c|B(t),\Theta] \\ &= 1-\Phi\left(\frac{c-B(t)-\frac{B(t)}{t}(1-t)}{\sqrt{1-t}}\right) \\ &= 1-\Phi\left(\frac{c-B(t)/t}{\sqrt{1-t}}\right). \end{aligned} \tag{5.3}$$

In practice, the situation is more complicated than the question listed above. Let $X_i\sim N(\mu_X,\sigma)$ be the new treatment group and g $Y_i\sim N(\mu_Y,\sigma)$ be the control group. The hypothesis we are interested is $H_0:\mu_X=\mu_Y$ versus $H_a:\mu_X>\mu_Y$, assuming positive difference indicates efficacy. N samples are needed for each group to reach the desired power, assuming $1:1$ randomization, which means that the treatment and control groups have the same sample size. The test statistic is:

$$T_N=\frac{\overline{X}_N-\overline{Y}_N}{s_N\sqrt{2/N}}=\frac{\sum_{i=1}^N X_i-\sum_{i=1}^N Y_i}{s_N\sqrt{2N}},$$

where s_N is the pooled standard deviation. At interim time t, suppose we observed n observations from each group. We assume observing equal number of patients per group. The interim t-statistic (Z-value) is

$$T_n=\frac{\overline{X}_n-\overline{Y}_n}{s_n\sqrt{2/n}}=\frac{\sum_{i=1}^n X_i-\sum_{i=1}^n Y_i}{s_n\sqrt{2n}}.$$

Then the corresponding B-value will be $B(t)=T_n\sqrt{t}$. For large n, the properties (5.1) are still valid with $\Theta=\sqrt{\frac{N}{2}}\frac{\mu_X-\mu_Y}{\sigma}$ and the estimated of Θ is $\hat{\Theta}=\sqrt{\frac{N}{2}}\frac{\overline{X}_n-\overline{Y}_n}{s_n}$.

In practice, it is common to use the binary endpoint as the primary endpoint for a clinical trial. In this case, let $X_1,X_2,\cdots,X_N\sim Bern(p_1)$ and $Y_1,Y_2,\cdots,Y_N\sim Bern(p_2)$, still consider the $1:1$ randomization. The hypothesis is $H_0:p_1=p_2$ versus $H_a:p_1>p_2$. Then the test statistic is

given by:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{2\bar{p}(1-\bar{p})}{N}}} = \frac{\frac{\sum_{i=1}^{N} X_i}{N} - \frac{\sum_{i=1}^{N} Y_i}{N}}{\sqrt{\frac{2\bar{p}(1-\bar{p})}{N}}} = \frac{\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} Y_i}{\sqrt{2N\bar{p}(1-\bar{p})}},$$

with $\bar{p} = \frac{1}{2}(p_1 + p_2)$. It can be decomposed as:

$$\begin{aligned} Z &= \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i + \sum_{i=n+1}^{N} X_i - \sum_{i=n+1}^{N} Y_i}{\sqrt{2N\bar{p}(1-\bar{p})}} \\ &= \sqrt{\frac{n}{N}} \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i}{\sqrt{2n\bar{p}(1-\bar{p})}} + \sqrt{\frac{N-n}{N}} \frac{\sum_{i=n+1}^{N} X_i - \sum_{i=n+1}^{N} Y_i}{\sqrt{2(N-n)\bar{p}(1-\bar{p})}} \\ &= \sqrt{t} Z_1 + \sqrt{1-t} Z_2, \end{aligned}$$

where Z_1 and Z_2 are the test statistics before and after the interim analysis respectively. At the interim analysis, Z_1 is observed and fixed. Z_2 remains random with $E(Z_2) = \frac{(N-n)(p_1-p_2)}{\sqrt{2(N-n)\bar{p}(1-\bar{p})}}$ and $Var(Z_2) = 1.$. The conditional power is:

$$\begin{aligned} CP &= P(\sqrt{t} Z_1 + \sqrt{1-t} Z_2 > c) \\ &= P\left(Z_2 > \frac{c - \sqrt{t} Z_1}{\sqrt{1-t}} \right) \\ &= P\left(Z > \frac{c - \sqrt{t} Z_1}{\sqrt{1-t}} - \frac{\sqrt{N-n}(p_1 - p_2)}{\sqrt{2\bar{p}(1-\bar{p})}} \right) \\ &= P\left(Z > \frac{\sqrt{N} c - \sqrt{n} Z_1 - \frac{(N-n)(p_1-p_2)}{\sqrt{2\bar{p}(1-\bar{p})}}}{\sqrt{N-n}} \right). \end{aligned}$$

At the interim analysis, we can replace p_1, p_2 and $\bar{p}$ by $\hat{p}_1 = \frac{\sum_{i=1}^{n} X_i}{n}$, $\hat{p}_2 = \frac{\sum_{i=1}^{n} Y_i}{n}$ and $\hat{\bar{p}} = \frac{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i}{2n}$ respectively to calculate the conditional power. The conditional power can be calculated using the B-value. The B-value is still the transformed Z-value at the interim. $B(t) = \sqrt{t} Z_1$. In this case, the drift parameter $\Theta = \sqrt{\frac{N}{2}} \frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})}}$. The conditional power simplifies to:

$$CP = P\left(Z > \frac{c - B(t) - (1-t)\Theta}{\sqrt{1-t}} \right).$$

At the interim analysis, replacing Θ by $\hat{\Theta} = B(t)/t$. The conditional power is:

$$CP = P\left(Z > \frac{c - B(t)/t}{\sqrt{1-t}}\right).$$

Remark: Note that the conditional power is calculated based on the assumption that the trial will continue under the current trend, which is a strong assumption. It is very sensitive to the interim estimates. The conditional power can also be calculated assuming the parameter remains the same as the design parameter or using other assumptions.

5.2 Multiple Co-Primary Endpoints

Hypotheses related to multiple co-primary endpoints, which is defined as clinical trials that need to declare the significance on more than one endpoints simultaneously, have received a lot of attention in recent years. Meyerson et al. (2007) provided the brief introduction. These problems are also called the *reverse multiplicity problems*. There are several diseases and therapeutic areas in which the regulatory agencies need the treatment to demonstrate statistical significance on multiple co-primary endpoints. Some examples of diseases where at least two primary endpoints may be of interest are (a) Migraine which is accompanied by nausea and photophobia; (b) Alzheimer's disease which is assessed by Alzheimer's disease assessment scale–cognitive and clinician interview-based impression of change; (c) Multiple Sclerosis is measured by relapse rate at 1 year and disability at 2 years and (d) Osteoarthritis which is evaluated by pain, patient global assessment, and quality of life. Hence, an efficacy in these diseases is evaluated with multiple endpoints. The study power may be considerably reduced depending on the correlation(s) among the endpoints.

Intersection-Union Test (IUT) is the standard technique for dealing with the question of reverse multiplicity question. More details about IUT can be found in Berger (1982), Casella and Berger (2001). Consider that $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_N}$ are independent random vectors drawn from the multivariate normal distribution with d dimensions $\mathbf{N_d}(\boldsymbol{\mu_d}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu_d} = (\mu_1, \mu_2, \cdots, \mu_d)$. Without loss of generality, let $\boldsymbol{\Sigma} = (1-\rho)\mathbf{I_d} + \rho\mathbf{J_d}$. In this case, the covariance matrix of $\mathbf{X}$ and the correlation

matrix are identical. We are interested in the following hypothesis:

$$H_0 : \mu_j \leq 0 \text{ for at least one } j,\ j = 1, 2, \ldots, d,$$
$$H_a : \mu_j > 0 \text{ for all } j. \tag{5.4}$$

This is the hypothesis of IUT. If we denote $H_{0,j} : \mu_j \leq 0$ versus $H_{a,j} : \mu_j > 0$,the Hypothesis (5.4) can be rewritten as:

$$H_0 : \bigcup_{j=1}^{d} H_{0,j},$$
$$H_a : \bigcap_{j=1}^{d} H_{a,j}. \tag{5.5}$$

The test statistic for $H_{0,j}$ versus $H_{a,j}$ is $Z_{N,j} = \sum_{i=1}^{N} X_{ij}/\sqrt{N}$. We reject $H_{0,j}$ if $Z_{N,j}$ is beyond the certain critical value c. Therefore we reject (5.5), and furthermore reject (5.4) if $Z_{N,j} > c$ for all j. It is straightforward to prove that for $\forall j$ and k, where $j, k = 1, 2, \ldots, d$, the covariance/correlation between the test statistics is the same as the covariance/correlation between the endpoints, i.e. $Corr(Z_{N,j}, Z_{N,k}) = \rho$.

In general, it is more difficult to achieve significance simultaneously with the increase of the number of multiple endpoints. The regulatory agencies requires a treatment to demonstrate statistically significant effect on multiple endpoints, each at the one-sided 2.5% level. It is a conservative test since the overall Type I error will be no greater than 2.5%. Chuang-Stein et al. (2007) introduced the concept of the average Type I error approach to adjust for the significance level. They assumed the parameters of effect sizes are uniformly distributed over the restricted null space and average the power function under the uniform distribution of the region of the restricted null space. In this case, the local significance level can be inflated more than 2.5%, but still controlling the average Type I error at the desired significance level. However, this method does not strictly control the overall Type I error.

Some methods for the calculation of the power and determining the sample size in clinical trials with more than one primary endpoint have been developed and published in the research literature. Xiong et al. (2005) introduced a formula to calculate the power with bivariate normal co-primary endpoints with known variance-covariance matrix. Sozu et al. (2006) extended it to include unknown variance-covariance matrix using the Wishart distribution. Sozu et al. (2010) proposed the closed form solution for the calculation of power and sample size with multiple co-primary binary endpoints. Sozu et al. (2011) provided the formulas to calculate power and sample size for some situations in superiority trials. However, there is lack of consideration in the monitoring of clinical trials with multiple co-primary endpoints.

Chapter 6

Group Sequential Design

Statistical methodologies of group sequential testing of interim analysis with one endpoint are widely used. However, the group sequential methods with multiple co-primary endpoints has not been considered. This chapter introduces the group sequential design with multiple co-primary normally distributed endpoints, which considers the correlation between the endpoints. Both the known and unknown variance-covariance matrix cases are considered. This chapter is organized as follows: Section 6.1 introduces the group sequential test procedures of clinical trials with multiple co-primary endpoints. Section 6.2 presents Theorem 6.1, which shows that the stopping boundary in single endpoint cases is applicable to multiple co-primary endpoints case. Section 6.3 discusses the power and sample size of the GSD with multiple co-primary endpoints. Section 6.4 discusses the cases when the correlation between the endpoints is unknown. Section 6.5 discusses the GSD with multiple co-primary binary endpoints. Section 6.6 concludes the chapter by using the multivariate regression to consider the analysis of multiple co-primary normal endpoints adjusted by some covariates.

6.1 Group Sequential Tests

Consider the simplest case of the hypothesis (5.4) where the study is a two-stage design ($K = 2$) with two co-primary endpoints ($d = 2$), EP1 and EP2, with known correlation between the two endpoints. Let EP1 and EP2 be normally distributed as:

$$\begin{pmatrix} EP1 \\ EP2 \end{pmatrix} \sim \mathbf{N_2} \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

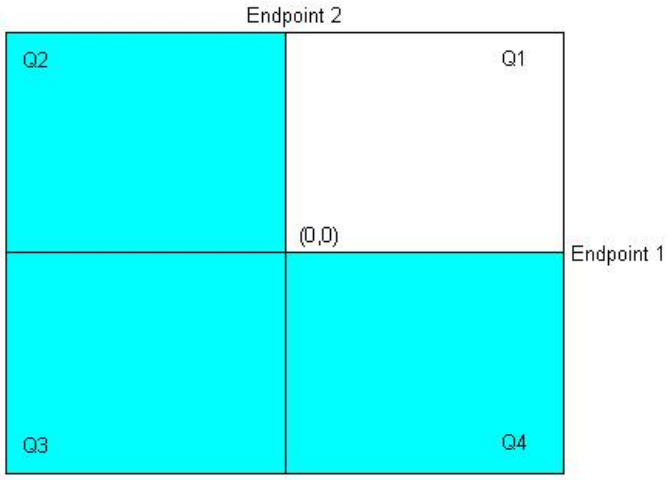

Figure 6.1: Null space of two co-primary endpoints

The hypothesis we are interested is:

$$H_0 : \mu_1 < 0 \text{ or } \mu_2 < 0;$$
$$H_a : \mu_1 > 0 \text{ and } \mu_2 > 0. \tag{6.1}$$

The null space of the Hypothesis (6.1) is shown in Figure 6.1. Define $t = n/N$ as the time of interim analysis. The formal group sequential test of (5.4) is:

1. At Stage 1, when $k = 1$

   if $Z_1(t) > Z_{\alpha_1}$ and $Z_2(t) > Z_{\alpha_1}$, stop and reject H_0;

   otherwise, continue the study;

2. At the Final Stage, when $k = 2$,

   if $Z_1(1) > Z_{\alpha_2}$ and $Z_2(1) > Z_{\alpha_2}$, stop and rejectH_0;

otherwise, stop and fail to reject H_0,

where $Z_j(t)$ and $Z_j(1)$ are the interim and final Z statistic for jth endpoint correspondingly and Z_{α_1} and Z_{α_2} are the critical values chosen for the two stages. The joint distribution of the test statistics of two endpoints and at the two stages is:

$$\begin{pmatrix} Z_1(t) \\ Z_2(t) \\ Z_1(1) \\ Z_2(1) \end{pmatrix} \sim \mathbf{N_4} \left(\begin{pmatrix} \sqrt{n}\mu_1 \\ \sqrt{n}\mu_2 \\ \sqrt{N}\mu_1 \\ \sqrt{N}\mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \sqrt{t} & \sqrt{t}\rho \\ \rho & 1 & \sqrt{t}\rho & \sqrt{t} \\ \sqrt{t} & \sqrt{t}\rho & 1 & \rho \\ \sqrt{t}\rho & \sqrt{t} & \rho & 1 \end{pmatrix} \right). \tag{6.2}$$

Then the power function of the hypothesis (5.4) is given by:

$$\begin{aligned} &P((Z_1(t) > Z_{\alpha_1} \cap Z_2(t) > Z_{\alpha_1}) \\ &\quad \cup ((Z_1(t) \leq Z_{\alpha_1} \cup Z_2(t) \leq Z_{\alpha_1}) \cap Z_1(1) > Z_{\alpha_2} \cap Z_2(1) > Z_{\alpha_2}) | \boldsymbol{\mu}). \end{aligned} \tag{6.3}$$

The power function (6.3) can be written as

$$\begin{aligned} &P((Z_1(t) > Z_{\alpha_1} \cap Z_2(t) > Z_{\alpha_1}) | \boldsymbol{\mu}) \\ &\quad + P(((Z_1(t) \leq Z_{\alpha_1} \cup Z_2(t) \leq Z_{\alpha_1}) \cap Z_1(1) > Z_{\alpha_2} \cap Z_2(1) > Z_{\alpha_2}) | \boldsymbol{\mu}). \end{aligned} \tag{6.4}$$

and can be simplified as:

$$P\left((Z_1(t) > Z_{\alpha_1} \cap Z_2(t) > Z_{\alpha_1}) \cup (Z_1(1) > Z_{\alpha_2} \cap Z_2(1) > Z_{\alpha_2}) | \boldsymbol{\mu}\right). \tag{6.5}$$

In general, if we have d endpoints and K stages, then the test statistics will have the following joint distribution:

Property 6.1 *The joint distribution of Z-statistics:*

1. $\mathbf{Z}(t_1), \mathbf{Z}(t_2), \cdots, \mathbf{Z}(t_K)$ *have a multivariate normal distribution with dimension* $d \times K$*;*

2. *For each dimension p,* $E(Z_p(t_k)) = \sqrt{t}\sqrt{N}\mu_j$ *and* $Var(Z_p(t_k)) = 1$ *for* $k = 1, 2, \cdots, K$ *and* $p = 1, 2, \cdots, d$*;*

3. *For each dimension p,* $Cov(Z_p(t_r), Z_p(t_s)) = \sqrt{t_r/t_s}$ *for* $t_r \leq t_s$*;*

4. *For any two dimensions r and s,* $Cov(Z_j(t_r), Z_k(t_s)) = \rho_{jk}\sqrt{t_r/t_s}$ *for* $t_r \leq t_s$*.*

The Property 6.1 can be easily transformed to the form of B-values with the following results:

Property 6.2 *The joint distribution of* B*-values:*

1. $\mathbf{B}(t_1), \mathbf{B}(t_2), \cdots, \mathbf{B}(t_K)$ *have a multivariate normal distribution with dimension* $d \times K$*;*

2. *For each dimension p,* $E(B_p(t_k)) = t\sqrt{N}\mu_j$ *and* $Var(B_p(t_k)) = t$ *for* $k = 1, 2, \cdots, K$ *and* $p = 1, 2, \cdots, d$*;*

3. *For each dimension d,* $Cov(B_p(t_r), B_p(t_s)) = t_r$ *for* $t_r \leq t_s$*;*

4. *For any two dimensions* j *and* $k, Cov(B_j(t_r), B_k(t_s)) = t_r\rho_{jk}$ *for* $t_r \leq t_s$*.*

6.2 Determining the Stopping Boundary

To control the overall Type I error of the group sequential design with multiple co-primary endpoints is not easy. It is known that the Type I error rate can be inflated due to performing the test at multiple interim looks. However, at each interim analysis, the IUT causes the drop of power. It is the combination of a multiplicity and reverse multiplicity problem.

The error spent form for the group sequential design of clinical trials with multiple co-primary endpoints is:

$$\pi_1 = P\{\mathbf{Z}(\mathbf{t_1}) > \mathbf{z}_{\alpha_1} | \mathbf{\Theta_0}\}$$
$$\pi_2 = P\{\mathbf{Z}(\mathbf{t_2}) > \mathbf{z}_{\alpha_2} \bigcap \mathbf{Z}(\mathbf{t_1}) \ngeq \mathbf{z}_{\alpha_1} | \mathbf{\Theta_0}\}$$
$$\vdots$$

ith stage:

$$\pi_i = P\{\mathbf{Z}(\mathbf{t_i}) > \mathbf{z}_{\alpha_i} \bigcap \mathbf{Z}(\mathbf{t_{i-1}}) \ngeq \mathbf{z}_{\alpha_{i-1}} \cdots \bigcap \mathbf{Z}(\mathbf{t_1}) \ngeq \mathbf{z}_{\alpha_1} | \mathbf{\Theta_0}\},$$

and the cumulative error is:

$$\pi_k^* = \sum_{i=1}^{k} \pi_i \text{ where } k = 1, 2, \cdots, K.$$

The Type I error control requires $\sup_{\boldsymbol{\theta} \in \mathbf{\Theta_0}} \pi_K^* \leq \alpha$(the nominal level). We want to control the maximum of π_K^* at the desired level.

Theorem 6.1 *π_K^* reaches the maximum value when one element of $\boldsymbol{\mu_p}$ is 0 while the other elements are $+\infty$.*

Proof of Theorem 6.1: First, we provide the conditional distribution $\mathbf{B}(\mathbf{t_i})|\mathbf{B}(\mathbf{t_{i-1}})$, following the derivation of equation in (7.2), which will be introduced in next chapter,

$$\mathbf{B}(\mathbf{t_i})|\mathbf{B}(\mathbf{t_{i-1}}) \sim \mathbf{N_p}(\mathbf{B}(\mathbf{t_{i-1}}) + \boldsymbol{\mu_p}\sqrt{N}(t_i - t_{i-1}), (t_i - t_{i-1})\mathbf{\Sigma}) \tag{6.6}$$

We prove the Theorem 6.1 by mathematical induction. We denote $\mathbf{b}_{\alpha_i} = \mathbf{z}_{\alpha_i}\sqrt{t_i}$. For $K = 2$,

$$\begin{aligned}
\pi_2^* &= \pi_1 + \pi_2 \\
&= \pi_1^* + P\{\mathbf{B}(\mathbf{t_2}) > \mathbf{b}_{\alpha_2} \bigcap \mathbf{B}(\mathbf{t_1}) \ngeq \mathbf{b}_{\alpha_1} | \Theta_0\} \\
&= \pi_1^* + P\{\mathbf{B}(\mathbf{t_2}) > \mathbf{b}_{\alpha_2} | \mathbf{B}(\mathbf{t_1}) \ngeq \mathbf{b}_{\alpha_1}, \Theta_0\}(1 - \pi_1^*) \\
&= \pi_1^* + \gamma_2(1 - \pi_1^*),
\end{aligned}$$

where $\gamma_2 = P\{\mathbf{B}(\mathbf{t_2}) > \mathbf{b}_{\alpha_2} | \mathbf{B}(\mathbf{t_1}) \ngeq \mathbf{b}_{\alpha_1}, \Theta_0\}$. From Equation (6.6), it is easy to verify that γ_2 reaches the maximum value under the domain of $\mathbf{\Theta}_0$ when one element of $\boldsymbol{\mu_d}$ is 0, while the others are $+\infty$. To simplify, let $\boldsymbol{\mu_{d,0}} = (0, +\infty, \cdots, +\infty)$. Also π_1^* reaches the maximum at $\boldsymbol{\mu_{d,0}}$. Note that $0 < \pi_1^*, \gamma_2 < 1$. π_2 is monotone increasing as π_1^* and γ_2 increase when $0 < \pi_1, \gamma_2 < 1$, since $\frac{\partial \pi_2^*}{\partial \pi_1^*} = 1 - \gamma_2 > 0$ and $\frac{\partial \pi_2^*}{\partial \gamma_2} = 1 - \pi_1^* > 0$. Therefore, π_2^* reaches the maximum value when both π_1^* and γ_2 reach the maximum values.

Assume Theorem 6.1 holds for $K = N$, i.e., π^*_N reaches the maximum at $\boldsymbol{\mu_{p,0}}$, for $K = N+1$,

$$\begin{aligned}
\pi^*_{N+1} &= \pi^*_N + \alpha_{N+1} \\
&= \pi^*_N + P\{\mathbf{B}(\mathbf{t_{N+1}}) > \mathbf{b}_{\alpha_{N+1}} \bigcap \mathbf{B}(\mathbf{t_N}) \ngeq \mathbf{b}_{\alpha_N} \cdots \bigcap \mathbf{B}(\mathbf{t_1}) \ngeq \mathbf{b}_{\alpha_1} | \Theta_0\} \\
&= \pi^*_N + P\{\mathbf{B}(\mathbf{t_{N+1}}) > \mathbf{b}_{\alpha_{N+1}} | \mathbf{B}(\mathbf{N}) \ngeq \mathbf{b}_{\alpha_N}, \cdots, \mathbf{B}(\mathbf{t_1}) \ngeq \mathbf{b}_{\alpha_1}, \Theta_0\}(1 - \pi^*_N) \\
&= \pi^*_N + \gamma_{N+1}(1 - \pi^*_N)
\end{aligned}$$

where $\gamma_{N+1} = P\{\mathbf{B}(\mathbf{t_{N+1}}) > \mathbf{b}_{\alpha_{N+1}} | \mathbf{B}(\mathbf{N}) \ngeq \mathbf{b}_{\alpha_N}, \cdots, \mathbf{B}(\mathbf{t_1}) \ngeq \mathbf{b}_{\alpha_1}, \Theta_0\}$. From Equation (6.6), γ_{N+1} achieves the maximum at $\boldsymbol{\mu_{p,0}}$. π^*_{N+1} reaches the maximum value at $\boldsymbol{\mu_{p,0}}$ by using similar argument where $K = 2$. Thus the theorem is proved. ■

This theorem shows that one can always reach the significance for the dimensions that have $\mu_j = +\infty$. It only needs to control the Type I error for the one at 0. Therefore the stopping boundaries of the multiple co-primary endpoints will be the same as the stopping boundaries for the single endpoint. Thus, even in the case of co-primary endpoints, one can use the stopping rules of the single endpoint that was reviewed in Section 5.1.1.

6.3 Power and Sample Size

In the two-stage design with two co-primary endpoints, EP1 and EP2, the correlation between the two endpoints is ρ. The joint distribution of the test statistics of two endpoints and at two stages is:

$$\begin{pmatrix} Z_1(t) \\ Z_2(t) \\ Z_1(1) \\ Z_2(1) \end{pmatrix} \sim \mathbf{N_4} \left(\begin{pmatrix} \sqrt{n}\mu_1 \\ \sqrt{n}\mu_2 \\ \sqrt{N}\mu_1 \\ \sqrt{N}\mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \sqrt{t} & \sqrt{t}\rho \\ \rho & 1 & \sqrt{t}\rho & \sqrt{t} \\ \sqrt{t} & \sqrt{t}\rho & 1 & \rho \\ \sqrt{t}\rho & \sqrt{t} & \rho & 1 \end{pmatrix} \right). \tag{6.7}$$

Table 6.1: Overall power under group sequential design with two stage and two endpoints when each marginal reaches 80% power at one-sided overall significant level 2.5%

correlation	overall power
-1	0.60
-0.8	0.61
-0.5	0.62
-0.2	0.63
0	0.64
0.2	0.66
0.5	0.69
0.8	0.73

When $t = 0.5$, the power is

$$P\left((Z_1(0.5) > Z_{\alpha_1} \cap Z_2(0.5) > Z_{\alpha_1}) \cup (Z_1(1) > Z_{\alpha_2} \cap Z_2(1) > Z_{\alpha_2})\right) \tag{6.8}$$

under the distribution (6.7).

For example, if we want to detect the effect size of 0.2 for each of the two normally distributed endpoints, 196 samples are needed in total so it can get the power of 80% on each endpoint. Under the group sequential design discussed above, the power with different correlations between endpoints are shown in Table 6.1. If the endpoints are less correlated, the overall power is smaller.

In a group sequential design, the maximum sample size needed is determined not only by the effect size, but also by the number of the stages in the study, the time of the interim analysis, and the choice of the stopping boundaries. Table 6.2 shows the various values of the maximum sample size needed to reach the overall 80% power if we want to detect the effect size 0.2 for both endpoints in two-stage design with two endpoints using different stopping boundaries. In general, the O'Brien-Flemming stopping rule can save more on the sample size than the Pocock approach.

From the distribution of (6.7), we can see the maximum sample size needed depends on the

Table 6.2: Maximum sample size needed under group sequential design with two stage and two endpoints reaches 80% overall power at one-sided overall significant level 2.5%

	maximum sample size needed			
correlation	Pocock	Pocock-like	O-F	O-F-like
0	288	292	260	260
0.2	284	288	256	256
0.5	274	276	248	248
0.8	256	260	232	232

interim time t, the choice of the stopping boundaries, the designed parameters including μ and $\mathbf{\Sigma}$, and the desired level of α and β. It is hard to give a closed form formula to calculate the maximum sample size needed. The numbers in Table 6.2 are from the numerical simulation. It is worth pointing out that in the industrial practice, when calculating the sample size, investigators assume all the endpoints are independent, as in most cases, the correlation among the endpoints are not easy to estimate. Overestimating the correlation could cause the underestimation of the sample size and make the study lack of power.

6.4 Unknown Correlation

In practice, it is very likely that the variance-covariance information between the endpoints is unknown. Still consider the case that $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_N}$ are independent random vectors that come from the multivariate normal distribution with 2 dimensions $\mathbf{N_2}(\boldsymbol{\mu_2}, \mathbf{\Sigma})$, where $\boldsymbol{\mu_2} = (\mu_1, \mu_2)$ and unknown variance-covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

Then the test statistic becomes the T-statistic.

$$T_j(t) = \frac{\bar{X}_{n,j}}{\frac{s_{j,n}}{\sqrt{n}}},$$

and

$$T_j(1) = \frac{\bar{X}_{N,j}}{\frac{s_{j,N}}{\sqrt{N}}}.$$

Using an error spending function, we decide to spend π_1 at Stage 1 and π_2 at stage 2. The group sequential t-test of (5.4) with unknown covariance matrix is:

1. At Stage 1, when $k = 1$

   if $T_1(t) > T_{n-1,\alpha_1}$ and $T_2(t) > T_{n-1,\alpha_1}$, stop and reject H_0;

   otherwise, continue the study.

2. At the Final Stage, when $k = 2$,

   if $T_1(1) > T_{N-1,\alpha_2}$ and $T_2(1) > Z_{N-1,\alpha_2}$, stop and reject H_0;

   otherwise, stop and fail to reject H_0.

If the sample size is large enough, the critical value of T-test will be approximately the same as the Z-test. It is difficult to find the joint distribution of the test T statistics for the two stage study with two endpoints, even though the marginal distribution is a T-distribution. It needs to simulate the entire trial to calculate the power and the maximum sample size needed. For example, if the trial is a two-stage design with two primary endpoints, and with true correlation $\rho = 0.2$ or $\rho = 0.5$. It is assumed the effect size is 0.4 for each endpoint. Table 6.3 provides the simulation results to compare the power with known and unknow variance-covariance matrix using the different sample sizes using the O'Brien-Fleming-like error spending function to determine the stopping boundaries. In Table 6.3, it is clear to see that when the sample size is small, the T-test provides a higher power, because with fewer degrees of freedom, the T-distribution has the heavier tail.

Table 6.3: Power comparison with known and unknown correlation

	$\rho = 0.2$		$\rho = 0.5$	
sample size	unknown	known	unknown	known
20	0.237	0.217	0.286	0.266
30	0.377	0.375	0.431	0.427
40	0.538	0.541	0.549	0.559
50	0.660	0.665	0.685	0.695
80	0.894	0.900	0.901	0.906
100	0.957	0.959	0.960	0.962

6.5 Binary Endpoints

In this section, we introduce the group sequential test procedures for multiple co-primary binary endpoints for both the single-arm and two-arm designs. For the binary endpoint, the asymptotic Z statistic is the sum of the random variable. Therefore, it is important to measure the association between two binary variables in order to determine the covariance between the two Z statistics.

6.5.1 Association Between Binary Variables

There are many ways to measure the association between the two binary random variables. In this section, we review the two most common approaches.

To measure the association between the two binary variables, we use the well known *Phi* correlation coefficient. It can be seen as a special case of the Pearson correlation coefficient. Denote $X_j, j = 1, 2, \cdots, P$ as the two binary endpoints with $E(X_j) = p_j$ and $Var(X_j) = p_j(1 - p_j)$. Denote $Cov(X_j, X_k) = \gamma_{jk}^X$, so the Pearson correlation between the two endpoints is

$$\phi_{jk}^X = \frac{\gamma_{jk}^X}{\sqrt{p_j(1-p_j)p_k(1-p_k)}}.$$

The data can be summarized into the following 2×2 table:

	$X_j = 1$	$X_j = 0$	Total
$X_k = 1$	n_{11}	n_{10}	$n_{1\cdot}$
$X_k = 0$	n_{01}	n_{00}	$n_{0\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 0}$	n

In this case, the estimated covariance

$$\hat{\gamma}_{jk}^{X} = \frac{n_{11}}{n} - \frac{n_{1\cdot}}{n}\frac{n_{\cdot 1}}{n}, \tag{6.9}$$

and the correlation coefficient

$$\begin{aligned} \hat{\phi}_{jk}^{X} &= \frac{\frac{n_{11}}{n} - \frac{n_{1\cdot}}{n}\frac{n_{\cdot 1}}{n}}{\sqrt{\frac{n_{1\cdot}}{n}\frac{n_{0\cdot}}{n}\frac{n_{\cdot 1}}{n}\frac{n_{\cdot 0}}{n}}} \\ &= \frac{n_{11}n_{00} - n_{01}n_{10}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 1}n_{\cdot 0}}}. \end{aligned}$$

In many situations, the binary responses $\mathbf{X} = (X_1, X_2, \cdots, X_K)$ are dichotomized from some underlining continuous responses $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_K)$, i.e, $\mathbf{X} = I(\mathbf{Z} > \mathbf{c})$, where $\mathbf{c} = (c_1, c_2, \cdots, c_K)$ are the cut off values of each response. $\mathbf{Z}$ follows the K variate normal distribution with mean $\boldsymbol{\mu_Z}$ and variance-covariance matrix

$$\boldsymbol{\Sigma_Z} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma_K^2 \end{pmatrix},$$

and its probability density function is $f_{\mathbf{Z}}$. In this case,

$$p_j = 1 - \phi\left(\frac{c_j - \mu_j}{\sigma_j}\right), \tag{6.10}$$

and

$$\gamma_{jk}^{X} = P(X_j = 1, X_k = 1) = \int_{-\infty}^{\infty} \cdots \int_{c_j}^{\infty} \cdots \int_{c_k}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{Z}}(\mathbf{z}) dz_1 \cdots dz_K. \tag{6.11}$$

In practice, if the binary endpoints are dichotomized from one multivariate normal distribution and if the multivariate normal parameters are known or can be estimated, it is ideal to use Equation (6.10) and Equation 6.11 to get the estimation of parameters of binary endpoints. This is because dichotomizing continuous random variables can lose information.

6.5.2 Single-arm Design

Consider that the d-dimensional multivariate Bernoulli random vector $\mathbf{X} = (X_1, X_2, \cdots, X_d) \in \{0,1\}^d$. $\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_N}$ are the independent samples from this distribution. Denote $\mathbf{p} = (p_1, \cdots, p_d)$ where $P(X_j = 1) = p_j$. We are interested in the hypothesis that

$$\begin{aligned} H_0 &: \mathbf{p} \ngeq \mathbf{p_0}; \\ H_a &: \mathbf{p} > \mathbf{p_0}, \end{aligned} \tag{6.12}$$

where $\mathbf{p_0} = (p_{0,1}, \cdots, p_{0,d})$ is the vector of proportions under the null hypothesis. We can also write (6.12) as:

$$\begin{aligned} H_0 &: p_j \leq p_{0,j} \text{ for any } j = 1, \cdots, d; \\ H_a &: p_j > p_{0,j} \text{ for all } j = 1, \cdots, d. \end{aligned} \tag{6.13}$$

At the end of the study, after observing N samples, the test statistic of each dimension j can be written as:

$$\begin{aligned} Z_{N,j} &= \frac{\hat{p}_j - p_{0,j}}{\sqrt{\frac{p_{0,j}(1-p_{0,j})}{N}}} \\ &= \frac{\sum_{i=1}^{N} X_{ij} - N p_{0,j}}{\sqrt{N p_{0,j}(1-p_{0,j})}}. \end{aligned}$$

$Z_{N,j}$ asymptotically follows the normal distribution with the expected value

$$E(Z_{N,j}) = \frac{N(p_j - p_{0,j})}{\sqrt{Np_{0,j}(1-p_{0,j})}},$$

and variance 1. At the interim analysis, after observing n samples, the test statistic is

$$\begin{aligned} Z_{n,j} &= \frac{\hat{p}_j - p_{0,j}}{\sqrt{\frac{p_{0,j}(1-p_{0,j})}{n}}} \\ &= \frac{\sum_{i=1}^{n} X_{ij} - np_{0,j}}{\sqrt{np_{0,j}(1-p_{0,j})}}. \end{aligned}$$

Thus we have

$$Cov(Z_{n,j}, Z_{N,j}) = \sqrt{t},$$

where $t = n/N$. Consider between any two dimensions j and k. The covariance between the two test statistics, which is denoted as φ_{jk} is:

$$\begin{aligned} \varphi_{jk} &= Cov(Z_{N,j}, Z_{N,k}) \\ &= Cov\left(\frac{\sum_{i=1}^{N} X_{ij} - Np_{0,j}}{Np_{0,j}(1-p_{0,j})}, \frac{\sum_{i=1}^{N} X_{ik} - Np_{0,k}}{Np_{0,k}(1-p_{0,k})}\right) \\ &= \frac{N\gamma_{jk}^{X}}{N\sqrt{p_{0,j}(1-p_{0,j})p_{0,k}(1-p_{0,k})}} \\ &= \frac{\gamma_{jk}^{X}}{\sqrt{p_{0,j}(1-p_{0,j})p_{0,k}(1-p_{0,k})}}. \end{aligned}$$

Also we can have:

$$\begin{aligned} Cov(Z_{n,j}, Z_{N,k}) &= \frac{n\gamma_{jk}^{X}}{\sqrt{nNp_{0,j}(1-p_{0,j})p_{0,k}(1-p_{0,k})}} \\ &= \sqrt{t}\frac{\gamma_{jk}^{X}}{\sqrt{p_{0,j}(1-p_{0,j})p_{0,k}(1-p_{0,k})}} \\ &= \sqrt{t}\varphi_{jk}. \end{aligned}$$

Consider the simplest case that the study has 2 co-primary endpoints ($p = 2$), E1 and E2, and 2 stage design ($K = 2$). Define $t = n/N$ as the time of the interim analysis. The group sequential test of the hypothesis (6.13) is:

1. At Stage 1, when $k = 1$

   if $Z_1(t) > Z_{\alpha_1}$ and $Z_2(t) > Z_{\alpha_1}$, stop and reject; H_0;

   otherwise, continue the study;

2. At the Final Stage, when $k = 2$,

   if $Z_1(1) > Z_{\alpha_2}$ and $Z_2(1) > Z_{\alpha_2}$, stop and reject; H_0;

   otherwise, stop and fail to reject H_0.

The joint distribution of the test statistics is asymptotically a 4-dimensional multivariate normal.

$$\begin{pmatrix} Z_1(t) \\ Z_2(t) \\ Z_1(1) \\ Z_2(1) \end{pmatrix} \sim \mathbf{N_4} \left(\begin{pmatrix} \frac{n(p_1-p_{0,1})}{\sqrt{np_{0,1}(1-p_{0,1})}} \\ \frac{n(p_2-p_{0,2})}{\sqrt{np_{0,2}(1-p_{0,2})}} \\ \frac{N(p_1-p_{0,1})}{\sqrt{Np_{0,1}(1-p_{0,1})}} \\ \frac{N(p_2-p_{0,2})}{\sqrt{Np_{0,2}(1-p_{0,2})}} \end{pmatrix}, \begin{pmatrix} 1 & \varphi_{12} & \sqrt{t} & \sqrt{t}\varphi_{12} \\ \varphi_{12} & 1 & \sqrt{t}\varphi_{12} & \sqrt{t} \\ \sqrt{t} & \sqrt{t}\varphi_{12} & 1 & \varphi_{12} \\ \sqrt{t}\varphi_{12} & \sqrt{t} & \varphi_{12} & 1 \end{pmatrix} \right). \quad (6.14)$$

In general, the joint distribution of the test statistics is asymptotically multivariate normal with $d \times K$ dimensions.

6.5.3 Two-arm Design

In this section, we compare the vector of proportions of two samples. $\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_N}$ are the samples from a d-dimensional multivariate Bernoulli distribution with the proportion vector $\mathbf{p_X} = (p_{X_1}, p_{X_2}, \cdots, p_{X_d})$ and $\mathbf{Y_1}, \mathbf{Y_2}, \cdots, \mathbf{Y_N}$ are the samples from another d-dimensional multivariate Bernoulli distribution with the proportion vector $\mathbf{p_Y} = (p_{Y_1}, p_{Y_2}, \cdots, p_{Y_d})$. The hy-

pothesis we are interested in is

$$
\begin{aligned}
&H_0 : p_{X_j} - p_{Y_j} \leq \delta \text{ for any } j = 1, \cdots, d; \\
&H_a : p_{X_j} - p_{Y_j} > \delta \text{ for all } j = 1, \cdots, d.
\end{aligned} \tag{6.15}
$$

In this dissertation, we consider $\delta = 0$ as it is the superiority clinical trial. N sample are needed for each group to have the desired level of power. Without loss of generality, For each dimension j, the test statistic is:

$$
\begin{aligned}
Z_j &= \frac{\hat{p}_{X_j} - \hat{p}_{Y_j}}{\sqrt{\frac{2}{N}\hat{\bar{p}}_j(1 - \hat{\bar{p}}_j)}} \\
&= \frac{\sum_{i=1}^{N} X_{ij} - \sum_{i=1}^{N} Y_{ij}}{\sqrt{2N\hat{\bar{p}}_j(1 - \hat{\bar{p}}_j)}}
\end{aligned}
$$

In general, the group sequential test procedure for the Hypothesis (6.15) is:

1. At Stage k, when $k = 1, 2, \cdots, K-1$

   if $Z_j(t_k) > Z_{\alpha_k}$ for all j, stop and reject H_0;

   otherwise, continue the study;

2. At Final Stage, when $k = K$,

   if $Z_j(1) > Z_{\alpha_K}$ for all j, stop and reject H_0;

   otherwise, stop and fail to reject H_0.

Similar to the normal endpoints with unknown covariance introduced in Section 6.4, it is difficult to find the joint distribution of the test statistics of a two sample test of proportions. The simulation of the entire trial is needed to determine the power and maximum sample size.

6.6 GSD with Covariates

The randomization in clinical trials is used to exclude the bias of the study. Even though the investigators randomized the patients and the use of the treatment, when analyzing the data, we still need to adjust the other covariates, which could be considered as confounders of the study. In the case with the normal response, the multiple linear regression is the commonly used model. The results with the known variance-covariance structure can be derived.

6.6.1 Single Endpoint Case

Recall that in the linear regression model with one response Y and a set of p predictor variables $X_1, X_2, \cdots, X_p$, the model is:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$. We assume the σ^2 is known.

In the setting of group sequential design, Let $\mathbf{Y}^{(k)}$ be the vector of n_k observations of response variable at kth interim analysis and $\mathbf{X}(k)$ be the corresponding design matrix. The least square estimate (LSE) at the kth interim analysis is:

$$\hat{\boldsymbol{\beta}}^{(k)} = (\mathbf{X}^{(k)\prime}\mathbf{X}^{(k)})^{-1}\mathbf{X}^{(k)\prime}\mathbf{Y}^{(k)}.$$

with $E(\hat{\boldsymbol{\beta}}^{(k)}) = \boldsymbol{\beta}$ and $Var(\hat{\boldsymbol{\beta}}^{(k)}) = \sigma^2(\mathbf{X}^{(k)\prime}\mathbf{X}^{(k)})^{-1}$ and

Theorem 6.2 $Cov(\hat{\boldsymbol{\beta}}^{(k_1)}, \hat{\boldsymbol{\beta}}^{(k_2)}) = Var(\hat{\boldsymbol{\beta}}^{(k_2)})$ *for* $k_1 \leq k_2$.

Jennison and Turnbull (1999) provided the details of the proof of this property. We recap this proof because we will follow a similar approach to prove the results in a later section.

Proof of Theorem 6.2: Each $\hat{\boldsymbol{\beta}}^{(k)}$ is a linear function of $\mathbf{Y}^{(k)}$. For $k_1 \leq k_2$, we can write

$$\hat{\boldsymbol{\beta}}^{(k_1)} = \mathbf{A}'\mathbf{Y}^{(k_2)},$$

for some matrix $\mathbf{A}$ with dimension $n_{K_2} \times (p+1)$ matrix. Since $\hat{\boldsymbol{\beta}}^{(k)}$ is an unbiased estimate of $\boldsymbol{\beta}$,

$$E(\mathbf{A}'\mathbf{Y}^{(k_2)}) = \mathbf{A}'\mathbf{X}^{(k_2)}\boldsymbol{\beta} \text{ for all } \boldsymbol{\beta},$$

and it follows that $\mathbf{A}'\mathbf{X}^{(k_2)} = \mathbf{I}_{p+1}$, the $p+1$-dimensional identity matrix. Therefore,

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}^{(k_1)}, \hat{\boldsymbol{\beta}}^{(k_2)}) &= Cov(\mathbf{A}'\mathbf{Y}^{(k_2)}, (\mathbf{X}^{(k_2)\prime}\mathbf{X}^{(k_2)})^{-1}\mathbf{X}^{(k_2)\prime}\mathbf{Y}^{(k_2)}) \\ &= \mathbf{A}'Var(\mathbf{Y}^{(k_2)})\mathbf{X}^{(k_2)\prime}(\mathbf{X}^{(k_2)}\mathbf{X}^{(k_2)})^{-1} \\ &= (\mathbf{X}^{(k_2)\prime}\mathbf{X}^{(k_2)})^{-1}\sigma^2. \end{aligned}$$

Thus Theorem 6.2 is proved. ■

Let X_1 be the treatment assignment. Then β_1 is the treatment effect. In general, we are interested in the hypothesis that

$$H_0 : \beta_1 = 0;$$

$$H_a : \beta_1 > 0.$$

We have $Var(\hat{\beta_1}^{(k)}) = \sigma^2(\mathbf{X}^{(k)\prime}\mathbf{X}^{(k)})^{-1}_{22}$. Note that $Var(\hat{\beta}_1^{(k)})$ is a constant. The test statistic is:

$$Z^{(k)} = \frac{\hat{\beta}_1^{(k)}}{\sqrt{Var(\hat{\beta}_1^{(k)})}} \sim N(\frac{\beta_1}{\sqrt{Var(\hat{\beta}_1^{(k)})}}, 1). \tag{6.16}$$

When H_0 is true, $Z^{(k)} \sim N(0,1)$. The group sequential design can be carried over.

6.6.2 Multivariate Regression Model

In an analysis when multiple endpoints are considered, multivariate regression model needs to be used. The multivariate regression models the d correlated responses $\mathbf{Y}' = (Y_1, Y_2, \cdots, Y_d)$ with assumed known $Var(\mathbf{Y}) = \mathbf{\Sigma}$ where

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix},$$

and a single set of the p predictor variables $X_1, X_2, \cdots, X_p$. Each response has the same as its own regression model, that is:

$$\begin{aligned} Y_1 &= \beta_{01} + \beta_{11}X_1 + \cdots + \beta_{p1}X_p + \epsilon_1 \\ Y_2 &= \beta_{02} + \beta_{12}X_1 + \cdots + \beta_{p2}X_p + \epsilon_2 \\ &\vdots \quad\quad \vdots \\ Y_d &= \beta_{0d} + \beta_{1d}X_1 + \cdots + \beta_{pd}X_p + \epsilon_d. \end{aligned} \tag{6.17}$$

The error term $\boldsymbol{\epsilon}'$ has $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\epsilon}) = \mathbf{\Sigma}$.

We present the model in matrix notation. With the sample size of n, the design matrix

$$\mathbf{X}_{n\times(p+1)} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}$$

is the same as the single response regression model. In the discuss of this chapter, we let X_1 be the

treatment assignment. The response matrix is

$$\mathbf{Y}_{n\times d} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1d} \\ Y_{21} & Y_{22} & \cdots & Y_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nd} \end{bmatrix} = [\mathbf{Y_{(1)}}|\mathbf{Y_{(2)}}|\cdots|\mathbf{Y_{(d)}}].$$

The model parameters matrix is

$$\boldsymbol{\beta}_{(p+1)\times d} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0d} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pd} \end{bmatrix} = [\boldsymbol{\beta_{(1)}}|\boldsymbol{\beta_{(2)}}|\cdots|\boldsymbol{\beta_{(d)}}].$$

In the second row of the $\boldsymbol{\beta}_{(p+1)\times d}$, the vector $\boldsymbol{\beta}_{1j} = (\beta_{11}, \beta_{12}, \cdots, \beta_{1d})$ is the treatment effect. β_{1j} is the treatment effect on the jth endpoint. And the error matrix is

$$\boldsymbol{\epsilon}_{n\times d} = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \cdots & \epsilon_{1d} \\ \epsilon_{21} & \epsilon_{22} & \cdots & \epsilon_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n1} & \epsilon_{n2} & \cdots & \epsilon_{nd} \end{bmatrix} = [\boldsymbol{\epsilon_{(1)}}|\boldsymbol{\epsilon_{(2)}}|\cdots|\boldsymbol{\epsilon_{(d)}}].$$

The multivariate regression model (6.17) can be written as:

$$\mathbf{Y}_{n\times d} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times d} + \boldsymbol{\epsilon}_{n\times d}, \tag{6.18}$$

with

$$E(\boldsymbol{\epsilon}_{(j)}) = \mathbf{0} \text{ and } Cov(\boldsymbol{\epsilon}_{(j)}, \boldsymbol{\epsilon}_{(k)}) = \sigma_{jk}\mathbf{I} \text{ where } j, k = 1, 2, \cdots, K.$$

The least square estimate is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{6.19}$$

$\hat{\boldsymbol{\beta}}$ is the unbiased estimator, thus $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. And

$$Cov(\hat{\boldsymbol{\beta}}_{(j)}, \hat{\boldsymbol{\beta}}_{(k)}) = \sigma_{jk}(\mathbf{X}'\mathbf{X})^{-1}. \tag{6.20}$$

Therefore,

$$Cov(\hat{\beta}_{1j}, \hat{\beta}_{1k}) = \sigma_{jk}(\mathbf{X}'\mathbf{X})^{-1}_{22}. \tag{6.21}$$

6.6.3 GSD with Multiple Endpoints

We still consider the example of the simplest case of a two-stage designed study with two normal endpoints Y_1 and Y_2 ($d = 2$), and two-stage study ($K = 2$) and a set of covariates, where the regression model is

$$\begin{aligned} Y_1 &= \beta_{01} + \beta_{11}trt + \beta_{21}X_2 + \cdots + \beta_{p1}X_p + \epsilon_1 \\ Y_2 &= \beta_{02} + \beta_{12}trt + \beta_{22}X_2 + \cdots + \beta_{p2}X_p + \epsilon_1. \end{aligned} \tag{6.22}$$

The hypothesis we are interested is

$$\begin{aligned} &H_0 : \beta_{11} \leq 0 \text{ or } \beta_{12} \leq 0; \\ &H_a : \beta_{11} > 0 \text{ and } \beta_{12} > 0. \end{aligned} \tag{6.23}$$

The test statistics for the two endpoints at two stages are:

$$Z_j^{(k)} = \frac{\hat{\beta}_{1j}^{(k)} - \beta_{1j}}{\sqrt{Var(\hat{\beta}_{1j}^{(k)})}} \sim N(0,1), \text{ for } j = 1, 2 \text{ and } k = 1, 2.$$

The joint distribution of the LSE has the following properties:

Property 6.3 *The LSE of the two stage tests and for two endpoints jointly follows the* $4\,(d \times K)$*-dimensional multivariate normal distribution with*

1. *Each* $\hat{\beta}_{1j}^{(k)}$ *for* $k = 1, 2$ *and* $j = 1, 2$ *is an unbiased estimator;*

2. $Cov(\hat{\beta}_{1j}^{(1)}, \hat{\beta}_{1j}^{(2)}) = Var(\hat{\beta}_{1j}^{(2)})$ *for* $j = 1, 2$*, as of Theorem 6.2;*

3. $Cov(\hat{\beta}_{11}^{(k)}, \hat{\beta}_{12}^{(k)}) = \sigma_{12}(\mathbf{X}^{(k)\prime}\mathbf{X}^{(k)})_{22}^{-1}$ *for* $k = 1, 2$*, as of* (6.21)*;*

4. $Cov(\hat{\beta}_{11}^{(k_1)}, \hat{\beta}_{12}^{(k_2)}) = \sigma_{12}(\mathbf{X}^{(k_2)\prime}\mathbf{X}^{(k_2)})_{22}^{-1}$ *for* $k_1 \leq k_2$.

Proof of Property 6.3: It is trivial to verify $1, 2$ and 3. The proof is very similar to the proof of Theorem 6.2. We have:

$$\hat{\boldsymbol{\beta}}_1^{(k_1)} = \mathbf{A}'\mathbf{Y_1}^{(k_2)}$$

for some matrix $\mathbf{A}$ with dimension $n_{K_2} \times (p+1)$ matrix. Since $\hat{\boldsymbol{\beta}}^{(k)}$ is an unbiased estimate of $\boldsymbol{\beta}$,

$$E(\mathbf{A}'\mathbf{Y_1}^{(k_2)}) = \mathbf{A}'\mathbf{X}^{(k_2)}\boldsymbol{\beta} \text{ for all } \boldsymbol{\beta},$$

and it follows that $\mathbf{A}'\mathbf{X}^{(k_2)} = \mathbf{I}_{p+1}$, the $p+1$-dimensional identity matrix. Therefore,

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}_1^{(k_1)}, \hat{\boldsymbol{\beta}}_2^{(k_2)}) &= Cov(\mathbf{A}'\mathbf{Y_1}^{(k_2)}, (\mathbf{X}^{(k_2)\prime}\mathbf{X}^{(k_2)})^{-1}\mathbf{X}^{(k_2)\prime}\mathbf{Y_2}^{(k_2)}) \\ &= \mathbf{A}'Cov(\mathbf{Y_1}^{(k_2)}, \mathbf{Y_2}^{(k_2)})\mathbf{X}^{(k_2)}(\mathbf{X}^{(k_2)\prime}\mathbf{X}^{(k_2)})^{-1} \\ &= (\mathbf{X}^{(k_2)\prime}\mathbf{X}^{(k_2)})^{-1}\sigma_{12}. \end{aligned}$$

Thus the property is proved. The theory is more complicated if the variance-covariance information is unknown. Similar to the group sequential test without adjusting the other covariates, the joint distribution of the LSE is not the multivariate T distribution, even though the marginal of the distribution is a T distribution.

Chapter 7

Multivariate B-value Tool

This chapter extends the B-value tool to multi-dimensions, thus named the multivariate B-value tool, so that we can calculate the conditional power of clinical trials with multiple co-primary endpoints. This chapter is organized as follows: Section 7.1 introduces the multivariate B-value tool. It follows the multi-dimensional Brownian motion distribution. Section 7.2 derives the formula to calculate the conditional power for multiple co-primary endpoints. Section 7.3 illustrates the use of the multivariate B-value tool in some common cases: the two sample test of multivariate mean vectors. Section 7.5 introduces the method of re-estimating the sample size based on the conditional power using the fixed weight approach.

7.1 Multivariate B-value Tool

This section introduces the multivariate B-value, which is a useful tool for the group sequential design and monitoring with multiple co-primary endpoints.

Consider the group sequential tests of the Hypothesis (5.4). At interim analysis, after observing n samples(n $\leq$ N), the B-value of each endpoint j is $B_j(t) = Z_{n,j}\sqrt{t}$, where $Z_{n,j} = \sum_{i=1}^{n} X_{ij}/\sqrt{n}$ is the Z-statistic at interim analysis time t. As described in Section 5.1.2, for a single endpoint, the B-value follows Brownian motion with drift parameter, $\sqrt{N}\mu$. The properties of Brownian motion are used to describe the distribution of conditional power under the assumption that the parameter estimated at interim analysis is equal to the true value. In the multi-dimension case, it has: $\forall j$ and k,

where $j, k = 1, 2, \ldots, d$

$$\begin{aligned} Cov(B_j(t), B_k(t)) &= Cov\left[\frac{\sum_{i=1}^{n} X_{ij}}{\sqrt{N}}, \frac{\sum_{i=1}^{n} X_{ik}}{\sqrt{N}}\right] \\ &= \frac{n}{N}\rho \\ &= t\rho. \end{aligned}$$

The multivariate B-value of each dimension has the same properties as the one-dimensional B-value described in Property (5.1) in Section 5.1.2. Therefore, the multivariate B-value has the consequence properties:

Property 7.1 *The distribution of multivariate B-value:*

1. $B_j(t)$ *follows Brownian motion with drift parameter* $\sqrt{N}\mu_j$*;*
2. $Cov(B_j(t), B_k(t)) = t\rho_{jk}$ *for* $j, k = 1, \ldots, d$*.*

Figure 7.1 shows one realization of the two-dimensional correlated Brownian Motion with the correlation between the two endpoints is 0.3. It is worthwhile to point out that if the two endpoints are normally distributed, at the end of the study, i.e., $t = 1$, the covariance between the test statistics (B-values) of any two dimensions is the same as the covariance of corresponding two endpoints.

7.2 Conditional Power

In this section, we consider the specifics of the two-stage design and provide the formula to calculate the conditional power of the two-stage group sequential design with multiple co-primary endpoints. First we derive the covariance of the test statistics among all the endpoints at the end of the study conditional on the interim B-values. For $\forall j$ and k, where $j, k = 1, 2, \ldots, d$, it it straightforward to see:

$$\begin{aligned} Cov(B_j(1), B_k(1)|B_j(t), B_k(t)) &= Cov[B_j(t) + (B_j(1) - B_j(t)), B_k(t) + (B_k(1) - B_k(t))] \\ &= Cov[B_j(1) - B_j(t), B_k(1) - B_k(t)] \end{aligned}$$

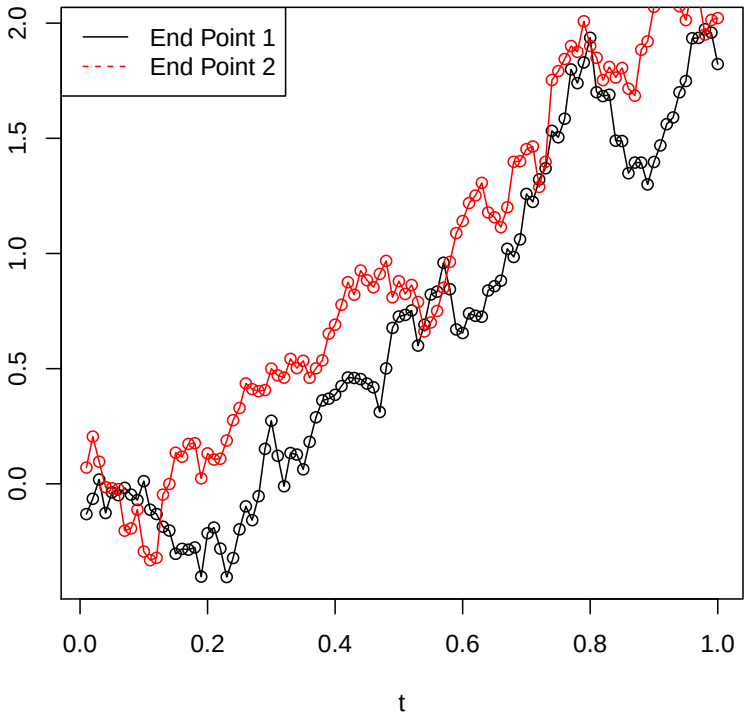

Figure 7.1: Correlated Brownian motion of bivariate b-value

$$
\begin{aligned}
&= Cov\left[\frac{\sum_{i=n+1}^{N} X_{ij}}{\sqrt{N}}, \frac{\sum_{i=n+1}^{N} X_{ik}}{\sqrt{N}}\right] \\
&= \frac{1}{N}(N-n)\rho = (1-t)\rho_{jk}. \qquad (7.1)
\end{aligned}
$$

Therefore,

$$\mathbf{B(1)}|\mathbf{B(t)} \sim \mathbf{N_d}(\mathbf{B(t)} + \mathbf{\Theta}(1-t), (1-t)\mathbf{\Sigma}). \qquad (7.2)$$

Using the $\hat{\mathbf{\Theta}} = \mathbf{B(t)}/t$ to replace $\mathbf{\Theta}$, we have:

$$\mathbf{B(1)}|\mathbf{B(t)} \sim \mathbf{N_d}(\mathbf{B(t)} + \frac{\mathbf{B(t)}}{t}(1-t), (1-t)\mathbf{\Sigma}). \qquad (7.3)$$

The probability we reject the Hypothesis (5.4) conditional on interim information is:

$$P(\mathbf{B(1)} > c\mathbf{1_d}|\mathbf{B(t)}, \mathbf{\Theta}) = \int_c^\infty \cdots \int_c^\infty \mathbf{f_{Z_N|B(t)}}(\mathbf{z})d\mathbf{z}, \qquad (7.4)$$

where $\mathbf{f_{Z_N|B(t)}}(\mathbf{z})$ is the probability density function of (7.3). The integration is carried out based on the simulation.

In particular that of $d = 2$, that is two random variables E1 and E2, (7.2) becomes

$$\left(\begin{array}{c|c} Z_{N,1} = B_1(1) & B_1(t) \\ Z_{N,2} = B_2(1) & B_2(t) \end{array}\right) \sim \mathbf{N_2}\left(\left(\begin{array}{c} B_1(t) + \Theta_1(1-t) \\ B_2(t) + \Theta_2(1-t) \end{array}\right), \left(\begin{array}{cc} 1-t & (1-t)\rho \\ (1-t)\rho & 1-t \end{array}\right)\right). \qquad (7.5)$$

Table 7.1 provides the results of overall conditional power, when the conditional power of each sub-hypothesis test achieves 85%, controlling each of the single tests at one-sided 2.5%. From Table 7.1, we observed that there are two main factors affecting the overall CP: the correlation between each pair of endpoints and number of co-primary endpoints. Controlling the test to achieve a conditional power of 85% on every single endpoint, we can see with the increase of the correlation, the overall CP increases as well. As the number of co-primary endpoints increases, the overall CP decreases. Intuitively, the overall conditional power will be less than the smallest conditional power of all the

H

Table 7.1: Overall conditional power when each marginal reaches 85% conditional power at one-sided significant level 2.5%

	number of co-primary endpoints				
correlation	2	3	4	5	9
0	0.73	0.61	0.52	0.44	0.23
0.2	0.74	0.65	0.58	0.51	0.35
0.5	0.76	0.70	0.65	0.61	0.51
0.8	0.79	0.76	0.73	0.72	0.66
1.0	0.85	0.85	0.85	0.85	0.85

individual endpoints. If all endpoints are mutually independent, the overall conditional power is the product of conditional power for all dimensions. If all endpoints are perfectly correlated, the conditional powers of individual sub-hypotheses will be the same as the overall conditional power. In general, the overall conditional power will be somewhere between the two extreme cases. The results in Table 7.1 validate this conclusion.

In practice, it is common that at the interim analysis, we observe different amounts of information for different endpoints. For example, if we have two endpoints, EP1 and EP2, we observed n_1 for EP1, n_2 for EP2 and $n_1 < n_2$. Thus $t_1 = n_1/N$ and $t_2 = n_2/N$. Then, (7.1) becomes

$$\begin{aligned} Cov(B_1(1), B_2(1)|B_1(t_1), B_2(t_2)) &= Cov[B_1(t_1) + (B_1(1) - B_1(t_1)), B_2(t_2) + (B_2(1) - B_2(t))] \\ &= Cov[B_1(1) - B_1(t_1), B_2(1) - B_2(t_2)] \\ &= Cov\left[\frac{\sum_{i=n_1+1}^{N} X_{i1}}{\sqrt{N}}, \frac{\sum_{i=n_2+1}^{N} X_{i2}}{\sqrt{N}}\right] \\ &= \frac{1}{N}(N - n_2)\rho = (1 - t_2)\rho_{12}. \end{aligned}$$

Thus, the conditional distribution (7.5) becomes

$$\left(\begin{array}{c|c} Z_{N,1} = B_1(1) & B_1(t_1) \\ Z_{N,2} = B_2(1) & B_2(t_2) \end{array}\right) \sim \mathbf{N_2}\left(\left(\begin{array}{c} B_1(t_1) + \Theta_1(1 - t_1) \\ B_2(t_2) + \Theta_2(1 - t_2) \end{array}\right), \left(\begin{array}{cc} 1 - t_1 & (1 - t_2)\rho \\ (1 - t_2)\rho & 1 - t_2 \end{array}\right)\right).$$

7.3 Two Sample Test of Means

In clinical trial applications, it is common to compare the new treatment group against the control group. It is applicable to use the multivariate B-value tool in this situation.

7.3.1 Two Sample Test of Means with Known Variance-Covariance

Now we illustrate the application of multivariate B-value to calculate the conditional power of two sample test on two co-primary endpoints. We are interested in the question of comparing the new drug to the placebo on two endpoints. The hypothesis is:

$$H_0 : \Delta_j \leq 0 \text{ for at least one } j,\ j = 1, 2;$$
$$H_a : \Delta_j > 0 \text{ for both } j = 1, 2.$$

Where Δ_j is the effect difference between the new treatment and placebo on the jth endpoint, while still assuming positive difference indicates efficacy. The standard test for each sub-hypothesis test is the two-sample test of mean. Patients are allocated into treatment and placebo groups. Let $\mathbf{X} \sim \mathbf{N}_2(\boldsymbol{\mu_X}, (1-\rho)\mathbf{I}_2 + \rho\mathbf{J}_2)$ and $\mathbf{Y} \sim \mathbf{N}_2(\boldsymbol{\mu_Y}, (1-\rho)\mathbf{I}_2 + \rho\mathbf{J}_2)$ be the variables of interest for the two groups respectively, where the I_2 is the 2-dimensional identity matrix and J_2 is the 2×2 matrix with all 1's. For simplicity, we assume the treatment group and control group have the same covariance structure between the two endpoints. The covariance between the test statistics conditional on interim B values is:

$$\begin{aligned}
& Cov(B_1(1), B_2(1)|B_1(t), B_2(t)) \\
= \ & Cov(B_1(t) + (B_1(1) - B_1(t)), B_2(t) + (B_2(1) - B_2(t))) \\
= \ & Cov(B_1(1) - B_1(t), B_2(1) - B_2(t)) \\
= \ & Cov\left(\frac{\sum_{i=n+1}^{N} X_{i1} - \sum_{i=n+1}^{N} Y_{i1}}{\sqrt{2N}}, \frac{\sum_{i=n+1}^{N} X_{i2} - \sum_{i=n+1}^{N} Y_{i2}}{\sqrt{2N}}\right) \\
= \ & \frac{1}{2N} 2(N-n)\rho \\
= \ & (1-t)\rho.
\end{aligned}$$

7.3.2 Two Sample Test of Means with Unknown Variance-Covariance

In most practical cases, we don't know the variance-covariance structure among the endpoints ahead of the study. In this case, we can use interim information to estimate the variance-covariance matrix. Let $\mathbf{X} \sim \mathbf{N}_2(\boldsymbol{\mu_X}, \boldsymbol{\Sigma})$ and $\mathbf{Y} \sim \mathbf{N}_2(\boldsymbol{\mu_Y}, \boldsymbol{\Sigma})$ be the variables of two groups respectively. Here we assume the treatment group and control group have the same covariance structure between the two endpoints with

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

Each marginal distribution will have the same properties as the ones stated in Section 5.1.2. For large n and N, we assume the interim estimated variance-covariance will remain the same for the following study. That is:

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} s_{1,n}^2 & s_{12,n} \\ s_{12,n} & s_{2,n}^2 \end{pmatrix}.$$

where $s_{j,n}, j = 1, 2$ are the sample variances of each endpoint and $s_{12,n}$ is the sample covariance between the two endpoints, based on the interim information. Further, the estimated correlation between the two endpoints is $\hat{\rho}_{1,2} = \frac{s_{12,n}}{s_{1,n}s_{2,n}}$. The covariance between the test statistics conditional on interim B values is

$$\begin{aligned}
& Cov(T_1(1), T_2(1)|B_1(t), B_2(t)) \\
= \;& Cov\left(\frac{\sum_{i=1}^N X_{i1} - \sum_{i=1}^N Y_{i1}}{\sqrt{2N}S_{1,N}}, \frac{\sum_{i=1}^N X_{i2} - \sum_{i=1}^N Y_{i2}}{\sqrt{2N}S_{2,N}}|B_1(t), B_2(t)\right) \\
\approx \;& Cov\left(\frac{\sum_{i=1+n}^N X_{i1} - \sum_{i=1+n}^N Y_{i1}}{\sqrt{2N}s_{1,n}}, \frac{\sum_{i=1+n}^N X_{i2} - \sum_{i=1+n}^N Y_{i2}}{\sqrt{2N}s_{2,n}}\right) \text{ using } s_{j,n} \text{ to replace } s_{j,N} \\
= \;& \frac{2(N-n)\sigma_{12}}{2Ns_{1,n}s_{2,n}} \\
\approx \;& (1-t)\frac{s_{12,n}}{s_{1,n}s_{2,n}} \\
= \;& (1-t)\hat{\rho}.
\end{aligned}$$

For example, if the assumed effect size is 0.4 (=$(\mu_{X,j} - \mu_{Y,j})/\sigma_j$) for both primary endpoints ($j = 1, 2$) with designed correlation 0.4, 130 patients are required for each group to reach the overall power 80%. At the interim analysis when $t = 0.5$, the information of 65 patients was collected for each group. The observed $B_1(0.5) = 1.46$ and $B_2(0.5) = 1.60$ and the estimated correlation between the two endpoints is $\hat{\rho} = 0.3$. Then $\hat{\Theta}_1 = \frac{1.46}{0.5} = 2.92$ and $\hat{\Theta}_2 = \frac{1.60}{0.5} = 3.20$. Therefore, we have

$$\left(\begin{array}{c|c} Z_{N,1} = B_1(1) & B_1(0.5) \\ Z_{N,2} = B_2(1) & B_2(0.5) \end{array} \right) \sim \mathbf{N_2} \left(\left(\begin{array}{c} 2.92 \\ 3.20 \end{array} \right), \left(\begin{array}{cc} 0.5 & 0.15 \\ 0.15 & 0.5 \end{array} \right) \right).$$

Using Monte Carlo integration for Equation (7.4), we get the conditional power of the Endpoint 1 is 89.9% and that of Endpoint 2 is 95.7%. The overall conditional power is 86.8%.

7.4 Binary Endpoints

This section extends the use of multivariate B-value tool to calculate the conditional power of clinical trials with multiple co-primary binary endpoints. As discussed in the previous sections, to calculate the conditional power of multiple co-primary endpoints, the key step is to determine the covariance/correlation between the test statistics of each endpoint conditional on the interim observed information. For the binary endpoint, the asymptotic Z statistic is the sum of the random variable. Therefore, it is important to measure the association between two binary variables in order to determine the covariance between the two Z statistics.

7.4.1 Single-arm Design

In the single-arm designed clinical trials, Let $\mathbf{X} = \{X_1, \cdots, X_d\}$ with $X_j \sim$ Bern(p_{X_j}). Denote $\mathbf{p} = (p_1, \cdots, p_d)$. The hypothesis test is:

$$\begin{aligned} &H_0 : p_j \leq p_{0,j} \text{ for any } j = 1, \cdots, d; \\ &H_a : p_j > p_{0,j} \text{ for all } j = 1, \cdots, d, \end{aligned} \tag{7.6}$$

where $\mathbf{p_0} = (p_{0,1}, \cdots, p_{0,d})$ is the proportion vector under the null hypothesis. For any dimension j, the test statistic at the end of the study is

$$Z_{N,j} = \frac{\hat{p}_j - p_{0,j}}{\sqrt{\frac{p_{0,j}(1-p_{0,j})}{N}}} = \frac{\sum_{i=1}^{N} X_{ij} - Np_{0,j}}{\sqrt{Np_{0,j}(1-p_{0,j})}}.$$

After observing n patients, we do the interim test. The test statistic

$$Z_{n,j} = \frac{\sum_{i=1}^{n} X_{ij} - np_{0,j}}{\sqrt{np_{0,j}(1-p_{0,j})}},$$

and the B-value is $B_j(t) = Z_{n,j}\sqrt{t}$ where $t = n/N$. As discussed in Section 6.5, for any two dimensions j and k,

$$Cov(Z_{N,j}, Z_{N,k}) \quad = \quad \varphi_{jk}.$$

As a reminder, γ_{jk} is the covariance between the two binary endpoints and $\varphi_{jk} = \frac{\gamma_{jk}}{\sqrt{p_{0,j}(1-p_{0,j})p_{0,k}(1-p_{0,k})}}$. In addition, it has

$$Cov(Z_j, Z_k | B_j(t), B_k(t)) = \varphi_{jk}(1-t)$$

Consider the trial with two co-primary binary endpoints, EP1 and EP2. Get the estimated $\hat{\phi}_{jk}$ by Equation (6.9). The joint distribution of the test statistics at the final stage, conditional on the interim B-values, is:

$$\left(\begin{array}{c|c} Z_{N,1} & B_1(t) \\ Z_{N,2} & B_2(t) \end{array} \right) \sim \mathbf{N}_2 \left(\left(\begin{array}{c} B_1(t) + (1-t)\Theta_1 \\ B_2(t) + (1-t)\Theta_2 \end{array} \right), \left(\begin{array}{cc} 1-t & \varphi_{12}(1-t) \\ \varphi_{12}(1-t) & 1-t \end{array} \right) \right),$$

with $\Theta = \frac{p_j - p_{0,j}}{\sqrt{Np_{0,j}(1-p_{0,j})}}$ At the interim analysis, the estimated $\hat{\gamma}_{12}$ and $\hat{p}_1$ and $\hat{p}_2$ can be obtained so that the conditional power can be further calculated.

7.4.2 Two-arm Design

The two-arm design clinical trial with single binary endpoint has been reviewed in Section 5.1.2. In the trial to compare the investigated treatment $\mathbf{X} = \{X_1, \cdots, X_d\}$ versus the control group $\mathbf{Y} = \{Y_1, \cdots, Y_d\}$. $X_j \sim$ Bern(p_{X_j}) and $Y_j \sim$ Bern(p_{Y_j}) with multiple co-primary binary endpoints. In the superiority trial, the hypothesis we are interested is:

$$\begin{aligned} H_0 &: p_{X_j} - p_{Y_j} \leq 0 \text{ for any } j = 1, \cdots, d \\ H_a &: p_{X_j} - p_{Y_j} > 0 \text{ for all } j = 1, \cdots, d \end{aligned} \tag{7.7}$$

N samples are needed for each treatment group to reach the desired level of power. For each dimension j, the test statistic is:

$$\begin{aligned} Z_j &= \frac{\hat{p}_{X_j} - \hat{p}_{Y_j}}{\sqrt{\bar{p}_j(1-\bar{p}_j)(\frac{1}{N} + \frac{1}{N})}} \\ &= \frac{\sum_{i=1}^N X_{ij} - \sum_{i=1}^N Y_{ij}}{\sqrt{2N\bar{p}_j(1-\bar{p}_j)}} \end{aligned}$$

For any two dimensions j and k,

$$\begin{aligned} Cov(Z_j, Z_k) &= Cov\left(\frac{\sum_{i=1}^N X_{ij} - \sum_{i=1}^N Y_{ij}}{\sqrt{2N\bar{p}_j(1-\bar{p}_j)}}, \frac{\sum_{i=1}^N X_{ik} - \sum_{i=1}^N Y_{ik}}{\sqrt{2N\bar{p}_k(1-\bar{p}_k)}}\right) \\ &= Cov\left(\frac{\sum_{i=1}^N X_{ij}}{\sqrt{2N\bar{p}_j(1-\bar{p}_j)}}, \frac{\sum_{i=1}^N X_{ik}}{\sqrt{2N\bar{p}_k(1-\bar{p}_k)}}\right) \\ &\quad + Cov\left(\frac{\sum_{i=1}^N Y_{ij}}{\sqrt{2N\bar{p}_j(1-\bar{p}_j)}}, \frac{\sum_{i=1}^N Y_{ik}}{\sqrt{2N\bar{p}_k(1-\bar{p}_k)}}\right) \\ &= \frac{N\gamma_{jk}^X}{2N\sqrt{\bar{p}_j(1-\bar{p}_j)\bar{p}_k(1-\bar{p}_k)}} + \frac{N\gamma_{jk}^Y}{2N\sqrt{\bar{p}_j(1-\bar{p}_j)\bar{p}_k(1-\bar{p}_k)}} \\ &= \frac{\gamma_{jk}}{\sqrt{\bar{p}_j(1-\bar{p}_j)\bar{p}_k(1-\bar{p}_k)}}. \end{aligned}$$

If $\gamma_{jk}^X = \gamma_{jk}^Y$, this means that the covariance between the endpoints for treatment and control groups are the same. Denote $\varphi_{jk} = \frac{\gamma_{jk}}{\sqrt{\bar{p}_j(1-\bar{p}_j)\bar{p}_k(1-\bar{p}_k)}}$. Furthermore, it has:

$$Cov(Z_j, Z_k | B_j(t), B_k(t)) = \varphi_{jk}(1-t).$$

In the case that the trial has two endpoints, EP1 and EP2. The joint distribution of test statistic conditional on the interim B-values is:

$$\left(\begin{array}{c|c} Z_{N,1} & B_1(t) \\ Z_{N,2} & B_2(t) \end{array} \right) \sim \mathbf{N}_2 \left(\left(\begin{array}{c} B_1(t) + \Theta_1(1-t) \\ B_2(t) + \Theta_2(1-t) \end{array} \right), \left(\begin{array}{cc} 1-t & \varphi_{12}(1-t) \\ \varphi_{12}(1-t) & 1-t \end{array} \right) \right).$$

where $\Theta_j = \sqrt{\frac{N}{2}} \frac{p_{X_j} - p_{Y_j}}{\sqrt{\bar{p}_j(1-\bar{p}_j)}}$ as discussed in Section 5.1.2. At the interim analysis, Θ can be replaced by $\hat{\Theta}$. The estimated $\hat{\gamma}_{jk}$ can be obtained by Equation (6.9) and thus $\hat{\varphi}_{jk} = \frac{\hat{\gamma}_{jk}}{\sqrt{\hat{\bar{p}}_j(1-\hat{\bar{p}}_j)\hat{\bar{p}}_k(1-\hat{\bar{p}}_k)}}$. Replace the parameters by the corresponding estimates, it gets:

$$\left(\begin{array}{c|c} Z_{N,1} & B_1(t) \\ Z_{N,2} & B_2(t) \end{array} \right) \sim \mathbf{N}_2 \left(\left(\begin{array}{c} B_1(t)/t \\ B_2(t)/t \end{array} \right), \left(\begin{array}{cc} 1-t & \hat{\varphi}_{12}(1-t) \\ \hat{\varphi}_{12}(1-t) & 1-t \end{array} \right) \right). \quad (7.8)$$

7.4.3 Example

In a two-arm designed clinical trial with two co-primary binary endpoints, we want to detect 0.1 difference for both endpoints at the one-sided significance level 2.5%. Let the study be two-stage (K=2) and the time of the interim analysis is $t = 0.5$. Using the O'Brien-Fleming-like error spending function to determine the stopping criterion, $c_1 = 2.963$ and $c_2 = 1.969$. Assume there is no information about the correlation between the two endpoints prior to the study. Then, to design the trial, we consider the most conservative case that assumes the two endpoints are independent. 510 samples are needed for each group to reach the overall power of 80% (the marginal power is $\sqrt{80\%} \approx 89.5\%$).

At the interim analysis, after observing 255 patients of each group, the observed $\hat{p}_{X_1} =$

$0.518, \hat{p}_{X_2} = 0.502, \hat{p}_{Y_1} = 0.463, \hat{p}_{Y_2} = 0.420$. Thus $\hat{\bar{p}}_1 = 0.490$ and $\hat{\bar{p}}_2 = 0.461$. Moreover, it has $\hat{\gamma} = 0.0821$ and $\hat{\varphi}_{12} = 0.330$. The interim test statistics are $Z_1(0.5) = 1.754$ and $Z_2(0.5) = 2.638$ for the two endpoints. This indicates that we do not stop the trial early for promising efficacy. The Distribution of (7.8) is:

$$\left(\begin{array}{c|c} Z_{N,1} & B_1(t) \\ Z_{N,2} & B_2(t) \end{array} \right) \sim \mathbf{N}_2 \left(\left(\begin{array}{c} 2.48 \\ 3.73 \end{array} \right), \left(\begin{array}{cc} 0.5 & 0.165 \\ 0.165 & 0.5 \end{array} \right) \right).$$

Using the numerical simulation, the overall conditional power is 76.1%.

7.5 Sample Size Re-estimation

Sample size re-estimation (SSR) provides a flexible tool for designing clinical trials. It allows one to adjust the sample size needed based on the results of interim analysis. Traditionally, the SSR can be classified into two groups: (1) blinded SSR and (2) unblinded SSR. Blinded SSR is based on nuisance designed parameter(s): usually the overall variability of continuous data or overall event rate of binary data. Lawrence Gould and Shih (1992) proposed a method to estimate the pooled unknown variance of the two samples of normally distributed data based on EM algorithm. It treats the randomization group as missing information.

The unblinded SSR is usually based on the observed effect size, the conditional error or conditional power, etc. There is potential issue of the inflation of Type I error. Chen et al. (2004) proved that if the conditional power at interim analysis is greater than 50%, the Type I error will not be inflated when increasing the sample size.

As discussed in Section 7.2, (7.4) provide the explicit formula to calculate the conditional power. It provides the basis of adjusting the sample size, if allowed, when the conditional power shows the more samples are needed. In this section, the fixed weight sample size re-estimation method to multiple co-primary endpoints based on the conditional power will be introduced.

7.5.1 Fixed-Weight Approach

In the case of testing the location of the mean of a normal population $N(\mu, \sigma^2)$, assume the variance is known with $\sigma = 1$. Consider the two-stage design ($K = 2$) and denote N_0 as the designed maximum sample size needed. The test statistic at the final stage can be decomposed as:

$$Z = \sqrt{t}Z_1 + \sqrt{1-t}Z_2 = \sqrt{\frac{n}{N_0}}Z_1 + \sqrt{\frac{N_0-n}{N_0}}\frac{\sum_{i=n+1}^{N_0} X_i}{N_0-n}.$$

The test statistic can be considered as the sum of the test statistics before and after the interim analysis by multiplying some weights $\sqrt{t}$ and $\sqrt{1-t}$. At the interim analysis, Z_1 is observed and fixed. Assume the sample size is adjusted from N_0 to N (N can increase or decreas). The modified fixed-weight test statistic is:

$$U = \sqrt{\frac{n}{N_0}}Z_1 + \sqrt{\frac{N_0-n}{N_0}}\frac{\sum_{i=n+1}^{N} X_i}{N-n}.$$

Cui et al. (1999) showed using the fixed-weight approach of the sample size re-estimation will not inflate the Type I error rate. This result holds for the co-primary endpoints. Still consider the example of hypothesis (5.4) with $d = 2$ co-primary endpoints and $K = 2$ stage design. The test statistic vector can be decomposed to

$$\begin{aligned}\mathbf{Z} &= \sqrt{t}\frac{\sum_{i=1}^{n}\mathbf{X_i}}{\sqrt{n}} + \sqrt{1-t}\mathbf{Z_{1-t}} \\ &= \sqrt{t}\frac{\sum_{i=1}^{n}\mathbf{X_i}}{\sqrt{n}} + \sqrt{1-t}\frac{\sum_{i=n+1}^{N_0}\mathbf{X_i}}{\sqrt{N_0-n}},\end{aligned}$$

where $t = n/N_0$. Denote N as the total sample size after adjustment, the weighted Z-statistic with the fixed rate, which is denoted as $\mathbf{U}$, is:

$$\mathbf{U} = \sqrt{t}\mathbf{Z_t} + \sqrt{1-t}\frac{\sum_{i=n+1}^{N}\mathbf{X_i}}{\sqrt{N-n}}, \tag{7.9}$$

where $\mathbf{U} = (U_1, U_2)$. Note that $N - n$ is the sample size needed for the second stage of the study. Denote it as n_2. It has

$$\begin{aligned} Cov(U_1, U_2|\mathbf{Z_t}) &= Cov\left(\sqrt{1-t}\frac{\sum_{i=n+1}^{N} X_{i1}}{\sqrt{N-n}}, \sqrt{1-t}\frac{\sum_{i=n+1}^{N} X_{i2}}{\sqrt{N-n}}\right) \\ &= (1-t)\rho. \end{aligned}$$

To control the Type I error rate at the worst case, consider $\boldsymbol{\mu_0} = (0, +\infty)$. In this case, it only needs to consider the EP1 as it can always declare the significance for the EP2. Cui et al. (1999) showed that for the single endpoint, when $\mu = 0$, $U_1|Z_{1,t}$ follows a normal distribution with

$$\begin{aligned} E(U_1|Z_{1,t}) &= \sqrt{t}Z_{1,t}; \\ Var(U_1|Z_{1,t}) &= (1-t). \end{aligned}$$

Thus $U_1|Z_{1,t}$ and $Z_1|Z_{1,t}$ will have the same distribution. It follows that $(Z_{1,t}, U_1)$ and $(Z_{1,t}, Z_1)$ have the same distribution. Therefore,

$$P_{\boldsymbol{\mu_0}}\left(\mathbf{Z_t} > c_1 \cup \mathbf{U} > c_2\right) = P_{\boldsymbol{\mu_0}}\left(\mathbf{Z_t} > c_1 \cup \mathbf{Z} > c_2\right).$$

Therefore, the maximum Type I error rate will be preserved. Let the conditional power be the target conditional power, $1 - \beta_0$, it has:

$$P(\sqrt{t}\mathbf{Z_t} + \sqrt{1-t}\frac{\sum_{i=n+1}^{N} \mathbf{X_i}}{\sqrt{n_2}} > \mathbf{z}_{\alpha_2}|\hat{\boldsymbol{\mu}}, \rho = \hat{\rho}) = 1 - \beta_0. \tag{7.10}$$

Re-writing (7.10), we can get:

$$P(\frac{\sum_{i=n+1}^{N} \mathbf{X_i}}{\sqrt{n_2}} > \frac{\mathbf{z}_{\alpha_2} - \sqrt{t}\mathbf{Z_t}}{\sqrt{1-t}}|\hat{\boldsymbol{\mu}}, \rho = \hat{\rho}) = 1 - \beta_0. \tag{7.11}$$

Note that

$$\frac{\sum_{i=n+1}^{N}\mathbf{X_i}}{\sqrt{n_2}}|\hat{\boldsymbol{\mu}}=\frac{\sum_{i=1}^{n}\mathbf{X_i}}{n},\rho=\hat{\rho}\sim\mathbf{N_2}\left(\begin{pmatrix}\sqrt{n_2}\hat{\mu}_1\\ \sqrt{n_2}\hat{\mu}_2\end{pmatrix},\begin{pmatrix}1 & \hat{\rho}\\ \hat{\rho} & 1\end{pmatrix}\right), \tag{7.12}$$

By combining (7.11) with (7.12), the n_2 can be solved using the numerical integration.

7.5.2 Hypothetical Example

Consider a one-arm two-stage study with two co-primary endpoints, EP1 and EP2. If we want to detect the effect size of 0.2 with both endpoints and the designed correlation is 0, 260 samples are needed to reach the power of 80%. At interim analysis, if the observed Z-values are $Z_1(0.5)=1.52$ and $Z_2(0.5)=1.83$, i.e, $B_1(0.5)=1.07$ and $B_2(0.5)=1.29$ and the estimated correlation is 0.35. Then, the conditional power of EP1 is 59.7% and that of EP2 is 80.7%, and the overall conditional power calculated using $\rho=\hat{\rho}$ is 52.0%. Plugging in all the information to (7.11) and (7.12) and letting $\beta_0=0.2$, we found 472 samples are needed for the second stage to reach the desired conditional power of 80%.

Using the fixed weight test statistic is controversial as it assigns different weights to different samples. It violates the "one patient, one vote" rule. Ideally, we should use equal weights. It is still prevalent in on-going research to find a reasonable approach of sample size re-estimation of co-primary endpoints.

Part III

Extension and Conclusion

Chapter 8

Blinded Interim Analysis Using Modality Inference

8.1 More on Two Component Normal Mixture Model

This section explores some interesting properties of the mixture of two d-dimensional multivariate normal distributions with equal mix proportions and the same variance-covariance matrix. The probability density function (PDF) of the mixture distribution is:

$$f(\mathbf{x}) = \frac{1}{2}\phi(\mathbf{x}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma}) + \frac{1}{2}\phi(\mathbf{x}, \boldsymbol{\mu_2}, \boldsymbol{\Sigma}) \tag{8.1}$$

where $\pi = 0.5$ and $\phi(\cdot, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the PDF of the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. For the PDF in (8.1), the ridgeline function (2.7) can be simplified to:

$$x(\alpha)^* = \alpha\boldsymbol{\mu_1} + (1-\alpha)\boldsymbol{\mu_2} \tag{8.2}$$

where $\alpha \in [0, 1]$. Let the constant

$$c = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}}$$

These are the following consequences of the results of the density in (8.1).

Result 8.1 *One critical value of the ridgeline is* $\mathbf{x_a} = \frac{1}{2}\boldsymbol{\mu_1} + \frac{1}{2}\boldsymbol{\mu_2}$*. In addition, if the Mahalanobis distance between* $\boldsymbol{\mu_1}$ *and* $\boldsymbol{\mu_2}$*, the* $D_M = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) \leq 4$*, Density* (8.1) *is unimodal. If* $D_M > 4$*, Density* (8.1) *is bimodal.*

Proof: It has

$$\begin{aligned} f(\alpha\boldsymbol{\mu_1} + (1-\alpha)\boldsymbol{\mu_2}) &= \frac{c}{2}exp\left\{-\frac{(1-\alpha)^2 D_M}{2}\right\} + \frac{c}{2}exp\left\{-\frac{\alpha^2 D_M}{2}\right\} \\ &\propto exp\left\{-\frac{(1-\alpha)^2 D_M}{2}\right\} + exp\left\{-\frac{\alpha^2 D_M}{2}\right\} \end{aligned}$$

which is a function of α, denote it as $g(\alpha)$. Differentiating $g(\alpha)$ and solve for 0, we have

$$g'(\alpha) \propto -exp\left\{-\frac{(1-\alpha)^2 D_M}{2}\right\}(1-\alpha) + exp\left\{-\frac{\alpha^2 D_M}{2}\right\}\alpha = 0 \tag{8.3}$$

$\alpha = \frac{1}{2}$ is one solution of the Equation (8.3). Thus we prove the first part of the results. If $\alpha \notin \{0, \frac{1}{2}, 1\}$, the Equation (8.3) can be simplified as

$$D_M = \frac{ln(1-\alpha) - ln\alpha}{\frac{1}{2} - \alpha} \tag{8.4}$$

The relationship between the α and D_M which satisfies the Equation (8.4) is shown in Figure 8.1. It is easy to verify the following properties of the Equation (8.4)

1. D_M is U-shape as the function of α;
2. As $\alpha \to \frac{1}{2}, D_M \to 4$;
3. D_M is monotone decreasing when $\alpha \in (0, \frac{1}{2})$ and monotone increasing when $\alpha \in (\frac{1}{2}, 1)$. Thus, if $D_M < 4$, there is no value of α to satisfy (8.4);
4. It has two solutions those are symmertic by $\frac{1}{2}$;
5. As $D_M \to \infty$, we have $\alpha \to 0$ or 1.

The last property shows that the modes of the Density (8.1) are not the means of the mixing normal distributions. When D_M is large enough, the modes converge to the means. Indeed, from Figure 8.1, it is clear to see that the values of α are very sensitive to the values of the Mahalanobis distance when α is in the ranges of the $(0, 0.2)$ and $(0.8, 1)$. We let $\boldsymbol{\mu_1} = (0,0)'$ and $\boldsymbol{\mu_2} = (0,m)'$. Note that in this case, $D_M = m^2$. The following table shows some values of α, D_M and m.

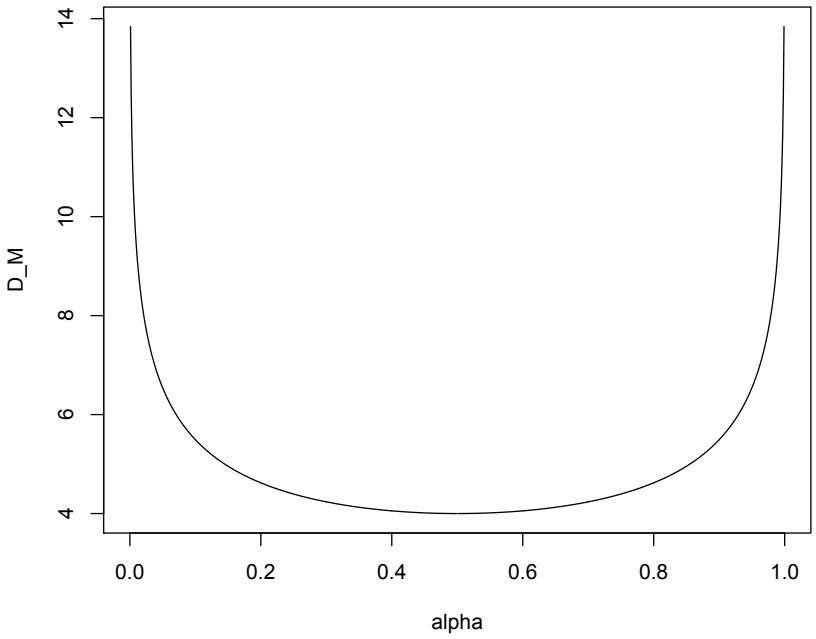

Figure 8.1: α vs Mahalanobis distance

α	0.00001	0.0001	0.001	0.01
D_M	23.026	18.424	13.841	9.378
m	4.799	4.292	3.720	3.062

Therefore, $g(\alpha)$ has one critical value when $D_M < 4$, and has three critical values when $D_M > 4$. Furthermore, the second-order derivative of $g(\alpha)$ evaluated at $\alpha = 1/2$ is:

$$g''(\frac{1}{2}) \propto exp\left\{-\frac{D_M}{8}\right\} - \frac{1}{4}D_M exp\left\{-\frac{D_M}{8}\right\}. \tag{8.5}$$

$g''(\frac{1}{2}) < 0$ if and only if $D_M < 4$. In summary, when $D_M < 4$, the probability density of (8.1) is unimodal. The mode is $\mathbf{x_a}$. When $D_M > 4$, the ridgeline of (8.1) has three critical values and it is bimodal. $\mathbf{x_a}$ is the antimode. Thus we prove the second part of the results. ■

It has been shown that for the density in (8.1) with a small D_M, the mean is not the mode. In fact, the density of the mean might be even lower than the density of the antimode.

Result 8.2 *$f(\boldsymbol{\mu_1}) - f(\mathbf{x_a}) \geq 0$ if and only if the Mahalanobis distance between $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$, the $D_M = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) \geq 4.875$.*

Proof:

$$\begin{aligned} f(\boldsymbol{\mu_1}) - f(\boldsymbol{x_a}) &\propto 1 + exp\left\{-\frac{D_M}{2}\right\} - 2exp\left\{-\frac{D_M}{8}\right\} \\ &= (a^4 - 2a + 1), \end{aligned}$$

where $a = exp\left\{-\frac{D_M}{8}\right\}$, and $0 < a < 1$. $f(\boldsymbol{\mu_1}) - f(\mathbf{x_a}) \geq 0$ implies $a \leq 0.5437$, and further implies $D_M \geq 4.875$. Thus we prove the result. ■

8.2 Blinded Interim Analysis

Consider the study of clinical trials with d multiple alternative primary normal endpoints. In a general case, the study has an equal or closed sample size per treatment and control groups. Assume

the variance-covariance structures are the same for the two treatment groups. Thus the distribution of the data is the mixture of two d-dimensional multivariate normal distributions with equal mixing proportions and the same variance-covariance matrix. The distribution of the probability density function is as (8.1):

$$f(\mathbf{x}) = \frac{1}{2}\phi(\mathbf{x}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma}) + \frac{1}{2}\phi(\mathbf{x}, \boldsymbol{\mu_2}, \boldsymbol{\Sigma})$$

where $\pi = 0.5$ and $\phi(\cdot, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the probability density function of the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Some interesting properties of (8.1) are studied in Section 8.1

One potential method of the blinded interim analysis is to apply the modality inference introduced in Chapter 3 on the data. Consider the blinded interim analysis of a clinical trial with multiple normal endpoints. The distribution of the multiple outcomes that are bimodal indicate that the treatment and control group have some different effects on any of the multiple endpoints. The hypothesis we are interested is:

H_0 : The distribution of $\mathbf{Y}$ is unimodal;

H_a : The distribution of $\mathbf{Y}$ is not unimodal.

As discussed in Section 8.1, in order to see the bimodal distribution, the Mahalanobis Distance between the two mean vectors needs to be greater than 4. This indicates the treatment effect needs to be large enough on at least one endpoint, which rarely happens in clinical trials. However, if the distribution is bimodal, there is enough confidence to claim that the mean vector of the two groups are quite separated. Note that this approach is not informative for the co-primary endpoints, but for the alternative multiple endpoints problem, in which the treatment needs to win on at least one endpoint. Some further research and practical applications need to be done to justify if it is applicable in the practice.

Chapter 9

Conclusion

In the first part of the dissertation, we developed the inference procedure to test the significance of a specific mode. The asymptotic distribution of the test statistic is derived based on the asymptotic normality of the kernel density estimates (KDE) to assess the significance of the mode. The traditional method to assess the significance of the modality of the data is to determine the test statistic and decide the reference distribution under the null hypothesis. Then, a large scale simulation is performed to simulate the reference data and compute the test statistic of the simulated reference data to form the null distribution of the test statistic. The method we introduced uses the asymptotic distribution of the statistic, thus, we can avoid the bootstrap testing, which could be computationally expensive.

Combined with the research work in Li et al. (2007), we provided a comprehensive mode hunting and inference tool for the investigated data set. The mode hunting and inference procedure is based on the KDE and using the normal density as the kernel function. It is important to select the bandwidth parameter h. There are two steps to select the bandwidth parameters. It is acknowledged that there is no best choice of h for estimating a density. For mode hunting, we chose to use the normal reference rule. For the inference, h has to satisfy the conditions of the asymptotic normality of the KDE. Due to the curse of the dimensionality, this method is limited to low to moderate dimensions.

We can apply this inference procedure on each pair of modes to assess how many modes the data has. In the MAC algorithm, the number of modes is the same as the number of clusters. It is

difficult but worthwhile to generate the automated algorithm to decide on how many clusters/modes of the data has, based on the modality inference procedure. The difficulty here is that the method is based on the KDE. The outliers of the data could easily form the spurious modes, especially for the high dimensional data, which makes it difficult to generate automated algorithm.

The parallel computing of MAC and its hierarchical version PHMAC is developed by using multiple processors simultaneously. It dramatically increases the computing speed. The R package *Modalclust* is created and is available on CRAN. One future direction from this stage is to increase computing speed, especially for relatively large data sets. From the discussion in Section 4.2, it is clear to see that parallel computing can dramatically increase the computing speed. This relies on the computing equipment. If one user has no multicore or only a few multicore processors available, it will take lot of the computing resources when clustering large data sets. One potential way to solve the computing speed problem is to use k-means or other faster clustering techniques initially, and using the HMAC from the centers of each cluster of initial clustering results. For example, if we have a data set with $20,000$ observations, we can use k-means clustering and choose a certain number of centers, e.g., 200 centers and perform k-means clustering first. Then, we start from the centers of 200 clusters and perform the consequent clustering by HMAC. Theoretically it is a sub-optimal way compared to running HMAC for all points. In practice, it is very useful to reduce the computing costs and still obtain the correct clustering.

In the second part of the dissertation, the method of group sequential design of clinical trials with multiple co-primary endpoints was introduced. It has been proved that the stopping boundaries of the group sequential trials with co-primary endpoints should be the same for a trial with a single endpoint. However, the power calculation is different and depends on the correlations among the endpoints (or the test statistics for the co-primary endpoints). It has been shown that the method gains power significantly when considering the correlation among the endpoints over the method in which multiple co-primary endpoints are considered as independent. Furthermore, the power increases as the correlations among endpoints increase and the number of endpoints decreases. The

design is α-exhausted as it considers the cumulative Type I error. Therefore, the family-wise type-I error is strongly controlled. The group sequential procedure with multiple co-primary binary endpoints has been discussed and it can be extended to continuous-binary endpoints. The key step is to evaluate the covariance between a continuous random variable and a binary random variable.

The multivariate B-value tool to calculate the conditional power for clinical trials with co-primary endpoints was discussed. Determining the conditional covariance between the test statistics is an important step in finding the joint distribution of test statistics of all the endpoints conditional on the interim multivariate B-value. It is worth pointing out that when considering the correlation between the endpoints, it indeed considers the covariance between the two endpoints. In most cases the test statistics are the sum of the independent random variables. The covariance is additive while the correlation is not. The multivariate B-value tool can also be extended to the study with co-primary binary and co-primary continuous-binary mixed endpoints.

The multiple co-primary time-to-event endpoints or time-to-event mixed with other types of endpoints is not considered in this dissertation. The problem with that is the definition of the information infraction is different for the time-to-event endpoint compared with the continuous or binary endpoint. For continuous or binary endpoint, the information infraction is the ratio of the sample size observed at interim analysis over the total sample size. For the time-to-event endpoint, the information infraction is the ratio of the number of the observed events over the expectation of the number of the events at the end of the study.

The stopping boundaries of the group sequential design procedure introduced in this dissertation only considers of early stopping for efficacy, but not for futility. However, based on the current regulatory practice of the non-binding futility rule, the stopping boundary and conditional power formulas can be conservatively applied to the adaptive trials with futility boundaries.

The results are based on the assumption that all the endpoints have the same effect size. In

practice, it is most often the case that the endpoints have different effect sizes. In this case, the marginal powers will be different. However, the conclusion is still valid and the method can be applied easily. The overall power has the same properties. When all endpoints are independent, the overall power will still be the product of the marginal ones. When all endpoints are perfectly correlated, the overall power will be the same as the minimum of the marginal ones. One potential research direction that this can apply to is distributing different errors on different endpoints, while still controlling the overall Type I error rate at the desired level.

It has been shown that the conditional power is very sensitive to the drift parameter, which cannot be well estimated at interim analysis, especially when t is small. Spiegelhalter et al. (1986) suggested to use the alternative Bayesian predictive power, which is defined as the average of the conditional power over the posterior distribution of the drift parameter conditional on the interim B-value. This idea can be generalized to multiple co-primary endpoints case. However, in practice, it might not be straightforward because choosing the proper prior is not an easy task.

Bibliography

R. Bellman. Adaptive control processes: a guided tour princeton university press. *Princeton, New Jersey, USA*, 1961.

R. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, pages 295–300, 1982.

P. Burman and W. Polonik. Multivariate mode hunting: Data analytic tools with measures of significance. *Journal of Multivariate Analysis*, 100(6):1198–1218, 2009.

G. Casella and R. Berger. *Statistical inference*. Duxbury Press, 2001.

Y. Chen, D. L. DeMets, and K. Gordon Lan. Increasing the sample size when the unblinded interim result is promising. *Statistics in medicine*, 23(7):1023–1038, 2004.

C. Chuang-Stein, P. Stryszak, A. Dmitrienko, and W. Offen. Challenge of multiple co-primary endpoints: a new approach. *Statistics in medicine*, 26(6):1181–1192, 2007.

L. Cui, H. Hung, and S.-J. Wang. Modification of sample size in group sequential clinical trials. *Biometrics*, 55(3):853–857, 1999.

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

A. Dmitrienko, A. C. Tamhane, and F. Bretz. *Multiple testing problems in pharmaceutical statistics*. CRC Press, 2010.

T. Duong and M. Hazelton. Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, 15(1):17–30, 2003.

B. Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 7(1):1–26, 1979.

B. Flury and H. Riedwyl. *Multivariate statistics: a practical approach*. Chapman & Hall, Ltd., 1988.

C. Fraley and A. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Pr, 1990.

C. Jennison and B. Turnbull. *Group sequential methods: applications to clinical trials*. Chapman & Hall/CRC, 1999.

K. Lan and D. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

K. Lan and J. Wittes. The b-value: a tool for monitoring data. *Biometrics*, pages 579–585, 1988.

A. Lawrence Gould and W. J. Shih. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics-Theory and Methods*, 21(10):2833–2853, 1992.

J. Li, S. Ray, and B. Lindsay. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8(8):1687–1723, 2007.

Q. Li and J. S. Racine. *Nonparametric econometrics: Theory and practice*. Princeton University Press, 2011.

B. Lindsay, M. Markatou, S. Ray, K. Yang, and S. Chen. Quadratic distances on probabilities: A unified foundation. *The Annals of Statistics*, 36(2):983–1006, 2008.

S. Lloyd. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2): 129–137, 1982.

G. McLachlan and D. Peel. *Finite mixture models*. Wiley. com, 2004.

L. Meyerson, R. Muirhead, P. Stryszak, A. Boddy, K. Chen, K. Copley-Merriman, W. Dere, S. Givens, D. Hall, D. Henry, et al. Multiple co-primary endpoints: Medical and statistical solutions a report from the multiple endpoints expert team of the pharmaceutical research and manufacturers of america. *Drug information journal*, 41:31–46, 2007.

M. Minnotte. Nonparametric testing of the existence of modes. *The Annals of Statistics*, pages 1646–1660, 1997.

P. O'Brien and T. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556, 1979.

S. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64 (2):191–199, 1977.

S. Ray and B. Lindsay. The topography of multivariate normal mixtures. *The Annals of Statistics*, 33(5):2042–2065, 2005.

S. Ray and S. Pyne. A computational framework to emulate the human perspective in flow cytometric data analysis. *PloS one*, 7(5):e35693, 2012.

S. Sain, K. Baggerly, and D. Scott. Cross-validation of multivariate densities. *Journal of the American Statistical Association*, 89(427):807–817, 1994.

D. Scott. Multivariate density estimation. *Multivariate Density Estimation, Wiley, New York, 1992*, 1, 1992.

B. Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 97–99, 1981.

T. Sozu, T. Kanou, C. Hamada, and I. Yoshimura. Power and sample size calculations in clinical trials with multiple primary variables. *Japanese Journal of Biometrics*, 27(2):83–96, 2006.

T. Sozu, T. Sugimoto, and T. Hamasaki. Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in medicine*, 29(21):2169–2179, 2010.

T. Sozu, T. Sugimoto, and T. Hamasaki. Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. *Journal of Biopharmaceutical Statistics*, 21(4): 650–668, 2011.

D. Spiegelhalter, L. Freedman, and P. Blackburn. Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials*, 7(1):8–17, 1986.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2): 411–423, 2001.

M. Wand and C. Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2): 97–116, 1994.

M. Wand and M. Jones. Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, 88(422):520–528, 1993.

C. Xiong, K. Yu, F. Gao, Y. Yan, and Z. Zhang. Power and sample size for clinical trials when efficacy is required in multiple endpoints: application to an alzheimer's treatment trial. *Clinical Trials*, 2(5):387–393, 2005.

X. Zhang, M. King, and R. Hyndman. A bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics & Data Analysis*, 50(11):3009–3031, 2006.

Curriculum Vitae