UNIVERSITY OF CALIFORNIA
RIVERSIDE

Nonparametric Methods for Combining Dependent Tests and Monitoring Count Data

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Linli Tang

December 2022

Dissertation Committee:

Dr. Jun Li, Chairperson
Dr. Subir Ghosh
Dr. Weixin Yao

The Dissertation of Linli Tang is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to express my sincere gratitude to my advisor, thesis and oral exam committee members, collaborators, family members, friends, and everyone who has helped me learn, grow, and improve throughout my academic journey at UC Riverside.

First and foremost, I would like to thank my advisor Dr. Jun Li for her support and guidance. I appreciate her for providing invaluable research ideas, discussion, and feedback on my research. This dissertation could not have been completed without her dedication.

My sincere thanks also go to the rest of my thesis committee, Dr. Subir Ghosh and Dr. Weixin Yao, for their support, encouragement, and insightful comments. I am also very grateful to have Dr. James Flegal and Dr. Soojin Park serve as my Ph.D. oral exam committee members.

Additionally, I would like to give special thanks to Dr. Zhiwei Zhang for offering me research opportunities and providing constant support for my academic endeavors.

I appreciate the teaching assistantships provided by the Department of Statistics and the Department of Economics, which has significantly eased my financial burdens during my Ph.D. studies. I am thankful to the Teaching Assistant Coordinators from both departments, Dr. Linda Penas, Dr. Analisa Flores, and Mr. Gary Kuzas, for their kind assistance.

Last but not least, I would like to thank my parents and my husband for their unconditional love and support. Without their support, I could not have had the opportunities to explore different fields of study and career opportunities. I enjoy every bit of the diverse experiences and am happy to see that they have turned into part of my skill set.

I especially thank my husband for sacrificing his time for family duties and providing me encouragement to overcome self-doubt.

The text appearing in Chapter 1, 2, 4 and the appendices of this dissertation, in part, is a reprint of the material as it appears in *Journal of Nonparametric Statistics* (2022), 34, 113-140. The coauthor Jun Li listed in that publication directed and supervised the research which forms the basis for this dissertation.

To my parents and my husband for all the support.

ABSTRACT OF THE DISSERTATION

Nonparametric Methods for Combining Dependent Tests and Monitoring Count Data

by

Linli Tang

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, December 2022
Dr. Jun Li, Chairperson

Combining multiple tests has many real world applications. However, most existing methods fail to directly take into account the underlying dependency among the tests. In the first project of this dissertation, we propose a novel procedure to combine dependent tests based on the notion of data depth. The proposed method can automatically incorporate the underlying dependency among the tests, and is nonparametric and completely data-driven. To demonstrate its application, we apply the proposed combining method to develop a new two-sample test for data of arbitrary types when the data can be metrizable and their information can be characterized by interpoint distances. Our simulation studies and real data analysis show that the proposed test based on the new combining method performs well across a broad range of settings and compares favorably with existing tests.

Count data monitoring has important applications in many fields. However, most of the existing control charts for monitoring count data are parametric. Parametric control charts can be problematic when the underlying parametric distributional assumption does not hold for the particular application. On the other hand, nonparametric control charts do

not require such distributional assumptions, and are more desirable in real-world situations where the underlying distribution cannot be easily described using a parametric distribution. In the second project of this dissertation, we extend the nonparametric control chart for continuous data monitoring in Li (2021) to count data monitoring. To guarantee a desired in-control performance, we further adopt the bootstrap procedure from Gandy and Kvaløy (2013) to help determine the control limit of our proposed control chart. Our simulation studies and real data analysis show that the proposed control chart performs well across a variety of settings, and compares favorably with other existing nonparametric control charts for count data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A data depth is a measure to depict the "depth" or "centrality" of a given point with respect to a multivariate data cloud or its underlying distribution, and it gives rise to a natural center-outward ordering of the points in a multivariate sample. Existing notions of data depth include: Mahalanobis depth (Mahalanobis 1936), halfspace depth (Hodges 1955; Tukey 1975), simplicial depth (Liu 1990), projection depth (Stahel 1981; Donoho 1982; Donoho and Gasko 1992; Zuo 2003), L_p depth (Zuo and Serfling 2000), zonoid depth (Koshevoy and Mosler 1997), spatial depth (Chaudhuri 1996; Vardi and Zhang 2000), onion depth (Barnett 1976; Eddy 1981), etc. For a more complete list of different notions of data depth, see Mosler and Mozharovskyi (2022).

Statistical process control (SPC) is a tool that applies statistical methods to measure, monitor and control a process. SPC keeps track of the output of a process over time and allows timely detection of abnormal variations of the process. It is a useful tool for mon-

itoring process performance and assuring process stability. In Chapter 3 of this dissertation, we focus on the control chart, which is the most commonly used SPC tool.

The control chart plots a statistic that measures a quality characteristic versus time or subgroup number, so that the process variations over time can be visualized. If the charting statistic goes beyond the predetermined control limit(s), an alarm will be signaled, indicating the process is out-of-control (OC). Otherwise, the state of the process is considered to be in-control (IC). The number of samples or subgroups collected before a chart first signals is a random variable called run length. The expected value of the run length distribution is known as the average run length (ARL).

In practice, control charts are implemented in two phases with different objectives (see Montgomery 2020). In Phase I, process data are collected and analyzed to ensure that the process is truly in a state of statistical control. The conventional practice for Phase I analysis is an iterative procedure, in which trial control limits are established to filter out possible OC samples. Through the detection of OC samples, any uncommon causes of variation are identified and eliminated. Once the IC state is established, a clean set of data that is representative of IC process performance is gathered. This data set is usually called Phase I sample or reference data. If the parameters of the underlying IC process distribution is unknown, the Phase I sample is used to estimate the IC distribution and construct reliable control limits of the control chart for Phase II process monitoring. In Phase II, the control chart is used to determine whether the process remains in control. At each time point, the charting statistic calculated based on successive observations drawn

from the process is compared to the predetermined control limit(s). An alarm will be triggered once the charting statistic goes beyond the control limit(s).

ARL is the most popular measure of the performance of control chart procedures. An ideal control chart would have a large IC ARL and a small OC ARL. However, this is difficult to achieve, because that a large IC ARL would result in a large OC ARL in most cases. In practice, quality practitioners usually fix the IC ARL at a given level, and the control chart that achieves the smallest OC ARL would be considered optimal.

The rest of the dissertation is organized as follows. In Chapter 2, we propose a novel nonparametric method of combining dependent tests based on data depth. We start this chapter with an introduction of background information and a literature review. We introduce our proposed method and its properties, and then demonstrate its application by developing a new two-sample test for data of arbitrary types. The performance of our proposed two-sample test is evaluated and compared with several existing tests in simulation studies and a real network data example. In Chapter 3, we develop a nonparametric control chart for detecting mean shifts for univariate count data. We firstly introduce the proposed charting statistic and a bootstrap-based algorithm for determining the control limits. Simulation studies and a real data analysis on crime statistic are conducted to demonstrate that the proposed control chart is more efficient than the existing nonparametric control charts. The concluding remarks are provided in Chapter 4.

Chapter 2

Combining Dependent Tests Based on Data Depth with Applications to the Two-Sample Problem for Data of Arbitrary Types

2.1 Introduction

Statistical hypothesis testing is a formal statistical procedure to determine whether or not to reject a given hypothesis based on data. In order to develop a powerful test, the key step is to identify a test statistic that can be used to assess the truth of the null hypothesis H_0. However, we often encounter situations where the hypothesis testing problem of interest is very complex and it is not easy to find an appropriate single overall test statistic. Often

in those situations, the null and alternative hypotheses, H_0 and H_1, can be properly broken down into a finite set of sub-hypotheses, H_{0i} and H_{1i}, $i = 1, ..., k$, each appropriate for a partial aspect of hypothesis of interest. The H_{0i} and H_{1i} are set up such that H_0 is true if all the H_{0i} are jointly true, and H_1 is true when at least one of the H_{1i} is true. For each of the sub-hypothesis testing problems, we assume that a partial test can be developed relatively easily. Denote T_i as the test statistic of the ith partial test that can be used to test the sub-hypothesis H_{0i} against H_{1i}, $i = 1, ..., k$. In order to provide an overall assessment of the original hypothesis H_0 versus H_1, we need combine those k partial tests based on T_1, ..., T_k.

When testing a single hypothesis, the Type-I error rate is simply the probability of a Type-I error. When testing multiple sub-hypotheses H_{0i} versus H_{1i}, $i = 1, ..., k$, there is a Type-I error associated with each partial test. Therefore, there can be different ways to define the overall Type-I error rate. One of the popular choices in the multiple hypothesis testing literature is the family-wise error rate (FWER), which is defined as the probability of making at least one Type-I error. Another popular choice for the overall Type-I error rate is the false discovery rate (FDR), which is defined as the expected proportion of Type-I errors among the rejected hypotheses. As described above, our goal is to test H_0 versus H_1, and a Type-I error from any of the k partial tests will lead to a Type-I error for testing H_0 versus H_1. Therefore, the FWER is a more appropriate overall Type-I error rate in this situation, and we will focus on combining the k partial tests to control the FWER in this paper.

When the k partial tests are assumed independent, there are many combining methods in the meta-analysis literature to control the FWER. The following lists some of the popular combining methods. Denote p_i as the p-value calculated from the test statistic T_i, $i = 1, ..., k$. Fisher's combining method (Fisher 1932) is based on the statistic $T_F = -2\sum_{i=1}^{k} \log(p_i)$. Lipták's combining method (Lipták 1958) uses the statistic $T_L = \sum_{i=1}^{k} \Phi^{-1}(1-p_i)$, where Φ is the standard normal cumulative distribution function (CDF). Tippett's combining method (Tippett 1931) is based on the statistic $T_T = \max_{1 \leqslant i \leqslant k}(1-p_i)$. If the k partial test statistics are independent and continuous, under H_0, T_F follows a chisquare distribution with $2k$ degrees of freedom, T_L follows a normal distribution with mean 0 and variance k, and T_T has the same distribution as the largest of k independent uniform random variables on (0,1). Based on the null distribution of the combined test statistic, an overall rejection rule for testing H_0 versus H_1 can be established accordingly.

In many situations, however, it might not be reasonable to assume complete independence among the k partial tests. In those cases, the underlying dependency among the k partial tests is usually unknown. Pesarin (2001) proposed a nonparametric procedure to combine those dependent tests. In his proposed procedure, Pesarin still uses the combined test statistics that are commonly used for independent tests, such as the above T_F, T_L, and T_T. When the partial tests are dependent, the null distributions of those combined test statistics no longer follow the distributions mentioned above, and they depend on the specific dependency among the partial tests. To control the FWER, Pesarin (2001) proposed using permutations to carry out the test. Although Pesarin's procedure provides one possible way to combine dependent tests nonparametrically, it does not incorporate the

underlying dependency directly into the combined test statistic. This can lead to efficiency loss of the overall test.

In the multiple hypothesis testing literature, various procedures have also been proposed to combine multiple partial tests to control the FWER. Most of those procedures also aim to identify which sub-null hypothesis H_{0i} should be rejected. Since our main objective is to determine whether or not to reject the global null hypothesis H_0, we can easily modify the existing FWER controlling procedures from the multiple hypothesis testing literature for our setting. For example, applying Holm's stepdown method (Holm 1979) to our setting, it rejects H_0 if $\min_{1\leqslant i\leqslant k} p_i \leqslant \alpha/k$. This is equivalent to the Bonferroni procedure, which is known to be very inefficient. Applying the stepdown method proposed in Romano and Wolf (2005) to our setting will lead to rejecting H_0 if $\max_{1\leqslant i\leqslant k} T_i > c$, where T_i is assumed to be large if H_{0i} is rejected and c is the critical value. To control the FWER, c is chosen to satisfy $P(\max_{1\leqslant i\leqslant k} T_i > c) = \alpha$ under H_0. It is easy to see that Romano and Wolf's procedure also fails to incorporate the underlying dependency of the test statistics T_1,...T_k into their combined test statistic $\max_{1\leqslant i\leqslant k} T_i$.

To address the above limitations of existing methods, we propose a novel way to combine T_1,...,T_k nonparametrically based on the notion of data depth. Our proposed procedure is capable of taking into account the underlying dependency among T_1,...,T_k. Furthermore, how T_1,...,T_k are combined in our proposed procedure is automatically determined by their underlying dependency structure, therefore our procedure is completely data-driven.

To demonstrate its application, we use our proposed combining method to develop a new test for the two-sample problem with data of arbitrary types. The two-sample problem is a fundamental problem in statistics. However, most existing two-sample tests in the literature have been limited to Euclidean data only. In this modern era, different types of data (discrete, functional, textual, image, graph, tree, etc.) are frequently collected in many disciplines. Effectively comparing samples of arbitrary types is a challenging but important problem. Since properly defined distance metrics are usually available for those data, their interpoint distances provide a promising approach to develop efficient two-sample tests for data of arbitrary types. In the literature, several nonparametric tests based on interpoint distances have been proposed for the two-sample problem, including the edge-count test based on the minimal spanning tree (MST) proposed by Friedman and Rafsky (1979). Chen and Friedman (2017) pointed out the lack of power in Friedman and Rafsky's edge-count test for detecting scale differences and further proposed a generalized edge-count test to make it sensitive to scale differences. Chen, Chen and Su (2018) developed a weighted edge-count test to take into account unequal sample sizes. To deal with possible ties in the distance matrix when constructing the MST, Chen and Zhang (2013) and Zhang and Chen (2022) proposed several modified versions of the original edge-count test and the generalized edge-count test. As their names suggest, all the above tests are based on the number of edges in the MST, so they do not directly use interpoint distances. Their failure to fully make use of interpoint distances leads to significant efficiency loss as shown in our simulation studies. In this paper, we propose a new two-sample test which utilizes all interpoint distances. In order to develop a test powerful for detecting both location and scale differences, our strategy

is first developing two tests, one for detecting location differences only and the other for detecting scale differences only, and then using our proposed combining method to combine the two tests into a single one that can be powerful for detecting both location and scale differences. In our simulation studies, the proposed test performs well under a variety of settings and has much better power than the existing MST-based tests for detecting both location and scale changes.

The rest of the chapter is organized as follows. We describe our proposed combining method in Section 2.2, and then use it to develop a two-sample test for data of arbitrary types in Section 2.3. In Section 2.4, we report several simulation studies to evaluate the performance of our proposed two-sample test. We demonstrate its application in a real data example in Section 2.5. All the proofs are deferred to the Appendix A.

2.2 The Proposed Method to Combine Dependent Tests

2.2.1 Data Depth

Since our proposed combining method is based on the notion of data depth, we use the halfspace depth to illustrate the general concept of data depth and its induced center-outward ordering.

Considering a random sample $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ from the distribution F in $\mathbb{R}^d$ ($d \geqslant 1$), the halfspace depth of $\boldsymbol{x}$ with respect to F is defined as

$$\begin{aligned} D_F(\boldsymbol{x}) &= \inf_H \{P_F(H) : H \text{ is a closed half-space in } \mathbb{R}^d \text{ and } \boldsymbol{x} \in H\} \\ &= \inf_{\|\boldsymbol{u}\|=1} P_F(\boldsymbol{u}'\boldsymbol{X} \geqslant \boldsymbol{u}'\boldsymbol{x}), \end{aligned}$$

and the halfspace depth of $\boldsymbol{x}$ with respect to $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ is then obtained by replacing F in $D_F(\boldsymbol{x})$ by its empirical distribution F_n,

$$\begin{aligned} D_{F_n}(\boldsymbol{x}) &= \inf_H \{P_{F_n}(H) : H \text{ is a closed half-space in } \mathbb{R}^d \text{ and } \boldsymbol{x} \in H\} \\ &= \inf_{\|\boldsymbol{u}\|=1} \#\left\{\boldsymbol{u}'\boldsymbol{X}_i \geqslant \boldsymbol{u}'\boldsymbol{x}\right\}/n. \end{aligned}$$

Based on the definition, we can see that a larger depth value indicates that $\boldsymbol{x}$ lies in a more central position with respect to the data cloud $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ or its underlying distribution, while a smaller depth value indicates a more outlying position.

Based on the notion of data depth, we can calculate the depth values $D_{F_n}(\boldsymbol{X}_i)$'s and then order the $\boldsymbol{X}_i$'s according to their descending depth values. This gives rise to a natural center-outward ordering of the sample points in a multivariate sample. Figure 2.1 helps demonstrate this feature of the depth ordering. Again we use the halfspace depth as an example. Each plot in Figure 2.1 shows a random sample of size 500 drawn from a particular bivariate distribution and its depth contours calculated using the halfspace depth. In Figures 2.1(a), (b) and (c), the two marginal distributions of the bivariate distribution are both the normal distributions. In Figures 2.1(d), (e) and (f), the two marginal distributions

are the chisquare distribution with 2 degrees of freedom and the normal distribution. In Figures 2.1(g), (h) and (i), the two marginal distributions are both the chisquare distributions with 4 degrees of freedom. The two variables are independent in Figures 2.1(a), (d) and (g), positively correlated in Figures 2.1(b), (e) and (h), and negatively correlated in Figures 2.1(c), (f) and (i). To generate a bivariate sample for the setting where the two variables are correlated and their marginal distributions are some pre-specified non-normal distributions F_1 and F_2, we first draw a sample from the bivariate normal distribution with mean 0 and variance 1 for both variables and the correlation coefficient between the two variables is set to be ρ. Let $\Phi(\cdot)$ be the CDF of the standard normal distribution $N(0,1)$. We then obtain our desired bivariate sample after applying the transformations $F_1^{-1}\{\Phi(\cdot)\}$ and $F_2^{-1}\{\Phi(\cdot)\}$ to the two variables of the above bivariate normal sample, respectively. Based on the probability integral transform and inverse probability integral transform, it is easy to see that the marginal distributions of the bivariate sample we obtain after the transformations are F_1 and F_2. Since the two variables in the bivariate normal sample are correlated, the two variables in our bivariate sample are also correlated. The correlation coefficient ρ used in the above procedure is 0.6 for Figures 2.1(b), (e) and (h), and -0.6 for Figures 2.1(c), (f) and (i). From the depth contours in Figure 2.1, we can see that the depth ordering is from the center outward, and the shape of the depth contours in those plots all closely follows their underlying probabilistic geometry, indicating the completely data-driven nature of the ordering induced by the halfspace depth.

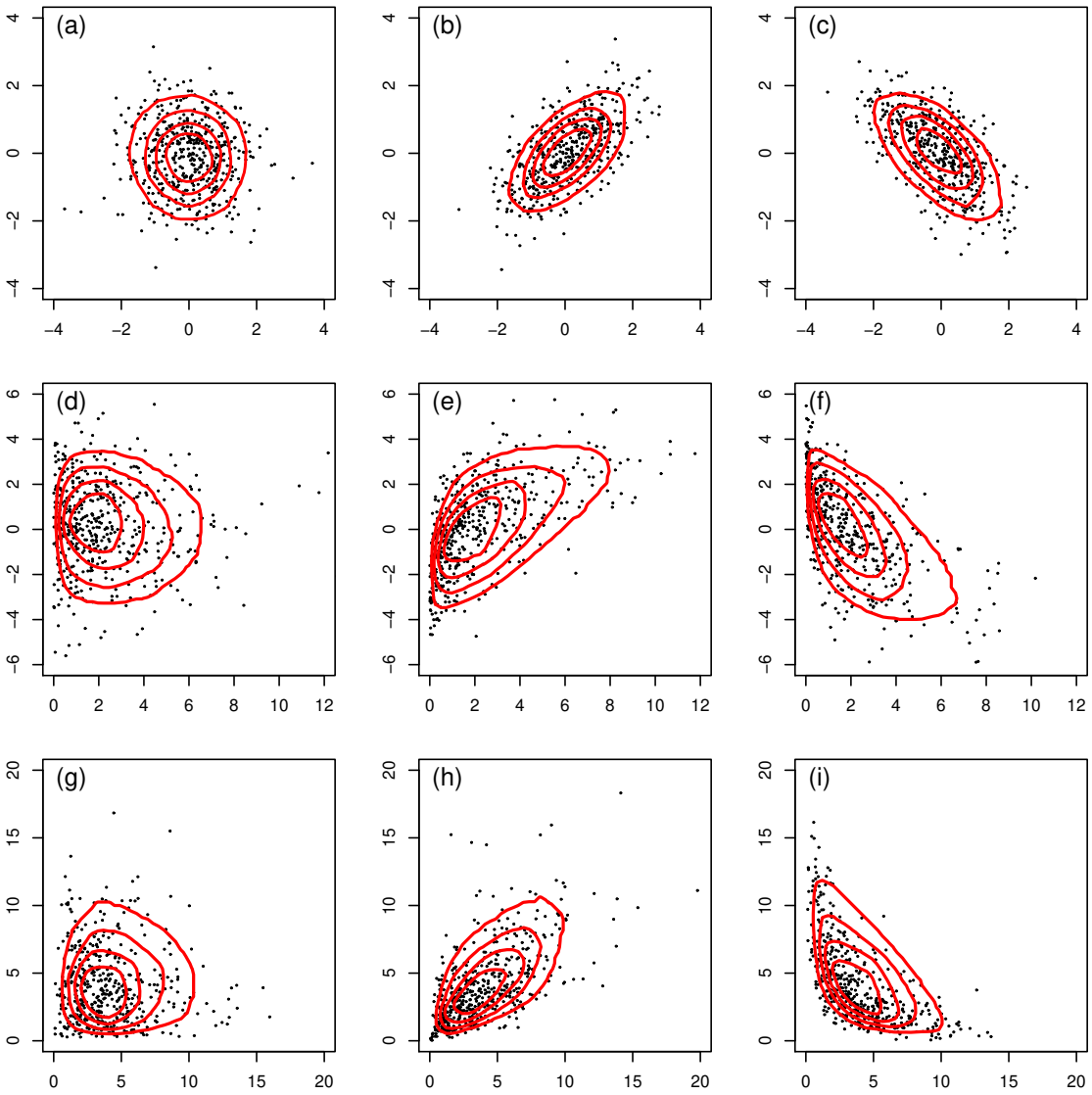

Figure 2.1: Depth contours for samples drawn from different bivariate distributions based on the half-space depth.

2.2.2 The Proposed Combining Method when all the Partial Tests are Two-sided

Let $\mathbf{X}$ denote the data under consideration and define $T_i = T_i(\mathbf{X})$. We first assume that all of the k partial tests are two-sided. To test the global hypothesis H_0 versus H_1, most of the existing methods consider some combining function which combines the k test statistics, T_1,...,T_k, into a scalar-valued test statistic. However, as mentioned in the Introduction, in many situations T_1,..., T_k are usually not independent and their dependency can be too complex to characterize. Therefore, finding a combining function that can incorporate the dependency among the T_i can be extremely difficult.

To circumvent this difficulty, we take a different approach. Instead of trying to combine T_1,...,T_k into a scalar-valued test statistic, we consider T_1,...,T_k simultaneously by putting them into a vector, denoted by $\boldsymbol{T} = (T_1, ..., T_k)'$. Then $\boldsymbol{T}$ is our proposed global test statistic, which is vector-valued instead of being scalar-valued.

To develop a testing procedure for the global hypothesis H_0 versus H_1 based on $\boldsymbol{T}$, we only need to find an appropriate way to set up the rejection rule for $\boldsymbol{T}$. In any hypothesis testing procedure, there are two ways to set up the rejection rule. One of them is called the critical value approach, in which the observed test statistic is compared with some critical value. If the observed test statistic is more extreme than the critical value, H_0 is rejected. When the test statistic is scalar-valued, it is easy to define extremeness and the corresponding critical value is some quantile of the null distribution of the test statistic in order to control the Type-I error rate. When the test statistic is vector-valued, it is not so obvious how to define extremeness since there is no natural ordering for vector-valued

data. However, as described in Section 2.2.1, based on the data depth, the vector-valued data can be ordered from the center outward, and data with smaller depth values are in more outlying/extreme positions. Therefore, based on this definition of extremeness, we can reject H_0 if our test statistic $\boldsymbol{T}$ has a smaller depth value than the depth value of some critical value. Similar to the scalar-valued test statistic case, to control the Type-I error rate, the critical value here should be taken as some depth-based quantile of the multivariate null distribution of $\boldsymbol{T}$. Such multivariate quantiles can be defined as follows. Let G denote the distribution of $\boldsymbol{T}$ under H_0. For the distribution G, the level-c depth inner region is given by

$$I(c, D, G) = \{\boldsymbol{x} \in \mathbb{R}^k : D_G(\boldsymbol{x}) \geqslant c\}, \tag{2.1}$$

and its boundary $\partial I(c, D, G)$ is called the level-c depth contour. For any $0 \leqslant p \leqslant 1$, let

$$c_p = \sup\{c : P(I(c, D, G)) \geqslant p\}.$$

Then the depth-based multivariate pth quantile of the distribution G is defined as $Q(p, D, G) = \partial I(c_p, D, G)$. In order to control the Type-I error rate of our test based on $\boldsymbol{T}$ at the level of α, the critical value in the rejection rule described earlier should be taken as the above multivariate $(1-\alpha)$th quantile, $Q(1-\alpha, D, G)$. Based on the definition of $Q(1-\alpha, D, G)$, the rejection rule for our test statistic $\boldsymbol{T}$ can be also stated as rejecting H_0 if $D_G(\boldsymbol{T}) < c_{1-\alpha}$.

The properties and applications of the above depth-based multivariate quantiles have been studied in Serfling (2002a, 2002b, 2010). There have been other efforts to extend the concept of quantile to the multivariate setting in the literature. For example, Hallin et

al. (2010) defined multivariate quantiles based on L_1 optimization, and the inner regions characterized by their proposed multivarite quantiles coincide with those defined in (2.1) using the halfspace depth. Recently, Chernozhukov et al. (2017) utilized the theory of optimal transport to define multivariate quantiles. Hallin et al. (2021) and Ghosal and Sen (2022) further studied the properties and applications of those multivariate quantiles. Recognizing the close relationship between the depth function and quantile function, Chernozhukov et al. (2017) also introduced a new data depth, called the Monge-Kantorovich depth, based on their proposed multivariate quantiles. The Monge-Kantorovich depth also gives rise to a natural center-outward ordering of multivariate data, therefore it can be also used in our proposed combining method.

Another approach to set up the rejection rule is called the p-value approach, in which a p-value is calculated and then compared to the significance level α. Based on its definition, the p-value is the probability of observing more extreme values of the test statistic than the one observed assuming that H_0 is true. According to the above definition of extremeness based on data depth, the p-value for our test statistic $\boldsymbol{T}$ can be calculated as

$$p = P\left(D_G(\boldsymbol{T}) \leqslant D_G(\boldsymbol{T}^{obs})\right), \tag{2.2}$$

where $\boldsymbol{T}^{obs}$ is the observed $\boldsymbol{T}$. We reject H_0 if $p < \alpha$. If the distribution of $D_G(\boldsymbol{T})$ is continuous, based on the probability integral transform, the p-value defined in (2.2) follows a uniform distribution on $(0, 1)$ under H_0. Therefore, the above testing procedure in which H_0 is rejected when $p < \alpha$ is a size-α test.

The critical value approach and p-value approach described above are equivalent and yield exactly the same conclusion. Due to its relatively simpler form, we use the p-value approach in the remaining of the paper.

The combining method we propose above uses the vector-valued test statistic $\boldsymbol{T} = (T_1, ..., T_k)'$ as our global test statistic, and makes use of all the information provided by T_1,...,T_k. By using the data depth, the dependency among T_1,...,T_k can be taken into account. The ability to automatically incorporate the dependency among T_1,...,T_k in our proposed combining method can help improve the efficiency of the resulting global test. In some cases, the improvement can be substantial and the resulting global test can be even asymptotically equivalent to the optimal one. For example, let $\boldsymbol{X}_1$,...,$\boldsymbol{X}_n$ be a random sample from a k-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (denoted by $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$), and one wants to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. In this setting, we know that the Hotelling's T^2 test is the most powerful test. Instead of directly testing H_0 versus H_1, we can also break it down into k sub-hypotheses, $H_{0i} : \mu_i = \mu_{0i}$ versus $H_{1i} : \mu_i \neq \mu_{0i}$, $i = 1, ..., k$, where μ_i and μ_{0i} are the i-th component of $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_0$, respectively. To test each sub-hypothesis H_{0i} versus H_{1i}, we can choose the t test as our partial test and $T_i = (\bar{X}_i - \mu_{0i})/\sqrt{s_{ii}/n}$, where $\bar{X}_i$ and s_{ii} are the sample mean and sample variance of the i-th component, respectively. Then the following establishes the asymptotic optimal property of the global test based on our proposed combining method.

Proposition 1 *If the data depth used in our combining method is affine invariant, the global test for $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ obtained by combining the above $T_1, ..., T_k$ based on our proposed method is asymptotically equivalent to the Hotelling's T^2 test.*

Remark 1: Many data depths, including the Mahalanobis depth, halfspace depth, simplicial depth, projection depth, zonoid depth and onion depth, satisfy the affine invariant requirement in the above proposition,

Remark 2: Cuesta-Albertos and Nieto-Reyes (2008) proposed an approximation algorithm based on a finite number of randomly selected one-dimensional projections to compute the halfspace depth. Let $\widehat{HD}_{m,G}(\boldsymbol{y})$ be the approximated halfspace depth of $\boldsymbol{y}$ with respect to G using m randomly selected one-dimensional projections based on the approximation method from Cuesta-Albertos and Nieto-Reyes (2008), and $HD_G(\boldsymbol{y})$ the exact halfspace depth of $\boldsymbol{y}$ with respect to G. Since G is an elliptical distribution from the proof of Proposition 1, according to Theorem 6 of Nagy et al. (2020),

$$\sup_{\boldsymbol{y}\in\mathbb{R}^k} |\widehat{HD}_{m,G}(\boldsymbol{y}) - HD_G(\boldsymbol{y})| \xrightarrow{a.s.} 0, \quad \text{as } m \to \infty.$$

Following this uniform convergence result, we have, as $m \to \infty$,

$$P\left(\widehat{HD}_{m,G}(\boldsymbol{T}) \leqslant \widehat{HD}_{m,G}(\boldsymbol{T}^{obs})\right) = \widehat{p}_m \xrightarrow{a.s.} p = P\left(HD_G(\boldsymbol{T}) \leqslant HD_G(\boldsymbol{T}^{obs})\right).$$

Therefore, if the approximated halfspace depth is used, Proposition 1 still holds almost surely as m goes to ∞.

In general, the null distribution of $\boldsymbol{T}$ is not easy to obtain. To circumvent this difficulty, we can resort to the resampling method. For example, if the data $\mathbf{X}$ consists of C independent random samples from distributions F_1,..., F_C, respectively, and the original null hypothesis is $H_0 : F_1 = \cdots = F_C$, then we can carry out our test using permutations.

More specifically, we randomly permute group labels of $\mathbf{X}$ and denote the permuted data from B random permutations by $\{\mathbf{X}_r^*\}_{r=1}^B$. Define our test statistic $\boldsymbol{T}$ calculated from the permuted data $\mathbf{X}_r^*$ as $\boldsymbol{T}_r^* = (T_1(\mathbf{X}_r^*), ..., T_k(\mathbf{X}_r^*))'$, $r = 1, ..., B$. Then the p-value of our test based on the above permutations is

$$\widehat{p}_B = \frac{1 + \sum_{r=1}^B I\left\{D_{G_B^*}(\boldsymbol{T}_r^*) \leqslant D_{G_B^*}(\boldsymbol{T}^{obs})\right\}}{B+1},$$

where G_B^* is the empirical distribution of $\boldsymbol{T}_1^*$,...,$\boldsymbol{T}_B^*$ (Ernst 2004). Similar to other permutation tests, our proposed test based on permutations can control the Type-I error rate at the nominal level (see Ernst (2004) for the probability basis of permutation methods).

If a permutation test can not apply, we can use the bootstrap method instead. For example, assume that the population distribution associated with $\mathbf{X}$ is F, and our hypothesis testing problem involves its k population parameters, θ_1,...,θ_k. For $i = 1, ..., k$, denote the reasonable estimator for θ_i by $\widehat{\theta}_i(\mathbf{X})$, which is a function of $\mathbf{X}$. To test the original null hypothesis, we assume that it is equivalent to testing the following k sub-hypotheses: $H_{0i} : \theta_i = \theta_{i,0}$ versus $H_{1i} : \theta_i \neq \theta_{i,0}$, where $\theta_{i,0}$ is some pre-specified constant, $i = 1, ..., k$. Suppose that we define $T_i = \{\widehat{\theta}_i(\mathbf{X}) - \theta_{i,0}\}/s.e\{\widehat{\theta}_i(\mathbf{X})\}$, where $s.e\{\widehat{\theta}_i(\mathbf{X})\}$ is the standard error of $\widehat{\theta}_i(\mathbf{X})$. Denote an estimator of F based on $\mathbf{X}$ by $F_n(\mathbf{X})$, where n is the sample size of $\mathbf{X}$. To approximate the null distribution of $\boldsymbol{T} = (T_1, ..., T_k)'$, generate B bootstrap resamples of $\mathbf{X}$ using $F_n(\mathbf{X})$, and denote them by $\mathbf{X}_{n,1}^*$,...,$\mathbf{X}_{n,B}^*$. Define T_i calculated from the bootstrap resample $\mathbf{X}_{n,r}^*$ as $T_{i,n,r}^* = \{\widehat{\theta}_i(\mathbf{X}_{n,r}^*) - \widehat{\theta}_i(\mathbf{X})\}/s.e\{\widehat{\theta}_i(\mathbf{X}_{n,r}^*)\}$, $i = 1, ..., k$, and $\boldsymbol{T}_{n,r}^* = (T_{1,n,r}^*, ..., T_{k,n,r}^*)'$, $r = 1, ..., B$. Then the null distribution of $\boldsymbol{T}$ can be approximated by the empirical distribution of $\boldsymbol{T}_{n,1}^*$,..., $\boldsymbol{T}_{n,B}^*$. As a result, the p-value defined in (2.2) can

be approximated by

$$\widehat{p}_{n,B} = \frac{1}{B}\sum_{r=1}^{B} I\left\{D_{G^*_{n,B}}(\boldsymbol{T}^*_{n,r}) \leqslant D_{G^*_{n,B}}(\boldsymbol{T}^{obs})\right\}, \tag{2.3}$$

where $G^*_{n,B}$ is the empirical distribution of $\boldsymbol{T}^*_{n,1}$,...,$\boldsymbol{T}^*_{n,B}$.

Proposition 2 *In the above bootstrap procedure, let G^*_n be the distribution of $\boldsymbol{T}^*_{n,r}$, $r = 1, ..., B$. If G^*_n converges weakly to G as $n \to \infty$, and the data depth $D(\cdot)$ used in (2.3) satisfies*

$$\sup_{\boldsymbol{y}\in\mathbb{R}^k} |D_{G^*_{n,B}}(\boldsymbol{y}) - D_{G^*_n}(\boldsymbol{y})| \xrightarrow{a.s.} 0, \quad \text{as } B \to \infty, \tag{2.4}$$

$$\sup_{\boldsymbol{y}\in\mathbb{R}^k} |D_{G^*_n}(\boldsymbol{y}) - D_{G}(\boldsymbol{y})| \xrightarrow{a.s.} 0, \quad \text{as } n \to \infty, \tag{2.5}$$

then $\widehat{p}_{n,B}$ converges to the p-value defined in (2.2) almost surely, as $B \to \infty$ and $n \to \infty$.

Remark 3: Most of the data depths satisfy (2.4) and (2.5) under proper conditions. Mosler and Mozharovskyi (2022) summarized those conditions for different data depths. For example, the Mahalanobis depth satisfies (2.4) and (2.5) if G and G^*_n have a regular covariance matrix. The zonoid depth satisfies (2.4) and (2.5) if G and G^*_n satisfy some regularity condition given in Cascos and López-Díaz (2016). The halfspace depth satisfies (2.4) and (2.5) if G is absolutely continuous (see Nagy et al. 2019).

Remark 4: Dyckerhoff and Mozharovskyi (2016) provided a theoretical framework for efficiently computing exact values of the halfspace depth. Pokotylo et al. (2020) implemented Dyckerhoff and Mozharovskyi's method in R-package "ddalpha" so that exact computation of the halfspace depth is available for any dimension. If the approximated

halfspace depth based on the algorithm from Cuesta-Albertos and Nieto-Reyes (2008) is used in the above bootstrap procedure, in order to prove that Proposition 2 still holds, we need the following result,

$$\sup_{\boldsymbol{y}\in\mathbb{R}^k} |\widehat{HD}_{m,G^*_{n,B}}(\boldsymbol{y}) - HD_{G^*_n}(\boldsymbol{y})| \xrightarrow{a.s.} 0, \quad \text{as } m \to \infty, B \to \infty \tag{2.6}$$

where $\widehat{HD}_{m,G^*_{n,B}}(\boldsymbol{y})$ is the approximated halfspace depth of $\boldsymbol{y}$ with respect to $G^*_{n,B}$ based on m randomly selected one-dimensional projections, and $HD_{G^*_n}(\boldsymbol{y})$ is the exact halfspace depth of $\boldsymbol{y}$ with respect to G^*_n. Using the results from Cuesta-Albertos and Nieto-Reyes (2008) and Nagy et al. (2020), we can prove (2.6) when G^*_n is absolutely continuous and has a bounded support, or G^*_n is an elliptical distribution or a p-symmetric distribution (Fang et al. 1990) with $p \in (0, 2]$. It remains an open problem whether (2.6) is still true for other G^*_n . We plan to investigate this further in our future research.

2.2.3 The Proposed Combining Method when Some of the Partial Tests are One-sided

From the previous section, we can see that, in our proposed method to combine dependent tests, we first consider the vector-valued test statistic $\boldsymbol{T}$, which consists of the test statistics from all the k partial tests, and then find a suitable measure of extremeness for $\boldsymbol{T}$ so that an appropriate rejection rule can be established. In the previous section, we consider the case where all of the k partial tests are two-sided. As we can see from the depth contours shown in Figure 2.1, the center-outward ordering induced by the data depth is a reasonable way to define extremeness for $\boldsymbol{T}$ in this situation. However, when some

of the partial tests are one-sided, this definition of extremeness might not be satisfactory. Therefore, we need to find another reasonable measure of extremeness in this situation.

To this end, we consider the halfspace depth described in Section 2.2.1. If the halfspace depth is used, $D_G(\boldsymbol{T}^{obs})$ is defined as

$$D_G(\boldsymbol{T}^{obs}) = \inf_{\|\boldsymbol{u}\|=1} P_G(\boldsymbol{u}'\boldsymbol{T} \geqslant \boldsymbol{u}'\boldsymbol{T}^{obs}).$$

When $k = 1$,

$$D_G(T^{obs}) = \inf_{|u|=1} P_G(u \cdot T \geqslant u \cdot T^{obs}) = \min\left\{P_G(T \geqslant T^{obs}), P_G(T \leqslant T^{obs})\right\}.$$

This implies that, when using the halfspace depth to define extremeness, we consider both tails of the null distribution G. This is the reason why the extremeness defined using any data depth is good only for a two-sided test. If the test based on T is one-sided, without loss of generality, we assume that H_0 is rejected when T is too large. Then in order to find an appropriate measure of extremeness in this case, we should modify the above $D_G(T^{obs})$ as

$$\widetilde{D}_G(T^{obs}) = \inf_{|u|=1, u>0} P_G(u \cdot T \geqslant u \cdot T^{obs}) = P_G(T \geqslant T^{obs}).$$

To generalize this modified definition to any number of partial tests, we can use the same projection idea as in the original halfspace depth. More specifically, if there are k_0 ($k_0 \leqslant k$) one-sided tests among the k partial tests, without loss of generality, we assume that those one-sided tests are for the first k_0 sub-hypothesis testing problems. Let

$\boldsymbol{u} = (u_1, ..., u_k)'$. Then the modified halfspace depth $\widetilde{D}_G^{(k_0)}(\boldsymbol{T}^{obs})$ is defined as

$$\widetilde{D}_G^{(k_0)}(\boldsymbol{T}^{obs}) = \inf_{\substack{\|\boldsymbol{u}\|=1 \\ u_i \geqslant 0, i=1,...,k_0}} P_G(\boldsymbol{u}'\boldsymbol{T} \geqslant \boldsymbol{u}'\boldsymbol{T}^{obs}),$$

where G is the distribution of $\boldsymbol{T}$.

Similar to the original halfspace depth, we can also establish the following uniform convergence results for the above modified halfspace depth.

Theorem 3 *Given a random sample $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ from the distribution F in $\mathbb{R}^d$ ($d \geqslant 1$), the modified halfspace depth of $\boldsymbol{x} \in \mathbb{R}^d$ with respect to F is defined as*

$$\widetilde{D}_F^{(k_0)}(\boldsymbol{x}) = \inf_{\substack{\|\boldsymbol{u}\|=1 \\ u_i \geqslant 0, i=1,...,k_0}} P_F(\boldsymbol{u}'\boldsymbol{X} \geqslant \boldsymbol{u}'\boldsymbol{x}),$$

and the modified halfspace depth of $\boldsymbol{x} \in \mathbb{R}^d$ with respect to $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ is defined as,

$$\widetilde{D}_{F_n}^{(k_0)}(\boldsymbol{x}) = \inf_{\substack{\|\boldsymbol{u}\|=1 \\ u_i \geqslant 0, i=1,...,k_0}} \#\{\boldsymbol{u}'\boldsymbol{X}_i \geqslant \boldsymbol{u}'\boldsymbol{x}\}/n,$$

where F_n is the empirical distribution of $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$. For any distribution F,

$$\sup_{\boldsymbol{x} \in \mathbb{R}^d} |\widetilde{D}_{F_n}^{(k_0)}(\boldsymbol{x}) - \widetilde{D}_F^{(k_0)}(\boldsymbol{x})| \xrightarrow{a.s.} 0, \quad \text{as } n \to \infty.$$

If the distribution F is absolutely continuous, for any sequence of distributions $\{F_\nu^*\}_{\nu=1}^{\infty}$ weakly convergent to F,

$$\sup_{\boldsymbol{x} \in \mathbb{R}^d} |\widetilde{D}_{F_\nu^*}^{(k_0)}(\boldsymbol{x}) - \widetilde{D}_F^{(k_0)}(\boldsymbol{x})| \longrightarrow 0, \quad \text{as } \nu \to \infty.$$

To see what kind of ordering our modified halfspace depth induces, we draw its depth contours in Figure 2.2 for the same bivariate samples we generate for Figure 2.1. The modified halfspace depth we use here is corresponding to the case when one of the two partial tests is one-sided, i.e., $k_0 = 1$ and $d = 2$ in the above definition of our modified halfspace depth. From the depth contours in Figure 2.2, we can see that the ordering based on our modified halfspace depth can reflect well the nature of the two partial tests, and the underlying probabilistic geometry automatically determines the shape of the depth contours in those plots, a data-driven feature inherited from the original halfspace depth.

From Figure 2.2, we can also see that the ordering derived from our modified halfspace depth can provide a reasonable measure of extremeness for the vector-valued test statistic $\boldsymbol{T}$ when some of the partial tests are one-sided. That is, data with smaller depth values based on the modified halfspace depth are in more extreme positions. With this definition of extremeness, the p-value for $\boldsymbol{T}$ when the first k_0 partial tests are one-sided can be calculated as

$$p = P\left(\widetilde{D}_G^{(k_0)}(\boldsymbol{T}) \leqslant \widetilde{D}_G^{(k_0)}(\boldsymbol{T}^{obs})\right), \tag{2.7}$$

where G is the distribution of $\boldsymbol{T}$ under H_0.

Similar to Section 2.2.2, if a permutation test can apply, our p-value can be calculated as

$$\widehat{p}_B = \frac{1 + \sum_{r=1}^{B} I\left\{\widetilde{D}_{G_B^*}^{(k_0)}(\boldsymbol{T}_r^*) \leqslant \widetilde{D}_{G_B^*}^{(k_0)}(\boldsymbol{T}^{obs})\right\}}{B+1},$$

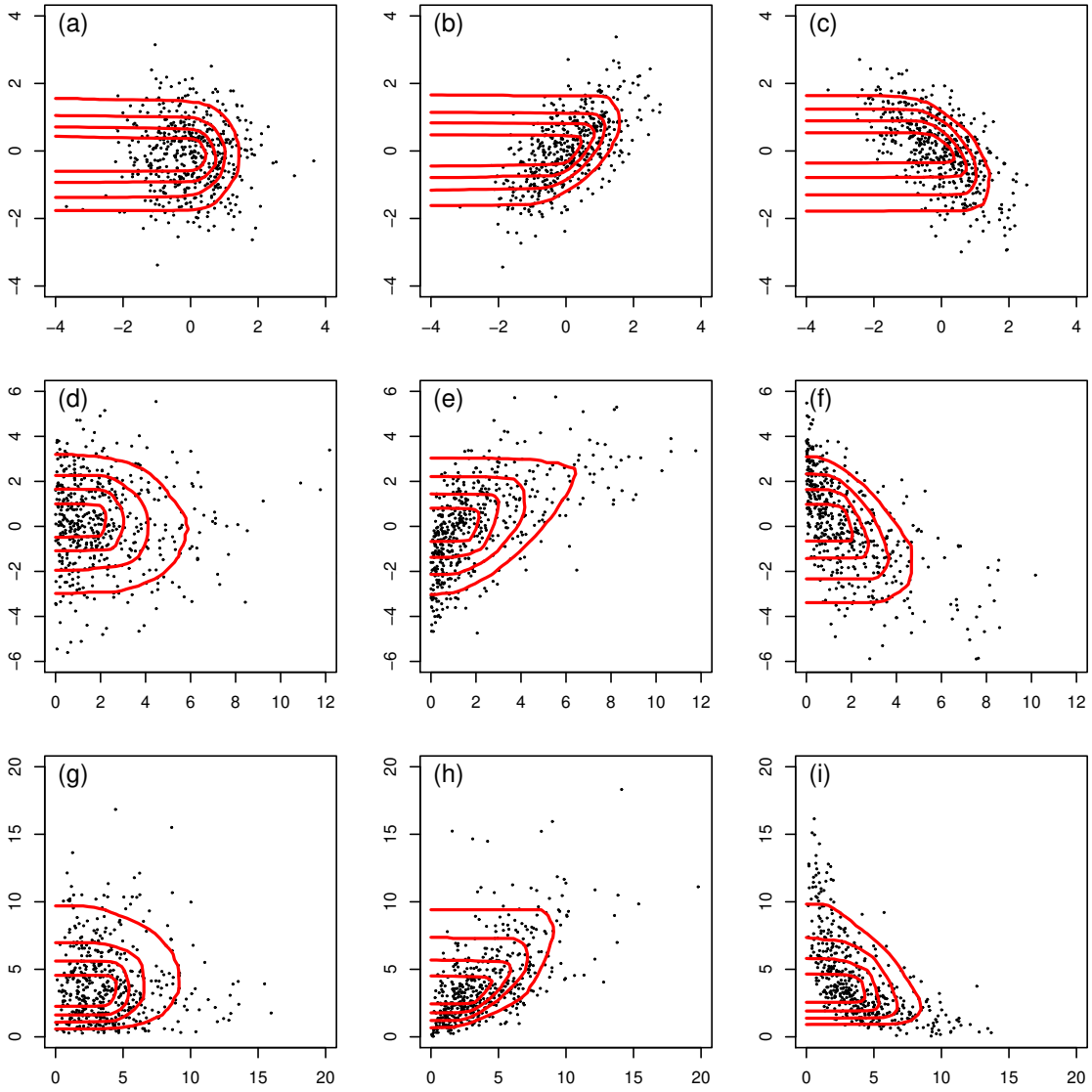

Figure 2.2: Depth contours for samples drawn from different bivariate distributions based on the modified halfspace depth corresponding to the case when one of the two partial tests is one-sided.

where $\boldsymbol{T}_1^*$,...,$\boldsymbol{T}_B^*$ are the test statistic $\boldsymbol{T}$ calculated from the permuted data, and G_B^* is their empirical distribution. Again using the above $\widehat{p}_B$ as the p-value can control the Type-I error rate at the nominal level.

If the permutation test can not apply, we can use the bootstrap method as described in Section 2.2.2 and the p-value is defined as

$$\widehat{p}_{n,B} = \frac{1}{B}\sum_{r=1}^{B} I\left\{\widetilde{D}^{(k_0)}_{G^*_{n,B}}(\boldsymbol{T}^*_{n,r}) \leqslant \widetilde{D}^{(k_0)}_{G^*_{n,B}}(\boldsymbol{T}^{obs})\right\}, \tag{2.8}$$

where $\boldsymbol{T}_{n,1}^*$,...,$\boldsymbol{T}_{n,B}^*$ are the test statistic $\boldsymbol{T}$ calculated from the bootstrap resamples, $G_{n,B}^*$ is their empirical distribution.

Proposition 4 *Let G_n^* be the distribution of $\boldsymbol{T}_{n,r}^*$, $r = 1, ..., B$. If G_n^* converges weakly to G as $n \to \infty$, then $\widehat{p}_{n,B}$ in (2.8) converges to p in (2.7) almost surely, as $B \to \infty$ and $n \to \infty$.*

Remark 5: As mentioned earlier, Dyckerhoff and Mozharovskyi (2016) provided a theoretical framework for efficiently computing exact values of the original halfspace depth. However, it is not clear how to extend Dyckerhoff and Mozharovskyi's method to our modified halfspace depth. For this reason, in order to compute our modified halfspace depth more efficiently, we adopt the approximation algorithm from Cuesta-Albertos and Nieto-Reyes (2008). More specifically, we approximate $\widetilde{D}^{(k_0)}_{F_n}(\boldsymbol{x}) = \inf_{\substack{\|\boldsymbol{u}\|=1 \\ u_i \geqslant 0, i=1,...,k_0}} \#\left\{\boldsymbol{u}'\boldsymbol{X}_i \geqslant \boldsymbol{u}'\boldsymbol{x}\right\}/n$ using a large number of $\boldsymbol{u}$'s randomly drawn from all the $\boldsymbol{u}$'s that satisfy $\|\boldsymbol{u}\| = 1$ and $u_i \geqslant 0, i = 1, ..., k_0$. When this approximated modified halfspace depth is used, if we can carry out our test using permutations, then it can still control the Type-I error rate at the

nominal level. If the permutation test can not apply and the bootstrap method needs to be used, similar to (2.6) in the approximated halfspace depth case, the uniform convergence result for the approximated modified halfspace depth remains an open problem. Therefore, it is currently unclear whether Proposition 3 still holds when the approximated modified halfspace depth is used in (2.8). We plan to investigate this further in our future research.

2.2.4 Differences Between Our Proposed Combining Method and Existing Combining Methods

To show the differences between our proposed combining method and existing combining methods, we first use Fisher's combining method as an example. Recall that Fisher's method uses the combined test statistic $T_F = -2\sum_{i=1}^{k}\log(p_i)$, where p_i is the p-value of the i-th partial test based on the test statistic T_i. The global p-value is then defined as

$$p = P_{H_0}(T_F \geqslant T_F^{obs}),$$

where T_F^{obs} is the observed T_F. The above definition of p-value implies that, in Fisher's combining method, we implicitly use $T_F = -2\sum_{i=1}^{k}\log(p_i)$ to order the vector-valued test statistic $\boldsymbol{T} = (T_1, ..., T_k)'$: the larger T_F is, the more extreme $\boldsymbol{T} = (T_1, ..., T_k)'$ is. To see what kind of ordering Fisher's method induces for $\boldsymbol{T} = (T_1, ..., T_k)'$, we draw the contours based on $T_F = -2\sum_{i=1}^{k}\log(p_i)$ in Figures 2.3 and 2.4 for the same bivariate samples we use for Figures 2.1 and 2.2. To draw those contours in Figures 2.3 and 2.4, for each point $\boldsymbol{T} = (T_1, T_2)' \in \mathbb{R}^2$, we first calculate its p_1 and p_2. Since the bivariate sample is assumed

to be generated from the null distribution, our p_i is calculated as the proportion of the ith component of the generated sample more extreme than T_i, $i = 1, 2$. Based on those p_1 and p_2, we can calculate the value of $T_F = -2\sum_{i=1}^{2} \log(p_i)$ for each point $\boldsymbol{T} \in \mathbb{R}^2$. The contours in Figures 2.3 and 2.4 are the points with the same T_F values.

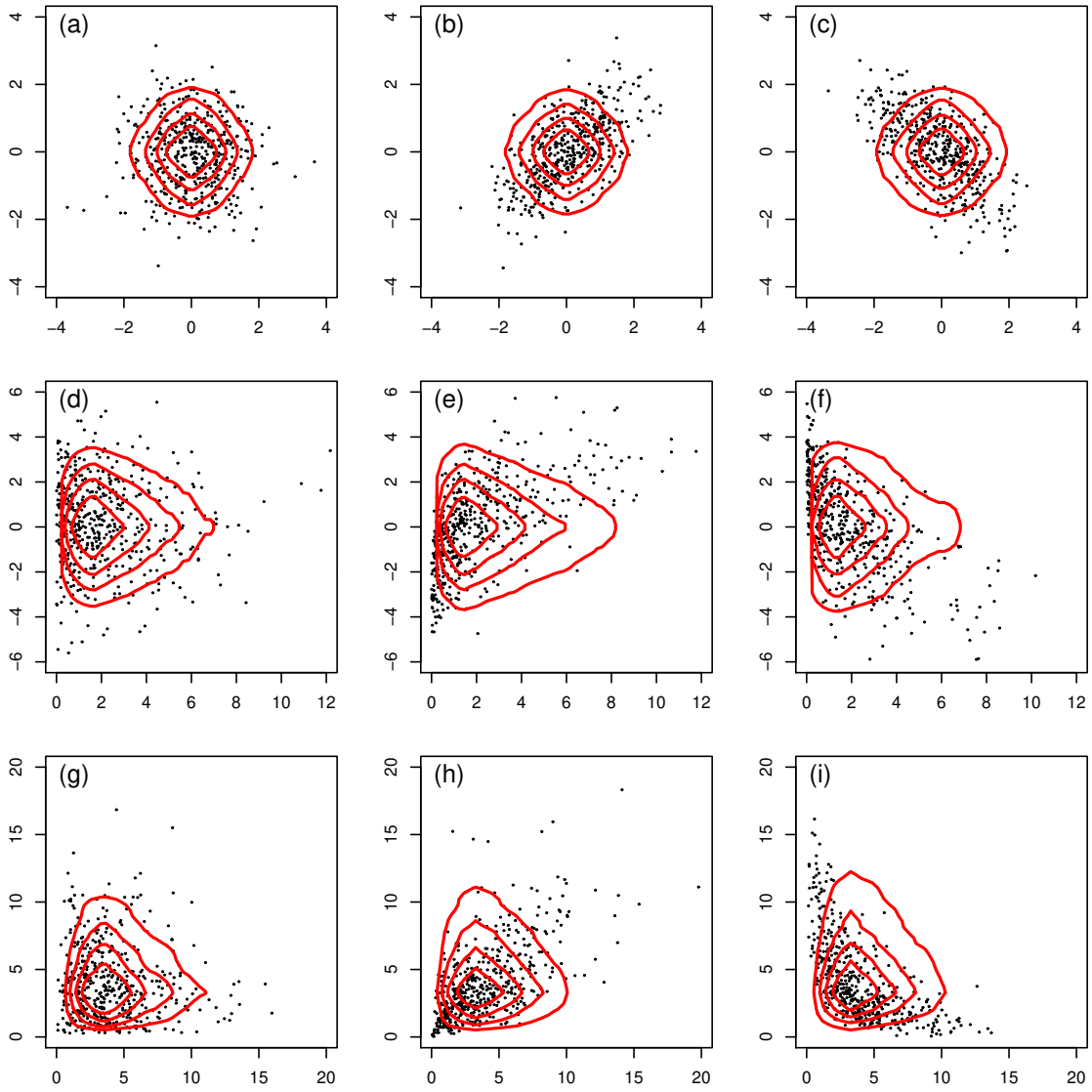

Figure 2.3: Contours for samples drawn from different bivariate distributions based on Fisher's combining function $T_F = -2\sum_{i=1}^{2}\log(p_i)$ when the two partial tests are both two-sided.

As we can see from Figures 2.3 and 2.4, the contours across each row based on Fisher's combining method remain similar, indicating that the ordering of $\boldsymbol{T} = (T_1, ..., T_k)'$ induced by Fisher's method largely ignores the underlying dependency among the partial

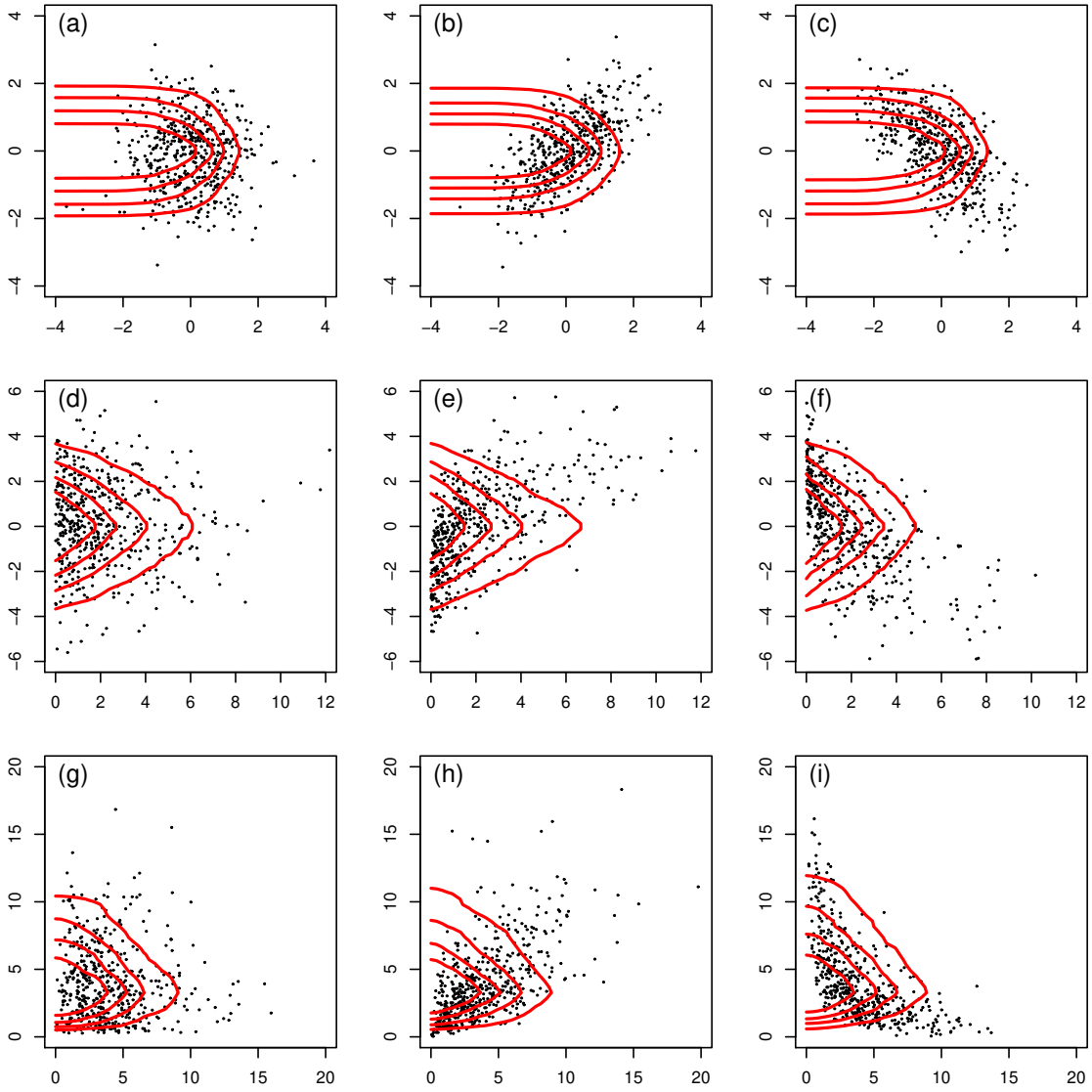

Figure 2.4: Contours for samples drawn from different bivariate distributions based on Fisher's combining function $T_F = -2\sum_{i=1}^{2}\log(p_i)$ when one of the two partial tests is one-sided.

tests. As a result, the combining method based on $T_F = -2\sum_{i=1}^{k}\log(p_i)$ fails to incorporate the underlying dependency, which can lead to efficiency loss. The same conclusion also holds for Lipták's combining method and Tippett's combining method. In contrast, as shown in Figures 2.1 and 2.2, our proposed combining method is capable of automatically taking into account the underlying dependency of the partial tests, therefore provides a completely data-driven approach for combining dependent tests.

2.3 New Two-sample Test for Data of Arbitrary Types

In this section, we demonstrate the application of our proposed combining method by developing a new two-sample test for data of arbitrary types. We first review the two-sample test proposed in Li (2018) for data in the Euclidean space.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be two independent random samples respectively drawn from distributions F and G, both in $\mathbb{R}^d$. The two-sample problem is then to test $H_0 : F = G$ versus $H_1 : F \neq G$. Instead of testing a general distributional difference between F and G, we focus on testing location and/or scale differences between F and G. Even after we focus only on these two types of differences, it is still difficult to develop a test that is efficient for both types of differences. To circumvent this difficulty, Li (2018) used interpoint Euclidean distances to develop two tests, one for detecting location differences only and the other for detecting scale differences only. More specifically, let $\|\boldsymbol{a} - \boldsymbol{b}\|$ denote the Euclidean distance between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. The two test statistics proposed in Li

(2018) are:

$$T_{\text{loc0}} = \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|\boldsymbol{X}_i-\boldsymbol{Y}_j\|^2-\binom{m}{2}^{-1}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\|\boldsymbol{X}_i-\boldsymbol{X}_j\|^2-\binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\|\boldsymbol{Y}_i-\boldsymbol{Y}_j\|^2, \tag{2.9}$$

and

$$T_{\text{scal0}} = \binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\|\boldsymbol{Y}_i-\boldsymbol{Y}_j\|^2-\binom{m}{2}^{-1}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\|\boldsymbol{X}_i-\boldsymbol{X}_j\|^2. \tag{2.10}$$

Let $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$ be the mean vector and dispersion matrix of F, and $\boldsymbol{\mu}_Y$ and $\boldsymbol{\Sigma}_Y$ be the mean vector and dispersion matrix of G. It is easy to verify that $E(T_{\text{loc0}}) = 2\|\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y\|^2$ and $E(T_{\text{scal0}}) = 2\{\text{tr}(\boldsymbol{\Sigma}_X) - \text{tr}(\boldsymbol{\Sigma}_Y)\}$. Therefore, T_{loc0} is good for detecting location differences and T_{scal0} is good for detecting scale differences.

As mentioned in the Introduction, the recent advances in computing and data acquisition technologies have made easy the collection of data of diverse types in all fields. Since properly defined distance metrics are usually available for many types of data, a promising approach for developing the two-sample test for those data is using their inter-point distances. Let O and E represent the original space of the data and the Euclidean space, respectively. For any $\boldsymbol{r}$ and $\boldsymbol{s}$ from O, define an appropriate distance metric between $\boldsymbol{r}$ and $\boldsymbol{s}$ to be $d(\boldsymbol{r}, \boldsymbol{s})$. Suppose there are n objects, $\boldsymbol{r}_1$, ..., $\boldsymbol{r}_n$, from O. Define matrix $\mathbf{A}$ as $[\mathbf{A}]_{ij} = a_{ij}$, where $a_{ij} = -d(\boldsymbol{r}_i, \boldsymbol{r}_j)^2/2$. Define matrix $\mathbf{B}$ as $\mathbf{B} = \mathbf{HAH}$, where $\mathbf{H} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ with $\mathbf{I}_n$ being the n-dimensional identity matrix and $\mathbf{1}_n$ being a vector of n ones. As shown in Cox and Cox (2001), if matrix $\mathbf{B}$ is positive semi-definite of rank r,

then there exists a mapping ϕ from O to $\mathbb{R}^r$ such that $d(\boldsymbol{r}_i, \boldsymbol{r}_j) = \|\phi(\boldsymbol{r}_i) - \phi(\boldsymbol{r}_j)\|$ for any $1 \leqslant i, j \leqslant n$. If matrix $\mathbf{B}$ is not positive semi-definite but its negative eigenvalues are small in magnitude, a mapping ϕ from O to E can be found by ignoring those negative eigenvalues such that $d(\boldsymbol{r}_i, \boldsymbol{r}_j) \approx \|\phi(\boldsymbol{r}_i) - \phi(\boldsymbol{r}_j)\|$ for any $1 \leqslant i, j \leqslant n$. If the negative eigenvalues of matrix $\mathbf{B}$ are large, we believe that the interpoint distances $d(\boldsymbol{r}_i, \boldsymbol{r}_j)$, $1 \leqslant i < j \leqslant n$, can be also used to characterize the underlying distribution similarly as their counterparts in E.

Based on the above connection between the interpoint distances in O and E, we propose two test statistics for our two-sample problem in O by replacing the interpoint Euclidean distances in (2.9) and (2.10) by their counterparts in O. More specifically, let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be two independent random samples in O. The two test statistics we propose for the two-sample problem in O are

$$T_{\text{loc}} = \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} d^2(\boldsymbol{X}_i, \boldsymbol{Y}_j) - \binom{m}{2}^{-1} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} d^2(\boldsymbol{X}_i, \boldsymbol{X}_j) - \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d^2(\boldsymbol{Y}_i, \boldsymbol{Y}_j),$$

and

$$T_{\text{scal}} = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d^2(\boldsymbol{Y}_i, \boldsymbol{Y}_j) - \binom{m}{2}^{-1} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} d^2(\boldsymbol{X}_i, \boldsymbol{X}_j).$$

Since T_{loc} and T_{scal} are built on T_{loc0} and T_{scal0}, they are also good at detecting location and scale differences respectively. Then our remaining task is to combine the two partial tests based on T_{loc} and T_{scal} into a single test for the two-sample problem. In Li (2018), asymptotic normality was established for the null distribution of $(T_{\text{loc0}}, T_{\text{scal0}})'$ when the dimension of the Euclidean data d goes to infinity. However, for low-dimensional Euclidean data and other non-Euclidean data, the null distribution of $(T_{\text{loc}}, T_{\text{scal}})'$ is not

known any more, neither is the dependency between T_{loc} and T_{scal}. Therefore, we can use our proposed combining method to combine the two partial tests.

According to Morgenstern (2001), when testing for location differences, we reject $H_0 : F = G$ only if T_{loc} is too large. When testing for scale differences, we reject H_0 if T_{scal} is too large or too small. Therefore, T_{loc} corresponds to a one-sided test, and T_{scal} corresponds to a two-sided test. As discussed in Section 2.2.3, the modified halfspace depth should be used to combined the two partial tests in this case. More specifically, we first calculate the observed vector of test statistics $\boldsymbol{T}_{obs} = (T_{\text{loc}}^{obs}, T_{\text{scal}}^{obs})'$ from the $\boldsymbol{X}$ and $\boldsymbol{Y}$ samples, and then randomly permute the observations between the two samples for B times, say $B = 1000$. For the B permuted samples, we calculate their corresponding vector of test statistics, denoted by $\boldsymbol{T}_1^*, \ldots, \boldsymbol{T}_B^*$. Then the p-value for our test is

$$\widehat{p}_B = \frac{1 + \sum_{i=1}^{B} I\left\{\widetilde{D}^{(1)}_{G_B^*}(\boldsymbol{T}_i^*) \leqslant \widetilde{D}^{(1)}_{G_B^*}(\boldsymbol{T}^{obs})\right\}}{B+1},$$

where $\widetilde{D}^{(1)}_{G_B^*}(\cdot)$ is our modified halfspace depth with respect to the data cloud $\boldsymbol{T}_1^*, \ldots, \boldsymbol{T}_B^*$. We then reject $H_0 : F = G$ if $\widehat{p}_B$ is smaller than the significance level α.

2.4 Simulation Studies

In this section, we report several simulation studies to evaluate the performance of our proposed two-sample test for data of arbitrary types. In particular, we compare our proposed test with those proposed in Chen and Friedman (2017) and Zhang and Chen (2022). In all the simulations, the Type-I error rate and power of each test are obtained

based on 1000 simulations. The R-codes for carrying out all the simulations in this section are available in our online supplementary materials.

2.4.1 Continuous Data

Our first simulation study evaluates the performance of the proposed test when data are from continuous distributions in $\mathbb{R}^d$. Since the MST is unique when distributions are continuous, the generalized edge-count test based on the MST in Chen and Friedman (2017) can also apply in this situation. Therefore, we follow the simulation settings used in Chen and Friedman (2017) to compare the powers of our proposed test and Chen and Friedman's test.

Before we compare the powers, we firstly evaluate the Type-I error rates of the two tests. To this end, the random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are generated independently from $N_d(\mathbf{0}_d, \mathbf{I}_d)$, where $\mathbf{0}_d$ is the vector of d zeros and $\mathbf{I}_d$ is the d-dimensional identity matrix. We set $m = 25$ and $n = 100$, and take d from $\{2, 10, 50, 100, 200\}$ so that the Type-I error rates of the tests can be evaluated in various dimensions. Table 2.1 shows the simulated Type-I error rates of Chen and Friedman's test and our proposed test when the nominal significance level $\alpha = 0.05$. As we can see from the table, the simulated Type-I error rates of both tests are close to the nominal level under all the settings.

Table 2.1: The simulated Type-I error rates of Chen and Friedman's test and our proposed test with $\alpha = 0.05$.

The simulated Type-I errors					
d	2	10	50	100	200
Chen and Friedman's test	0.051	0.052	0.048	0.046	0.044
Proposed	0.054	0.051	0.049	0.044	0.047

To compare the powers of the two tests, the random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are first drawn from $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and $N_d(\Delta/\sqrt{d} \times \mathbf{1}_d, \mathbf{I}_d)$, respectively, where $\mathbf{1}_d$ is the vector of d ones and $\Delta/\sqrt{d}$ is the mean shift size in all the components. The two distributions differ in location only in this setting. In the second simulation setting of our power comparison study, the two random samples are drawn from $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and $N_d(\mathbf{0}_d, \sigma\mathbf{I}_d)$, respectively. The two distributions differ in scale only. In the third simulation setting, we first independently draw $\{\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_m^*\}$ and $\{\boldsymbol{Y}_1^*, \ldots, \boldsymbol{Y}_n^*\}$ from $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and $N_d(\Delta/\sqrt{d} \times \mathbf{1}_d, \mathbf{I}_d)$, respectively. Then we set $\boldsymbol{X}_i = \mathbf{exp}(\boldsymbol{X}_i^*)$ and $\boldsymbol{Y}_i = \mathbf{exp}(\boldsymbol{Y}_i^*)$, where $\mathbf{exp}(\boldsymbol{b})$ is a mapping from $\mathbb{R}^d$ to $\mathbb{R}^d$ that applies the function $\exp(\cdot)$ to each component of $\boldsymbol{b}$. Therefore, $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ can be considered as being drawn from some multivariate lognormal distributions, and the two distributions differ in both location and scale.

Similar to our Type-I error study, we set $m = 25$ and $n = 100$, and take d from $\{2, 10, 50, 100, 200\}$ in all the three settings so that the powers of the tests can be evaluated and compared in various dimensions. Since the two distributions F and G are continuous in $\mathbb{R}^d$, the Euclidean distance is used as $d(\cdot, \cdot)$ in our T_{loc} and T_{scal} as well as in building the MST for Chen and Friedman's test. The significance levels of all the tests are set at $\alpha = 0.05$. Figure 2.5 shows the powers of Chen and Friedman's test and our proposed test for detecting different types of distributional differences.

As we can see from Figure 2.5, our proposed test outperforms Chen and Friedman's test in all the settings. As mentioned in the Introduction, Chen and Friedman's test is based on the number of edges in the MST, so it does not fully make use of interpoint distances.

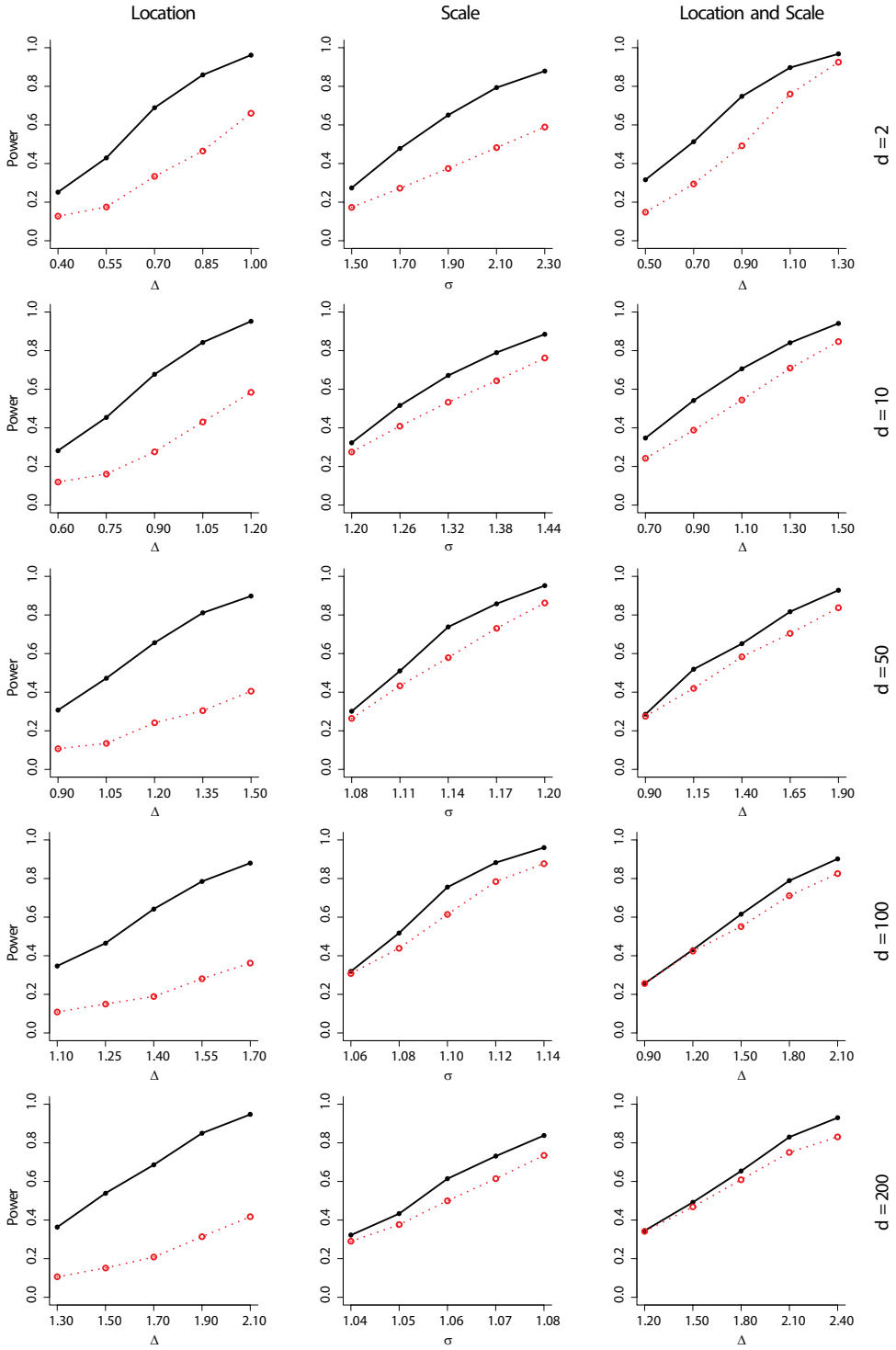

Figure 2.5: The simulated powers of Chen and Friedman's test (··○··) and our proposed test (–●–) at $\alpha = 0.05$ for detecting location differences, scale differences, and both location and scale differences.

This explains its inferior performance comparing with our proposed test which directly uses interpoint distances.

2.4.2 Preference Ranking Data

In our second simulation study, we consider the preference ranking data used in Zhang and Chen (2022). In the preference ranking data, each observation is the rankings of multiple objects by some subject. Let N be the number of objects and Ξ be the set of all permutations of $\{1, ..., N\}$. Then the preference ranking data are drawn from Ξ according to some probability model. A commonly used probability model to generate the preference ranking data is the following Mallows model (Mallows 1957):

$$\mathbf{P}_{\theta,\eta}(\zeta) = \frac{1}{\psi(\theta)} \exp\{-\theta d(\zeta, \eta)\}, \quad \zeta, \eta \in \Xi, \quad \theta \in \mathbf{R}.$$

where $d(\cdot, \cdot)$ is a distance metric suitable for the ranking data and $\psi(\theta)$ is a normalizing constant. In this Mallows model, there are two parameters (η, θ). The parameter η can be considered as the "center" of the distribution, while the parameter θ controls the "spread" of the distribution – the larger θ is, the less the distribution spreads.

In this simulation study, the random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are generated by the above Mallows model with parameters (η_1, θ_1) and (η_2, θ_2), respectively. For both samples, $N = 6$, so the two samples are rankings of six objects by two groups of subjects. By testing the equality of their underlying distributions F and G, we can determine whether the two groups have the same preference over the six objects. In our simulations, following Zhang and Chen (2022), we choose $d(\cdot, \cdot)$ to be Spearman's distance.

To evaluate the Type-I error rate of our proposed two-sample test, we set $\eta_1 = \eta_2 = \eta$ and $\theta_1 = \theta_2 = \theta$ and consider the following four settings for (η, θ).

- Setting 1:

  $\eta = \{1, 2, 3, 4, 5, 6\}$, $\theta = 4$.

- Setting 2:

  $\eta = \{1, 2, 3, 4, 5, 6\}$, $\theta = 5.5$.

- Setting 3:

  $\eta = \{1, 2, 5, 4, 3, 6\}$, $\theta = 4$.

- Setting 4:

  $\eta = \{1, 2, 5, 4, 3, 6\}$, $\theta = 5.5$.

To compare the powers of our proposed test with the ones based on the MST, we consider the following settings for (η_1, θ_1) and (η_2, θ_2).

- Setting 1 (Location difference):

  $\eta_1 = \{1, 2, 3, 4, 5, 6\}$, $\eta_2 = \{1, 2, 5, 4, 3, 6\}$, $\theta_1 = \theta_2 = 5$.

- Setting 2 (Scale difference with $\theta_1 > \theta_2$):

  $\eta_1 = \eta_2 = \{1, 2, 3, 4, 5, 6\}$, $\theta_1 = 5.5$, $\theta_2 = 4$.

- Setting 3 (Scale difference with $\theta_1 < \theta_2$):

  $\eta_1 = \eta_2 = \{1, 2, 3, 4, 5, 6\}$, $\theta_1 = 4$, $\theta_2 = 5.5$.

- Setting 4 (Location and scale differences with $\theta_1 > \theta_2$):

  $\eta_1 = \{1, 2, 3, 4, 5, 6\}$, $\eta_2 = \{1, 2, 5, 4, 3, 6\}$, $\theta_1 = 5.5$, $\theta_2 = 4$.

- Setting 5 (Location and scale differences with $\theta_1 < \theta_2$):

  $\eta_1 = \{1, 2, 3, 4, 5, 6\}$, $\eta_2 = \{1, 2, 5, 4, 3, 6\}$, $\theta_1 = 4$, $\theta_2 = 5.5$.

We also use Spearman's distance as the distance metric in our T_{loc} and T_{scal} as well as in building the MST. However, by doing so, it is common to have ties in the distance matrix for the above preference ranking data, which can lead to multiple solutions when constructing the MST. To address this issue, Zhang and Chen (2022) proposed several two-sample tests based on either the statistics from the union of all MSTs or the average of the statistics from all MSTs. Therefore, in our simulations, we compare our proposed test with their tests, which are based on the test statistics $S_{(a)}$, $M_{(a)}(1.14)$, $S_{(u)}$ and $M_{(u)}(1.14)$.

Table 2.2 shows that the simulated Type-I error rates for Zhang and Chen's four tests and our proposed test under different settings when the nominal significance level $\alpha = 0.05$. As we can see from the table, the simulated Type-I error rates of all the five tests are close to the nominal level under all the settings.

Table 2.2: The simulated Type-I error rates of the four tests from Zhang and Chen (2022) and our proposed test under different settings with $\alpha = 0.05$.

Setting	Sample Size	$S_{(a)}$	$M_{(a)}(1.14)$	$S_{(u)}$	$M_{(u)}(1.14)$	Proposed
1	$n_1 = n_2 = 50$	0.054	0.068	0.045	0.058	0.046
	$n_1 = 25, n_2 = 100$	0.052	0.065	0.050	0.068	0.036
2	$n_1 = n_2 = 50$	0.037	0.053	0.047	0.058	0.050
	$n_1 = 25, n_2 = 100$	0.043	0.048	0.053	0.054	0.041
3	$n_1 = n_2 = 50$	0.048	0.053	0.040	0.045	0.053
	$n_1 = 25, n_2 = 100$	0.057	0.066	0.057	0.066	0.038
4	$n_1 = n_2 = 50$	0.040	0.052	0.053	0.056	0.046
	$n_1 = 25, n_2 = 100$	0.059	0.062	0.059	0.065	0.041

Table 2.3 shows the simulated powers of different tests under various settings. Again the significance levels of all the tests are set at $\alpha = 0.05$. As we can see from Table

2.3, in the settings considered here, Zhang and Chen's tests based on $S_{(u)}$ and $M_{(u)}(1.14)$ are generally better than their $S_{(a)}$ and $M_{(a)}(1.14)$ counterparts. However, they are still significantly outperformed by our proposed test. Again this is because Zhang and Chen's tests are all based on the number of edges in the MST, while our proposed test is constructed directly on interpoint distances.

Table 2.3: The simulated powers of the four tests from Zhang and Chen (2022) and our proposed test at $\alpha = 0.05$ under different settings.

Setting	Sample Size	$S_{(a)}$	$M_{(a)}(1.14)$	$S_{(u)}$	$M_{(u)}(1.14)$	Proposed
1	$n_1 = n_2 = 50$	0.464	0.539	0.488	0.545	0.698
	$n_1 = 25, n_2 = 100$	0.373	0.437	0.386	0.449	0.589
2	$n_1 = n_2 = 150$	0.112	0.140	0.425	0.421	0.591
	$n_1 = 200, n_2 = 400$	0.134	0.179	0.726	0.729	0.833
3	$n_1 = n_2 = 150$	0.097	0.130	0.421	0.427	0.551
	$n_1 = 100, n_2 = 500$	0.138	0.163	0.495	0.506	0.674
4	$n_1 = n_2 = 50$	0.476	0.530	0.552	0.586	0.760
	$n_1 = 50, n_2 = 100$	0.573	0.642	0.637	0.675	0.858
5	$n_1 = n_2 = 50$	0.481	0.542	0.548	0.581	0.764
	$n_1 = 25, n_2 = 75$	0.345	0.404	0.456	0.452	0.690

2.4.3 Haplotype Association Data

Our third simulation study uses the haplotype association data considered in Chen and Zhang (2013). The data consists of haplotypes of 400 subjects at 11 single nucleotide polymorphisms (SNPs). For each subject, their haplotypes at these 11 SNPs are coded as an 11-dimensional binary vector. As a result, there are $2^{11} = 2048$ possible haplotypes. The haplotype data are binary vectors in nature, and the Hamming distance is a popular choice for measuring the interpoint distances. Therefore, we use it as the distance metric to construct our T_{loc} and T_{scal} as well as the MST. Again, there are ties in the distance matrix

when constructing the MST. As a result, the MST is not unique. Similar to the previous simulation study for the preference ranking data, we compare our proposed test with the four tests from Zhang and Chen (2022) in our simulations.

Before we compare the powers of the five tests, we first conduct a simulation study to assess their Type-I error rates. For this purpose, the haplotypes of 400 subjects are uniformly generated from the 2048 possible haplotypes in each simulation run. Regardless of his or her haplotype, we assume that each subject has a 30% chance of getting a certain disease. Subjects are then separated into the "patient" group and the "normal" group depending on whether they get the disease. Our proposed test along with Zhang and Chen's four tests are applied to the resulting haplotype data to see if there is a difference between the two groups. Since getting the disease is independent of the haplotype in this setting, the haplotypes of the "patient" group have the same distribution as those of the "normal" group. Therefore, we can use this setting to evaluate the Type-I error rates of the five tests. Table 2.4 presents the simulated Type-I error rates of the five tests at different nominal levels. As we can see from the table, the simulated Type-I error rates of the five tests are all close to their corresponding nominal levels.

Table 2.4: The simulated Type-I error rates of the four tests from Zhang and Chen (2022) and our proposed test.

The simulated Type-I errors										
α	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
$S_{(a)}$	0.005	0.025	0.029	0.057	0.048	0.063	0.070	0.087	0.084	0.110
$M_{(a)}(1.14)$	0.011	0.026	0.035	0.053	0.054	0.054	0.072	0.089	0.091	0.112
$S_{(u)}$	0.007	0.017	0.038	0.054	0.061	0.070	0.067	0.084	0.081	0.106
$M_{(u)}(1.14)$	0.014	0.026	0.035	0.059	0.054	0.074	0.080	0.096	0.089	0.116
Proposed	0.011	0.023	0.026	0.034	0.044	0.048	0.063	0.090	0.100	0.104

To compare the powers of the five tests, we assume that among the 11 SNPs, four are informative and seven are non-informative. The probability of the subject getting the disease is determined by the number of the four informative SNP positions at which the subject's haplotype agrees with a target haplotype, that is,

$$P(\text{Disease}) = 0.3 + 0.1 \times (\text{Number of positions in agreement}).$$

Again, in each simulation run, the haplotypes of 400 subjects are uniformly generated from the 2048 possible haplotypes. Subjects are divided into the "patient" group and the "normal" group according to the above disease model. Our proposed test along with Zhang and Chen's four tests are then applied to the resulting haplotype data to see if they are able to detect the differences between the two groups. Figure 2.6 shows the simulated powers of different tests as the significance level α changes. As we can see from Figure 2.6, although $M_{(u)}(1.14)$ is the best among the four MST-based tests from Zhang and Chen (2022), our proposed test has significantly better power than $M_{(u)}(1.14)$. The reason is the same as those in the previous two simulation studies.

From all the simulations presented in this section, we can see that the failure to fully utilize the interpoint distances in the MST-based tests can lead to significant efficiency loss. Our proposed test which makes good use of both the interpoint distances and the underlying dependency between T_{loc} and T_{scal} has the best performance in all the simulation settings considered here.

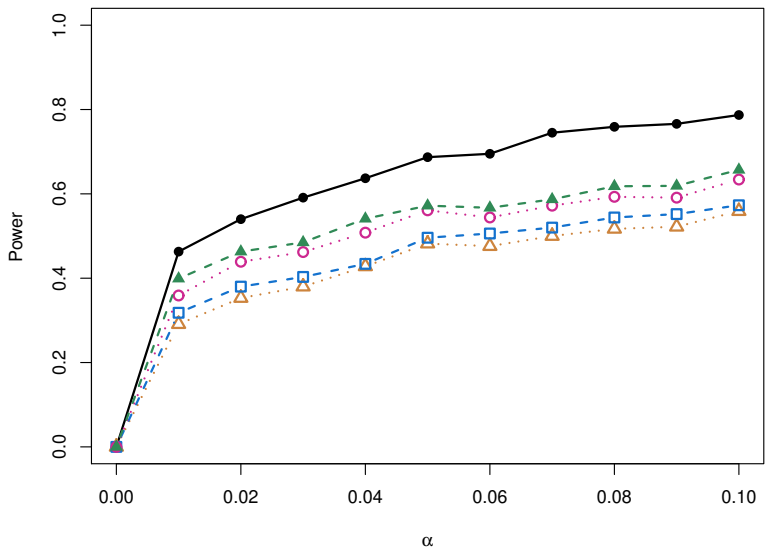

Figure 2.6: The simulated powers of our proposed test (–●–), and the four tests from Zhang and Chen (2022) ($S_{(a)}$ (··△··), $S_{(u)}$ (-⊟-), $M_{(a)}(1.14)$ (··○··) and $M_{(u)}(1.14)$ (-▲-)).

2.5 A Real Data Example

In this section, we demonstrate the application of our proposed test on a real phone-call network data set, which was collected by the MIT Media Laboratory and is available at `http://realitycommons.media.mit.edu/realitymining.html`. The data set contains call logs from 106 students and staff in an institute from July 2004 to June 2005 (Eagle et al. 2009). During that period of time, there were 31 days when no call was made among the 106 subjects, therefore those days are removed from our analysis. For the remaining 299 days, 19 subjects did not made any calls with the 106 subjects, therefore are also excluded from our analysis. For the remaining data, one question of interest is whether phone call patterns on weekdays are different from those on weekends.

To answer this question, we first construct a directed phone-call network for each day with the 87 subjects as nodes and a directed edge pointing from subject i to subject j if subject i made at least one call to subject j on that day. We then divide the networks into two groups according to whether the day is a weekday or weekend. Then our task is essentially to compare the underlying distributions of the phone-call networks between these two groups.

To apply our proposed test to this setting, we need to find an appropriate distance metric $d(\cdot, \cdot)$ for the phone-call network data. For this purpose, we follow Zhang and Chen (2022) and use the number of different directed edges between two directed networks as our distance metric $d(\cdot, \cdot)$. When this distance metric is used, there are many ties in the distance matrix, making the MST non-unique. Therefore, in the following, we compare the performance of our proposed test to that of the four tests from Zhang and Chen (2022) on this phone-call network data.

If we use the whole data set, all the five tests yield significant results at the significance level 0.05. To show the superior performance of our proposed tests over Zhang and Chen's tests, we randomly sample subsets of different sizes from the whole data set, keeping the proportions of observations from the two groups as close as they are in the original data set, and then use them to test the differences between the two groups based on the five tests. We repeat this procedure 100 times. The simulated powers of the five tests based on the 100 replicates using subsets of different sizes are shown in Figure 2.7. As we can see from Figure 2.7, our proposed test has significantly higher power than Zhang and Chen's four MST-based tests, which once again shows the substantial efficiency gain

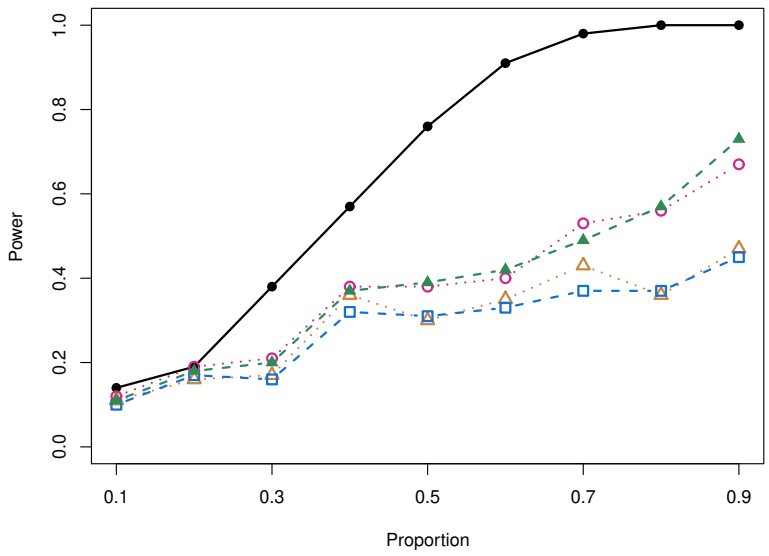

Figure 2.7: Powers of our proposed test (–●–), and the four tests from Zhang and Chen (2022) ($S_{(a)}$ (···△···), $S_{(u)}$ (-⊟-), $M_{(a)}(1.14)$ (···○···) and $M_{(u)}(1.14)$ (-▲-)) for comparing phone-call patterns on weekdays and on weekends at $\alpha = 0.05$.

of our proposed test due to its capability of fully utilizing the information contained in the interpoint distances.

Chapter 3

Nonparametric Control Chart for Count Data

3.1 Introduction

Count data monitoring arises from various applications across many industries, including defect detection in manufacturing, fraud detection in financial services, disease outbreak surveillance, network traffic monitoring and others. In the statistical process control (SPC) literature, many parametric control charts have been proposed to monitor count data. See, for example, Woodall (1997) for an overview on this topic. The binomial distribution and Poisson distribution are two classic probability distributions to model count data. The binomial distribution is usually used when people are interested in monitoring the proportion of nonconforming or defective items observed in a sample. Besides the most widely used Shewhart p-chart and np-chart, many other control charts have been developed

for monitoring binomial data. See, for example, Gan (1990, 1993), Khoo (2004), Wu et al. (2008). When control charts are needed to detect a change in the rate of occurrence of an event, we deal with count data that represent the number of times an event occurs in a given time interval. In such situations, the Poisson distribution is commonly used for describing data of this type. There is a considerable literature on control charts for monitoring independent and identically distributed Poisson random variables, for example, the traditional Shewhart c-chart and u-chart, the CUSUM charts developed in Lucas (1985), Lai (1995) and White and Keats (1996), and the EWMA charts from Gan (1990).

Parametric control charts introduced above are based on the assumption that the underlying process follows a particular discrete probability distribution. In practice, these assumptions may not be valid. Many researchers have investigated the impact of violations of such assumptions and offered their solutions. For instance, when data exhibit over-dispersion (i.e., the variance is greater than the mean), which is a violation of the equi-dispersion (i.e., the variance equals the mean) assumption for the Poisson distribution, the false alarm rate will be higher than the nominal level. To overcome this drawback of control charts developed from the Poisson distribution, Sheaffer and Leavenworth (1976) and Kaminsky et al. (1992) proposed control charts based on the negative binomial distribution or its special case, the geometric distribution. Sellers (2012) developed the generalized version of the traditional Shewhart charts by using the Conway-Maxwell-Poisson distribution, which is flexible enough to model count data that are either over- or under-dispersed. Other researchers considered compound distributions such as the Poisson-gamma mixture

(Cheng and Yu (2013)) and the shifted (or zero-truncated) generalized Poisson distribution (Famoye (1994)) to develop control charts for count data.

Despite the availability of various discrete probability distributions, Wang and Qiu (2018) argued that it is usually very difficult to find a proper parametric distribution to model count data in real-world settings. This is partly due to the fact that count data can be easily affected by confounding factors that are possibly unknown or unmeasurable. Motivated by the above limitation, Wang and Qiu (2018) developed two nonparametric CUSUM charts for detecting mean shifts for count data. In their proposed procedure, count data are firstly categorized, and then the Pearson's Chi-squared test and likelihood ratio test are used to develop two CUSUM charts for monitoring the resulting categorical data. However, their control charts are directly based on the categorized data, which fail to preserve the ordering information of the original data. As a result, as shown in our simulation studies, their control charts can suffer efficiency loss when detecting mean shifts. Additionally, Wang and Qiu (2018) proposed to use a bisection searching algorithm based on bootstrap to find the control limits. Based on some simulation studies that we conducted, their bootstrap algorithm requires a substantial amount of Phase-I sample to achieve the desired in-control performance.

To address the above limitations that the existing control charts have, we introduce a new nonparametric control chart for detecting mean shifts for count data. Our proposed control chart is also based on the idea of data categorization similar to the one used in Wang and Qiu (2018), but borrows the idea from Li (2021) to incorporate the ordering information of the original data, which leads to significant gain in efficiency as shown in our

simulation studies. In order to ensure a desired in-control performance with modest amount of Phase-I sample, we adopt the bootstrap procedure from Gandy and Kvaløy (2013) to help determine the control limit of our proposed control chart. Our simulation studies and real data analysis show that the proposed control chart performs well across a variety of settings, and significantly outperforms the two nonparametric CUSUM charts developed in Wang and Qiu (2018).

The remainder of the chapter is organized as follows. In Section 3.2, we introduce our proposed nonparametric control chart and describe the procedure to determine its control limit. The performance of the proposed control chart is evaluated and compared with the two nonparametric CUSUM charts from Wang and Qiu (2018) in several simulation studies and a real data example in Section 3.3 and Section 3.4 respectively.

3.2 Methodology

3.2.1 The Proposed Nonparametric Control Chart

The typical setup we consider in this project is the following. There are m independent and identically distributed Phase-I count data, denoted by $X_{-m+1}, \ldots, X_0$ from some in-control process. Let $X_1, X_2, \ldots$ be the Phase-II count data collected over time from the process. Denote the support of X_i by $\mathcal{S}$. Then the probability mass functions of the in-control (IC) and out-of-control (OC) distributions of X_i are denoted by $p_{0,X}(x)$ and $p_{1,X}(x)$, respectively, $x \in \mathcal{S}$, At any time t, the task of any control chart is to determine whether the process has changed from the IC distribution $p_{0,X}(x)$ to the OC distribution $p_{1,X}(x)$ based on all the available observations $X_1, ..., X_t$. This is equivalent to a test of the

following hypothesis:

$$H_0 : X_1, \ldots, X_t \text{ follow } p_{0,X}(x), x \in \mathcal{S}$$

versus

$$H_1 : \exists\, \tau \in [1, t] \text{ such that } X_1, \ldots, X_{\tau-1} \text{ follow } p_{0,X}(x) \text{ and } X_\tau, \ldots, X_t \text{ follow } p_{1,X}(x), x \in \mathcal{S},$$

where τ is the change point.

If we assume that the IC and OC distributions $p_{0,X}(x)$ and $p_{1,X}(x)$ are both completely known, to test the above hypothesis, the charting statistic based on the likelihood ratio method is $S_t = \max_{1 \leqslant \tau \leqslant t} \sum_{i=\tau}^{t} \log\{p_{1,X}(X_i)/p_{0,X}(X_i)\}$, and it can be also computed recursively by:

$$S_t = \max\left(0, S_{t-1} + \log\left\{\frac{p_{1,X}(X_t)}{p_{0,X}(X_t)}\right\}\right). \tag{3.1}$$

To implement the above CUSUM chart, both $p_{0,X}(x)$ and $p_{1,X}(x)$ need to be completely specified. However, in our nonparametric setting, both $p_{0,X}(x)$ and $p_{1,X}(x)$ are unknown. To get around this difficulty, we use the same data categorization idea in Wang and Qiu (2018) to categorize the data so that any unknown $p_{0,X}(x)$ and $p_{1,X}(x)$ can be converted into a multinomial distribution. More specifically, let $0 < q_1 < q_2 < \ldots < q_{d-1} < \infty$ be $d-1$ boundary points, and then the support of X_i, $\mathcal{S}$, can be partitioned into the following d intervals,

$$A_1 = [0, q_1), A_2 = [q_1, q_2), \ldots, A_d = [q_{d-1}, \infty).$$

Let

$$Y_{t,j} = I(X_t \in A_j), \text{ for } j = 1, \ldots, d,$$

where $I(u)$ is the indicator function that takes the value of 1 if u is true and the value of 0 otherwise. In other words, $Y_{t,j}$ indicates whether X_t falls within the interval A_j, $j = 1, ..., d$. Let $\boldsymbol{Y}_t = (Y_{t,1}, ..., Y_{t,d})'$. It is easy to see that $\boldsymbol{Y}_t$ follows a multinomial distribution with $n = 1$ and category probabilities $c_j = P(X_t \in A_j)$, $j = 1, \ldots, d$, denoted by Multi(1; $c_1, \ldots, c_d$). Therefore, based on the above data categorization, the original count data X_t with any arbitrary distribution is converted into the multinomial random variable $\boldsymbol{Y}_t$, and detecting a mean shift in X_t can be achieved by detecting a mean shift in $\boldsymbol{Y}_t$.

We denote the IC distribution of $\boldsymbol{Y}_t$ by Multi(1; $c_1^{(0)}, \ldots, c_d^{(0)}$), and the OC distribution by Multi(1; $c_1^{(1)}, \ldots, c_d^{(1)}$), where $\sum_{j=1}^d c_j^{(0)} = \sum_{j=1}^d c_j^{(1)} = 1$ and $(c_1^{(0)}, \ldots, c_d^{(0)}) \neq (c_1^{(1)}, \ldots, c_d^{(1)})$. Replacing the IC and OC distributions of X_t in (3.1) with the ones of $\boldsymbol{Y}_t$, the charting statistic becomes

$$S_t = \max\left(0, S_{t-1} + \sum_{j=1}^{d} Y_{t,j} \log\left\{c_j^{(1)}/c_j^{(0)}\right\}\right).$$

As pointed out in Li (2021), the above charting statistic is constructed based on the multinomial distributions Multi(1; $c_1^{(0)}, \ldots, c_d^{(0)}$) and Multi(1; $c_1^{(1)}, \ldots, c_d^{(1)}$), which do not make use of the ordering information of the d intervals, A_1,...,A_d. This loss of ordering information can lead to efficiency loss when detecting mean shifts in $\boldsymbol{Y}_t$. In order to incorporate the ordering information of A_1,...,A_d, following the approached proposed in Li (2021), we construct cumulative unions of the d intervals and then develop a charting

statistic based on cumulative sums of $Y_{t,j}, j = 1, \ldots, d$. This is similar to what is usually done to incorporate the ordering information of the categories in the ordered logistic regression (versus the regular multinomial logistic regression). More specifically, we define the cumulative unions of the intervals $A_1, \ldots, A_d$ as follows,

$$A_1, A_1 \cup A_2, A_1 \cup A_2 \cup A_3, \ldots, A_1 \cup \ldots \cup A_d.$$

and denote the cumulative sums of $Y_{t,1}, ..., Y_{t,d}$ as

$$Z_{t,j} = \sum_{l=1}^{j} Y_{t,l}, \text{ for } j = 1, \ldots, d,$$

where $Z_{t,j}$ indicates whether X_t falls within the interval $A_1 \cup \ldots \cup A_j$. It is easy to see that $Z_{t,j}$, $j = 1, ..., d-1$, is a Bernoulli random variable and the log-likelihood ratio based on $Z_{t,j}$ is

$$Z_{t,j} \log \left(\frac{\sum_{l=1}^{j} c_l^{(1)}}{\sum_{l=1}^{j} c_l^{(0)}} \right) + (1 - Z_{t,j}) \log \left(\frac{1 - \sum_{l=1}^{j} c_l^{(1)}}{1 - \sum_{l=1}^{j} c_l^{(0)}} \right).$$

Similar to the ordered logistic regression, if we use an appropriate test statistic based on all $Z_{t,j}$, $j = 1, ..., d-1$, in our charting statistic, the ordering information of A_1,..., A_d can be incorporated. Following Li (2021), to incorporate all $Z_{t,j}$, $j = 1, ..., d-1$, we consider a weighted sum of the above log-likelihood ratios from each $Z_{t,j}$, $j = 1, ..., d-1$, as the test statistic and the resulting test statistic is

$$\sum_{j=1}^{d-1} \omega(j) \Big\{ Z_{t,j} \log \left(\frac{\sum_{l=1}^{j} c_l^{(1)}}{\sum_{l=1}^{j} c_l^{(0)}} \right) + (1 - Z_{t,j}) \log \left(\frac{1 - \sum_{l=1}^{j} c_l^{(1)}}{1 - \sum_{l=1}^{j} c_l^{(0)}} \right) \Big\},$$

where $\omega(j)$ is the weight function, and following Li (2021), we choose $\omega(j) = (j/d)^{-1}(1 - j/d)^{-1}$ to give more weights to the tail areas. Using the above test statistic based on all $Z_{t,j}$, $j = 1, ..., d-1$, our proposed charting statistic is given by

$$S_t = \max\left(0, S_{t-1} + \sum_{j=1}^{d-1} \frac{d^2}{j(d-j)} \left\{ Z_{t,j} \log\left(\frac{\sum_{l=1}^{j} c_l^{(1)}}{\sum_{l=1}^{j} c_l^{(0)}}\right) + (1 - Z_{t,j}) \log\left(\frac{1 - \sum_{l=1}^{j} c_l^{(1)}}{1 - \sum_{l=1}^{j} c_l^{(0)}}\right) \right\}\right). \tag{3.2}$$

To implement the control chart based on the above charting statistic, we need to first choose the boundary points $\{q_1, q_2, ..., q_{d-1}\}$. As mentioned earlier, through data categorization, we detect a mean shift in X_t by detecting a mean shift in $\boldsymbol{Y}_t$, which is equivalent to detecting a change in $(c_1^{(0)}, \ldots, c_d^{(0)})$. To help detect changes in $(c_1^{(0)}, \ldots, c_d^{(0)})$, research in the categorical data analysis literature suggests to choose $\{q_1, q_2, ..., q_{d-1}\}$ such that $c_1^{(0)}, \ldots, c_d^{(0)}$ are roughly the same (see, for example, Agresti (2002), Section 1.5). Following this suggestion, we choose $\{q_1, q_2, ..., q_{d-1}\}$ such that the estimated IC category probabilities based on the Phase-I sample, $\widehat{c}_1^{(0)}, ...\widehat{c}_d^{(0)}$, are as close to one another as possible.

Once we have chosen the boundary points $\{q_1, q_2, ..., q_{d-1}\}$ as described above, next we need to determine the IC and OC category probabilities, $\boldsymbol{c}^{(0)} = (c_1^{(0)}, \ldots, c_d^{(0)})'$ and $\boldsymbol{c}^{(1)} = (c_1^{(1)}, \ldots, c_d^{(1)})'$ used in (3.2). Similar to Wang and Qiu (2018), we use the estimated IC category probabilities based on the Phase-I sample, $\widehat{\boldsymbol{c}}_0 = (\widehat{c}_1^{(0)}, ...\widehat{c}_d^{(0)})'$, to substitute $\boldsymbol{c}^{(0)} = (c_1^{(0)}, \ldots, c_d^{(0)})'$ in (3.2).

To find the OC category probabilities $\boldsymbol{c}^{(1)}$, if we know the change point τ, we can simply use the observations collected after the change point τ to estimate $\boldsymbol{c}^{(1)}$. However,

the change point τ is usually unknown in practice. When both the IC and OC distributions are the normal distributions but with different means, to achieve the optimal performance of the standard CUSUM chart, the OC mean also needs to be estimated. Sparks (2000) proposed using an exponentially weighted moving average of all the past observations to estimate the OC mean and simulation studies have shown that this method of estimating the OC mean works well. Following the same idea, we also propose to estimate the OC category probabilities $\boldsymbol{c}^{(1)}$ by the exponentially weighted moving average of all the past observations. More specifically, the estimate of $\boldsymbol{c}^{(1)} = (c_1^{(1)}, \ldots, c_d^{(1)})'$ at time t is given by $\widehat{\boldsymbol{c}}_t^{(1)} = (\widehat{c}_{t,1}^{(1)}, ..., \widehat{c}_{t,d}^{(1)})'$, and for $t > 1$,

$$\widehat{\boldsymbol{c}}_t^{(1)} = \lambda \boldsymbol{Y}_{t-1} + (1-\lambda)\widehat{\boldsymbol{c}}_{t-1}^{(1)}, \tag{3.3}$$

where $\widehat{\boldsymbol{c}}_1^{(1)} = \widehat{\boldsymbol{c}}_0$ and $\lambda \in (0, 1]$ is a weighting parameter.

Substituting $\boldsymbol{c}^{(0)} = (c_1^{(0)}, \ldots, c_d^{(0)})'$ and $\boldsymbol{c}^{(1)} = (c_1^{(1)}, \ldots, c_d^{(1)})'$ in (3.2) by their respective estimates described above, the charting statistic in (3.2) becomes,

$$\begin{aligned} S_t = \max \Bigg(0, S_{t-1} + \sum_{j=1}^{d-1} \frac{d^2}{j(d-j)} \Bigg\{ Z_{t,j} \log \left(\frac{\sum_{l=1}^{j} \widehat{c}_{t,l}^{(1)}}{\sum_{l=1}^{j} \widehat{c}_l^{(0)}} \right) \\ + (1 - Z_{t,j}) \log \left(\frac{1 - \sum_{l=1}^{j} \widehat{c}_{t,l}^{(1)}}{1 - \sum_{l=1}^{j} \widehat{c}_l^{(0)}} \right) \Bigg\} \Bigg). \end{aligned}$$

Then our proposed control chart is to plot the above S_t over the time t and it raises an alarm if $S_t > h$, where h is the predetermined control limit. In the next section, we describe how to determine h to guarantee certain IC performance.

3.2.2 Determining the Control Limit

As described in the previous section, in the charting statistic of our proposed control chart, the unknown IC category probabilities $\boldsymbol{c}^{(0)} = (c_1^{(0)}, \ldots, c_d^{(0)})'$ are estimated by the Phase-I sample, $X_{-m+1}, \ldots, X_0$. The resulting average run length (denoted by ARL) of our proposed control chart is a random variable whose distribution depends on the estimated IC category probabilities $\widehat{\boldsymbol{c}}^{(0)} = (\widehat{c}_1^{(0)}, \ldots \widehat{c}_d^{(0)})'$ from the Phase-I sample, therefore it is usually referred to as the conditional ARL. In the SPC literature, control limits are often determined to control the unconditional IC ARL (the average of the conditional IC ARL over a large number of Phase-I samples) at the nominal level. However, as shown in the literature (see, for example, Jensen et al. (2006), Psarakis et al. (2014), Saleh et al. (2016), Capizzi and Masarotto (2020)), when we only consider controlling the unconditional IC ARL at the nominal level, the conditional IC ARL can be much lower than the nominal level in a large proportion of cases. To avoid this undesirable consequence, Jones and Steiner (2012) and Gandy and Kvaløy (2013) proposed to find the control limit so that a large proportion of the conditional IC ARL is close to the nominal level. Since this new criterion can guarantee the desired performance of the conditional IC ARL, we will adopt this criterion in determining our control limit h. More specifically, denote condi-$\text{ARL}_0(h)$ as the conditional IC ARL of our proposed control chart using control limit h, and ARL_0 as the nominal level of the unconditional IC ARL. Then we want to find h so that

$$\text{Prob}\Big(\text{condi-ARL}_0(h) > \text{ARL}_0\Big) = 1 - \alpha, \tag{3.4}$$

where α is some small tolerance selected by users.

To find h to satisfy the above requirement, we resort to the bootstrap method proposed in Gandy and Kvaløy (2013). Recall that $p_{0,X}$ is the IC distribution of X_i, and $\widehat{\boldsymbol{c}}^{(0)} = (\widehat{c}_1^{(0)}, ..., \widehat{c}_d^{(0)})'$ are the estimated IC category probabilities of $\boldsymbol{Y}_t$ based on the Phase-I sample $X_{-m+1}, \ldots, X_0$. Given the Phase-I sample $X_{-m+1}, \ldots, X_0$, the conditional IC ARL of our proposed control chart depends on (i) $\widehat{\boldsymbol{c}}^{(0)}$, since the charting statistic of our proposed control chart involves $\widehat{\boldsymbol{c}}^{(0)}$; and (ii) $p_{0,X}$, since the future IC count data X_1, X_2,... are drawn from $p_{0,X}$. Due to this dependence, we define $g(h; \widehat{\boldsymbol{c}}^{(0)}, p_{0,X})$ as the conditional IC ARL of our proposed control chart based on the control limit h and the Phase-I sample $X_{-m+1}, \ldots, X_0$. Then its inverse function $g^{-1}(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, p_{0,X}) \equiv q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, p_{0,X})$ gives the desired control limit such that the conditional IC ARL of our proposed control chart based on the Phase-I sample $X_{-m+1}, \ldots, X_0$ is ARL_0. However, $p_{0,X}$ is unknown in practice, so $q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, p_{0,X})$ can not be easily obtained. To circumvent this difficulty, we can use the popular plug-in principle, i.e, replace the unknown $p_{0,X}$ by its estimate based on the Phase-I sample $X_{-m+1}, \ldots, X_0$, denoted by $\widehat{p}_{0,X}$. Since $\widehat{\boldsymbol{c}}^{(0)}$ and $\widehat{p}_{0,X}$ are both known for the given Phase-I sample, by applying some numerical algorithm (e.g. the bisection searching algorithm), we can find $q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X})$.

Based on $q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X})$, the $100(1-\alpha)\%$ lower one-sided confidence interval for $q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, p_{0,X})$ can be constructed as $\Big(-\infty,\, q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X}) - p_\alpha \Big)$, where p_α satisfies

$$\text{Prob}\Big(q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X}) - p_\alpha > q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, p_{0,X}) \Big) = 1 - \alpha. \tag{3.5}$$

According to (3.5), if we set our control limit at $h = q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X}) - p_\alpha$, then (3.4) will be automatically satisfied. Therefore, our remaining task is to find the above p_α.

To this end, we first notice that (3.5) can be written as

$$\text{Prob}\Big(q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X}) - q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, p_{0,X}) > p_\alpha\Big) = 1 - \alpha.$$

This implies that p_α is actually the α-th quantile of the sampling distribution of $q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X}) - q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, p_{0,X})$. However, the sampling distribution of $q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X}) - q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, p_{0,X})$ is usually unknown. Gandy and Kvaløy (2013) proposed to approximate it by bootstrap. More specifically, let $\boldsymbol{X} = (X_{-m+1}, \ldots, X_0)'$. Randomly generate B bootstrap resamples of $\boldsymbol{X}$, and denote them by $\mathbf{X}_1^*, \ldots, \mathbf{X}_B^*$. Let $\widehat{p}_{0,X,r}^*$ and $\widehat{\boldsymbol{c}}_r^{(0)*}$ be the counterparts of $\widehat{p}_{0,X}$ and $\widehat{\boldsymbol{c}}^{(0)}$, respectively, calculated from the bootstrap resample $\mathbf{X}_r^*$, $r = 1, \ldots, B$. Then the sampling distribution of $q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X}) - q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, p_{0,X})$ can be approximated by the empirical distribution of $q(\text{ARL}_0; \widehat{\boldsymbol{c}}_r^{(0)*}, \widehat{p}_{0,X,r}^*) - q(\text{ARL}_0; \widehat{\boldsymbol{c}}_r^{(0)*}, \widehat{p}_{0,X})$, $r = 1, \ldots, B$. As a result, we can use the α-th quantile of this empirical distribution, denoted by $\widehat{p}_\alpha^*$, to estimate p_α. Following the results from Gandy and Kvaløy (2013), if we set our control limit at $h = q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X}) - \widehat{p}_\alpha^*$, then (3.4) will hold asymptotically.

As suggested by Gandy and Kvaløy (2013), replacing $q(\text{ARL}_0; \cdot, \cdot)$ by its log transformation at each step of the above procedure can help improve the coverage probability. Following this suggestion, we summarize below the details of the algorithm we use for finding the control limit h of our proposed control chart.

1. Obtain $\widehat{p}_{0,X}$ and $\widehat{\boldsymbol{c}}^{(0)}$ using the Phase-I sample $X_{-m+1}, \ldots, X_0$.

2. Find $q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X})$ using some numerical algorithm (e.g., the bisection searching algorithm), and calculate $\log\Big(q(\text{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X})\Big)$.

3. Randomly generate B bootstrap resamples $\mathbf{X}_1^*, \ldots, \mathbf{X}_B^*$ from the Phase-I sample $\boldsymbol{X} = (X_{-m+1}, \ldots, X_0)'$. For the bootstrap resample $\mathbf{X}_r^*$, $r = 1, \ldots, B$, first calculate $\widehat{p}_{0,X,r}^*$ and $\widehat{\boldsymbol{c}}_r^{(0)*}$, the counterparts of $\widehat{p}_{0,X}$ and $\widehat{\boldsymbol{c}}^{(0)}$, respectively. Then use the same numerical algorithm used in Step 2 to find $q(\mathrm{ARL}_0; \widehat{\boldsymbol{c}}_r^{(0)*}, \widehat{p}_{0,X,r}^*)$ and $q(\mathrm{ARL}_0; \widehat{\boldsymbol{c}}_r^{(0)*}, \widehat{p}_{0,X})$, and compute $\log\Big(q(\mathrm{ARL}_0; \widehat{\boldsymbol{c}}_r^{(0)*}, \widehat{p}_{0,X,r}^*)\Big) - \log\Big(q(\mathrm{ARL}_0; \widehat{\boldsymbol{c}}_r^{(0)*}, \widehat{p}_{0,X})\Big)$. Then calculate $\widehat{p}_\alpha^*$, the α quantile of the empirical distribution of $\log\Big(q(\mathrm{ARL}_0; \widehat{\boldsymbol{c}}_r^{(0)*}, \widehat{p}_{0,X,r}^*)\Big) - \log\Big(q(\mathrm{ARL}_0; \widehat{\boldsymbol{c}}_r^{(0)*}, \widehat{p}_{0,X})\Big)$, $r = 1, \ldots, B$.

4. Finally, the control limit h that can guarantee the desired performance of the conditional IC ARL as in (3.4) is calculated as $\exp\Big\{\log\Big(q(\mathrm{ARL}_0; \widehat{\boldsymbol{c}}^{(0)}, \widehat{p}_{0,X})\Big) - \widehat{p}_\alpha^*\Big\}$.

3.3 Simulation Studies

In this section, we present several simulation studies to evaluate the performance of our proposed control chart. In particular, we compare our control chart with the two existing nonparametric CUSUM charts for count data, the P-CUSUM and L-CUSUM charts, from Wang and Qiu (2018). Following the simulation settings used in Wang and Qiu (2018), we consider the following four discrete distributions throughout our simulation studies:

- Binomial distribution: Suppose there are n independent Bernoulli trials, and the probability of a success is constant from trial to trial, denoted by p. The random variable X which represents the number of successes obtained in the n trials follows a binomial distribution with parameters n and p, denoted by $\mathrm{Bin}(n, p)$.

- Negative binomial distribution: Suppose there are a sequence of independent Bernoulli trials. The probability of a success for each trial is equal to p. The random variable X which represents the number of failures until the rth success has a negative binomial distribution, denoted by $\mathrm{NB}(r, p)$.

- Discrete uniform distribution: Let r be a specified integer. The random variable X which takes any integer value in $\{0, 1, 2, ..., r\}$ with equal probability $1/(r+1)$ has a discrete uniform distribution, denoted by $\mathrm{DU}(r)$.

- Generalized Poisson distribution: If X is a random variable following a generalized Poisson distribution with parameters η and θ, denoted by $\mathrm{GP}(\eta, \theta)$, then the probability mass function of X is given by:

$$p(x; \eta, \theta) = \begin{cases} \eta(\eta + \theta x)^{x-1} e^{-\eta - \theta x}/x!, & x = 0, 1, 2, \ldots, \\ 0, & \text{for } x > m \text{ if } \theta < 0, \end{cases}$$

  where $\eta > 0$ and $\max(-1, -\eta/m) \leqslant \theta < 1$, and $m \geqslant 4$. When $\theta < 0$, m is the largest positive integer that satisfies $\eta + m\theta > 0$. It is easy to see that, if $\theta = 0$, then $\mathrm{GP}(\eta, 0)$ reduces to the standard Poisson distribution with mean $\lambda = \eta$.

In the literature, the IOD is defined as the ratio of the variance to the mean, and is commonly used to measure the dispersion of count data. Poisson distribution is one of the most widely used distributions for count data. With identical mean and variance, Poisson distribution has the IOD value of 1, so it is said to be equi-dispersed. If the IOD of a distribution is smaller than 1 (i.e., the variance is smaller than the mean), the distribution

is said to be under-dispersed. If the IOD is larger than 1 (i.e., the variance is larger than the mean), the distribution is considered to be over-dispersed. For some distributions, the IOD could be either less than or greater than 1, depending on the value of the parameters, then they are considered as being mixed-dispersed. Table 3.1 lists the IOD and type of dispersion as well as the mean and variance for the above four distributions. In reality, count data often exhibit under-dispersion or over-dispersion compared to the standard Poisson distribution. From Table 3.1, we can see that the four distributions considered in our simulation studies cover a variety of under-dispersion and over-dispersion scenarios for count data.

Table 3.1: The mean, variance, IOD and type of dispersion for the four distributions used in our simulation studies.

Distribution	Mean	Variance	IOD	Type of dispersion
Bin(n, p)	np	$np(1-p)$	$0 < 1-p < 1$	under-dispersed
NB(r, p)	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$	$\frac{1}{p} > 1$	over-dispersed
DU(r)	$\frac{r}{2}$	$\frac{r(r+2)}{12}$	$\frac{r+2}{6}$	mixed-dispersed
GP(η, θ)	$\frac{\eta}{1-\theta}$	$\frac{\eta}{(1-\theta)^3}$	$\frac{1}{(1-\theta)^2}$	mixed-dispersed

3.3.1 IC Performance Evaluation

In this section, we evaluate the IC performance of our proposed control chart. We also assess the IC performance of the P-CUSUM and L-CUSUM charts to ensure a fair comparison of OC performance among those control charts in the subsequent section.

Since the P-CUSUM and L-CUSUM charts involve tuning parameters k_P and k_L respectively, we first discuss how these parameters are chosen in our simulation studies. According to Hawkins and Olwell (1998), the IC run-length distribution of a control chart

is considered to be satisfactory if it is close to the geometric distribution. Based on the properties of the geometric distribution, in order for the control chart to have satisfactory IC performance, we should expect the average and standard deviation of IC run-length (denoted by IC ARL and IC SDRL, respectively) to be roughly the same. In Wang and Qiu (2018), they recommended choosing k_P in the range of 0.001 to 0.01 for the P-CUSUM chart. According to some simulation study we conducted, the IC SDRL of the P-CUSUM chart or L-CUSUM chart is significantly different from their respectively IC ARL when k_P or k_L is very small. Therefore, to have a fair comparison, we try different choices of k_P and k_L for the P-CUSUM and L-CUSUM charts and choose the ones that yield similar IC ARL and IC SDRL. The smallest value of k_P and k_L that satisfy this criteria is 0.3. To help detect relatively large mean shifts, we also consider $k_P = 0.5$ and $k_L = 0.5$. To study how λ (the weighting parameter used in (3.3)) affects the performance of our proposed control chart, we consider the following choices for λ: $\lambda = 0.05, 0.1$ and 0.2. They are popular choices for λ in practice and also satisfy the criteria given in Hawkins and Olwell (1998).

Following Wang and Qiu (2018), we consider the following four IC distributions in our simulation studies: Bin$(20, 0.75)$, NB$(20, 0.75)$, DU(10), GP$(5, 0.25)$. We standardize the data generated from GP$(5, 0.25)$, so that the IC distribution has mean 0 and standard deviation 1. For each distribution, we randomly generate $m = 500$ identically and independently distributed Phase-I count data. For all the control charts, the number of categories d is chosen to be 5. For each control chart, we apply the procedure described in Section 3.2.2 to find the control limit so that its conditional IC ARL satisfies (3.4) with $\text{ARL}_0 = 200$ and $\alpha = 0.1$. When implementing the procedure from Section 3.2.2 to find the control limit,

we set $B = 1000$ and calculate all the conditional ARLs by averaging over 1000 sample paths. After obtaining the control limit for each control chart, the true conditional IC ARL of each control chart is calculated by averaging over 10000 sample paths. In order to evaluate the distribution of the conditional IC ARL over different Phase-I samples, the above procedure is repeated over 500 Phase-I samples. The boxplots of the 500 conditional IC ARLs from those 500 Phase-I samples for different distributions are displayed in Figure 3.1, and the proportions of the conditional IC ARL values that are greater than $\text{ARL}_0 = 200$ are presented in Table 3.2. Both Figure 3.1 and Table 3.2 show that approximate 90% of the conditional IC ARLs are at least the nominal level ($\text{ARL}_0 = 200$) for all the control charts under various settings. This indicates that, using the control limits obtained by the approach described in Section 3.2.2, all the control charts can satisfy (3.4) under all the settings considered here.

Table 3.2: The simulated proportions of the conditional IC ARLs that are at least the nominal level ($\text{ARL}_0 = 200$) for the P-CUSUM chart, L-CUSUM chart, and our proposed control chart.

	P-CUSUM		L-CUSUM		Proposed		
Distribution	$k_P = 0.3$	$k_P = 0.5$	$k_L = 0.3$	$k_L = 0.5$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$
Bin(20,0.75)	0.922	0.926	0.898	0.910	0.896	0.890	0.890
NB(20,0.75)	0.924	0.918	0.900	0.904	0.908	0.904	0.898
DU(10)	0.908	0.926	0.896	0.906	0.900	0.896	0.896
GP(5,0.25)	0.904	0.914	0.890	0.884	0.892	0.886	0.902

3.3.2 OC Performance Comparison

In this section, we report the results of simulation studies that are conducted to compare the detection power of our proposed control chart with the P-CUSUM and L-

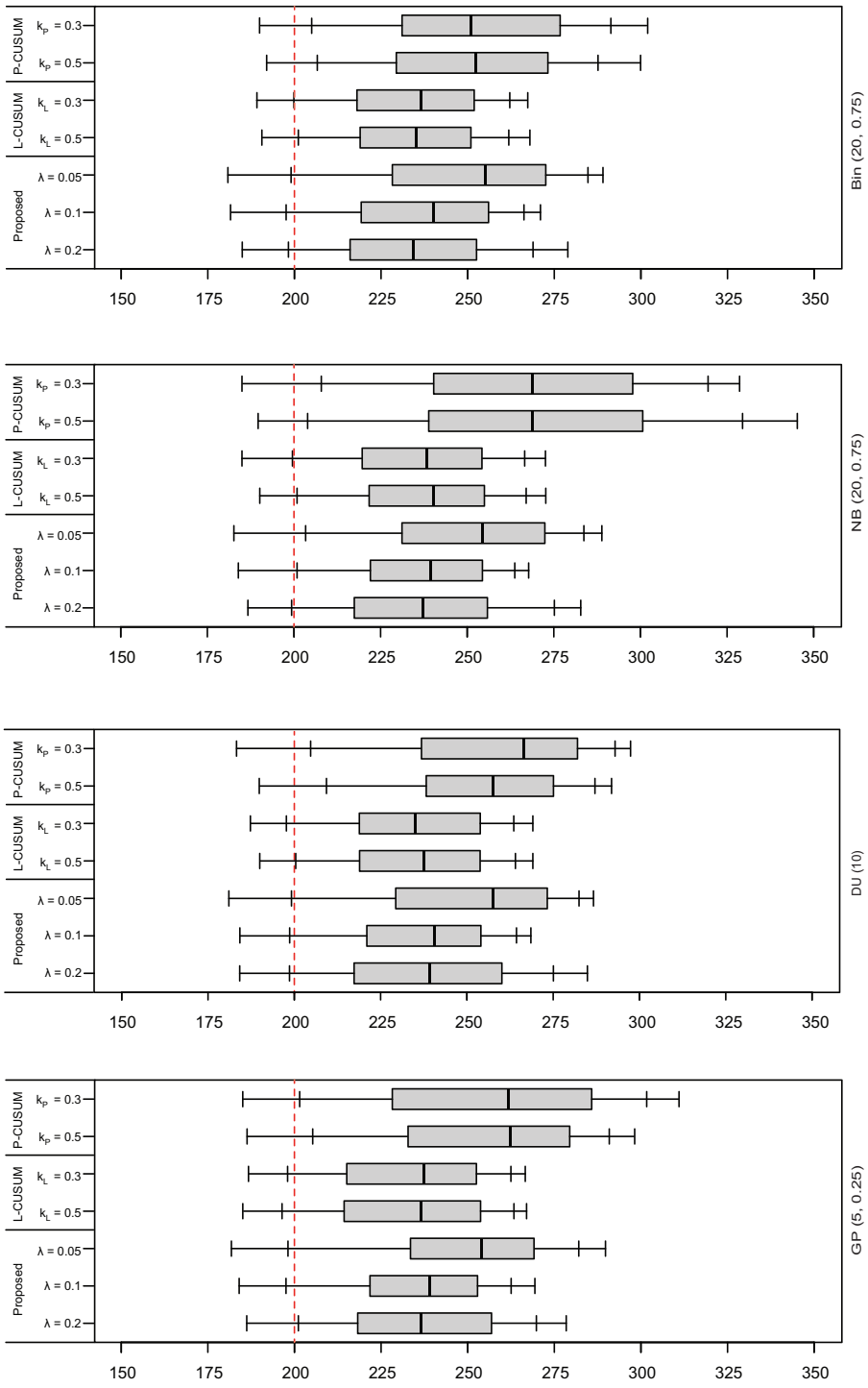

Figure 3.1: The boxplots of the conditional IC ARLs from 500 Phase-I samples of the P-CUSUM chart, L-CUSUM chart, and our proposed control chart. The boxplots show the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles of the conditional IC ARL distribution.

CUSUM charts for detecting mean shifts. Similar to Section 3.3.1, we assume that each Phase-I sample consists of $m = 500$ identically and independently distributed count data drawn from one of the four discrete distributions considered in Section 3.3.1, and for each Phase-I sample, the control limits for our proposed control chart, the P-CUSUM and L-CUSUM charts are determined by the procedure from Section 3.2.2 so that their conditional IC ARLs all satisfy (3.4) with $\text{ARL}_0 = 200$ and $\alpha = 0.1$.

To evaluate the detection power of different control charts for mean shifts, we need to simulate a certain type of mean shifts in Phase-II count data. To this end, we set the change point $\tau = 50$. For each simulated sample path, $X_1, \ldots, X_{\tau-1}$, the Phase-II observations collected before time τ, are generated from the IC distribution, and $X_\tau, X_{\tau+1}, \ldots$, the observations collected after time τ, are generated from some OC distribution. The OC distributions of $\text{Bin}(20, 0.75)$ and $\text{NB}(20, 0.75)$ are chosen to be $\text{Bin}(20, p)$ and $\text{NB}(20, p)$ respectively, with p ranging from 0.70 to 0.80 in increments of 0.01. For the OC distribution of $\text{DU}(10)$, the sample space remains as $\{0, 1, 2, \ldots, 10\}$, but the probability associated with each outcome is replaced by $P(X = i) = (i + 1)/66, i = 0, \ldots, 10$. For the OC distribution of $\text{GP}(5, 0.25)$, we simply add a shift of δ to the count data simulated from the standardized $\text{GP}(5, 0.25)$ after the change point τ, with δ ranging from -1.0 to 1.0 in increments of 0.2.

Under the above settings, the conditional OC ARL of each control chart for a particular Phase-I sample is calculated by the average run length from 10000 sample paths. This procedure is repeated over 500 Phase-I samples and the average of the 500 conditional OC ARLs is the simulated unconditional OC ARL. Tables 3.3 - 3.6 show the simulated unconditional OC ARLs along with their corresponding standard errors (in parentheses)

for the P-CUSUM, L-CUSUM charts and our proposed control chart under different OC settings. The smallest OC ARL value in each row, representing the best detection power under each setting, is marked in bold. From Tables 3.3 - 3.6, we can see that, consistent with EWMA-type control charts, our proposed control chart is more powerful in detecting small mean shifts when small values of λ are used, and is more powerful in detecting large mean shifts when large values of λ are used. Comparing our proposed control chart with the P-CUSUM and L-CUSUM charts, we can see that our proposed control chart significantly outperforms the P-CUSUM and L-CUSUM charts in all the scenarios except for the case when the IC distribution is $\text{GP}(5, 0.25)$ and the OC observations experience an upward mean shift with magnitude 0.2. However, all the control charts seem to be ineffective in detecting such a small mean shift in this scenario, and the slightly better performance of the L-CUSUM chart might be due to its smaller unconditional IC ARLs as shown in Figure 3.1.

From both the IC and OC performance reported above, it is clear that our proposed control chart can achieve the desired control of the conditional IC ARL under different distributions, and is more efficient than the two existing nonparametric control charts, the P-CUSUM and L-CUSUM charts, for detecting various mean shifts.

3.4 Real Data Application

In compliance with the Jeanne Clery Act, the website of the University of Florida Public Safety at `https://publicsafety.ufl.edu/clery/` maintains the Crime Log and

Table 3.3: The simulated unconditional OC ARLs along with their corresponding standard errors (in parentheses) for the P-CUSUM, L-CUSUM charts and our proposed control chart when the IC distribution is Bin$(20, 0.75)$.

	P-CUSUM		L-CUSUM		Proposed		
p	$k_P = 0.3$	$k_P = 0.5$	$k_L = 0.3$	$k_L = 0.5$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$
0.70	54.75	68.32	39.12	42.38	26.15	**25.94**	28.90
	(0.79)	(1.11)	(0.32)	(0.39)	(0.22)	(0.21)	(0.24)
0.71	83.11	100.52	54.80	61.03	**35.48**	36.59	42.18
	(1.28)	(1.62)	(0.53)	(0.63)	(0.35)	(0.36)	(0.41)
0.72	127.19	143.52	82.52	92.15	**53.27**	56.85	66.49
	(1.88)	(2.03)	(0.89)	(1.01)	(0.65)	(0.67)	(0.74)
0.73	182.99	191.34	129.16	140.18	**92.28**	98.45	112.19
	(2.17)	(2.09)	(1.36)	(1.43)	(1.38)	(1.30)	(1.31)
0.74	235.41	232.74	189.89	196.83	174.16	**174.03**	183.54
	(1.90)	(1.74)	(1.47)	(1.44)	(2.25)	(1.84)	(1.66)
0.76	246.71	246.34	188.05	198.24	**170.20**	171.41	182.24
	(1.87)	(1.77)	(1.45)	(1.44)	(2.24)	(1.88)	(1.74)
0.77	197.57	211.38	124.84	139.51	**87.04**	93.87	108.14
	(2.25)	(2.27)	(1.29)	(1.42)	(1.25)	(1.24)	(1.30)
0.78	136.09	160.37	77.22	88.70	**49.04**	52.58	61.88
	(2.02)	(2.33)	(0.80)	(0.98)	(0.57)	(0.60)	(0.69)
0.79	85.12	109.38	49.84	56.51	**32.06**	33.04	37.93
	(1.36)	(1.89)	(0.45)	(0.58)	(0.30)	(0.31)	(0.36)
0.80	52.85	70.26	34.86	37.87	23.31	**23.02**	25.28
	(0.77)	(1.24)	(0.26)	(0.33)	(0.18)	(0.18)	(0.20)

Table 3.4: The simulated unconditional OC ARLs along with their corresponding standard errors (in parentheses) for the P-CUSUM, L-CUSUM charts and our proposed control chart when the IC distribution is NB$(20, 0.75)$.

	P-CUSUM		L-CUSUM		Proposed		
p	$k_P = 0.3$	$k_P = 0.5$	$k_L = 0.3$	$k_L = 0.5$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$
0.70	36.17	42.31	30.77	32.89	21.02	**20.40**	21.90
	(0.44)	(0.61)	(0.22)	(0.29)	(0.16)	(0.15)	(0.17)
0.71	56.95	67.03	43.09	48.84	**28.52**	28.77	32.37
	(0.82)	(1.07)	(0.38)	(0.51)	(0.25)	(0.25)	(0.30)
0.72	96.09	109.71	66.86	78.29	**42.98**	45.24	52.86
	(1.49)	(1.79)	(0.70)	(0.90)	(0.47)	(0.49)	(0.58)
0.73	161.85	173.88	112.66	129.21	**76.37**	81.91	95.36
	(2.26)	(2.41)	(1.25)	(1.43)	(1.07)	(1.08)	(1.19)
0.74	240.08	242.95	182.89	195.34	**158.78**	160.81	173.20
	(2.33)	(2.30)	(1.56)	(1.56)	(2.20)	(1.88)	(1.78)
0.76	241.90	242.35	179.77	193.36	**159.61**	161.35	177.15
	(2.52)	(2.72)	(1.58)	(1.54)	(2.23)	(1.88)	(1.79)
0.77	168.93	176.38	109.29	127.07	**75.96**	81.17	97.33
	(2.43)	(2.63)	(1.22)	(1.40)	(1.06)	(1.05)	(1.18)
0.78	103.81	114.65	63.99	76.10	**42.44**	44.17	53.06
	(1.64)	(1.87)	(0.67)	(0.87)	(0.45)	(0.46)	(0.57)
0.79	61.46	71.42	40.77	46.71	28.04	**27.78**	31.86
	(0.91)	(1.13)	(0.35)	(0.47)	(0.24)	(0.23)	(0.28)
0.80	37.82	44.70	28.88	30.91	20.62	**19.52**	21.12
	(0.47)	(0.63)	(0.20)	(0.26)	(0.15)	(0.14)	(0.15)

Table 3.5: The simulated unconditional OC ARLs along with their corresponding standard errors (in parentheses) for the P-CUSUM, L-CUSUM charts and our proposed control chart when the IC distribution is DU(10).

P-CUSUM		L-CUSUM		Proposed		
$k_P = 0.3$	$k_P = 0.5$	$k_L = 0.3$	$k_L = 0.5$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$
53.28	69.41	30.84	34.57	22.11	**21.60**	24.44
(0.77)	(1.14)	(0.21)	(0.31)	(0.16)	(0.16)	(0.20)

Table 3.6: The simulated unconditional OC ARLs along with their corresponding standard errors (in parentheses) for the P-CUSUM, L-CUSUM charts and our proposed control chart when the IC distribution is GP$(5, 0.25)$.

	P-CUSUM		L-CUSUM		Proposed		
δ	$k_P = 0.3$	$k_P = 0.5$	$k_L = 0.3$	$k_L = 0.5$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$
-1.0	9.47	9.16	13.53	11.19	9.31	8.08	**7.42**
	(0.07)	(0.07)	(0.05)	(0.05)	(0.05)	(0.04)	(0.03)
-0.8	14.45	14.85	17.42	15.68	12.56	11.27	**10.91**
	(0.12)	(0.14)	(0.08)	(0.08)	(0.08)	(0.06)	(0.06)
-0.6	14.45	14.84	17.42	15.67	12.55	11.26	**10.91**
	(0.12)	(0.14)	(0.08)	(0.08)	(0.08)	(0.06)	(0.06)
-0.4	29.65	33.25	28.30	29.06	20.43	**19.53**	20.56
	(0.34)	(0.47)	(0.19)	(0.23)	(0.15)	(0.14)	(0.15)
-0.2	105.42	116.74	77.86	88.33	**53.00**	55.68	63.21
	(1.71)	(2.01)	(0.84)	(1.01)	(0.64)	(0.64)	(0.69)
0.2	265.74	257.83	**221.62**	225.51	241.06	227.57	230.06
	(1.70)	(1.53)	(1.15)	(1.16)	(1.62)	(1.20)	(1.25)
0.4	137.13	148.93	76.18	93.12	**51.79**	56.71	71.75
	(1.96)	(2.08)	(0.80)	(1.09)	(0.63)	(0.68)	(0.90)
0.6	40.89	52.53	26.84	29.25	19.41	**18.65**	20.62
	(0.54)	(0.82)	(0.17)	(0.25)	(0.14)	(0.13)	(0.16)
0.8	40.87	52.47	26.85	29.26	19.40	**18.65**	20.61
	(0.54)	(0.81)	(0.17)	(0.25)	(0.14)	(0.13)	(0.16)
1.0	17.56	19.95	17.01	15.33	11.92	10.66	**10.32**
	(0.15)	(0.22)	(0.07)	(0.08)	(0.07)	(0.06)	(0.05)

Fire Log to record criminal incidents and fires within residential housing reported to the University of Florida. To demonstrate the application of our proposed control chart, we consider monitoring the daily counts of narcotics violations. Based on the data provided at `https://publicsafety.ufl.edu/clery/`, the mean daily counts of narcotics violations between 2016 to 2017 is 0.49, and it decreases to 0.30 in 2018, which indicates that there is a downward mean shift in 2018. Therefore, we use the observations from January 1st, 2016 to December 31st, 2017 as our Phase-I sample, and use the ones from January 1st, 2018 to December 31st, 2018 for Phase-II monitoring. Our proposed control chart along with the P-CUSUM and L-CUSUM charts from Wang and Qiu (2018) are applied to this data set to evaluate the efficiency of those control charts in detecting the downward mean shift.

Since our proposed control chart, the P-CUSUM and L-CUSUM charts all require choosing the boundary points $q_1, ..., q_{d-1}$. In Section 3.2.1, we recommend choosing $q_1, ..., q_{d-1}$ such that the estimated IC category probabilities are as uniform as possible. Following this recommendation, for this data set, we set $d = 3$ and choose the boundary points to be $\{1, 2\}$. Similar to our simulation studies reported in Section 3.3, the control limits for our proposed control chart, the P-CUSUM and L-CUSUM charts are all determined by the procedure from Section 3.2.2 so that their conditional IC ARLs satisfy (3.4) with $\text{ARL}_0 = 200$ and $\alpha = 0.1$.

Figure 3.2 shows the P-CUSUM chart, the L-CUSUM chart and our proposed control chart when they are applied to monitor the narcotics violation data in 2018. In each plot, the curve represents the charting statistic of one given control chart and the horizontal dashed line denotes the given control chart's control limit. In all the plots, when

the curve is below the corresponding horizontal line, it implies that the underlying process is in control. When the curve crosses the corresponding horizontal line, the given control chart signals an alarm, indicating that the underlying process has changed. As we can see from Figure 3.2, when $k_P = 0.3$ and $k_L = 0.5$, the P-CUSUM and L-CUSUM charts fail to detect the mean shift. In contrast, our proposed chart with different choices of λ can all successfully detect the mean shift.

Among the control charts that can detect the mean shift, the time when the alarm first goes off (referred to as the first alarm time) varies. The earlier the first alarm time, the more sensitive the corresponding control chart is to the change. Table 3.7 shows the first alarm time for all the control charts. As we can see from Table 3.7, our proposed control chart detects the mean shift much faster than the P-CUSUM and L-CUSUM charts. This is consistent with the simulation results reported in Section 3.3 and demonstrates once again the superiority of our proposed control chart over the P-CUSUM and L-CUSUM charts.

Table 3.7: The first alarm time of different control charts for monitoring the narcotics violation data

P-CUSUM		L-CUSUM		Proposed		
$k_P = 0.3$	$k_P = 0.5$	$k_L = 0.3$	$k_L = 0.5$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$
—	278th	120th	—	69th	**67th**	**67th**

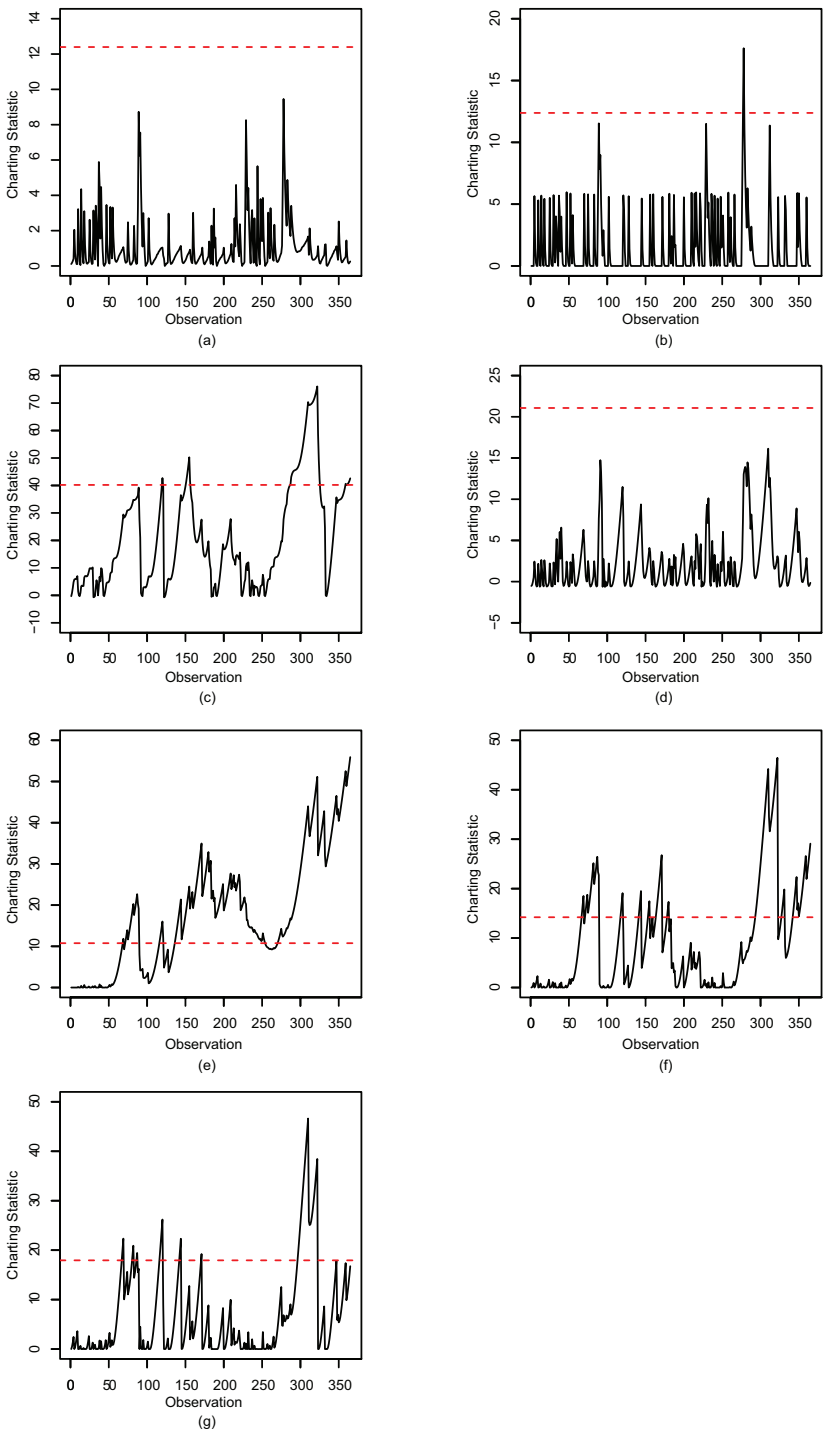

Figure 3.2: Different control charts for monitoring the narcotics violation data. The P-CUSUM chart with (a) $k_P = 0.3$, (b) $k_P = 0.5$; the L-CUSUM chart with (c) $k_L = 0.3$, (d) $k_L = 0.5$; the proposed chart with (e) $\lambda = 0.05$, (f) $\lambda = 0.1$, (g) $\lambda = 0.2$. The horizontal dashed line in each plot denotes the control limit.

Chapter 4

Concluding Remarks

4.1 Concluding Remarks for the Nonparametric Method of Combining Dependent Tests Based on Data Depth

In Chapter 2, we propose a novel nonparametric combining method for dependent tests. Through the use of data depth, our proposed method can automatically incorporate the underlying dependency among the partial tests. When all the partial tests are two-sided, any data depth can be used to combine the tests in our proposed method. Similar to other applications of data depth, when different data depths are used, different results may arise. In our case, the resulting global tests using different data depths may have different performance. In general, we recommend using the data depth that can follow closely the true geometry of the underlying distribution, since it can better incorporate the dependency among the partial tests. In our online supplementary materials, we show the depth contours based on different data depths for the same bivariate samples used in Figure 1. As we can

see there, the depth contours generated from the halfspace depth, simplicial depth, zonoid depth and onion depth can reflect better the underlying probabilistic geometry of the data than the other data depths. For other considerations when deciding which data depth to use, including their computational feasibility, we refer to Mosler and Mozharovskyi (2021) for a thorough discussion on this topic.

When some of the partial tests are one-sided, in principle our proposed combining method can also use any data depth instead of our modified halfspace depth. However, as pointed out in Section 2.2.3, if we combine those partial tests using any regular data depth, it implies that we use two-sided tests for the partial tests that are supposed to be one-sided. This will lead to some efficiency loss. To demonstrate this point, we also carry out all the simulations in Section 2.4 and real data analysis in Section 2.5 using the halfspace depth, simplicial depth, zonoid depth and onion depth. The results using those regular data depths are reported in our online supplementary materials. As we can see there, our proposed combining method can still control the Type-I error rates at the nominal level if the regular data depth is used. However, the powers of those tests are mostly lower than the powers of our test based on the modified halfspace depth. This is because one of the partial tests is based on T_{loc}, which is one-sided.

In comparison with the existing combining methods, our proposed method has higher computational cost due to the computation of data depth. However, in the last few years, efficient algorithms have been developed for computing exact and approximate values of many data depths and several software packages which implement those algorithms have been made available to practitioners (see Mosler and Mozharovskyi (2021) for a survey

of those existing software packages). Although we only deal with two partial tests in the application of our combining method in Section 2.3, with today's computing power and efficient software packages, it is still computationally feasible to implement our method when the number of partial tests gets larger.

4.2 Concluding Remarks for the Nonparametric Control Chart for Count Data

In this project, we develop a nonparametric control chart for detecting mean shifts for count data. It can be used to monitor count data generated from any distributions. By adopting the bootstrap procedure from Gandy and Kvaløy (2013) to determine its control limit, our proposed control chart can guarantee a desired performance of the conditional IC ARL. Our simulation studies and real data application show that the proposed control chart significantly outperforms the two existing nonparametric control charts, the P-CUSUM and L-CUSUM charts, for detecting various types of mean shifts. All these properties make our proposed control chart a flexible and efficient monitoring tool for count data.

Our proposed control chart, the P-CUSUM and L-CUSUM charts all use the same idea of data categorization. However, as described in Section 3.2.1, our proposed control chart is capable of incorporating the ordering information of the data into its charting statistic. In contrast, the P-CUSUM and L-CUSUM charts are both directly based on the categorized data and their multinomial distributions. As a result, both of them fail to make use of the ordering information of the data. Therefore, the results from our simulation

studies and real data application show the importance of preserving the ordering information of the data when designing nonparametric control charts through data categorization.

Bibliography

Agresti, A. (2002). *Categorical data analysis (2nd ed.)*. John Wiley & Sons.

Barnet, V. (1976). The ordering of multivariate data (with discussion). *Journal of the Royal Statistics Society, Series A*, 139:318–354.

Capizzi, G. and Masarotto, G. (2020). Guaranteed in-control control chart performance with cautious parameter learning. *Journal of Quality Technology*, 52:385–403.

Cascos, I. and López-Díaz, M. (2016). On the uniform consistency of the zonoid depth. *Journal of Multivariate Analysis*, 143:394–397.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91:862–872.

Chen, H., Chen, X., and Su, Y. (2018). A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 113:1146–1155.

Chen, H. and Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 112:397–409.

Chen, H. and Zhang, N. R. (2013). Graph-based tests for two-sample comparisons of categorical data. *Statistica Sinica*, 23:1479–1503.

Cheng, S.-S. and Yu, F.-J. (2013). A CUSUM control chart to monitor wafer quality. *International Journal of Industrial and Manufacturing Engineering*, 7:1183–1188.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45:223–256.

Cox, T. F. and Cox, M. A. A. (2001). *Multidimensional Scaling (2nd ed.)*. Boca Raton: Chapman and Hall/CRC.

Cuesta-Albertos, J. A. and Nieto-Reyes, A. (2008). The random tukey depth. *Computational Statistics & Data Analysis*, 52:4979–4988.

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Ph.D Thesis, Harvard University.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20:1803–1827.

Dümbgen, L. (1992). Limit theorems for the simplicial depth. *Statistics & Probability Letters*, 14:119–128.

Dyckerhoff, R. and Mozharovskyi, P. (2016). Exact computation of the halfspace depth. *Computational Statistics & Data Analysis*, 98:19–30.

Eagle, N., Pentland, A., and Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106:15274–15278.

Eddy, W. F. (1981). Graphics for the multivariate two-sample problem: Comment. *Journal of the American Statistical Association*, 76:287–289.

Ernst, M. D. (2004). Permutation methods: a basis for exact inference. *Statistical Science*, 19:676–685.

Famoye, F. (1994). Statistical control charts for shifted generalized Poisson distribution. *Journal of the Italian Statistical Society*, 3:339–354.

Fang, K. T., Kotz, S., and Ng, K. W. (1990). *Symmetric multivariate and related distributions*. Boca Raton: Chapman and Hall/CRC.

Fisher, R. A. (1932). *Statistical methods for research workers (4th ed.)*. Edinburgh: Oliver and Boyd.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, 7:697–717.

Gan, F. F. (1990a). Monitoring observations generated from a binomial distribution using modified exponentially weighted moving average control chart. *Journal of Statistical Computation and Simulation*, 37:45–60.

Gan, F. F. (1990b). Monitoring Poisson observations using modified exponentially weighted moving average control charts. *Communications in Statistics-Simulation and Computation*, 19:103–124.

Gan, F. F. (1993). An optimal design of CUSUM control charts for binomial counts. *Journal of Applied Statistics*, 20:445–460.

Gandy, A. and Kvaløy, J. T. (2013). Guaranteed conditional performance of control charts via bootstrap methods. *Scandinavian Journal of Statistics*, 40:647–668.

Ghosal, P. and Sen, B. (2022). Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *The Annals of Statistics*, 50:1012–1037.

Hallin, M., Del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49:1139–1165.

Hallin, M., Paindaveine, D., and Šiman, M. (2010). Multivariate quantiles and multiple-output regression quantiles: From L_1 optimization to halfspace depth. *The Annals of Statistics*, 38:635–669.

Hawkins, D. M. and Olwell, D. H. (1998). *Cumulative sum charts and charting for quality improvement*. Springer Science & Business Media.

Hodges, J. L. (1955). A bivariate sign test. *The Annals of Mathematical Statistics*, 26:523–527.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Jensen, W. A., Jones-Farmer, L. A., Champ, C. W., and Woodall, W. H. (2006). Effects of parameter estimation on control chart properties: a literature review. *Journal of Quality technology*, 38:349–364.

Jones, M. A. and Steiner, S. H. (2012). Assessing the effect of estimation error on risk-adjusted CUSUM chart performance. *International journal for quality in health care*, 24:176–181.

Kaminsky, F. C., Benneyan, J. C., Davis, R. D., and Burke, R. J. (1992). Statistical control charts based on a geometric distribution. *Journal of Quality Technology*, 24:63–69.

Khoo, M. B. (2004). A moving average control chart for monitoring the fraction non-conforming. *Quality and Reliability Engineering International*, 20:617–635.

Koshevoy, G. and Mosler, K. (1997). Zonoid trimming for multivariate distributions. *The Annals of Statistics*, 25:1998–2017.

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57:613–644.

Li, J. (2018). Asymptotic normality of interpoint distances for high-dimensional data with applications to the two-sample problem. *Biometrika*, 105:529–546.

Li, J. (2021). Nonparametric adaptive CUSUM chart for detecting arbitrary distributional changes. *Journal of Quality Technology*, 53:154–172.

Lipták, T. (1958). On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*, 3:171–197.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18:405–414.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by liu and singh). *The Annals of Statistics*, 27:783–858.

Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88:252–260.

Lucas, J. M. (1985). Counted data CUSUM's. *Technometrics*, 27:129–144.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 12:49–55.

Mallows, C. L. (1957). Non-null ranking models. I. *Biometrika*, 44:114–130.

Montgomery, D. C. (2020). *Introduction to statistical quality control*. John Wiley & Sons.

Morgenstern, D. (2001). Proof of a conjecture by walter deuber concerning the distances between points of two types in R^d. *Discrete Mathematics*, 226:347–349.

Mosler, K. and Mozharovskyi, P. (2022). Choosing among notions of multivariate depth statistics. *Statistical Science*, 37:348–368.

Nagy, S., Dyckerhoff, R., and Mozharovskyi, P. (2020). Uniform convergence rates for the approximated halfspace and projection depth. *Electronic Journal of Statistics*, 14:3939–3975.

Nagy, S., Schütt, C., and Werner, E. M. (2019). Halfspace depth and floating body. *Statistics Surveys*, 13:52–118.

Pesarin, F. (2001). *Multivariate permutation tests: with applications in biostatistics*. Wiley Chichester.

Pokotylo, O., Mozharovskyi, P., Dyckerhoff, R., Nagy, S., and Pokotylo, M. O. (2020). ddalpha: Depth-based classification and calculation of data depth. R package version 1.3.11.

Psarakis, S., Vyniou, A. K., and Castagliola, P. (2014). Some recent developments on the effects of parameter estimation on control charts. *Quality and Reliability Engineering International*, 30:1113–1129.

Qiu, P. and Li, Z. (2011). On nonparametric statistical process control of univariate processes. *Technometrics*, 53:390–405.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100:94–108.

Saleh, N. A., Zwetsloot, I. M., Mahmoud, M. A., and Woodall, W. H. (2016). CUSUM charts with controlled conditional performance under estimated parameters. *Quality Engineering*, 28:402–415.

Sellers, K. F. (2012). A generalized statistical control chart for over-or under-dispersed data. *Quality and Reliability Engineering International*, 28:59–65.

Serfling, R. (2002a). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, 56:214–232.

Serfling, R. (2002b). Generalized quantile processes based on multivariate depth functions, with applications in nonparametric multivariate analysis. *Journal of Multivariate Analysis*, 83:232–247.

Serfling, R. (2010). Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardisation. *Journal of Nonparametric Statistics*, 22:915–936.

Sheaffer, R. and Leavenworth, R. (1976). The negative binomial model for counts in units of varying size. *Journal of Quality Technology*, 8:158–163.

Sparks, R. S. (2000). Cusum charts for signalling varying location shifts. *Journal of Quality Technology*, 32:157–171.

Stahel, W. A. (1981). Robust schaetzungen: Infinitesmale optimalitaet und schaetzungen von kovarianzmatrizen (Robust estimation: Infinitesimal optimality and covariance matrix estimators). Ph.D. Thesis, ETH, Zurich.

Tippett, L. H. C. (1931). *The Methods of Statistics. An introduction mainly for workers in the biological sciences.* London: Williams and Norgate.

Tukey, J. W. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians, Vancouver*, 2:523–531.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97:1423–1426.

Wang, Z. and Qiu, P. (2018). Count data monitoring: Parametric or nonparametric? *Quality and Reliability Engineering International*, 34:1763–1774.

White, C. H. and Keats, J. B. (1996). ARLs and higher-order run-length moments for the Poisson CUSUM. *Journal of Quality Technology*, 28:363–369.

Woodall, W. H. (1997). Control charts based on attribute data: bibliography and review. *Journal of Quality Technology*, 29:172–183.

Wu, Z., Jiao, J., and Liu, Y. (2008). A binomial CUSUM chart for detecting large shifts in fraction nonconforming. *Journal of Applied Statistics*, 35:1267–1276.

Zhang, J. and Chen, H. (2022). Graph-based two-sample tests for data with repeated observations. *Statistica Sinica*, 32:391–415.

Zuo, Y. (2003). Projection-based depth functions and associated medians. *The Annals of Statistics*, 31:1460–1490.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28:461–482.

Appendix A

Proofs

Proof of Proposition 1: As $n \to \infty$, it is easy to see that $\boldsymbol{T} = (T_1, ..., T_k)'$ asymptotically follows $N_k(\boldsymbol{0}, \boldsymbol{R})$ under H_0, where $\boldsymbol{R}$ is the correlation matrix corresponding to $\boldsymbol{\Sigma}$. Since $N_k(\boldsymbol{0}, \boldsymbol{R})$ is an elliptical distribution, if the data depth $D_G(\boldsymbol{t})$ is affine invariant, $D_G(\boldsymbol{t}) = h(\boldsymbol{t}'\boldsymbol{R}^{-1}\boldsymbol{t})$ for some nonincreasing function h (see, Mosler and Mozharovskyi 2021). Then

$$D_G(\boldsymbol{T}) \leqslant D_G(\boldsymbol{T}^{obs}) \quad \Longleftrightarrow \quad \boldsymbol{T}'\boldsymbol{R}^{-1}\boldsymbol{T} \geqslant (\boldsymbol{T}^{obs})'\boldsymbol{R}^{-1}\boldsymbol{T}^{obs}.$$

Let $\overline{\boldsymbol{X}}$ and $\boldsymbol{S}$ be the sample mean and sample covariance matrix of $\boldsymbol{X}_1$,...,$\boldsymbol{X}_n$. The Hotelling's T^2 test statistic is defined as $T^2_{Hotelling} = n(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{S}^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)$. It is not difficult to see that, as $n \to \infty$, $\boldsymbol{T}'\boldsymbol{R}^{-1}\boldsymbol{T}$ and $T^2_{Hotelling}$ are equal. Therefore, the p-value calculated from (2.2) is equal to the p-value calculated from the Hotelling's T^2 test as $n \to \infty$. This completes the proof.

Proof of Proposition 2: Since the data depth $D(\cdot)$ used in (2.3) satisfies (2.4) and (2.5), the conditions of Lemma 6.1 from Liu and Singh (1993) are satisfied. Based on

the proof of Lemma 6.1 in Liu and Singh (1993), we can conclude that

$$\widehat{p}_{n,B} \xrightarrow{a.s.} \widehat{p}_n = P(D_{G_n^*}(\boldsymbol{T}_n^*) \leqslant D_{G_n^*}(\boldsymbol{T}^{obs})), \quad \text{as } B \to \infty,$$

where $\boldsymbol{T}_n^*$ is a random vector drawn from G_n^*, and

$$\widehat{p}_n \xrightarrow{a.s.} p = P(D_G(\boldsymbol{T}) \leqslant D_G(\boldsymbol{T}^{obs})), \quad \text{as } n \to \infty.$$

This completes the proof.

Proof of Theorem 1: For any $\boldsymbol{x}, \boldsymbol{u} \in \mathbb{R}^d$, let $H_{\boldsymbol{u},\boldsymbol{x}}$ denote the half-space $\{\boldsymbol{y} \in \mathbb{R}^d : \boldsymbol{u}'\boldsymbol{y} \geqslant \boldsymbol{u}'\boldsymbol{x}\}$. For any measurable set $S \subset \mathbb{R}^d$, define $F(S) = P(\boldsymbol{X} \in S\}$ and $F_n(S) = \#\{i : \boldsymbol{X}_i \in S\}$. Note that

$$\sup_{\substack{\boldsymbol{x},\boldsymbol{u}\in\mathbb{R}^d \\ u_i \geqslant 0, i=1,\ldots,k_0}} |F_n(H_{\boldsymbol{u},\boldsymbol{x}}) - F(H_{\boldsymbol{u},\boldsymbol{x}})| \leqslant \sup_{\boldsymbol{x},\boldsymbol{u}\in\mathbb{R}^d} |F_n(H_{\boldsymbol{u},\boldsymbol{x}}) - F(H_{\boldsymbol{u},\boldsymbol{x}})|.$$

According to the result in Donoho and Gasko (1992) (p. 1816), for any distribution F,

$$\sup_{\boldsymbol{x},\boldsymbol{u}\in\mathbb{R}^d} |F_n(H_{\boldsymbol{u},\boldsymbol{x}}) - F(H_{\boldsymbol{u},\boldsymbol{x}})| \xrightarrow{a.s.} 0, \quad \text{as } n \to \infty.$$

Therefore,

$$\sup_{\substack{\boldsymbol{x},\boldsymbol{u}\in\mathbb{R}^d \\ u_i \geqslant 0, i=1,\ldots,k_0}} |F_n(H_{\boldsymbol{u},\boldsymbol{x}}) - F(H_{\boldsymbol{u},\boldsymbol{x}})| \xrightarrow{a.s.} 0, \quad \text{as } n \to \infty.$$

Based on the definitions of $\widetilde{D}_F^{(k_0)}(\boldsymbol{x})$ and $\widetilde{D}_{F_n}^{(k_0)}(\boldsymbol{x})$, we have the following inequality:

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widetilde{D}_{F_n}^{(k_0)}(\boldsymbol{x}) - \widetilde{D}_F^{(k_0)}(\boldsymbol{x})| \leqslant \sup_{\substack{\boldsymbol{x},\boldsymbol{u}\in\mathbb{R}^d \\ u_i\geqslant 0, i=1,\ldots,k_0}} |F_n(H_{\boldsymbol{u},\boldsymbol{x}}) - F(H_{\boldsymbol{u},\boldsymbol{x}})|.$$

This implies that, for any distribution F,

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widetilde{D}_{F_n}^{(k_0)}(\boldsymbol{x}) - \widetilde{D}_F^{(k_0)}(\boldsymbol{x})| \xrightarrow{a.s.} 0, \quad \text{as } n \to \infty.$$

Define $\mathcal{H} = \{H_{\boldsymbol{u},\boldsymbol{x}} : \boldsymbol{x}, \boldsymbol{u} \in \mathbb{R}^d, u_i \geqslant 0, i = 1, ..., k_0\}$. Note that $\mathcal{H} \subset \mathcal{A}$ for $\mathcal{A}$ being a collection of sets defined in the proof of Corollary 2 in Dümbgen (1992). Based on the proof there, if F is absolutely continuous, for any sequence of distributions $\{F_\nu^*\}_{\nu=1}^\infty$ weakly convergent to F,

$$\sup_{H_{\boldsymbol{u},\boldsymbol{x}}\in\mathcal{H}} |F_\nu^*(H_{\boldsymbol{u},\boldsymbol{x}}) - F(H_{\boldsymbol{u},\boldsymbol{x}})| \longrightarrow 0, \quad \text{as } \nu \to \infty.$$

Again based on the definitions of $\widetilde{D}_F^{(k_0)}(\boldsymbol{x})$ and $\widetilde{D}_{F_\nu^*}^{(k_0)}(\boldsymbol{x})$, we have

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widetilde{D}_{F_\nu^*}^{(k_0)}(\boldsymbol{x}) - \widetilde{D}_F^{(k_0)}(\boldsymbol{x})| \leqslant \sup_{H_{\boldsymbol{u},\boldsymbol{x}}\in\mathcal{H}} |F_\nu^*(H_{\boldsymbol{u},\boldsymbol{x}}) - F(H_{\boldsymbol{u},\boldsymbol{x}})|.$$

Therefore,

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widetilde{D}_{F_\nu^*}^{(k_0)}(\boldsymbol{x}) - \widetilde{D}_F^{(k_0)}(\boldsymbol{x})| \longrightarrow 0, \quad \text{as } \nu \to \infty.$$

Proof of Proposition 3: If G is absolutely continuous, based on Theorem 1, our modified halfspace depth satisfies (2.4) and (2.5). If the bootstrap distribution G_n^* converges weakly to G, following Proposition 2, $\widehat{p}_{n,B} \xrightarrow{a.s.} p$, as $n \to \infty$ and $B \to \infty$.

Appendix B

Supplementary Materials for Combining Dependent Tests Based on Data Depth with Applications to the Two-sample Problem for Data of Arbitrary Types

B.1 Depth Contours Based on Different Data Depths

Here we show the depth contours based on the Mahalanobis depth, L_2 depth, simplicial depth, simplicial volume depth, zonoid depth, spatial depth, lens depth, onion depth and projection depth for the same bivariate samples used in Figure 2.1. For the

definitions of those data depths, see Mosler and Mozharovskyi (2022). As we can see from Figures B.1 - B.9 and Figure 2.1, the depth contours generated from the halfspace depth, simplical depth, zonoid depth and onion depth can reflect better the underlying probabilistic geometry of the data than the other depths.

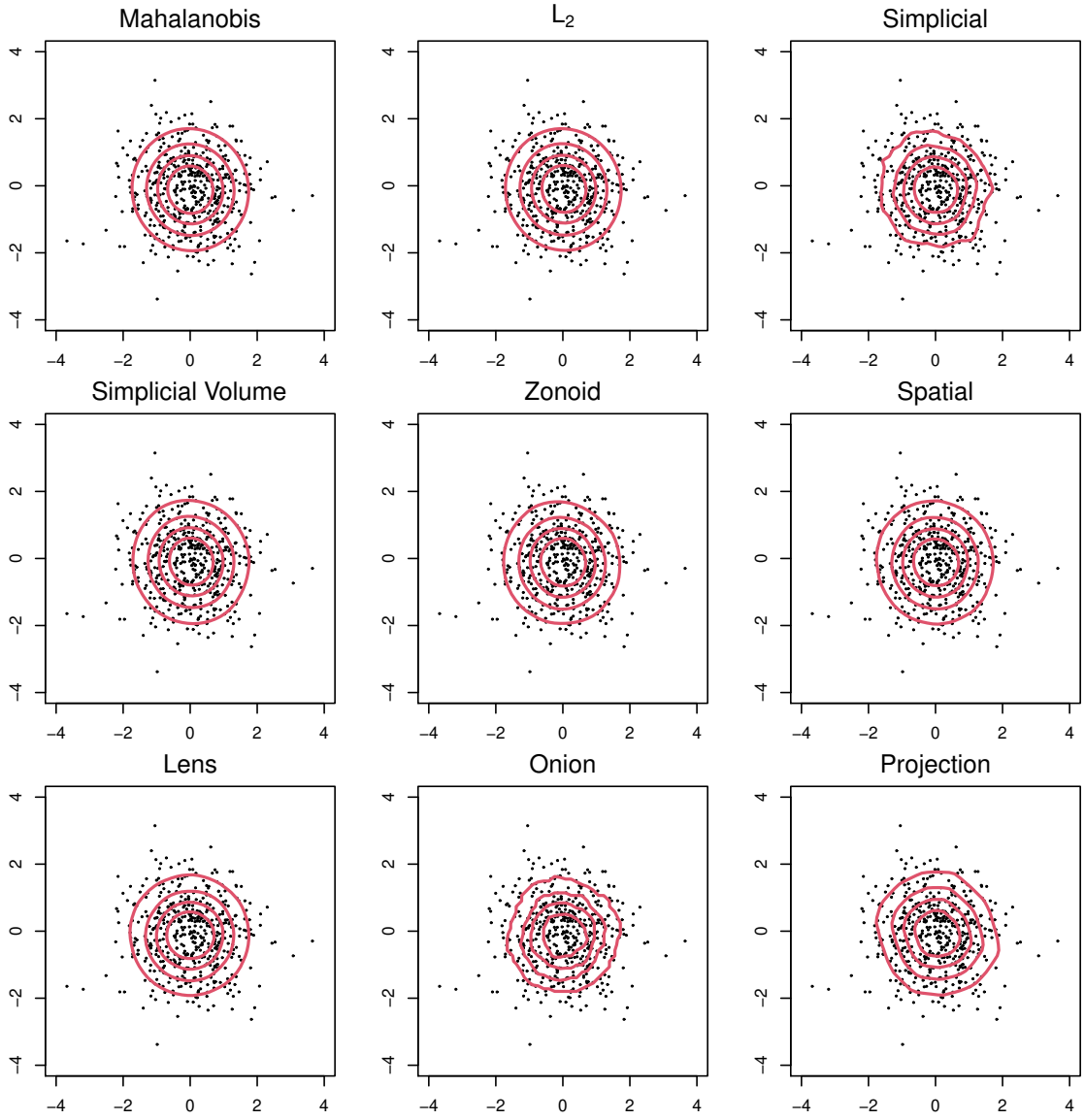

Figure B.1: Depth contours for a random sample drawn from a bivariate distribution based on nine notions of data depth. The two marginal distributions are both the normal distributions. The two variables are independent.

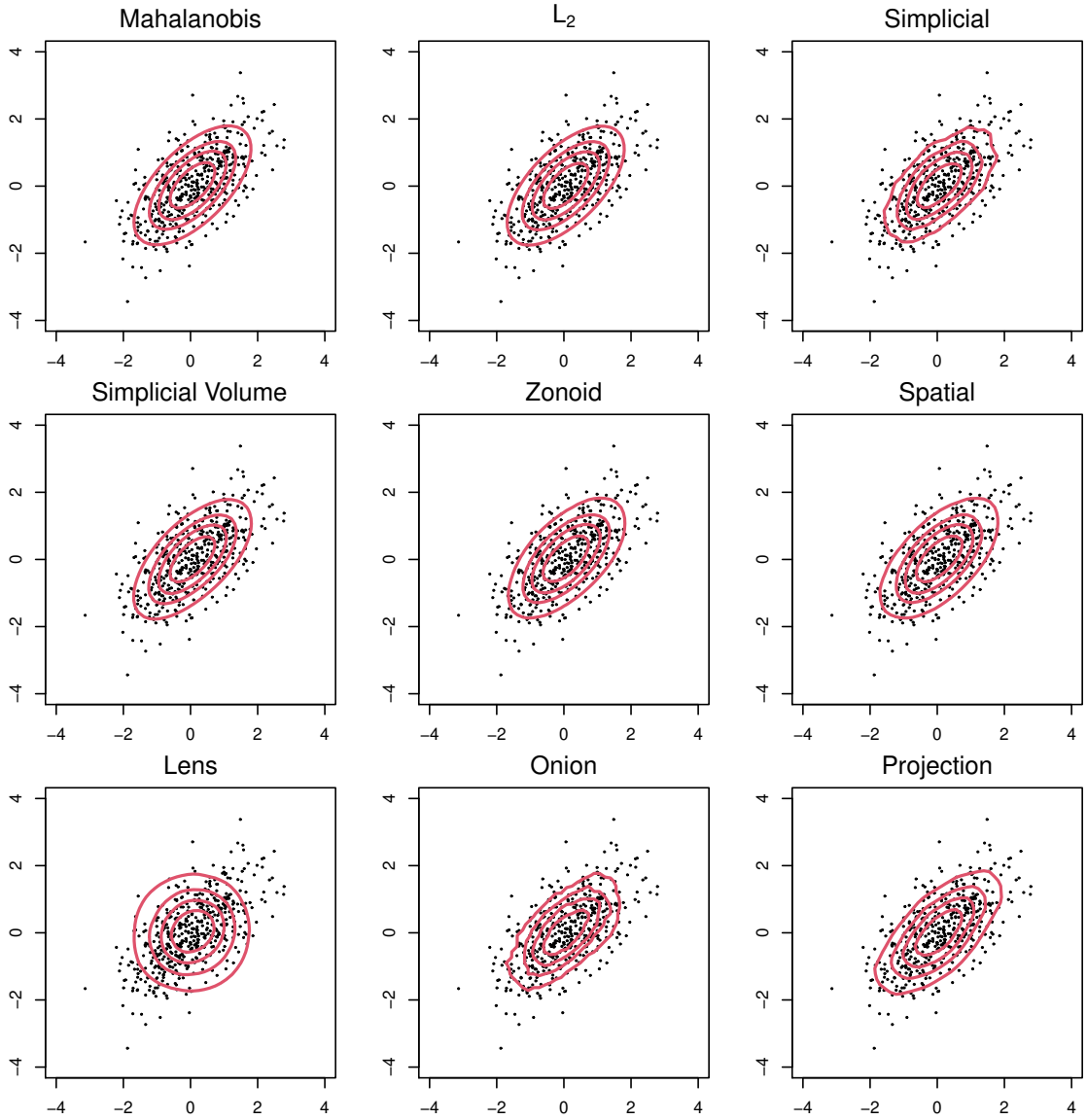

Figure B.2: Depth contours for a random sample drawn from a bivariate distribution based on nine notions of data depth. The two marginal distributions are both the normal distributions. The two variables are positively correlated.

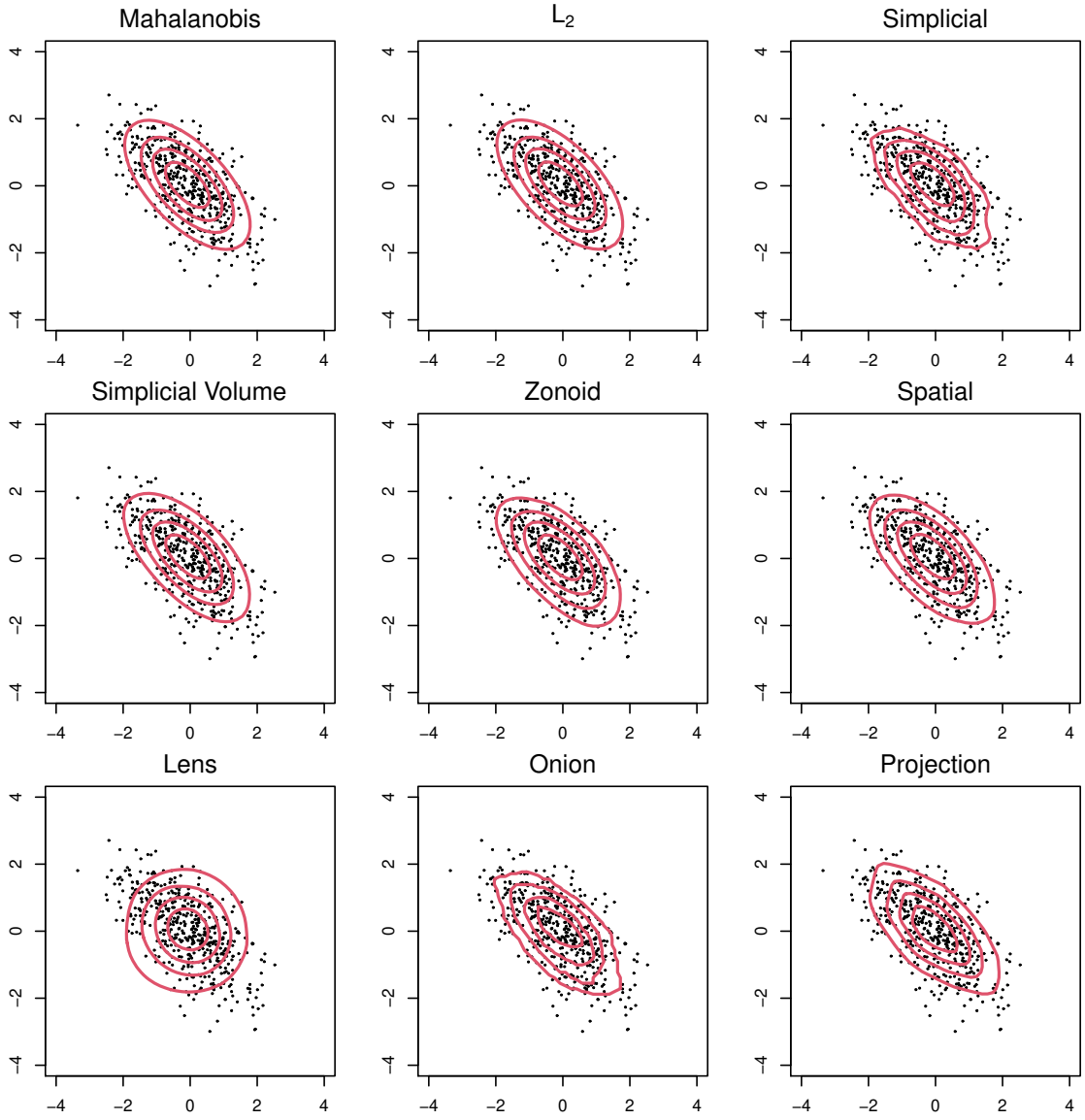

Figure B.3: Depth contours for a random sample drawn from a bivariate distribution based on nine notions of data depth. The two marginal distributions are both the normal distributions. The two variables are negatively correlated.

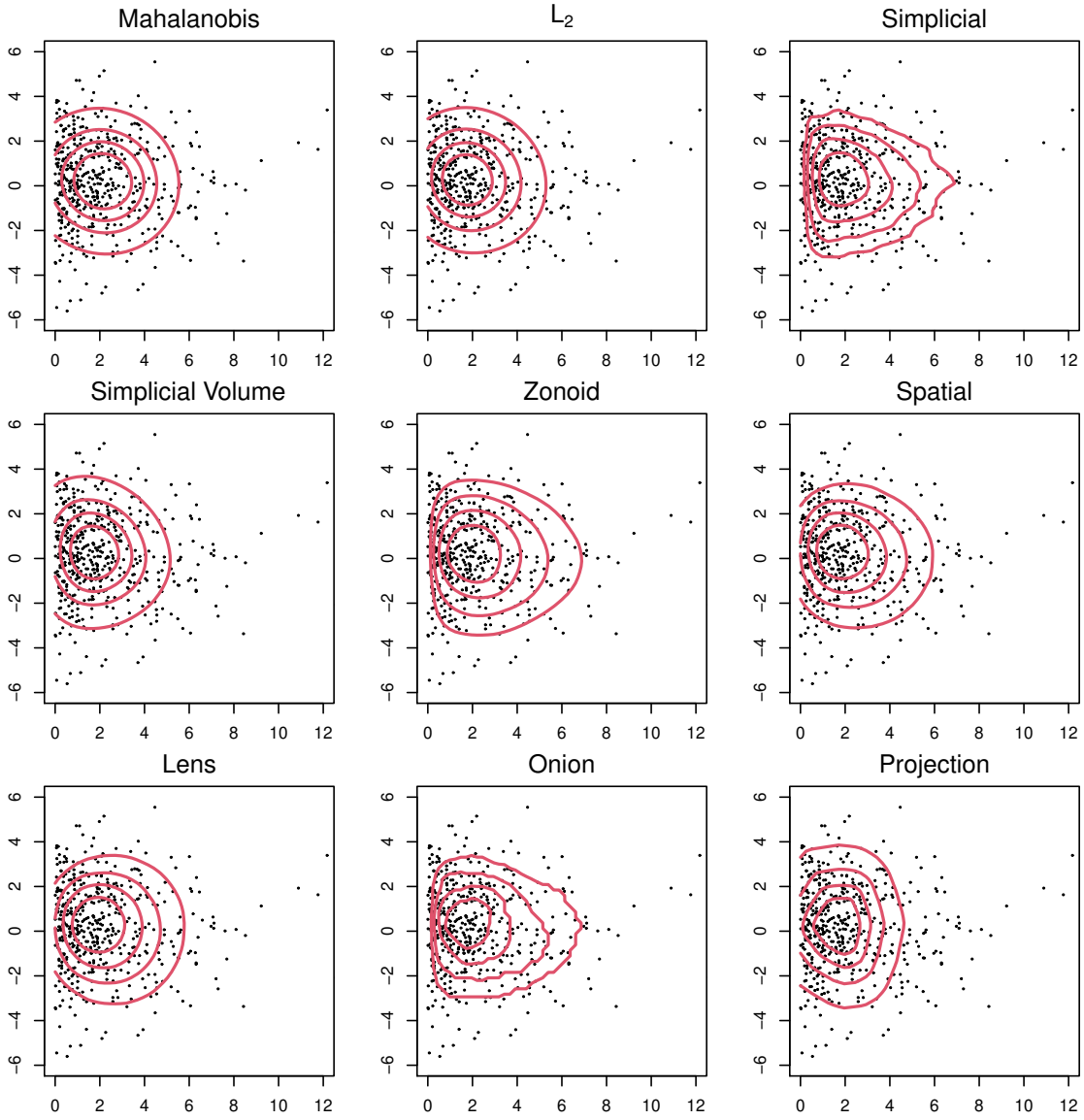

Figure B.4: Depth contours for a random sample drawn from a bivariate distribution based on nine notions of data depth. The two marginal distributions are the chisquare distribution with 2 degrees of freedom and the normal distribution. The two variables are independent.

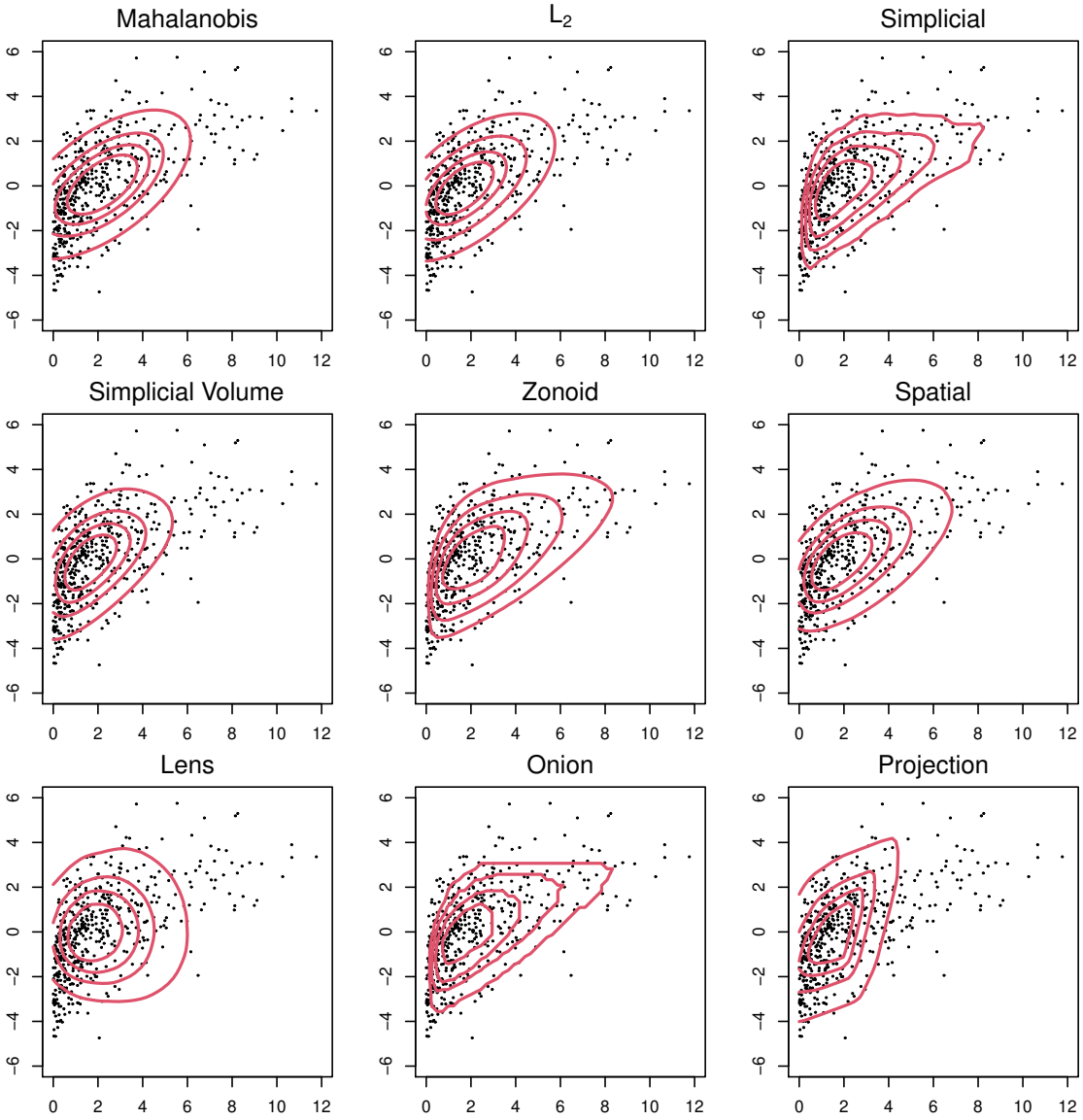

Figure B.5: Depth contours for a random sample drawn from a bivariate distribution based on nine notions of data depth. The two marginal distributions are the chisquare distribution with 2 degrees of freedom and the normal distribution. The two variables are positively correlated.

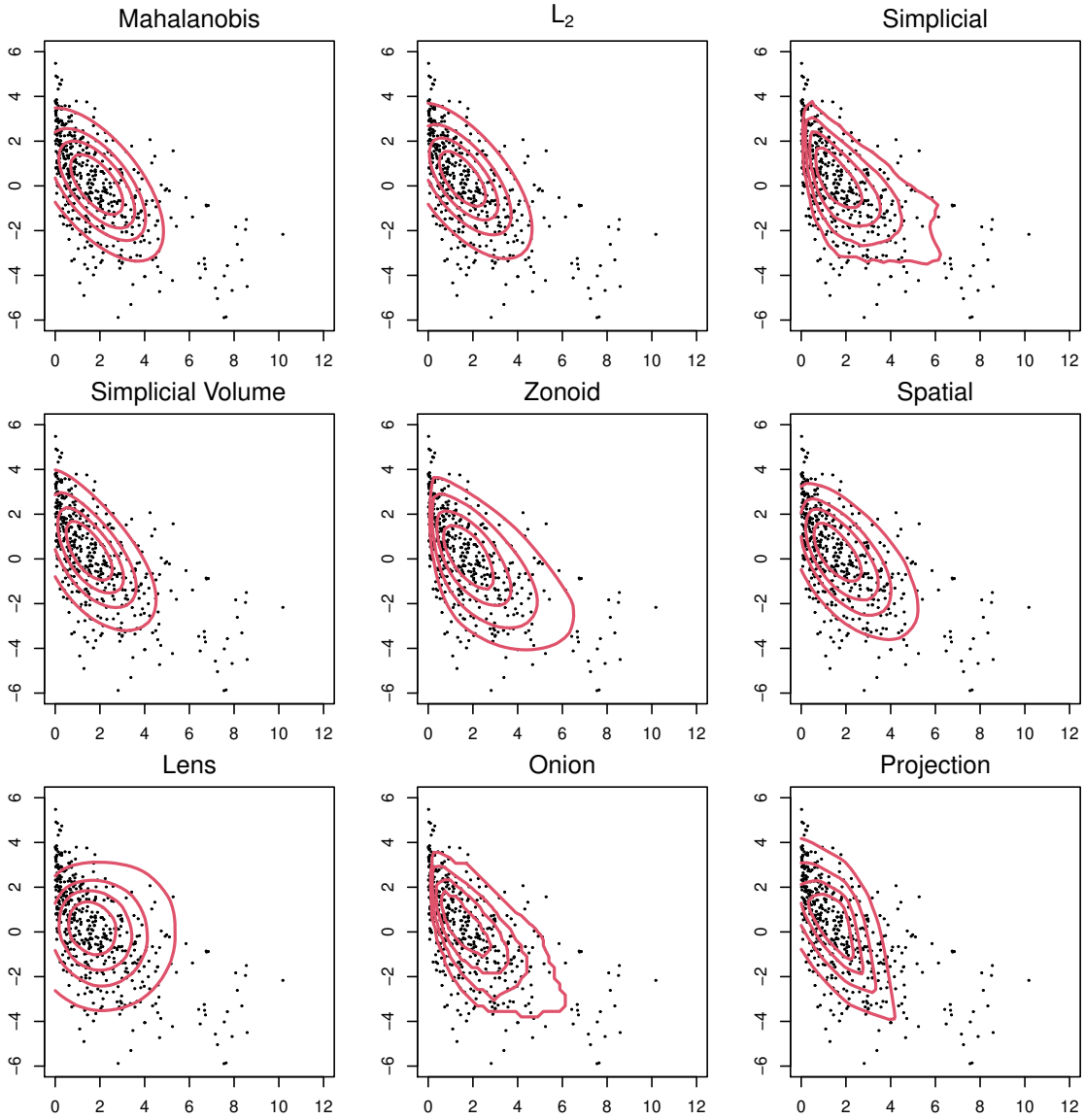

Figure B.6: Depth contours for a random sample drawn from a bivariate distribution based on nine notions of data depth. The two marginal distributions are the chisquare distribution with 2 degrees of freedom and the normal distribution. The two variables are negatively correlated.

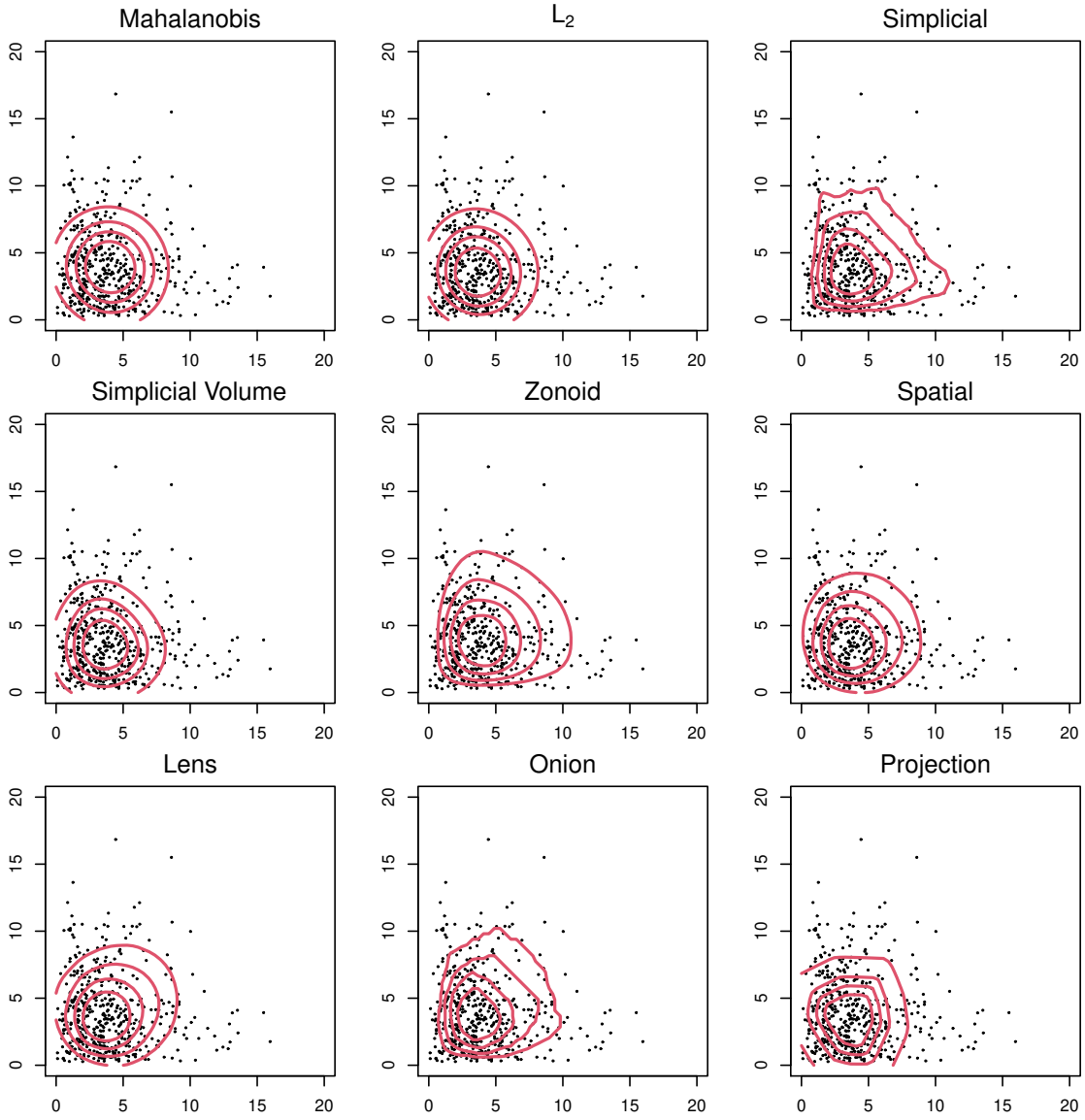

Figure B.7: Depth contours for a random sample drawn from a bivariate distribution based on nine notions of data depth. The two marginal distributions are both the chisquare distributions with 4 degrees of freedom. The two variables are independent.

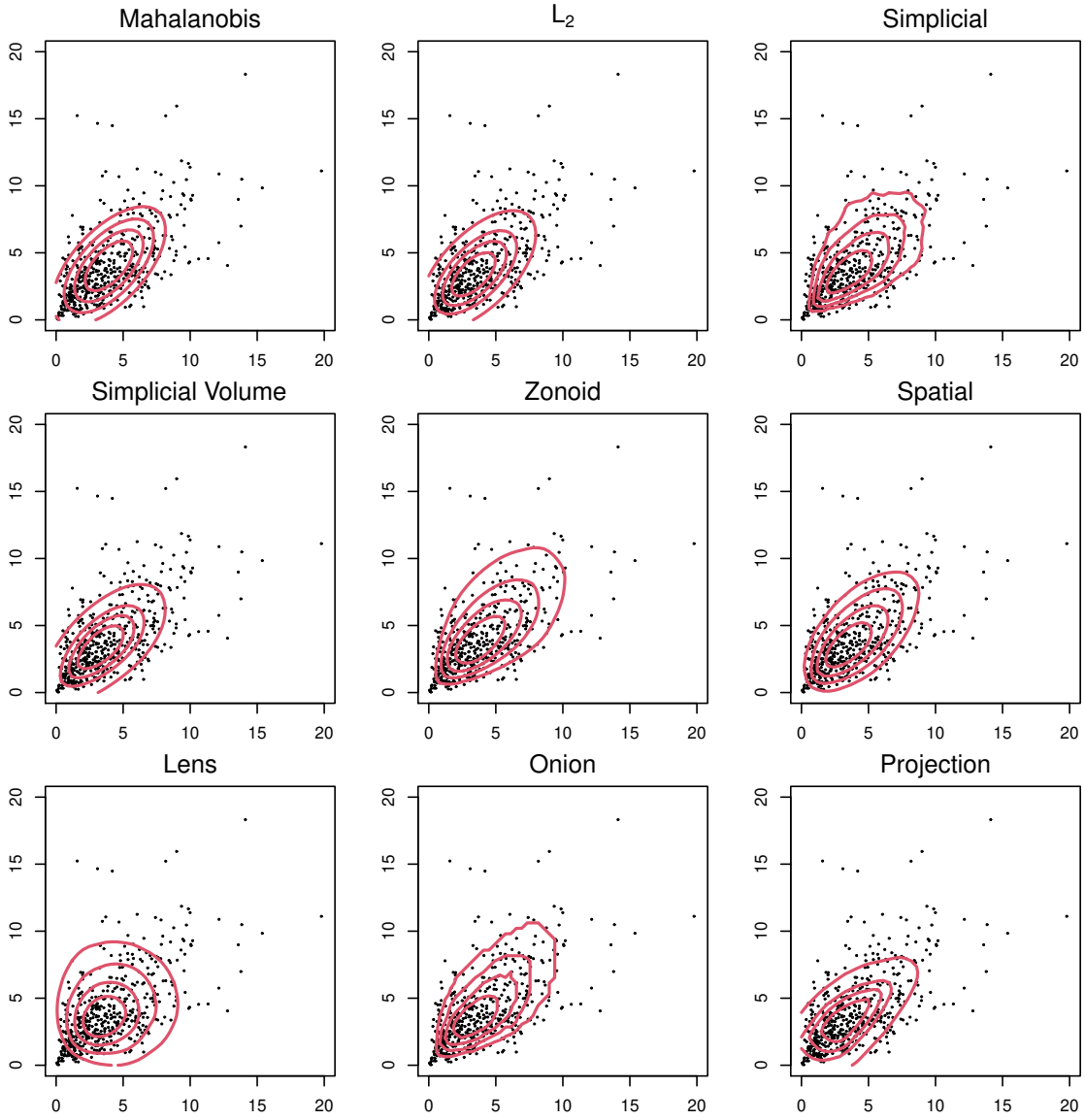

Figure B.8: Depth contours for a random sample drawn from a bivariate distribution based on nine notions of data depth. The two marginal distributions are both the chisquare distributions with 4 degrees of freedom. The two variables are positively correlated.

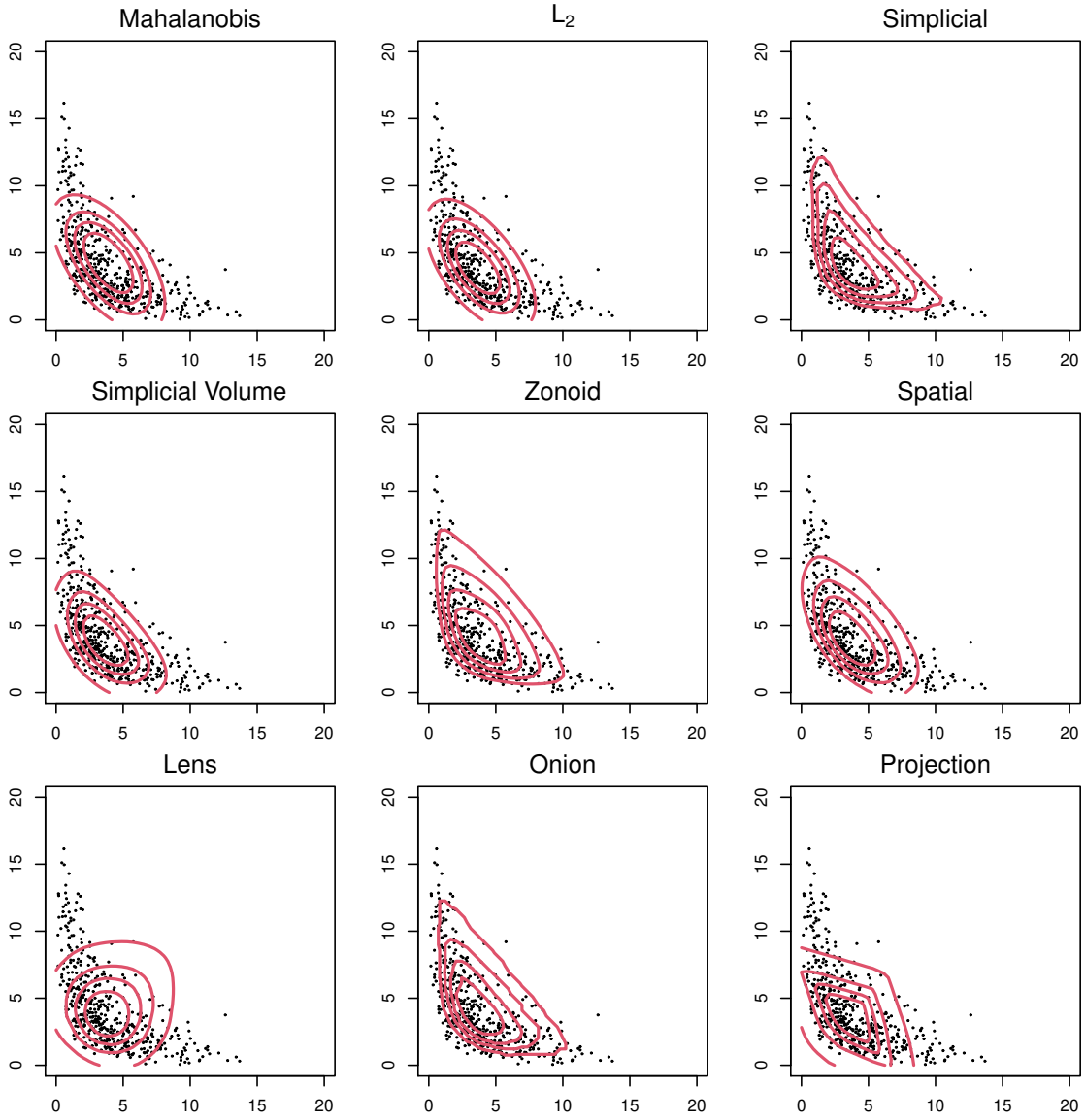

Figure B.9: Depth contours for a random sample drawn from a bivariate distribution based on nine notions of data depth. The two marginal distributions are both the chisquare distributions with 4 degrees of freedom. The two variables are negatively correlated.

B.2 Additional Simulation Results

Here we report the additional simulation results when we carry out all the simulations in Section 2.4 and real data analysis in Section 2.5 using the simplicial depth, zonoid depth, onion depth and halfspace depth. As we can see from Tables B.1 -B.9, our proposed combining method can still control the Type-I error rates at the nominal level if the regular depth is used. However, the powers of those tests are mostly lower than the powers of our test based on the modified halfspace depth.

Table B.1: The simulated Type-I error rates with $\alpha = 0.05$ for data from continuous distributions.

The simulated Type-I errors					
d	2	10	50	100	200
Chen and Friedman's test	.051	.052	.048	.046	.044
Simplicial depth	.048	.061	.057	.040	.050
Zonoid depth	.044	.053	.054	.041	.053
Onion depth	.035	.046	.040	.036	.046
Halfspace depth	.042	.046	.049	.038	.050
Proposed	.054	.051	.049	.044	.047

Table B.2: The simulated powers for detecting location differences at $\alpha = 0.05$ for data from continuous distributions.

Location differences ($d = 2$)					
Δ	.40	.55	.70	.85	1.00
Chen and Friedman's test	.127	.175	.333	.465	.661
Simplicial depth	.166	.302	.568	.755	.902
Zonoid depth	.166	.328	.607	.786	.924
Onion depth	.148	.309	.573	.766	.917
Halfspace depth	.161	.323	.599	.781	.922
Proposed	**.252**	**.429**	**.689**	**.859**	**.962**
Location differences ($d = 10$)					
Δ	.60	.75	.90	1.05	1.20
Chen and Friedman's test	.119	.161	.276	.431	.584
Simplicial depth	.220	.345	.591	.770	.905
Zonoid depth	.226	.369	.614	.789	.924
Onion depth	.214	.348	.592	.769	.911
Halfspace depth	.223	.361	.609	.785	.922
Proposed	**.282**	**.454**	**.677**	**.843**	**.952**
Location differences ($d = 50$)					
Δ	.90	1.05	1.20	1.35	1.50
Chen and Friedman's test	.107	.135	.241	.304	.405
Simplicial depth	.231	.368	.568	.738	.851
Zonoid depth	.247	.386	.587	.751	.853
Onion depth	.227	.351	.560	.736	.843
Halfspace depth	.240	.374	.581	.749	.850
Proposed	**.307**	**.472**	**.656**	**.811**	**.898**
Location differences ($d = 100$)					
Δ	1.10	1.25	1.40	1.55	1.70
Chen and Friedman's test	.109	.150	.189	.281	.362
Simplicial depth	.273	.393	.551	.718	.826
Zonoid depth	.292	.403	.576	.728	.836
Onion depth	.260	.372	.546	.700	.822
Halfspace depth	.276	.398	.569	.714	.834
Proposed	**.347**	**.465**	**.642**	**.785**	**.880**
Location differences ($d = 200$)					
Δ	1.30	1.50	1.70	1.90	2.10
Chen and Friedman's test	.106	.152	.208	.314	.417
Simplicial depth	.280	.444	.611	.792	.911
Zonoid depth	.297	.457	.622	.814	.924
Onion depth	.266	.424	.598	.781	.920
Halfspace depth	.291	.439	.614	.800	.921
Proposed	**.363**	**.539**	**.686**	**.850**	**.947**

Table B.3: The simulated powers for detecting scale differences at $\alpha = 0.05$ for data from continuous distributions.

Scale differences ($d = 2$)					
σ	1.50	1.70	1.90	2.10	2.30
Chen and Friedman's test	.173	.272	.374	.482	.589
Simplicial depth	.234	.415	.603	.724	.831
Zonoid depth	.263	.472	**.656**	**.794**	.866
Onion depth	.238	.429	.617	.756	.834
Halfspace depth	.255	.456	.630	.774	.859
Proposed	**.274**	**.478**	.651	.793	**.879**
Scale differences ($d = 10$)					
σ	1.20	1.26	1.32	1.38	1.44
Chen and Friedman's test	.275	.409	.533	.644	.762
Simplicial depth	.298	.475	.623	.762	.868
Zonoid depth	.309	.504	.657	.788	**.889**
Onion depth	.285	.477	.633	.767	.869
Halfspace depth	.297	.493	.639	.770	.879
Proposed	**.323**	**.516**	**.671**	**.790**	.885
Scale differences ($d = 50$)					
σ	1.08	1.11	1.14	1.17	1.20
Chen and Friedman's test	.264	.433	.579	.731	.862
Simplicial depth	.298	.498	.725	.844	.937
Zonoid depth	**.310**	.508	**.748**	**.867**	.949
Onion depth	.284	.479	.721	.845	.939
Halfspace depth	.302	.492	.734	.859	.945
Proposed	.301	**.510**	.738	.858	**.952**
Scale differences ($d = 100$)					
σ	1.06	1.08	1.10	1.12	1.14
Chen and Friedman's test	.308	.439	.614	.784	.877
Simplicial depth	.318	.509	.745	.865	.952
Zonoid depth	**.327**	**.537**	.754	.878	**.966**
Onion depth	.302	.498	.725	.860	.954
Halfspace depth	.317	.516	.744	.873	.960
Proposed	.319	.518	**.755**	**.882**	.960
Scale differences ($d = 200$)					
σ	1.04	1.05	1.06	1.07	1.08
Chen and Friedman's test	.290	.376	.500	.614	.735
Simplicial depth	.307	.442	.607	.711	.837
Zonoid depth	**.324**	**.445**	**.630**	.726	**.856**
Onion depth	.295	.417	.598	.702	.834
Halfspace depth	.315	.440	.617	.711	.842
Proposed	.322	.433	.614	**.731**	.838

Table B.4: The simulated powers for detecting both location and scale differences at $\alpha = 0.05$ for data from continuous distributions.

Location and scale differences ($d = 2$)					
Δ	.50	.70	.90	1.10	1.30
Chen and Friedman's test	.148	.294	.492	.760	.925
Simplicial depth	.160	.315	.588	.808	.918
Zonoid depth	.204	.393	.640	.843	.926
Onion depth	.164	.339	.592	.822	.919
Halfspace depth	.189	.380	.631	.840	.927
Proposed	**.316**	**.513**	**.748**	**.897**	**.968**
Location and scale differences ($d = 10$)					
Δ	.70	.90	1.10	1.30	1.50
Chen and Friedman's test	.242	.388	.545	.709	.847
Simplicial depth	.228	.435	.618	.764	.901
Zonoid depth	.249	.452	.629	.782	.900
Onion depth	.235	.428	.611	.770	.888
Halfspace depth	.247	.444	.620	.779	.902
Proposed	**.347**	**.542**	**.706**	**.841**	**.941**
Location and scale differences ($d = 50$)					
Δ	.90	1.15	1.40	1.65	1.90
Chen and Friedman's test	.275	.420	.583	.704	.837
Simplicial depth	.218	.429	.566	.756	.885
Zonoid depth	.232	.432	.577	.765	.889
Onion depth	.209	.418	.565	.756	.879
Halfspace depth	.224	.426	.565	.763	.887
Proposed	**.285**	**.519**	**.651**	**.817**	**.928**
Location and scale differences ($d = 100$)					
Δ	.90	1.20	1.50	1.80	2.10
Chen and Friedman's test	**.256**	.424	.551	.711	.826
Simplicial depth	.194	.365	.541	.725	.867
Zonoid depth	.205	.367	.555	.744	.884
Onion depth	.190	.350	.527	.726	.883
Halfspace depth	.197	.362	.542	.738	.884
Proposed	**.256**	**.431**	**.615**	**.789**	**.902**
Location and scale differences ($d = 200$)					
Δ	1.20	1.50	1.80	2.10	2.40
Chen and Friedman's test	.341	.468	.609	.750	.831
Simplicial depth	.286	.417	.600	.783	.904
Zonoid depth	.297	.435	.610	.786	.912
Onion depth	.273	.409	.591	.767	.902
Halfspace depth	.282	.421	.601	.777	.905
Proposed	**.345**	**.492**	**.654**	**.830**	**.930**

Table B.5: The simulated Type-I error rates with $\alpha = 0.05$ for the preference ranking data.

	Setting 1		Setting 2	
Sample Size	$n_1 = 50$, $n_2 = 50$	$n_1 = 25$, $n_2 = 100$	$n_1 = 50$, $n_2 = 50$	$n_1 = 25$, $n_2 = 100$
$S_{(a)}$	.054	.052	.037	.043
$M_{(a)}(1.14)$	.068	.065	.053	.048
$S_{(u)}$	.045	.050	.047	.053
$M_{(u)}(1.14)$	.058	.068	.058	.054
Simplicial depth	.050	.041	.057	.038
Zonoid depth	.051	.037	.054	.041
Onion depth	.041	.037	.054	.038
Halfspace depth	.047	.036	.053	.037
Proposed	.046	.036	.050	.041
	Setting 3		Setting 4	
Sample Size	$n_1 = 50$, $n_2 = 50$	$n_1 = 25$, $n_2 = 100$	$n_1 = 50$, $n_2 = 50$	$n_1 = 25$, $n_2 = 100$
$S_{(a)}$	.048	.057	.040	.059
$M_{(a)}(1.14)$	.053	.066	.052	.062
$S_{(u)}$	.040	.057	.053	.059
$M_{(u)}(1.14)$	.045	.066	.056	.065
Simplicial depth	.050	.047	.058	.045
Zonoid depth	.049	.051	.055	.040
Onion depth	.046	.044	.048	.036
Halfspace depth	.046	.050	.053	.038
Proposed	.053	.038	.046	.041

Table B.6: The simulated powers at $\alpha = 0.05$ for the preference ranking data.

	Setting 1		Setting 2	
Sample Size	$n_1 = 50$, $n_2 = 50$	$n_1 = 25$, $n_2 = 100$	$n_1 = 150$, $n_2 = 150$	$n_1 = 200$, $n_2 = 400$
$S_{(a)}$	.464	.373	.112	.134
$M_{(a)}(1.14)$	.539	.437	.140	.179
$S_{(u)}$	.488	.386	.425	.726
$M_{(u)}(1.14)$	.545	.449	.421	.729
Simplicial depth	.584	.481	.512	.764
Zonoid depth	.595	.497	.532	.791
Onion depth	.579	.475	.496	.762
Halfspace depth	.591	.488	.510	.780
Proposed	**.698**	**.589**	**.591**	**.833**
	Setting 3		Setting 4	
Sample Size	$n_1 = 150$, $n_2 = 150$	$n_1 = 100$, $n_2 = 500$	$n_1 = 50$, $n_2 = 50$	$n_1 = 50$, $n_2 = 100$
$S_{(a)}$	.097	.138	.476	.573
$M_{(a)}(1.14)$	.130	.163	.530	.642
$S_{(u)}$	.421	.495	.552	.637
$M_{(u)}(1.14)$	.427	.506	.586	.675
Simplicial depth	.494	.597	.660	.786
Zonoid depth	.509	.611	.677	.803
Onion depth	.475	.588	.642	.775
Halfspace depth	.491	.599	.665	.788
Proposed	**.551**	**.674**	**.760**	**.858**
	Setting 5			
Sample Size	$n_1 = 50$, $n_2 = 50$	$n_1 = 25$, $n_2 = 75$		
$S_{(a)}$	.481	.345		
$M_{(a)}(1.14)$	.542	.404		
$S_{(u)}$	.548	.456		
$M_{(u)}(1.14)$	.581	.452		
Simplicial depth	.653	.533		
Zonoid depth	.664	.555		
Onion depth	.647	.523		
Halfspace depth	.654	.540		
Proposed	**.764**	**.690**		

Table B.7: The simulated Type-I error rates for the haplotype association data.

The simulated Type-I errors										
α	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10
$S_{(a)}$	.005	.025	.029	.057	.048	.063	.070	.087	.084	.110
$M_{(a)}(1.14)$	.011	.026	.035	.053	.054	.054	.072	.089	.091	.112
$S_{(u)}$	.007	.017	.038	.054	.061	.070	.067	.084	.081	.106
$M_{(u)}(1.14)$	.014	.026	.035	.059	.054	.074	.080	.096	.089	.116
Simplicial depth	.012	.023	.032	.036	.044	.060	.071	.082	.097	.101
Zonoid depth	.012	.023	.031	.040	.049	.072	.073	.080	.101	.097
Onion depth	.008	.012	.029	.026	.042	.051	.062	.074	.093	.089
Halfspace depth	.011	.015	.030	.033	.039	.064	.064	.075	.100	.093
Proposed	.011	.023	.026	.034	.044	.048	.063	.090	.100	.104

Table B.8: The simulated powers for the haplotype association data.

The simulated powers										
α	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10
$S_{(a)}$	.291	.353	.380	.428	.482	.476	.500	.517	.522	.559
$M_{(a)}(1.14)$	.359	.439	.462	.508	.561	.544	.572	.593	.591	.634
$S_{(u)}$	.318	.380	.403	.434	.496	.506	.520	.544	.552	.573
$M_{(u)}(1.14)$	.399	.463	.485	.541	.572	.567	.587	.618	.619	.657
Simplicial depth	.346	.444	.491	.502	.587	.613	.642	.663	.677	.701
Zonoid depth	.347	.459	.492	.527	.593	.612	.652	.658	.676	.698
Onion depth	.297	.380	.455	.475	.575	.585	.639	.641	.661	.683
Halfspace depth	.334	.397	.464	.513	.589	.611	.643	.661	.674	.703
Proposed	**.463**	**.540**	**.591**	**.637**	**.687**	**.695**	**.745**	**.759**	**.766**	**.787**

Table B.9: The simulated powers for comparing phone-call patterns on weekdays and on weekends at $\alpha = 0.05$.

The simulated powers									
Proportion	.1	.2	.3	.4	.5	.6	.7	.8	.9
$S_{(a)}$	.11	.16	.17	.36	.30	.35	.43	.36	.57
$M_{(a)}(1.14)$	.12	**.19**	.21	.38	.38	.40	.53	.56	.80
$S_{(u)}$	.10	.17	.16	.32	.31	.33	.37	.37	.58
$M_{(u)}(1.14)$	.11	.18	.20	.37	.39	.42	.49	.57	.81
Simplicial depth	.06	.15	.34	.43	.70	.86	.92	**1.00**	**1.00**
Zonoid depth	.07	.15	.31	.48	.69	.88	.96	**1.00**	**1.00**
Onion depth	.06	.13	.30	.46	.66	.85	.94	**1.00**	**1.00**
Halfspace depth	.06	.12	.32	.49	.70	.87	.97	**1.00**	**1.00**
Proposed	**.14**	**.19**	**.38**	**.57**	**.76**	**.91**	**.98**	**1.00**	**1.00**