Title
New Kernel-based Methods for High-dimensional Inferences

Author
Song, Hoseung

Publication Date
2021

New Kernel-based Methods for High-dimensional Inferences

By

HOSEUNG SONG
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Hao Chen, Chair

Alexander Aue

Miles Lopes

Committee in Charge

2021

To my lovely wife Jungyoon and my adorable daughters Zion and Irene

Contents

New Kernel-based Methods for High-dimensional Inferences

Abstract

As we are entering the big data era with technological advances of data collection, high-dimensional and complex data is becoming prevalent and the development of effective analysis is gaining more attention to researchers in statistics and data science. Many approaches are usually parametric, but they are highly context specific.

Kernel-based methods are widely used as a nonparametric approach and they have the potential to capture changes in the distribution. This dissertation aims to develop novel kernel-based methods for high-dimensional data on two problems: (i) two-sample testing and (ii) change-point analysis.

Kernel two-sample tests have been widely used for high-dimensional data as an elegant nonparametric framework of testing equal distribution. However, existing tests based on kernel embeddings of probability distributions into reproducing kernel Hilbert spaces (RKHS) do not work well for a wide rage of alternatives when the dimension of the data is moderate to high due to the curse of dimensionality. We propose a new test statistic that makes use of patterns under high dimension and achieves substantial power improvement over existing kernel two-sample tests for general alternatives. We also propose an alternative testing procedure that maintains high power with little computational cost, offering easy off-the-shelf tools for large datasets.

We also consider the testing and estimation of change-points, locations where the distribution abruptly changes in a sequence. Compared with two-sample testing problems, kernel-based methods in change-point analysis have not been well explored. We propose a new kernel-based framework that exhibits high power in detecting and estimating the location of the change-point under general alternatives. Analytic approximations to the significance of the new test statistics for both single change-point and changed-interval alternatives are derived and fast tests are proposed, offering easy off-the-shelf tools for large datasets.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Hao Chen, for her constant guidance, support, and encouragement. Her expertise and extensive experience have inspired me to pursue my academic interests and seek a better insight when I was struggling. Not to mention her inspiration for research, I would like to pay homage to her attitude and personality towards research as a successful researcher and there are so many things I learned. The times I spent with her at every step of my journey will be invaluable presents throught my academic career. Without her patience and thoughtful advice, this dissertation could never have been possible.

I wish to show my gratitude to my dissertation committee members, Dr. Alexander Aue and Dr. Miles Lopes for their valuable feedback and participation on my thesis. Their insightful comments steer my thesis in the right direction. I am grateful to our department staff for their professional supports. I would also like to acknowledge Dr. Jayaraman J. Thiagarajan and Dr. Bhavya Kailkhura for their collaborative efforts and guidance since I did an internship at LLNL.

I would like to thank many friends and colleagues in my department for the fruitful and delightful academic and non-academic conversations. Particularly, I deeply thank to many Korean friends in Davis, neighbors at Solano Park, and the community at the Davis Korean Church. They have made my five years in Davis wonderful and I am deeply indebted to them whenever my family encounters difficulties. There is no doubt that I owe them a debt and I will never forget this.

I wish to thank my parents for their prayers and unconditional love. Without their endless supports, I would not be here now and nothing would have done. I am proud to be the son of my parents.

Last but not at least, I mostly want to acknowledge my beloved wife Jungyoon and my daughters, Zion and Irene, who stand by me as always. Especially, I owe the most to my wife Jungyoon and nothing would have been possible without her great love and dedication. She deserves my heartfelt thanks and she is my true hero.

CHAPTER 1

Introduction

1.1. Problem statements

Statistics has become one of the fastest growing fields among the sciences and recent developements in statistics have been motivated by the need to analyze increasingly large and complex data. Advanced technologies in producing and collecting vast amoung of data leads to rapidly growing demand for new statistical methodology. For example, many fundamental topics in statistics are re-adressed in the face of new types of data. Motivated by these challenges, this dissertation aims to develop novel methods and practical tools for two testing problems: two-sample testing and change-point analysis.

Two-sample hypothesis testing plays a significant role in a variety of scientific applications, such as bioinformatics, social sciences, and image analysis. Formally speaking, given samples $X_1, X_2, \ldots, X_m \overset{iid}{\sim} P$ and $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} Q$ where P and Q are distributions in $\mathcal{R}^d$, one wants to test $H_0 : P = Q$ against $H_1 : P \neq Q$.

Change-point analysis is regaining attention as we enter the big data era. High-dimensional complex data sequences are becoming prevalent and the development of efficient change-point detection method is gaining more attention for this new setting. Given a sequence of independent observations $\{y_i\}_{1,\ldots,n}$, we consider testing the null hypothesis

$$H_0 :\ y_i \sim F_0,\ i = 1, \ldots, n \tag{1.1}$$

against the single change-point alternative

$$H_1 : \exists\, 1 \leq \tau < n,\ y_i \sim \begin{cases} F_0, & i \leq \tau \\ F_1, & \text{otherwise} \end{cases} \tag{1.2}$$

or the changed interval alternative

$$H_2 : \exists\, 1 \leq \tau_1 < \tau_2 < n,\ y_i \sim \begin{cases} F_0, & i = \tau_1 + 1, \ldots, \tau_2 \\ F_1, & \text{otherwise} \end{cases} \tag{1.3}$$

where F_0 and F_1 are two different disbtributions.

1.2. Overview

1.2.1. Generalized kernel two-sample tests. Kernel two-sample tests have been widely used for multivariate data in testing equal distribution. However, existing tests based on mapping distributions into a reproducing kernel Hilbert space are mainly targeted at specific alternatives and do not work well for some scenarios when the dimension of the data is moderate to high due to the curse of dimensionality. We propose a new test statistic that makes use of a common pattern under moderate and high dimensions and achieves substantial power improvements over existing kernel two-sample tests for a wide range of alternatives. We also propose alternative testing procedures that maintain high power with low computational cost, offering easy off-the-shelf tools for large datasets. The new approaches are compared to other state-of-the-art tests under various settings and show good performance. The new approaches are illustrated on two applications: The comparison of musks and non-musks using the shape of molecules, and the comparison of taxi trips started from John F.Kennedy airport in consecutive months. All proposed methods are implemented in an `R` package `kerTests`.

1.2.2. New kernel-based change-point detection. Change-point analysis plays a significant role in various fields as it can reveal the discrepancies in the relational information in the sequence. While many algorithms have been proposed, kernel-based methods have not been well explored due to difficulties in offering false positive controls and mediocre performance. In this paper, we propose a new kernel-based framework that makes use of an important pattern of data in high dimensions to boost power. Analytic approximations to the significance of the new statis- tics are derived and fast tests based on the asymptotic results are proposed, offering easy off-the-shelf tools for large datasets. The new tests show superior performance for a wide range of alternatives when comparing with other state-of-the-art methods. We illustrate these new approaches through an analysis of a phone-call network data.

CHAPTER 2

Generalized Kernel Two-Sample Tests

2.1. Introduction

2.1.1. Background. Nonparametric two-sample hypothesis testing received a lot of attention as challenging data, both in dimension and size, are produced in many fields. Formally speaking, given samples $X_1, X_2, \dots, X_m \overset{iid}{\sim} P$ and $Y_1, Y_2, \dots, Y_n \overset{iid}{\sim} Q$ where P and Q are distributions in $\mathcal{R}^d$, one wants to test $H_0 : P = Q$ against $H_1 : P \neq Q$. When d is large, such as in hundreds or thousands or even more, it is common that one has little or no clue of P or Q, which makes parametric tests unrealistic in many applications. Several nonparametric tests have been proposed for high-dimensional data, including rank-based tests [**3**, **33**, **44**, **50**], inter-point distances-based tests [**4**, **38**, **60**], graph-based tests [**10**, **20**, **32**, **49**, **55**], and kernel-based tests [**17**, **26**, **27**, **29**]. They all have succeeded in many applications. In this paper, we focus on kernel-based tests.

The most well-known kernel two-sample test was proposed by [**26**]. They first map the observations into a reproducing kernel Hilbert space (RKHS) generated by a given kernel $k(\cdot, \cdot)$ and consider the maximum mean discrepancy (MMD) between two probability distributions P and Q,

$$\text{MMD}^2(P, Q) = E_{X,X'}[k(X, X')] - 2E_{X,Y}[k(X, Y)] + E_{Y,Y'}[k(Y, Y')], \tag{2.1}$$

where X and X' are independent random variables drawn from P and Y and Y' are independent random variables drawn from Q. [**26**] considered two empirical estimates of $\text{MMD}^2(P, Q)$:

$$\text{MMD}_u^2 = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} k(X_i, X_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k(Y_i, Y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(X_i, Y_j),$$

$$\text{MMD}_b^2 = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(X_i, X_j) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(Y_i, Y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(X_i, Y_j).$$

Here, MMD_u^2 is an unbiased estimator of $\text{MMD}^2(P, Q)$ and is in general preferred over MMD_b^2. When the kernel k is characteristic, such as the Gaussian kernel or the Laplacian kernel, the MMD behaves as a metric [**59**].

[**26**] studied asymptotic behaviors of MMD_u^2 and found that MMD_u^2 degenerated under the null hypothesis of equal disrtribution. They then considered $m\text{MMD}_u^2$ when $m = n$ and showed that $m\text{MMD}_u^2$ converged to $\sum_{l=1}^{\infty} \lambda_l(z_l^2 - 2)$ under H_0. Here $z_l \overset{iid}{\sim} N(0, 2)$ and λ_l's are the solutions of the eigenvalue equation

$$\int_{\mathcal{X}} \tilde{k}(X, X')\psi_l(X)dP(X) = \lambda_l \psi_l(X'),$$

with $\tilde{k}(X_i, X_j) = k(X_i, X_j) - E_X k(X_i, X) - E_X k(X, X_j) + E_{X,X'} k(X, X')$ the centred RKHS kernel. Since the limiting distribution $\sum_{l=1}^{\infty} \lambda_l(z_l^2 - 2)$ is an infinite sum, a few approaches were proposed to approximate it: a moment matching approach using Pearson curves [**26**], a spectrum approximation approach, and a Gamma approximation approach [**29**]. However, they have some drawbacks. For example, [**29**] mentioned that the performance of the tests based on the moment matching method and the Gamma approximation are not guaranteed. In addition, all these approaches only work for the balanced sample design, i.e, the sample sizes of the two samples are the same. Hence, in terms of guaranteed performance of the test and for possibly unbalanced sample sizes, a bootstrap approach is usually preferred in many applications to approximate the p-value, despite a high computational cost.

[**28**] studied the choice of the kernel and the bandwidth parameter to maximize the power of the test from the set of a linear combination of Gaussian kernels in a training set. More recently, [**47**] found that the power of the test based on the Gaussian kernel is independent of the kernel bandwidth, when the bandwidth is greater than the median of all pairwise distances among observations. Therefore, in the following, without further specification, we use the most popular characterstic kernel, the Gaussian kernel, with the median heuristic as the bandwidth parameter.

2.1.2. A problem of MMD_u^2. Even though MMD_u^2 works well under many settings, it has some weird behaviors under some common alternatives. Consider a toy example for Gaussian data: $X_1, \ldots, X_{50} \overset{iid}{\sim} N_d(\mathbf{0}_d, \Sigma)$; $Y_1, \ldots, Y_{50} \overset{iid}{\sim} N_d(a\mathbf{1}_d, b\Sigma)$, where the (i, j)th element of Σ is $\Sigma_{i,j} = 0.4^{|i-j|}$, $\mathbf{0}_d$ is a d dimensional vector of zeros, $\mathbf{1}_d$ is a d dimensional vector of ones, and $d = 50$. Three settings are considered:

- Setting 1: $a = 0.21, b = 1$.

- Setting 2: $a = 0.21, b = 1.04$.
- Setting 3: $a = 0, b = 1.1$.

Table 2.1 presents the estimated power of the MMD_u^2 test based on 1,000 simulation runs. In each simulation run, 10,000 bootstrap replicates are used to approximate the p-value. We refer to this test 'MMD-Bootstrap' for simplicity. We see that MMD-Boostrap performs well for the mean difference in setting 1, but it has slightly lower power in setting 2 than in setting 1, despite the additional variance difference in setting 2. When the difference is only in the variance (setting 3), MMD-Boostrap performs poorly.

TABLE 2.1. Estimated power (by 1,000 trials) of MMD-Bootstrap at 0.05 significance level

Setting 1	Setting 2	Setting 3
0.912	0.886	0.071

To explore the underlying reason why this happens, we examine the empirical distributions of $\alpha - \gamma$ and $\beta - \gamma$, where $\alpha = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} k(X_i, X_j)$, $\beta = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} k(Y_i, Y_j)$, $\gamma = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(X_i, Y_j)$ ($\text{MMD}_u^2 = \alpha + \beta - 2\gamma$). This is shown in Figure 2.1. We see that, in setting 1, the distributions of $\alpha - \gamma$ and $\beta - \gamma$ shift to the right compared to those under the null. Hence, MMD_u^2 tends to be large in setting 1, and the power of the test in setting 1 in 0.912. In setting 2, with the additional variance change, we see that the empirical distribution of $\alpha - \gamma$ indeed shift to further right. However, the empirical distribution of $\beta - \gamma$ is similar to that under the null. As a result, the effects of $\alpha - \gamma$ and $\beta - \gamma$ offset in setting 2, and the power of setting 2 is lower than that under setting 1. This phenomenon gets severer in setting 3 where $\beta - \gamma$ is mainly negative and almost completely offsets $\alpha - \gamma$. From Figure 2.1, in setting 3, $\alpha - \gamma$ and $\beta - \gamma$ do display their derivations from the null (purple versus pink). The amount of derivations in setting 3 is larger than that in setting 1 for $\alpha - \gamma$ and $\beta - \gamma$. It is just that the derivations are in oppposite directions that the test statistic MMD_u^2 cannot capture the signal.

2.1.3. Our contribution. With the observations in Section 2.1.2, we explore further the behavior of α and β under the permutation null distribution and propose a new statistic (GPK) that takes into account derivations in both directions. This new test works for a wider range of alternatives that are common in high dimesions than MMD_u^2. We also work out a test statistic (fGPK) that works similar to GPK but with fast type I error control. Using a similar technique, we further work out fGPK_M that has power on par and sometimes much better than prevailing MMD-based tests and at the same time with fast type I error control. All these

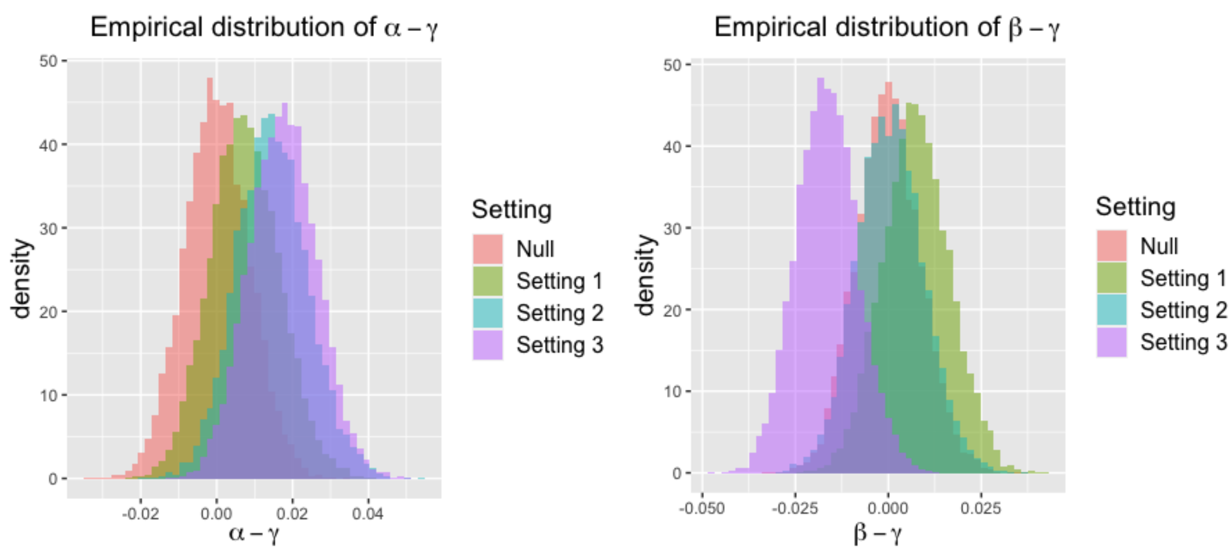


FIGURE 2.1. Empirical distributions of $\alpha - \gamma$ and $\beta - \gamma$ based on 10,000 simulation runs under settings 1,2,3 and the null of no distribution difference $(a = 0, b = 1)$.

new tests, GPK, fGPK, and fGPK_M, work for both equal and unequal sample sizes. The new methods are implemented in an `R` package `kerTests`.

2.2. A New Test Statistic

2.2.1. A pattern under moderate/high dimension. To better understand the behavior of MMD_u^2 under setting 2 in Section 2.1.2, we explore more on α and β. We compare them with their expected values under the permutation null distribution, which places $1/\binom{N}{m}$ probability on each of the $\binom{N}{m}$ permutations of the sample lables $(N = m + n)$. With no further specification, P, E, Var, and Cov denote the probability, the expectation, the variance, and the covaraince, repectively, under the permutation null distribution.

Figure 2.2 shows boxplots of $\alpha - \mathsf{E}(\alpha)$ and $\beta - \mathsf{E}(\beta)$ from 10,000 simulated datasets under the three settings in Section 2.1.2 as well as under the null hypothesis $(a = 0, b = 1)$. In setting 1, we see that both α and β tend to be larger than their null expectations, which is consistent with MMD_u^2 being large. In setting 2, α still tends to be larger than its null expectation, while β tends to be smaller than its null expectation, which could cause the effect of α and β in MMD_u^2 to offset. This phenomenon gets severer in setting 3. The reason this happens lies in the curse of dimensionality: The volume of a d-dimensional space increases exponentially in d. Then, many observations from the distribution with a larger variance can be spasely separated and they tend to be closer to the observations from the distribution with a smaller variance, which

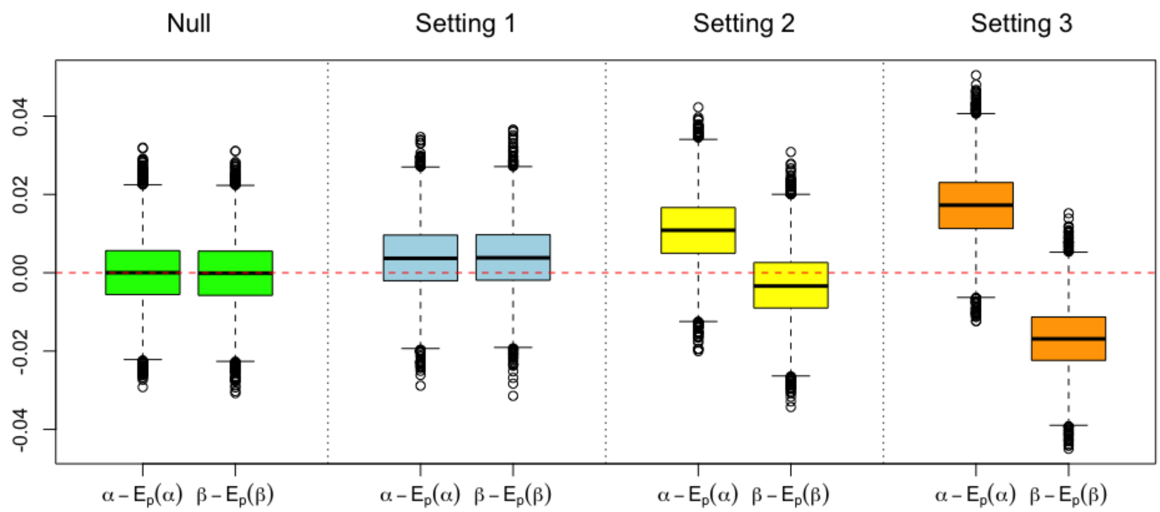


FIGURE 2.2. Boxplots of $\alpha - \mathsf{E}(\alpha)$ and $\beta - \mathsf{E}(\beta)$ of 10,000 simulated datasets under null $(a = 0, b = 1)$, setting 1 $(a = 0.21, b = 1)$, setting 2 $(a = 0.21, b = 1.04)$, and setting 3 $(a = 0, b = 1.1)$.

could lead to one of α or β smaller than its expectation under the null, depending on which sample has a smaller variance.

2.2.2. A generalized permutation-based kernel two-sample test statistic. Based on the findings in Section 2.2.1, we segregate α and β and propose the following statistic:

$$\text{GPK} = \big(\alpha - \mathsf{E}(\alpha), \beta - \mathsf{E}(\beta)\big)\Sigma_{\alpha,\beta}^{-1}\begin{pmatrix} \alpha - \mathsf{E}(\alpha) \\ \beta - \mathsf{E}(\beta) \end{pmatrix}, \tag{2.2}$$

where $\Sigma_{\alpha,\beta} = \mathsf{Var}((\alpha,\beta)^T)$. The analytic expressions of $\mathsf{E}(\beta)$, $\mathsf{E}(\beta)$, and $\Sigma_{\alpha,\beta}$ can be derived and are provided in Theorem 2.2.1. The new test statistic designed in this way aggregates deviations of α and β from their expectations under the permutation null in both directions, so it can cover more general alternatives than MMD_u^2.

We briefly check how GPK works for Gaussian data $N_d(\mathbf{0}_d, \Sigma)$ vs. $N_d(a\mathbf{1}_d, b\Sigma)$, where $\Sigma_{i,j} = 0.4^{|i-j|}$ and $m = n = 50$, under location and/or scale alternatives. The estimated power and empirical sizes of GPK and MMD-Bootstrap are presented in Figure 2.3 and Table 2.2, respectively. We see that GPK has comparable power to MMD_u^2 for location alternatives. However, when the change is in scale, MMD-Bootstrap performs poorly and GPK has much higher power. When both the mean and the variance differ, GPK in general outperforms MMD-Bootstrap. We also see that GPK controls the type I error well (Table 2.2). Here, we briefly check the performance of GPK for illustration and more simulation studies are in Section 2.4.

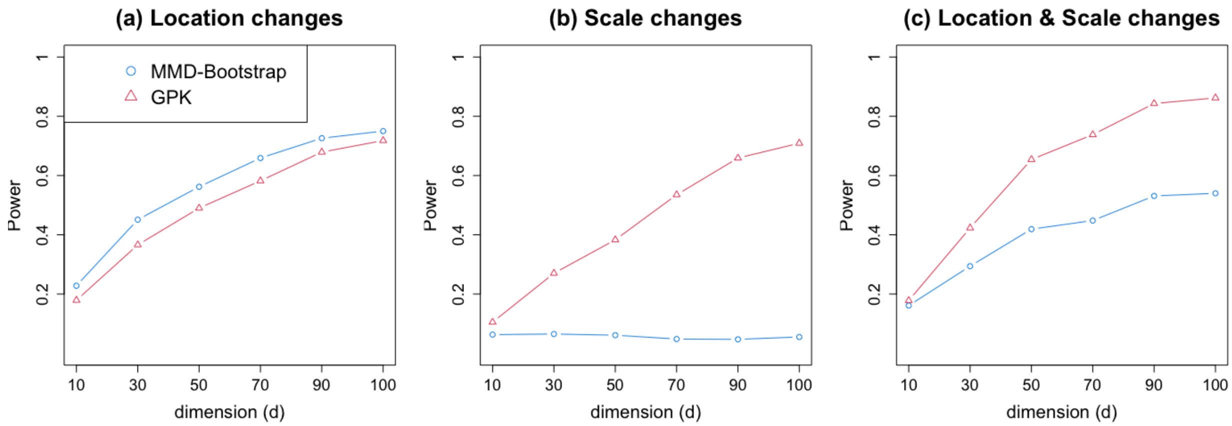


FIGURE 2.3. Estimated power (by 1,000 trials) of MMD-Bootstrap (o) and GPK ($\triangle$) at 0.05 significance level for multivariate Gaussian data: (a) $a = 0.15$, $b = 1$, (b) $a = 0$, $b = 1.1$, (c) $a = 0.1$, $b = 1.1$.

TABLE 2.2. Empirical size at 0.05 significance level estimated by 1,000 trials for MMD-Bootstrap and GPK under different dimensions

d	10	30	50	70	90	100
MMD-Bootstrap	0.045	0.044	0.038	0.043	0.026	0.028
GPK	0.045	0.045	0.056	0.060	0.051	0.051

For notation simplicity, we pool observations from the two samples together and denote them by $z_1, \ldots, z_N$. Let $k(z_i, z_j) = k_{ij}$ for $i, j = 1, \ldots, N$. Then, the analytic formulas for $\mathsf{E}(\alpha)$, $\mathsf{E}(\beta)$, and $\Sigma_{\alpha,\beta(i,j)}$, the (i, j) element of $\Sigma_{\alpha,\beta}$, are provided in the following theorem.

THEOREM 2.2.1. *Under the permutation null distribution, we have*

$$\mathsf{E}(\alpha) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} k_{ij}, \quad \mathsf{E}(\beta) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} k_{ij},$$

$$\Sigma_{\alpha,\beta(1,1)} = \frac{1}{m^2(m-1)^2} \bigg(2A \frac{m(m-1)}{N(N-1)} + 4B \frac{m(m-1)(m-2)}{N(N-1)(N-2)}$$

$$+ C \frac{m(m-1)(m-2)(m-3)}{N(N-1)(N-2)(N-3)} \bigg) - \mathsf{E}(\alpha)^2,$$

$$\Sigma_{\alpha,\beta(2,2)} = \frac{1}{n^2(n-1)^2}\Bigg(2A\frac{n(n-1)}{N(N-1)} + 4B\frac{n(n-1)(n-2)}{N(N-1)(N-2)} + C\frac{n(n-1)(n-2)(n-3)}{N(N-1)(N-2)(N-3)}\Bigg) - \mathsf{E}(\beta)^2,$$

$$\Sigma_{\alpha,\beta(1,2)} = \Sigma_{\alpha,\beta(2,1)} = \frac{C}{N(N-1)(N-2)(N-3)} - \mathsf{E}(\alpha)\mathsf{E}(\beta),$$

where

$$A = \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}^2, \quad B = \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\sum_{u=1,u\neq j,u\neq i}^{N} k_{ij}k_{iu},$$

$$C = \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\sum_{u=1,u\neq j,u\neq i}^{N}\sum_{v=1,v\neq u,v\neq j,v\neq i}^{N} k_{ij}k_{uv}.$$

To prove this theorem, we rewrite α and β in the following way. For each $z_1,\dots,z_N$, let $g_i = 0$ if observation z_i is from sample X and $g_i = 1$ if observation z_i is from sample Y. Then,

$$\alpha = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m} k(X_i,X_j) = \frac{1}{m(m-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k(z_i,z_j)I_{g_i=g_j=0}, \tag{2.3}$$

$$\beta = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} k(Y_i,Y_j) = \frac{1}{n(n-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k(z_i,z_j)I_{g_i=g_j=1}. \tag{2.4}$$

Hence, computing $\mathsf{E}(\alpha)$ boils down to $\mathsf{E}(I_{g_i=g_j=0})$ and similar for $\mathsf{E}(\beta)$. Computing the variance of α and the covariance of α and β under the permutation null distribution needs more careful analysis on different combinations. The detailed proof of the theorem is in Appendix A.1.

To assure that the new test statistic is well-defined, the covariance matrix $\Sigma_{\alpha,\beta}$ needs to be invertible.

THEOREM 2.2.2. *For $m,n \geq 2$, the proposed statistic GPK is well-defined when k_{ij}'s do not satisfy either of the following two corner cases:*

(C1) $\sum_{j=1,j\neq i}^{N} k_{ij}$ *are all the same for* $i = 1,\dots,N$.

(C2) $\sum_{j=1,j\neq i}^{N} k_{ij} - (N-2)k_{iN}$ *are all the same for* $i = 1,\dots,N-1$.

This theorem can be proved through mathematical induction. The complete proof is in Appendix A.2. It is difficult to simplify the descriptions of (C1) and (C2) further. We briefly illustrate this and its simulation results are provided in Appendix A.3.

2.3. Asymptotics and Alternative Tests

2.3.1. Outline. Given the new test statistic GPK, the next question is to compute the p-value of the test. In Figure 2.3 and Table 2.2 in Section 3.2, we use 10,000 random permutations to approximate the p-value, but this is time consuming. To this end, we attempt to study the asymptotic distribution of GPK under the permutation null distribution. We first notice that GPK can be decomposed to the squares of two uncorrelated quantities with one quantity asymptotically normally distributed under some mild conditions and the other quantity closely related to MMD_u^2. Moreover, the quantity closely related to MMD_u^2 after some modifications is also asymptotically normally distributed under some mild conditions. Based on these findings, we propose two tests, fGPK and fGPK_M, whose p-values can be approximated by analytic formulas with the former closely related to the test based on GPK and the latter related to the test based on MMD_u^2.

2.3.2. A decomposition of GPK and asymptotic results.

THEOREM 2.3.1. *The statistic GPK can be decomposed as*

$$GPK = Z_W^2 + Z_D^2,$$

where

$$Z_W = \frac{W - \mathsf{E}(W)}{\sqrt{\mathsf{Var}(W)}}, \quad Z_D = \frac{D - \mathsf{E}(D)}{\sqrt{\mathsf{Var}(D)}}, \tag{2.5}$$

with $W = \frac{m}{N}\alpha + \frac{n}{N}\beta$ *and* $D = m(m-1)\alpha - n(n-1)\beta$.

The proof to this theorem is in Appendix A.4.

REMARK 2.3.1. *The quantity Z_W is closely related to MMD_u^2. Notice that*

$$m(m-1)\alpha + n(n-1)\beta + 2mn\gamma = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} k_{ij}.$$

Hence,

$$\begin{aligned} MMD_u^2 = \alpha + \beta - 2\gamma &= \alpha + \beta - \frac{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij} - m(m-1)\alpha - n(n-1)\beta}{mn} \\ &= \alpha\left(1 + \frac{m-1}{n}\right) + \beta\left(1 + \frac{n-1}{m}\right) - \frac{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}}{mn} \end{aligned}$$

$$= \frac{N(N-1)}{mn}\Big(\frac{m}{N}\alpha + \frac{n}{N}\beta\Big) - \frac{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}}{mn}$$
$$= \frac{N(N-1)}{mn}W - \frac{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}}{mn}.$$

Here, the quantity $\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}/mn$ does not change under permutation. Therefore, Z_W is equivalent to MMD-Permutation, the MMD$_u^2$ test with its p-value computed under the permutation null distribution.

Since Z_W is closely related to MMD_u^2, GPK could in general deal with the alternatives that MMD_u^2 covers. In addition, Z_D covers a new region of alternatives that could be missed by MMD_u^2, making GPK work for more general alternatives.

REMARK 2.3.2. *From the proof of Theorem 2.2.2, the determinant of $\Sigma_{\alpha,\beta}$ can be expressed as*

$$|\Sigma_{\alpha,\beta}| = \frac{1}{mn(m-1)^2(n-1)^2N(N-1)(N-2)}\Big((4A+4B) - \frac{4(2A+4B+C)}{N}\Big)$$
$$\times \Big((N-2)2A + \frac{2}{N-1}(2A+4B+C) - (4A+4B)\Big).$$

Also, in the proof of Theorem 2.3.1, we have that

$$\textit{Var}(W) = \frac{mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)} \times \frac{(N-2)}{N^2(m-1)^2(n-1)^2}$$
$$\times \Big((N-2)2A + \frac{2}{N-1}(2A+4B+C) - (4A+4B)\Big),$$
$$\textit{Var}(D) = \frac{mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)} \times \frac{(N-2)(N-4)}{(m-1)(n-1)}$$
$$\times \Big((4A+4B) - \frac{4(2A+4B+C)}{N}\Big).$$

When the corner case (C1) happens, the first product term in $|\Sigma_{\alpha,\beta}|$ is zero and Z_D is not well defined; when the corner case (C2) happens, the second product term in $|\Sigma_{\alpha,\beta}|$ is zero and Z_W is not well defined.

We now examine the asymptotic permutation null distribution in the *usual limiting regime*, which is defined as $N \to \infty$, $m/N \to p$ with p a constant and $0 < p < 1$. Let $k_{i\cdot} = \sum_{j=1,j\neq i}^{N} k_{ij}$ for $i = 1, \dots, N$ and $\dot{k} = \sum_{i=1}^{N} k_{i\cdot}/N$.

THEOREM 2.3.2. *With the characteristic kernels, when* $k_{ij} = O(1)\ \forall i, j$ *and* $\sum_{i=1}^{N} k_{i\cdot}^2 - N\dot{k}^2 = O\big(\sum_{i=1}^{N} k_{i\cdot}^2\big)$*, in the usual limiting regime, under the permutation null,*

$$Z_D \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

The proof to this theorem is in Appendix A.5.

REMARK 2.3.3. $\sum_{i=1}^{N} k_{i\cdot}^2 - N\dot{k}^2 = \sum_{i=1}^{N}(k_{i\cdot} - \dot{k})^2$ *can be seen as the variability of* $k_{i\cdot}$ *and the condition* $\sum_{i=1}^{N} k_{i\cdot}^2 - N\dot{k}^2 = O\big(\sum_{i=1}^{N} k_{i\cdot}^2\big)$ *ensures GPK to be well-defined in the usual limiting regime.*

Let $W_r = r\frac{m}{N}\alpha + \frac{n}{N}\beta$ be an weighted version of W, where r is a constant. Note that $W_1 = W$.

THEOREM 2.3.3. *With the characteristic kernels, when* $k_{ij} = O(1)\ \forall i, j$ *and* $\sum_{i=1}^{N} k_{i\cdot}^2 - N\dot{k}^2 = O\big(\sum_{i=1}^{N} k_{i\cdot}^2\big)$*, in the usual limiting regime, under the permutation null,*

$$Z_{W,r} \triangleq \frac{W_r - \mathsf{E}(W_r)}{\sqrt{\mathsf{Var}(W_r)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

when $r \neq 1$.

The proof to this theorem is in Appendix A.6.

REMARK 2.3.4. *From Remark 2.3.1 and 2.3.2, it is easy to see that* Z_W *is of the same order of* $m\mathit{MMD}_u^2$*. [**26**] showed that* $m\mathit{MMD}_u^2$ *converges to the intractable distribution under the true null and they thus relied on other approaches to approximate it. Similarly,* Z_W *may also converge to some distributions, but instead we propose to use* $Z_{W,r}$ *which is applicable in practice.*

Figure 2.4 shows the normal quantile-quantile plots for Z_D, $Z_{W,1.0}$, $Z_{W,1.1}$, and $Z_{W,1.2}$ from 10,000 permutations under different choices of m and n for Gaussian data with $d = 100$. We see that, when m,n are in hundreds, the permutation distributions can already be well approximated by the standard normal distributions for Z_D and for $Z_{W,r}$, when r is away from 1, such as $r = 1.2$. Similarly, for $r < 1$, such as $r = 0.8$, the permutation distributions can be well approximated by the standard normal distributions, when m,n are in hundreds or more.

2.3.3. Fast test: fGPK. Although $Z_{W,r}$, $r \neq 1$, converges to the standard normal distribution under mild conditions, the performance of the test decreases as r goes away from 1 under the location alternative.

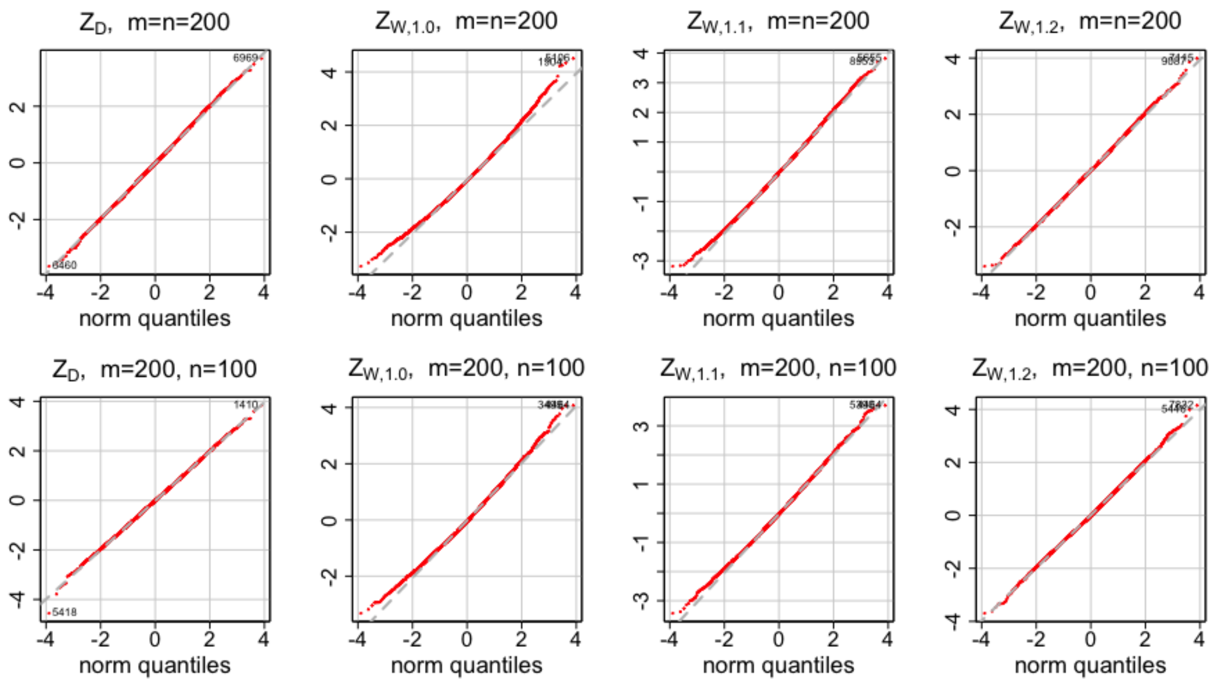


FIGURE 2.4. Normal quantile-quantile plots (red dots) of Z_D, $Z_{W,1.0}$, $Z_{W,1.1}$, $Z_{W,1.2}$ with the gray dashed line the baseline goes through the origin and of slope 1.

TABLE 2.3. Estimated power (by 100 simulation runs) of $Z_{W,r}$ at 0.05 significance level. $m = n = 100$ and $\Delta = \|\mu_1 - \mu_2\|_2$

Location Alternatives						
d	10	30	50	70	90	100
Δ	0.3	0.5	0.7	0.8	0.9	1.0
$r = 1.4$	0.10	0.21	0.29	0.21	0.38	0.37
$r = 1.3$	0.11	0.24	0.36	0.36	0.49	0.50
$r = 1.2$	0.15	0.28	0.43	0.50	0.68	0.63
$r = 1.1$	0.10	0.42	0.55	0.70	0.83	0.84
$r = 1.0$	**0.25**	**0.52**	**0.60**	**0.77**	**0.90**	**0.86**
$r = 0.9$	0.22	0.47	0.41	0.77	0.76	0.78
$r = 0.8$	0.16	0.36	0.27	0.49	0.57	0.54
$r = 0.7$	0.15	0.23	0.20	0.37	0.32	0.33

Table 2.3 shows the estimated power of $Z_{W,r}$ for Gaussian data $N_d(\mu_1, I_d)$ vs. $N_d(\mu_2, I_d)$, with the mean difference $\Delta = \|\mu_1 - \mu_2\|_2$. The p-value of each test is approximated by 10,000 permutations for fair comparison. We see that the power of the test decreases as r goes away from 1.

To make use of asymptotic results and maximize power, we propose to use a Bonferroni test on $Z_{W,1.2}$, $Z_{W,0.8}$, and Z_D. We choose $Z_{W,1.2}$ and $Z_{W,0.8}$ since they cover different regions of alternatives. Let $p_{W,1.2}$, $p_{W,0.8}$, and p_D be the approximated p-value of the test that rejects for large values of $Z_{W,1.2}$, $Z_{W,0.8}$, and $|Z_D|$, respectively, based on their limiting distributions, i.e., if the values of $Z_{W,1.2}$, $Z_{W,0.8}$, and Z_D are $b_{W,1.2}$, $b_{W,0.8}$, and b_D, respectively, then $p_{W,1.2} = 1 - \Phi(b_{w,1.2})$, $p_{W,0.8} = 1 - \Phi(b_{w,0.8})$, and $p_D = 2\Phi(-|b_D|)$. Then, we propose the fast test, fGPK, that is defined to reject the null if $3\min(p_D, p_{W,1.2}, p_{W,0.8})$ is less than the significance level. Hence, as long as $Z_{W,1.2}$, $Z_{W,0.8}$, and Z_D are computed, the p-value of fGPK can be obtained instantly.

REMARK 2.3.5. $r = 1.2$ *and 0.8 are determined empirically based on the fact that r's are away from 1 enough so that the normal approximation is reasonable and not too away to maintain a good power.*

REMARK 2.3.6. *We adopt the Bonferroni procedure for the fast test to combine the advantages of each test statistic. To improve the power of the fast test, other global testing methods, such as the Simes procedure, can be used since the Bonferroni procedure is a bit conservative. (see Section 2.6.2).*

2.3.4. Fast test: fGPK$_\mathrm{M}$. Based on the limiting distribution of $Z_{W,r}$, $r \neq 1$, the same technique can also be applied to approximate the MMD test. That is, we propose the fast test, fGPK$_\mathrm{M}$, that is defined to reject the null if $2\min(p_{W,1.2}, p_{W,0.8})$ is less than the significance level.

Since the test of Z_W is equivalent to MMD-Permutation, we expect fGPK$_\mathrm{M}$ to be powerful for locational alternatives. Moreover, according to the simulation results, it turns out that fGPK$_\mathrm{M}$ can also detect variance differences to some extent as $r = 1.2, 0.8$ cover more types of alternatives than $r = 1$.

To check the effectiveness of fGPK$_\mathrm{M}$, we compare fGPK$_\mathrm{M}$ with tests based on MMD. According to the simulation results, fGPK$_\mathrm{M}$ exhibits comparably high power for location alternatives, while fGPK$_\mathrm{M}$ shows better performance than other MMD-based tests for scale alternatives in high dimensions. The details of the simulation results are provided in Appendix A.7.

Additionally, we compare the computational cost of these tests as well as fGPK. Both samples are drawn from the standard 100-dimensional Gaussian distribution and the sample sizes are the same $(m = n)$. Table 2.4 reports the time cost of the methods implemented in `Matlab`. For MMD-Pearson and MMD-Bootstrap, we use the `Matlab` codes released by Arthur Gretton, publicly available at `http://www.gatsby.ucl.ac.uk/~gretton/mmd/mmd.htm`. Time comparison for these methods implemented in R is in Appendix A.8. We see that fGPK_M is the fastest to run as its computation cost is $O(N^2d)$, while MMD-Bootstrap is the slowest as it is $O(N^2dR)$ time test with R bootstrap replicates. fGPK is also faster than the existing tests. The Pearson approximation test is fairly fast in the `Matlab` implementation, but it would become slow when the sample size is quite large as it costs $O(N^3d)$ [**26**]. To sum, compared to other tests of MMD_u^2, fGPK_M exhibits good performance with better computational efficiency.

TABLE 2.4. Average computation time in seconds (standard deviation) from 10 simulations for each m. All experiments were run by `Matlab` on 2.2 GHz Intel Core i7

m	50	100	250	500	1000
fGPK_M	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.03 (0.00)	0.09 (0.00)
fGPK	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.03 (0.00)	0.11 (0.01)
MMD-Pearson	0.00 (0.00)	0.00 (0.00)	0.09 (0.00)	0.76 (0.05)	12.70 (0.31)
MMD-Bootstrap	0.43 (0.02)	1.51 (0.05)	8.39 (0.22)	37.44 (5.13)	251.9 (16.1)

2.4. More Numerical Experiments

In this section, we compare the three new tests (GPK, fGPK, fGPK_M) with two commonly used MMD-based tests (MMD-Pearson and MMD-Bootstrap) on more diverse examples in moderate/high dimensions. We also include other nonparametric tests using the ball divergence (BT) [**46**], classifier (CT) [**41**], and graphs (GT) [**10**], which can be implemented by R packages `ball`, `Ecume`, and `gTests`, respectively. Here, we use a 5-MST (minimum spanning tree) for GT. We consider the following settings:

- Multivariate Gaussian data: $N_d(\mathbf{0}_d, \Sigma)$ vs. $N_d(a\mathbf{1}_d, \sigma^2\Sigma)$, where $\Delta = \|a\mathbf{1}_d\|_2$ and $\Sigma_{i,j} = 0.4^{|i-j|}$.
- Multivariate t-distributed data: $t_{20}(\mathbf{0}_d, \Sigma)$ vs. $t_{20}(a\mathbf{1}_d, \sigma^2\Sigma)$, where $\Delta = \|a\mathbf{1}_d\|_2$ and $\Sigma_{i,j} = 0.4^{|i-j|}$.
- Chi-square data: $\Sigma^{1/2}u^{\chi^2_{3,d}}$ vs. $(\sigma^2\Sigma)^{1/2}u^{\chi^2_{3,d}} + a\mathbf{1}_d$, where $\Delta = \|a\mathbf{1}_d\|_2$, $\Sigma_{i,j} = 0.4^{|i-j|}$, and $\chi^2_{3,d}$ is a length-d vector with each component i.i.d. from the centered χ^2_3 distribution.

For the multivariate Gaussian data, we also compare the tests under the unbalanced setting ($m \neq n$). Notice that MMD-Pearson cannot be applied to the unbalanced setting.

In each simulation setting, we consider various dimensions. For each dimension, we simulate 1,000 datasets. The parameters of the distributions are chosen so that the tests are of moderate power to be comparable. The significance level is set to be 0.05 for all tests.

TABLE 2.5. Estimated power of the tests for multivariate Gaussian data. The balanced two-sample sizes setting ($m = n = 50$). Top 1 method and those higher than 95% of the top 1 are in bold

	Location Alternatives (Δ)				Scale Alternatives (σ^2)			
d	50	100	500	1000	50	100	500	1000
$\Delta \mid \sigma^2$	1.13	1.50	2.23	2.84	1.11	1.09	1.05	1.04
MMD-Pearson	0.177	0.155	0.006	0.002	0.001	0.001	0.000	0.000
MMD-Bootstrap	**0.651**	**0.801**	0.516	0.334	0.065	0.042	0.001	0.000
GPK	0.567	**0.761**	**0.772**	**0.891**	0.472	0.611	0.843	**0.913**
fGPK	0.527	0.704	0.747	**0.868**	0.460	0.605	**0.848**	**0.900**
fGPK_M	0.578	0.749	**0.800**	**0.905**	0.317	0.432	0.612	0.702
BT	0.362	0.384	0.216	0.222	**0.534**	**0.686**	**0.890**	**0.941**
CT	0.367	0.464	0.525	0.635	0.074	0.040	0.023	0.018
GT	0.193	0.282	0.303	0.388	0.370	0.418	0.659	0.706

TABLE 2.6. Estimated power of the tests for multivariate Gaussian data. The unbalanced two-sample sizes setting ($m = 100,\ n = 50$)

	Location Alternatives (Δ)				Scale Alternatives (σ^2)			
d	50	100	500	1000	50	100	500	1000
$\Delta \mid \sigma^2$	0.98	1.30	2.01	2.84	1.11	1.09	1.04	1.04
MMD-Pearson	-	-	-	-	-	-	-	-
MMD-Bootstrap	**0.612**	0.632	0.132	0.085	0.044	0.014	0.000	0.001
GPK	**0.620**	**0.733**	**0.817**	**0.979**	**0.624**	**0.761**	**0.867**	**0.980**
fGPK	0.529	0.673	0.770	**0.964**	**0.604**	**0.747**	**0.863**	**0.972**
fGPK_M	**0.592**	**0.731**	**0.832**	**0.980**	0.451	0.574	0.710	0.875
BT	0.316	0.342	0.190	0.303	**0.628**	**0.773**	**0.887**	**0.982**
CT	0.271	0.309	0.395	0.617	0.055	0.050	0.029	0.014
GT	0.162	0.249	0.302	0.516	0.372	0.442	0.522	0.745

TABLE 2.7. Estimated power of the tests for multivariate t-distributed data ($m = n = 50$)

	Location Alternatives (Δ)				Scale Alternatives (σ^2)			
d	50	100	500	1000	50	100	500	1000
$\Delta \mid \sigma^2$	0.8	1.2	1.9	2.5	1.15	1.13	1.08	1.08
MMD-Pearson	0.075	0.121	0.685	0.829	0.006	0.007	0.024	0.095
MMD-Bootstrap	**0.454**	**0.721**	**0.993**	**1.000**	0.131	0.248	0.249	0.564
GPK	0.397	**0.690**	**1.000**	**1.000**	0.359	0.581	0.641	**0.883**
fGPK	0.238	0.341	0.654	0.683	0.356	0.573	0.633	**0.875**
fGPK_M	0.292	0.430	0.772	0.801	0.380	0.613	**0.677**	**0.900**
BT	0.101	0.079	0.082	0.078	**0.460**	**0.689**	**0.690**	**0.910**
CT	0.243	0.408	0.787	0.796	0.062	0.017	0.010	0.000
GT	0.164	0.301	0.932	**0.980**	0.272	0.376	0.292	0.408

TABLE 2.8. Estimated power of the tests for chi-square data ($m = n = 50$)

	Location Alternatives (Δ)				Scale Alternatives (σ^2)			
d	50	100	500	1000	50	100	500	1000
$\Delta \mid \sigma^2$	2.05	2.90	5.36	7.90	1.12	1.11	1.06	1.06
MMD-Pearson	0.072	0.043	0.006	0.011	0.042	0.029	0.001	0.000
MMD-Bootstrap	**0.352**	**0.467**	0.450	0.633	0.247	0.369	0.068	0.013
GPK	0.330	0.437	**0.738**	**0.988**	0.344	0.563	0.657	**0.919**
fGPK	0.224	0.280	0.543	0.912	0.338	0.557	**0.681**	**0.932**
fGPK_M	0.265	0.347	0.615	**0.952**	**0.375**	**0.605**	**0.698**	**0.939**
BT	0.131	0.104	0.091	0.120	0.344	0.547	**0.698**	**0.937**
CT	0.206	0.294	0.444	0.665	0.149	0.130	0.054	0.034
GT	0.150	0.164	0.259	0.586	0.193	0.272	0.372	0.565

TABLE 2.9. Empirical size of the tests at 0.05 significance level ($m = n = 50$)

	Multivariate Gaussian				Chi-square			
d	50	100	500	1000	50	100	500	1000
MMD-Pearson	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000
MMD-Bootstrap	0.045	0.029	0.002	0.000	0.042	0.022	0.002	0.000
GPK	0.044	0.051	0.048	0.046	0.046	0.040	0.044	0.054
fGPK	0.042	0.038	0.041	0.043	0.042	0.025	0.038	0.044
fGPK_M	0.047	0.043	0.056	0.054	0.048	0.039	0.050	0.055
BT	0.043	0.047	0.050	0.047	0.049	0.046	0.050	0.055
CT	0.054	0.055	0.075	0.059	0.055	0.056	0.044	0.058
GT	0.045	0.053	0.048	0.041	0.045	0.052	0.045	0.044

Tables 2.5 and 2.6 show results for multivariate Gaussian distributions with different means and/or variances. We see that MMD-Pearson has considerably lower power than other tests in all settings. We thus compare the other seven tests in more details. Under the location alternatives, when $d = 50$ or 100, MMD-Bootstrap does very well and followed immediately by fGPK_M and GPK, and then by fGPK; when d is larger ($d = 500$ or 1000), MMD-Bootstrap is outperformed by the new tests with fGPK_M exhibiting the hightest power. Under the unbalanced sample design, both GPK and fGPK_M exhibit high power. Under scale alternatives, MMD-Bootstrap has much lower power than the new tests. Among the new tests, GPK and fGPK are doing similar and they are both better than fGPK_M. BT exhibits high power for scale alternatives and the new tests also have comparable power, while BT is outperformed by the new tests under location alternatives.

Table 2.7 shows results for multivariate t-distributed data. We see that MMD-Bootstrap and GPK are very sensitive to the mean change and fGPK_M also shows good performance. However, MMD-Bootstrap performs poorly for the scale alternatives, while the new tests still perform well. CT and GT exhibit high power for the location alternatives, but they lose power for the scale alternatives. BT shows the opposite pattern.

Tables 2.8 shows results for chi-square data. Similar to results of multivariate Gaussian data, the new tests with GPK and fGPK_{M} dominate in power for the location alternatives when d is larger ($d = 500$ or 1000). Under scale alternatives, when $d = 50$ or 100, fGPK_{M} outperforms other tests, while BT and fGPK also exhibit high power when d is larger ($d = 500$ or 1000). These results show that the new tests work well for both symmetric and asymmetric distributions under moderate to high dimensions.

Table 2.9 shows empirical size of the tests at 0.05 significance level for the multivariate Gaussian and chi-square data. We see that the new tests control the type I error rate well.

The overall pattern of the power tables shows that the new tests exhibit good performance for a wide range of alternatives. GPK performs well for a wide range of alternatives and fGPK maintains high power with computational advantage. Unlike MMD tests, fGPK_{M} is computationally efficient and can also capture the variance difference to some extent. In practice, fGPK and fGPK_{M} would be preferred as they are fast and highly effective to a wide range of alternatives. If further investigation is needed, the permutation test based on GPK would also be useful.

2.5. Real Data Examples

2.5.1. Musk data. We first illustrate the new tests on Musk data [**5**], which is publicly available at `https://archive.ics.uci.edu/ml/datasets.php`. The Musk dataset consists of molecule structure data. The features indicate the shape of the molecule constructed by the rotation of bonds. This dataset describes a set of 476 molecules of which 269 are judged by human experts to be musks and the remaining 207 molecules are judged to be non-musks, where $d = 166$.

We utilize this dataset to illustrate how the new tests distinguish musks versus non-musks from the shape of the molecule. To this end, we conduct the testing procedures on subsets of the whole data to compare their empirical power. For each m, we randomly draw m observations from these 269 musk observations and m observations from these 207 non-musk observations. We repeat this for 1,000 times and conduct the test with the significance level set to be 0.01.

TABLE 2.10. Estimated power of the tests

m	30	40	50	60	70
MMD-Pearson	0.058	0.121	0.190	0.270	0.402
MMD-Bootstrap	0.091	0.167	0.275	0.388	0.568
GPK	0.133	0.265	0.434	0.606	0.780
fGPK	0.260	0.445	0.618	0.742	0.865
fGPK_M	0.077	0.215	0.301	0.437	0.639

The results are shown in Table 2.10. We see that the new tests in general outperform the existing tests for any m, indicating the consistent improvement of the new test.

2.5.2. New York City taxi data. We illustrate the new tests on New York City taxi data, which is publicly available on the NYC Taxi & Limousine Commission (TLC) website (`https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`). The data contains latitude and longitude coordinates of pickup and drop-off locations, taxi pickup and drop-off date, driver-reported passenger counts, fares, and so on. The data is very rich, and we use it to illustrate the three new tests by testing travel patterns in consecutive months. In particular, we consider the trips that start from the John F.Kennedy (JFK) international airport. We preprocessed the data in the same way as in [**13**] such that we set the boundary of JFK airport to be 40.63 to 40.66 latitude and -73.80 to -73.77 longitude. We split this area into a 30×30 grid with equal size and count the number of trips whose drop-off location fall in each cell for each day. Then, we use these 30×30 matrices to test whether there is a difference in travel patterns between January and February in 2015. To do this, we let the distance of two matrice be the Frobenius norm of the difference of the two matrices, and use the Gaussian kernel with the median of all pairwise distances as the bandwidth.

Table 2.11 shows the results of the tests. Notice that MMD-Pearson cannot be applied due to the unbalanced sample sizes. We see that the new tests reject the null hypothesis of equal distributions at 0.05 significance level, while MMD-Bootstrap does not.

TABLE 2.11. p-values of the tests

	MMD-Bootstrap	GPK	fGPK	fGPK_M
Jan vs. Feb	0.141	**0.027**	**0.027**	**0.020**

We investigate the test statistics in more detail for this comparison where the four tests have different conclusions. Table 2.12 shows $\alpha - \gamma$ and $\beta - \gamma$ values and p-values of the test based on $Z_{W,1.2}$, $Z_{W,0.8}$, and Z_D. For this testing, $\alpha - \gamma$ is negative, so it offsets with $\beta - \gamma$ and MMD-Boostrap is insignificant. Also, the amount of $|\alpha - \gamma|$ and $|\beta - \gamma|$ is relatively large. This implies that there is a significant variance difference. We see that p_D is very small. $p_{W,0.8}$ is also very small as it covers this specific alternative here. As a result, GPK, fGPK, and fGPK_M are all significant.

TABLE 2.12. Values and p-values of each test

Jan vs. Feb	$\alpha - \gamma$	$\beta - \gamma$	MMD-Bootstrap	$Z_{W,1.2}$	$Z_{W,0.8}$	Z_D
Value	-0.061	0.070	0.009	-1.164	2.481	-2.547
p-value	-	-	0.141	0.904	**0.010**	**0.009**

2.6. Discussion

2.6.1. A Brief Discussion on Bandwidth. In this section, we briefly discuss the bandwidth choice in Gaussian kernels. MMD behaves as a metric when the kernel is characteristic ([**59**]) and the most popular characterstic kernel is the Gaussian kernel with the median heuristic as a bandwidth parameter [**56**]. [**47**] found that the performance of the test based on MMD using Gaussian kernel is independent of the bandwidth when the bandwidth is greater than the median heuristic. We used the median heuristic in the earlier implements of the new tests, and we here briefly check whether this heuristic is reasonable for the new tests through numerical studies.

The simulation setup is as follows: we use Gaussian data and examine the average median heuristic in each setting by 100 trials (the averaged median heuristic is around 10 when $d = 100$ and 14 when $d = 200$ in our settings). We then choose 8 bandwidths that differ by 2 from each other, starting from the averaged median heuristic -8 to the averaged median heuristic +8 so as to check bandwidths in a wide range. We then check the performance of GPK for each bandwidth choice for four different settings (Figure 2.5). The results of fGPK and fGPK_M are provided in Appendix A.9.

We see that there is no significant difference in the performance unless the bandwidth is too small. This result coincides with argument in [**47**] that the power of the test is independent of the kernel bandwidth, as long as it is greater than the choice made by the median heuristic. Through this numerical study, we see that the median heuristic would be a reasonable choice for our new tests under the permutation null distribution.

2.6.2. The fast tests with the Simes procedure. Instead of the Bonferroni test, the Simes test may be used to improve the performance of the fast tests. It has been shown that the Simes procedure is exact

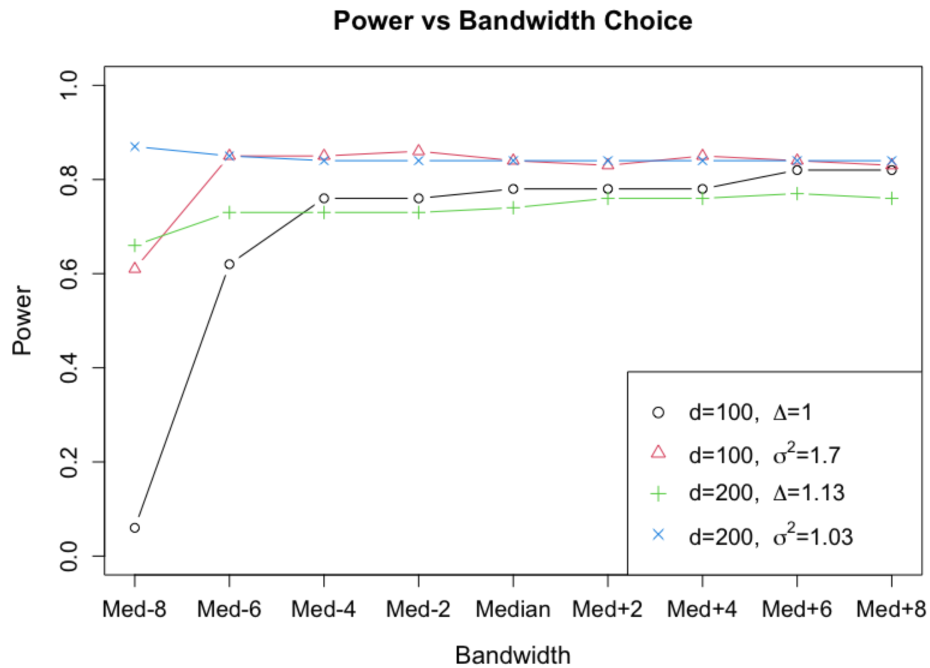


FIGURE 2.5. Estimated power based on 100 trials of GPK vs. bandwidth when 100 samples are generated from $N_d(\mathbf{0}, I_d)$ and 100 samples are drawn from $N_d(\mu, \sigma^2 I_d)$ with $\Delta = \|\mu\|_2$, $\alpha = 0.05$. 'Median' on the X axis indicates the averaged median heruistic in each simulation run.

under independent distributions, while it becomes conservative under positively dependent distributions and slightly liberal under negatively dependency. There have been a lot of works to prove the validity of the Simes test under dependency [**6**, **7**, **19**, **24**, **25**, **34**, **52**, **53**, **54**], but they are restricted to special cases. Nevertheless, the Simes test is widely used in many applications. [**48**] proved that the overall relative deviation of the Simes p-value from the true p-value is strongly bounded and showed that, although the Simes procedure may be liberal, it cannot be consistently. It is therefore reasonably expected that the Simes p-value will be asymptotically valid in most practical cases.

Let $p_{(1)} \leq p_{(2)} \leq p_{(3)}$ be the ordered p-values of $p_{W,1.2}$, $p_{W,0.8}$, and p_D. Then, the fast test, fGPK-Simes, is defined to reject the null if $\min(3p_{(1)}, 1.5p_{(2)}, p_{(3)})$ is less than the significance level. Similarly, let $p'_{(1)} \leq p'_{(2)}$ be the ordered p-values of $p_{W,1.2}$ and $p_{W,0.8}$. Then, the fast test, fGPK$_{\mathrm{M}}$-Simes, is defined to reject the null if $\min(2p'_{(1)}, p'_{(2)})$ is less than the significance level.

Table 2.13 shows the empirical size of the tests for Gaussian data and chi-square data used in Section 2.4. We see that the Simes procedure also controls the type I error well. Hence, if we want to focus on the performance of the test and improve the power of the fast tests, the Simes procedure would be useful for the fast tests.

TABLE 2.13. Empirical size of the tests at 0.05 significance level ($m = n = 50$)

	Multivariate Gaussian				Chi-square			
d	50	100	500	1000	50	100	500	1000
fGPK	0.051	0.045	0.034	0.044	0.052	0.053	0.045	0.037
fGPK-Simes	0.052	0.046	0.039	0.049	0.048	0.050	0.042	0.035
fGPK$_{\text{M}}$	0.055	0.045	0.042	0.051	0.055	0.058	0.041	0.043
fGPK$_{\text{M}}$-Simes	0.055	0.045	0.042	0.052	0.055	0.058	0.041	0.043

CHAPTER 3

New Kernel-Based Change-Point Detection

3.1. Introduction

3.1.1. Problem statement. Recent technological advances have facilitated the collection of high-dimensional data sequence in various high-impact applications, including social sciences, neuroscience, molecular chemistry, and computer graphics. High-dimensional complex data sequences are becoming prevalent and the development of efficient change-point detection method is gaining more attention for this new setting. In this paper, we consider the offline change-point detection problem where a sequence of independent observations $\{y_i\}_{1,\ldots,n}$ is given at the time when data analysis is conducted. Specifically, we consider testing the null hypothesis

$$H_0 :\ y_i \sim F_0,\, i = 1, \ldots, n \tag{3.1}$$

against the single change-point alternative

$$H_1 : \exists\, 1 \leq \tau < n,\ y_i \sim \begin{cases} F_0, & i \leq \tau \\ F_1, & \text{otherwise} \end{cases} \tag{3.2}$$

or the changed interval alternative

$$H_2 : \exists\, 1 \leq \tau_1 < \tau_2 < n,\ y_i \sim \begin{cases} F_0, & i = \tau_1 + 1, \ldots, \tau_2 \\ F_1, & \text{otherwise} \end{cases} \tag{3.3}$$

where F_0 and F_1 are two different disbtributions. Following are some motivating examples.

- *Abnormality in biological system*: Abrupt changes in biological activity by internal or external events may cause serious health problems and it is important to detect such changes as early as possible. For example, detecting changes in gene expression, normally called differential gene expression (DGE), can reveal chromosomal aberrations in the genomic DNA and this is critical to the early diagnosis of cancer [**65**]. Change-point detection can also be used in neurophysiological studies of the brain to analyze the synchrony between neural activity and external event [**61**] and hypothesis testing on changes in dynamic brain networks [**37**].

- *Image Analysis*: The detection of changes in a sequence of images is an important task as image data is becoming increasingly prevalent in many applications, including video surveillance [**1**,**22**], driver assistance systems [**18**], and medical diagnosis and treatment [**8**]. In these applications, each observation consists of a number of pixels and detecting unusual events in the sequence is very challenging.
- *Speech Recognition*: Speech recognition represents the process of converting spoken language to words or text and it is often of interest to capture abrupt changes in the audio composition. For example, audio segmentation is an important task in many audio processing applications and this significantly impacts on the performance of speech recognition [**51**]. An audio signal is superimposed and of high dimensions [**66**]. Change-point detection methods can be applied for audio segmentation by recognizing boundaries between sentences, silence, and noise.

Many approaches for change-point detection are usually parametric [**12**, **63**, **64**, **68**], but they rely on strong assumptions on the sequence. The challenging problems of parametric methods have been addressed in the nonparametric setting, including the methods using marginal rankings [**42**], interpoint distances [**39**, **43**], similarity graphs [**11**,**13**], and Fréchet mean changes [**15**]. For example, [**43**] proposed the method based on U-statistics and a divisive hierarchical estimation algorithm, but it is time-consuming due to an intensive permutation use. [**11**] and [**13**] proposed nonparametirc methods using graphs, such as a k-MST (minimum spanning tree), which is the union of the 1st, . . ., kth MSTs, where a kth MST does not contain edges in the 1st, . . ., (k-1)th MSTs. They provided analytical p-value approximations and showed good performance for general alternatives. [**15**] recently proposed test statistics for detecting change-points in Fréchet mean and/or Fréchet variance. However, this approach is not versatile to use since it needs to compute Fréchet mean and variance changes differently depending on the data structure.

3.1.2. Kernel change-point detection methods and their limitations. Recently, kernel methods are widely used in the two-sample comparison as a nonparametric approach and they have the potential to capture any change in the distribution. This framework began with the test based on mapping distributions into a reproducing kernel Hilbert space (RKHS) generated by a given kernel $k(\cdot,\cdot)$ and the most well-known kernel two-sample test statistic, the maximum mean discrepancy (MMD), proposed by [**26**].

Compared with kernel methods in two-sample testing problems, kernel-based change-point analysis has not been well explored since it is in general difficult to conduct theoretical analysis, such as controlling the type I error, and it is computationally inefficient to use. The first practical offline change-point detection method using kernels was proposed by [**30**]. They incorporated kernels into dynamic programming algorithms for detecting jumps in the sequence. A kernel-based test statistic, called the maximum kernel Fisher discriminant ratio, was also proposed by [**31**]. However, the test statistic has $O(n^3 d)$ time complexity and the approach relies on the bootstrap resampling method for computing the decision threshold, making the test impractical in reality. [**40**] proposed MMD-based test statistic by adopting a strategy developed by [**67**]. Though it is computationally efficient when the amount of data is large, the method specifies a block that could have a potential change-point in the sequence and requires a large amount of reference data before the changes happens. Some other kernel-based methods also do not guarantee on estimating the location of change-points when detected [**9, 35**]. Recently, [**2**] developed a kernel change-point detection procedure (KCP) that extends the method proposed by [**30**]. KCP utilizes a model-selection penalty that allows to select the number of change-points, while the work of [**30**] assumes a fixed known number of change-points. However, a major disadvantage is that KCP heavily depends on the penalty constant and it is very difficult to control the type I error. Table 3.1 shows the empirical size of KCP under different dimensions and penalty constants for Gaussian data when $n = 200$. We see that the empirical size of the test is very sensitive to the penalty constant, particularly for high-dimensional data.

3.1.3. Generalized kernel two-sample tests. The most well-known kernel two-sample test is based on an unbiased estimate of MMD^2 defined as

$$\mathrm{MMD}_u^2 = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m} k(X_i,X_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} k(Y_i,Y_j)$$

$$- \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(X_i,Y_j) \tag{3.4}$$

$$\triangleq \alpha + \beta - 2\gamma, \tag{3.5}$$

TABLE 3.1. Empirical size of KCP under different dimensions and penalty constants for Gaussian data

Penalty constants	0.360	0.355	0.350	0.345	0.340	0.335	0.330	0.325	0.320	0.315
$d = 100$	0.010	0.014	0.029	0.041	0.056	0.084	0.117	0.159	0.204	0.207

Penalty constants	0.0605	0.0600	0.0595	0.0590	0.0585	0.0580	0.0575	0.0570	0.0565
$d = 500$	0.000	0.008	0.019	0.028	0.051	0.081	0.132	0.186	0.242

Penalty constants	0.0302	0.0297	0.0292	0.0287	0.0282	0.0277	0.0272	0.0267	0.0262
$d = 1000$	0.000	0.000	0.001	0.009	0.036	0.159	0.392	0.716	0.932

Penalty constants	0.0139	0.0138	0.0137	0.0136	0.0135
$d = 2000$	0.006	0.027	0.071	0.135	0.252

where $X_1, X_2, \ldots, X_m \overset{iid}{\sim} F_0$ and $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} F_1$. Asymptotic behaviors of MMD_u^2 were studied in [**26**] and it was revealed that MMD_u^2 degenerated under the null hypothesis of equal distributions. Several approaches for approximating the limiting distribution of MMD_u^2 were provided in [**26**], [**29**], and [**27**]. [**28**] and [**47**] studied the choice of the kernel and the bandwidth parameter, and the most popular characterstic kernel, the Gaussian kernel, with the median heuristic (the median of all pairwise distances among observations) is recommended.

Though MMD is widely used and it works well under many settings, it could have low power under some common alternatives for high-dimensional data [**58**]. Moreover, existing testing procedures for MMD_u^2 have other drawbacks, such as the large computational cost or the assumption for the balanced sample design, i.e, the two samples are of the same size. To address this, [**58**] proposed generalized kernel two-sample tests which allow a great versatility in use. They also proposed testing procedures that maintain high power with low computational cost.

3.1.4. Our contribution. To the best of our knowledge, all existing kernel change-point detection methods are either restricted to specific settings and/or computationally intensive. Starting from the motivation of the two-sample version in [**58**], we propose new kernel-based test statistics for the single change-point alternative (3.2) and the changed-interval alternative (3.3). The new tests are easy to implement and have no tuning parameter. The new test statistic performs well for a wide range of alternatives and achieves high power in detecting and estimating change-points in the high-dimensional sequence compared to other state-of-the-art methods. We also derive analytic formulas for type I error control and propose fast tests based on the asymptotic results, making the tests instantly applicable for large datasets.

The organization of the paper is as follows. In Section 3.2, we propose new scan statistics for the single change-point and changed-interval alternatives. The asymptotic behavior of the new test statistics, the analytical p-value approximations, and fast tests are provided in Section 3.3. Section 3.4 examines the performance of the new tests under various simulation settings. The new approaches are illustrated by a real data application on a phone-call network data in Section 3.5. We conclude with discussion in Section 3.6.

3.2. New scan statistics

In this section, the new scan statistics for testing the null H_0 (3.1) against the single change-point alternative H_1 (3.2) (Section 3.2.1) and the changed-interval alternative H_2 (3.3) (Section 3.2.2) are presented. We work under the permutation null distribution, which places $1/n!$ probability on each of the $n!$ permutations of $\{y_i\}_{1,\dots,n}$. With no further specification, P, E, Var, and Cov denote the probability, expectation, variance, and covaraince, repectively, under the permutation null distribution. In addition, without further specification, we use the Gaussian kernel with the median heuristic as the bandwidth parameter.

3.2.1. Scan statistics for the single change-point alternative. Let $g_i(t) = I_{i>t}$, I_x be the indicator function and $k_{ij} = k(y_i, y_j)$ $(i, j = 1, \dots, n)$. The quantity of α, β in testing the two samples $\{y_1, \dots, y_t\}$ and $\{y_{t+1}, \dots, y_n\}$ can be written as

$$\alpha(t) = \frac{1}{t(t-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k_{ij} I_{g_i(t)=g_j(t)=0}, \tag{3.6}$$

$$\beta(t) = \frac{1}{(n-t)(n-t-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k_{ij} I_{g_i(t)=g_j(t)=1}. \tag{3.7}$$

[**58**] proposed the new kernel two-sample test statistic that makes use of a common pattern under moderate and high dimensions. They showed that the new test statistic consists of two uncorrelated quantities and they cover different regions of alternatives, making the new test versatile for a wide range alternatives.

Starting from the motivation of the two-sample version in [**58**], we consider two basic quantities for detecting the change-point as follows:

$$D(t) = t(t-1)\alpha(t) - (n-t)(n-t-1)\beta(t), \tag{3.8}$$

$$W(t) = \frac{n-t}{n} t(t-1)\alpha(t) + \frac{t}{n}(n-t)(n-t-1)\beta(t). \tag{3.9}$$

It is expected that $D(t)$ would be sensitive to scale alternatives and $W(t)$ would be sensitive to location alternatives.

Since the null distributions of $D(t)$ and $W(t)$ depend on t, we standardize $D(t)$ and $W(t)$ so that they are comparable across t. Let $\mathsf{E}(D(t))$, $\mathsf{Var}(D(t))$, $\mathsf{E}(W(t))$, and $\mathsf{Var}(W(t))$ be the expectations and variances of $D(t)$ and $W(t)$, respectively, under the permutation null distribution. Then,

$$Z_D(t) = \frac{D(t) - \mathsf{E}(D(t))}{\sqrt{\mathsf{Var}(D(t))}}, \quad Z_W(t) = \frac{W(t) - \mathsf{E}(W(t))}{\sqrt{\mathsf{Var}(W(t))}}. \tag{3.10}$$

Finally, we define the test statistic for detecting and estimating the change-point in the sequence as

$$\text{GKCP}(t) = Z_D^2(t) + Z_W^2(t), \tag{3.11}$$

to utilize the advantages of $Z_D(t)$ and $Z_W(t)$.

Under the permutation null, the analytic expressions for the expectation and the variance of $D(t)$ and $W(t)$ can be calculated through combinatorial analysis. They are provided in Theorem 3.2.1 (proof in Appendix B.2).

THEOREM 3.2.1. *Under the permutation null, we have*

$$\begin{aligned}
\mathsf{E}(\alpha(t)) = \mathsf{E}(\beta(t)) &= \frac{1}{n(n-1)} R_0, \\
\mathsf{Var}(\alpha(t)) &= \frac{1}{t^2(t-1)^2}\left(2R_1 p_1(t) + 4R_2 p_2(t) + R_3 p_3(t)\right) - \mathsf{E}(\alpha(t))^2, \\
\mathsf{Var}(\beta(t)) &= \frac{1}{(n-t)^2(n-t-1)^2}\left(2R_1 q_1(t) + 4R_2 q_2(t) + R_3 q_3(t)\right) - \mathsf{E}(\beta(t))^2, \\
\mathsf{Cov}(\alpha(t), \beta(t)) &= \frac{R_3}{n(n-1)(n-2)(n-3)} - \mathsf{E}(\alpha(t))\,\mathsf{E}(\beta(t)),
\end{aligned}$$

where

$$\begin{aligned}
R_0 &= \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k_{ij}, \quad R_1 = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k_{ij}^2, \quad R_2 = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \sum_{u=1, u\neq j, u\neq i}^{n} k_{ij} k_{iu}, \\
R_3 &= \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \sum_{u=1, u\neq j, u\neq i}^{n} \sum_{v=1, v\neq u, v\neq j, v\neq i}^{n} k_{ij} k_{uv},
\end{aligned}$$

$$p_1(t) = \frac{t(t-1)}{n(n-1)}, \quad p_2(t) = p_1(t)\frac{t-2}{n-2}, \quad p_3(t) = p_2(t)\frac{t-3}{n-3},$$
$$q_1(t) = \frac{(n-t)(n-t-1)}{n(n-1)}, \quad q_2(t) = q_1(t)\frac{n-t-2}{n-2}, \quad q_3(t) = q_2(t)\frac{n-t-3}{n-3}.$$

To test H_0 (3.1) versus H_1 (3.2), the following scan statistic is used:

$$\max_{n_0 \le t \le n_1} \text{GKCP}(t), \tag{3.12}$$

where n_0 and n_1 are pre-specified constraints on the region where the change-point τ is searched. By default, we can set $n_0 = [0.05n]$[1] and $n_1 = n - n_0$. If there are prior information on the range of the potential change-point, then n_0 and n_1 can be specified accordingly. The null hypothesis H_0 (3.1) is rejected if the scan statistic is greater than a threshold. Detail strategies about how to choose the threshold to control the type I error are discussed in Section 3.3.

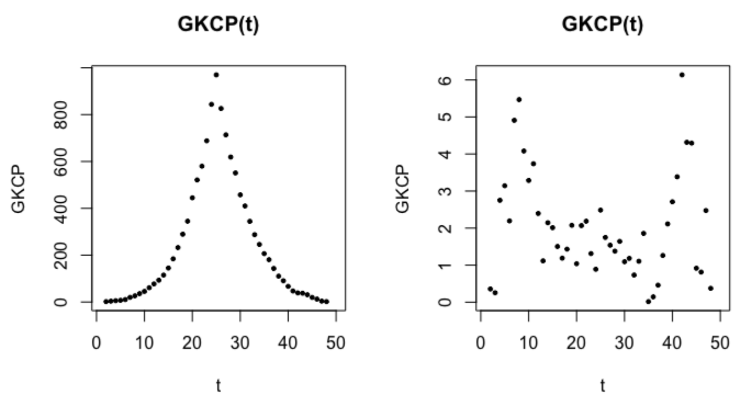


FIGURE 3.1. The profile of GKCP(t) against t for Gaussian data. In the left panel, the first 25 observations are generated from $N_d(\mathbf{0}_d, I_d)$, $d = 100$, and the second 25 observations are drawn from $N_d(\mu, I_d)$ with $\|\mu\|_2 = 3$. In the right panel, all 50 observations are generated from $N_d(\mathbf{0}_d, I_d)$.

Figure 3.1 illustrates the GKCP(t) process for a toy example, where the left panel consists of the first 25 observations generated from $N_d(\mathbf{0}_d, I_d)$, $d = 100$, and the second 25 observations drawn from $N_d(\mu, I_d)$ with $\|\mu\|_2 = 3$ (a change-point in the middle of the sequence) and the right panel consists of 50 observations generated from $N_d(\mathbf{0}_d, I_d)$ (no change-point). It is clear that $\max$ GKCP(t) in the left panel is much larger than that in the right panel and GKCP(t) peaks at the true change-point 25 in the left panel.

[1] $[x]$ denotes the largest integer that is no larger than x.

3.2.2. Scan statistics for the changed-interval alternative. Here, we define the test statistic for testing H_0 (3.1) against the changed-interval alternative H_2 (3.3). Similar to the singe change-point alternative, each possible interval $(t_1, t_2]$ divides the data sequence into two groups. Then, for any candidate interval $(t_1, t_2]$, the test statistics $Z_D(t_1, t_2)$ and $Z_W(t_1, t_2)$ can be defined in the similar manner to the single change-point alternative. Under the permutation null, the analytic expression for $\mathsf{E}(D(t_1, t_2))$, $\mathsf{E}(W(t_1, t_2))$, $\mathsf{Var}(D(t_1, t_2))$, and $\mathsf{Var}(W(t_1, t_2))$ can be obtained similary as in the single change-point setting and details are provided in Appendix B.1. The scan statistic involves a maximization over t_1 and t_2,

$$\max_{\substack{1<t_1<t_2\leq n \\ n_0\leq t_2-t_1\leq n_1}} \text{GKCP}(t_1, t_2), \tag{3.13}$$

where n_0 and n_1 are constraints on the window size.

3.3. Analytical p-value approximations and fast tests

Given the test statistic, the next step is to determine how large the test statistic needs to provide sufficient evidence to reject the null hypothesis of homogeneity. That is, we are concerned with the tail probabilities of the scan statistic under H_0 (3.1),

$$\mathsf{P}\left(\max_{n_0\leq t\leq n_1} \text{GKCP}(t) > b\right) \tag{3.14}$$

for the single change-point alternative, and

$$\mathsf{P}\left(\max_{\substack{1<t_1<t_2\leq n \\ n_0\leq t_2-t_1\leq n_1}} \text{GKCP}(t_1, t_2) > b\right) \tag{3.15}$$

for the changed-interval alternative.

The threshold can be approximated by drawing random permutations of the sequence. This is, however, time consuming. We thus aim to obtain analytical expressions to approximate these tail probabilities.

[**58**] showed that the test of the two-sample version of $Z_W(t)$ is equivalent to the test based on $m\text{MMD}_u^2$ when $m = n$, but [**26**] showed that $m\text{MMD}_u^2$ converges to very complicated distibution under the true null. Hence, we first define an weighted version of $W(t)$ to obtain the tractable asymptotic results:

$$W_r(t) = r\frac{n-t}{n}t(t-1)\alpha(t) + \frac{t}{n}(n-t)(n-t-1)\beta(t). \tag{3.16}$$

$Z_{W,r}(t)$ is the standardized $W_r(t)$, where r is a constant. Note that $W_1(t) = W(t)$.

In the rest of this chapter, we first study the asymptotic properties of the stochastic processes $\{Z_D([nu]) : 0 < u < 1\}$, $\{Z_{W,r}([nu]) : 0 < u < 1\}$, $\{Z_D([nu],[nv]) : 0 < u < v < 1\}$, and $\{Z_{W,r}([nu],[nv]) : 0 < u < v < 1\}$, and then make adjustments for finite samples (Section 3.3.1). We then derive analytic approximations to the tail probabilities under the Gaussian field approximation (Section 3.3.2). We improve our approximations by correcting the skewness in the marginal distributions (Section 3.3.3) and these approximations are checked by numerical studies in Section 3.3.4. Finally, we propose fast tests based on the asymptotic results in Section 3.3.5.

3.3.1. Asymptotic distributions of the basic processes. In this section, we derive the limiting distributions of $\{Z_D([nu]) : 0 < u < 1\}$ and $\{Z_{W,r}([nu]) : 0 < u < 1\}$ for the single change-point alternative and $\{Z_D([nu],[nv]) : 0 < u < v < 1\}$ and $\{Z_{W,r}([nu],[nv]) : 0 < u < v < 1\}$ for the changed-interval alternative.

In the following, we write $a_n = o(b_n)$ when a_n has order smaller than b_n. Let $k_{i\cdot} = \sum_{j=1,j\neq i}^{n} k_{ij}$ for $i = 1, \dots, n$ and $\dot{k} = \sum_{i=1}^{n} k_{i\cdot}/n$.

THEOREM 3.3.1. *With the characteristic kernels, when $k_{ij} = O(1)\ \forall i,j$ and $\sum_{i=1}^{N} k_{i\cdot}^2 - N\dot{k}^2 = O\big(\sum_{i=1}^{N} k_{i\cdot}^2\big)$, as $n \to \infty$,*

(1) $\{Z_D([nu]) : 0 < u < 1\}$ converges to a Gaussian process in finite dimensional distributions, which we denote as $\{Z_D^*(u) : 0 < u < 1\}$.

(2) $\{Z_D([nu],[nv]) : 0 < u < v < 1\}$ converges to a two-dimensional Gaussian random field in finite dimensional distributions, which we denote as $\{Z_D^*(u,v) : 0 < u < v < 1\}$.

The proof for this theorem is in Appendix B.2.

THEOREM 3.3.2. *With the characteristic kernels, when $k_{ij} = O(1)\ \forall i,j$ and $\sum_{i=1}^{N} k_{i\cdot}^2 - N\dot{k}^2 = O\big(\sum_{i=1}^{N} k_{i\cdot}^2\big)$ and $r \neq 1$, as $n \to \infty$,*

(1) $\{Z_{W,r}([nu]) : 0 < u < 1\}$ converges to a Gaussian process in finite dimensional distributions, which we denote as $\{Z_{W,r}^*(u) : 0 < u < 1\}$.

(2) $\{Z_{W,r}([nu],[nv]) : 0 < u < v < 1\}$ converges to a two-dimensional Gaussian random field in finite dimensional distributions, which we denote as $\{Z_{W,r}^*(u,v) : 0 < u < v < 1\}$.

The proof for this theorem is in Appendix B.2.

REMARK 3.3.1. *To prove the convergence of stochastic processes indexed by a continuous variable, we also need a tightness condition of the process. Here, we do not explicitly show the tightness theoretically. However, our simulation results show that the approximation seems quite accurate in practice, particularly for high-dimesional cases (see Tables 3.2, 3.3, and 3.4). We also briefly check this by simulations for Gaussian data used in Figure 3.1 when* $n = 1000$ *and* $d = 500$*. Figure 3.2 illustrates 10 processes of* $Z_D(t)$ *under the null (left pannel) and 10 permuted sequences of one case of the null (right pannel). We see that huge spikes are rare to occur in the middle of the sequences except at the two ends. We plan to study the tightness of the processes theoretically and leave this for post-graudate works.*

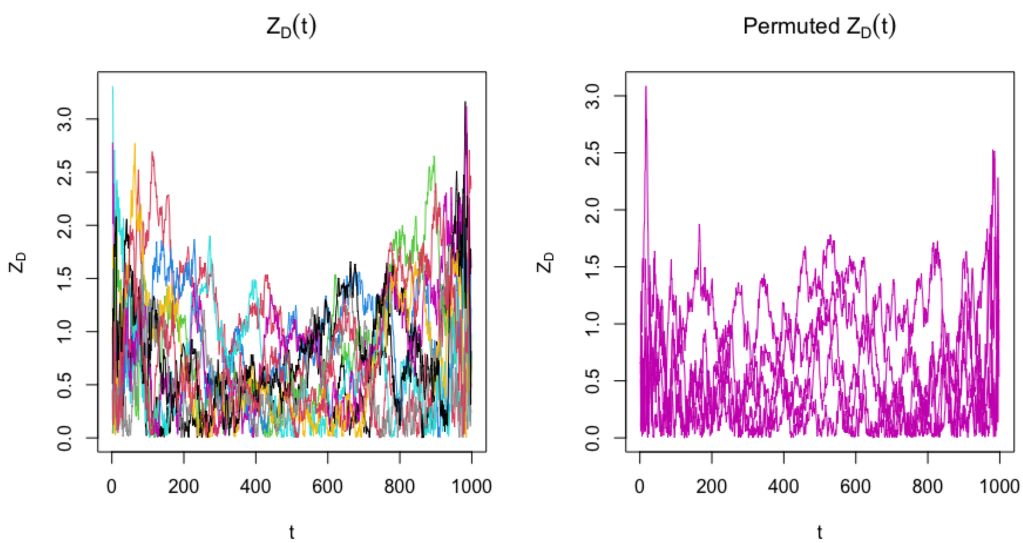


FIGURE 3.2. The profile of $Z_D(t)$ against t for Gaussian data.

Let $\rho^*_D(u,v) = \mathsf{Cov}\left(Z^*_D(u), Z^*_D(v)\right)$ and $\rho^*_{W,r}(u,v) = \mathsf{Cov}\big(Z^*_{W,r}(u), Z^*_{W,r}(v)\big)$. The explicit covariance functions of the limiting Gaussian processes, $\{Z^*_D(u) : 0 < u < 1\}$ and $\{Z^*_{W,r}(u) : 0 < u < 1\}$ are stated in the following theorem.

THEOREM 3.3.3. *The exact expression for* $\rho^*_D(u,v)$ *and* $\rho^*_{W,r}(u,v)$ *are*

$$\rho^*_D(u,v) = \frac{(u \wedge v)\,(1 - (u \vee v))}{\sqrt{u(1-u)v(1-u)}},$$

$$\rho^*_{W,r}(u,v) = \frac{2R_1\{r^2(u\wedge v)\,(1-(u\wedge v))\,(1-(u\vee v))^2\}}{(u\vee v)\,(1-(u\wedge v))\,\sigma^*_{W,r}(u)\sigma^*_{W,r}(v)}$$
$$+ \frac{2R_1\{r(u\vee v)\,(1-(u\wedge v))\,(u+v-uv)+uv\,(2-(u\wedge v))\,(1-(u\vee v))\}}{(u\vee v)\,(1-(u\wedge v))\,\sigma^*_{W,r}(u)\sigma^*_{W,r}(v)}$$
$$+ \frac{4R_2uv(1-u)(1-v)}{(u\vee v)\,(1-(u\wedge v))\,\sigma^*_{W,r}(u)\sigma^*_{W,r}(v)},$$

where $u\wedge v = \min(u,v)$, $u\vee v = \max(u,v)$, *and*

$$\sigma^*_{W,r}(u) = \sqrt{2R_1\{r(1-u)+u\}^2 + (4R_1+4R_2)u(1-u)(r-1)^2}.$$

The above theorem is proved through combinatorial analysis and the details are in Appendix B.2. From the above theorem, we see that the limiting process $\{Z^*_D(u) : 0 < u < 1\}$ does not depend on kernel values, while $\{Z^*_{W,r}(u) : 0 < u < 1\}$ depends on kernel values.

3.3.2. Asymptotic p-value approximations. We now examine the asymptotic behavior of two tail probabilities (3.14) and (3.15). Following similar arguments in the proof for Proposition 3.4 in [**11**], when $n_0, n_1, n, b \to \infty$ in a way such that for some $0 < x_0 < x_1 < 1$ and $b_0 > 0$, $n_0/n \to x_0$, $n_1/n \to x_1$, and $b/\sqrt{n} \to b_0$, as $n \to \infty$, we have

$$\mathsf{P}\left(\max_{n_0\le t\le n_1} |Z^*_D(t/n)| > b\right) \sim 2b\phi(b)\int_{x_0}^{x_1} h^*_D(x)\nu\Big(b_0\sqrt{2h^*_D(x)}\Big)dx, \tag{3.17}$$

$$\mathsf{P}\left(\max_{n_0\le t_2-t_1\le n_1} |Z^*_D(t_1/n, t_2/n)| > b\right) \sim 2b^3\phi(b)\int_{x_0}^{x_1}\left(h^*_D(x)\nu\Big(b_0\sqrt{2h^*_D(x)}\Big)\right)^2(1-x)dx, \tag{3.18}$$

$$\mathsf{P}\left(\max_{n_0\le t\le n_1} Z^*_{W,r}(t/n) > b\right) \sim b\phi(b)\int_{x_0}^{x_1} h^*_{W,r}(x)\nu\Big(b_0\sqrt{2h^*_{W,r}(x)}\Big)dx, \tag{3.19}$$

$$\mathsf{P}\left(\max_{n_0\le t_2-t_1\le n_1} Z^*_{W,r}(t_1/n, t_2/n) > b\right) \sim b^3\phi(b)\int_{x_0}^{x_1}\left(h^*_{W,r}(x)\nu\Big(b_0\sqrt{2h^*_{W,r}(x)}\Big)\right)^2(1-x)dx, \tag{3.20}$$

where the function $\nu(\cdot)$ can be numerically estimated as

$$\nu(s) \approx \frac{(2/s)\,(\Phi(s/2)-0.5)}{(s/2)\Phi(s/2)+\phi(s/2)}$$

[57] with $\Phi(\cdot)$ and $\phi(\cdot)$ being the standard normal cumulative density function and probability density function, respectively, and

$$h_D^*(x) = \lim_{s \nearrow x} \frac{\partial \rho_D^*(s,x)}{\partial s} = -\lim_{s \searrow x} \frac{\partial \rho_D^*(s,x)}{\partial s},$$
$$h_{W,r}^*(x) = \lim_{s \nearrow x} \frac{\partial \rho_{W,r}^*(s,x)}{\partial s} = -\lim_{s \searrow x} \frac{\partial \rho_{W,r}^*(s,x)}{\partial s}.$$

REMARK 3.3.2. *In practice, when using (3.17)–(3.20) for finite sample, we use*

$$\mathsf{P}\left(\max_{n_0 \le t \le n_1} |Z_D(t)| > b\right) \sim 2b\phi(b) \sum_{n_0 \le t \le n_1} C_D(t)\nu\Big(b\sqrt{2C_D(t)}\Big),$$
$$\mathsf{P}\left(\max_{n_0 \le t_2 - t_1 \le n_1} |Z_D(t_1,t_2)| > b\right) \sim 2b^3\phi(b) \sum_{n_0 \le t \le n_1} \Big(C_D(t)\nu\Big(b\sqrt{2C_D(t)}\Big)\Big)^2 (n-t),$$
$$\mathsf{P}\left(\max_{n_0 \le t \le n_1} Z_{W,r}(t) > b\right) \sim b\phi(b) \sum_{n_0 \le t \le n_1} C_{W,r}(t)\nu\Big(b\sqrt{2C_{W,r}(t)}\Big),$$
$$\mathsf{P}\left(\max_{n_0 \le t_2 - t_1 \le n_1} Z_{W,r}(t_1,t_2) > b\right) \sim b^3\phi(b) \sum_{n_0 \le t \le n_1} \Big(C_{W,r}(t)\nu\Big(b\sqrt{2C_{W,r}(t)}\Big)\Big)^2 (n-t),$$

where

$$C_D(t) = \frac{\partial \rho_D(s,t)}{\partial s}\Big|_{s=t}, \quad C_{W,r}(t) = \frac{\partial \rho_{W,r}(s,t)}{\partial s}\Big|_{s=t}$$

with $\rho_D(u,v) = \mathsf{Cov}(Z_D(u), Z_D(v))$ *and* $\rho_{W,r}(u,v) = \mathsf{Cov}(Z_{W,r}(u), Z_{W,r}(v))$*. The explicit expressions for* $C_D(t)$ *and* $C_{W,r}(t)$ *can be calculated in the similar manner to the proof of Theorem 3.3.3.*

3.3.3. Skewness correction. The analytical p-value approximations based on the asymptotic results provide a practical tool for large datasets. However, they become less precise if we set n_0 and n_1 close to the two ends since the convergence of $Z_D(t)$ and $Z_{W,r}(t)$ to the Gaussian process is slow as t/n is close to 0 or 1. As illustrated in Figure 3.3, the skewness is a bit severe when t is close to the two ends.

Hence, we improve the accuracy of the analytical p-value approximations for finite sample sizes by skewness correction. As the skewness depends on the value of t, we adopt a similar treatment discussed in [11] and we add extra terms in the analytic formulas to correct skewness.

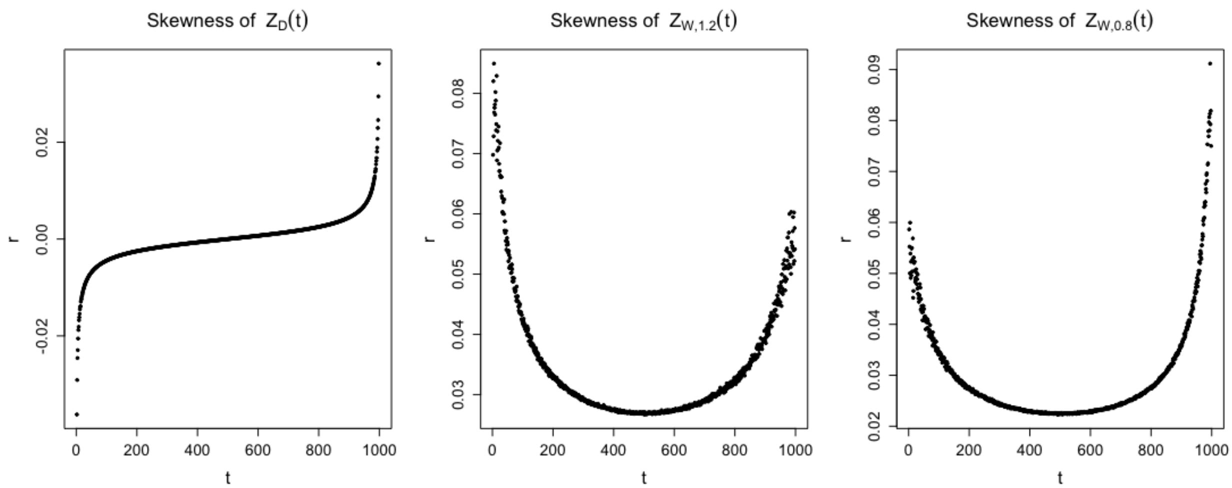


FIGURE 3.3. Plots of skewness of $Z_D(t)$, $Z_{W,1.2}(t)$ and $Z_{W,0.8}(t)$ against t for a sequence of 1000 points generated from $N_{1000}(\mathbf{0}, I_{1000})$.

After skewness correction, the analytical p-value approximations are

$$\mathsf{P}\left(\max_{n_0 \le t \le n_1} |Z_D^*(t/n)| > b\right) \sim 2b\phi(b)\int_{x_0}^{x_1} S_D(x)h_D^*(x)\nu\left(b_0\sqrt{2h_D^*(x)}\right)dx, \tag{3.21}$$

$$\mathsf{P}\left(\max_{n_0 \le t_2 - t_1 \le n_1} |Z_D^*(t_1/n, t_2/n)| > b\right) \sim 2b^3\phi(b)\int_{x_0}^{x_1} S_D(x)\left(h_D^*(x)\nu\left(b_0\sqrt{2h_D^*(x)}\right)\right)^2(1-x)dx, \tag{3.22}$$

$$\mathsf{P}\left(\max_{n_0 \le t \le n_1} Z_{W,r}^*(t/n) > b\right) \sim b\phi(b)\int_{x_0}^{x_1} S_{W,r}(x)h_{W,r}^*(x)\nu\left(b_0\sqrt{2h_{W,r}^*(x)}\right)dx, \tag{3.23}$$

$$\mathsf{P}\left(\max_{n_0 \le t_2 - t_1 \le n_1} Z_{W,r}^*(t_1/n, t_2/n) > b\right) \sim b^3\phi(b)\int_{x_0}^{x_1} S_{W,r}(x)\left(h_{W,r}^*(x)\nu\left(b_0\sqrt{2h_{W,r}^*(x)}\right)\right)^2(1-x)dx, \tag{3.24}$$

where

$$S_D(t) = \frac{\exp\left\{\frac{1}{2}(b - \hat{\theta}_{b,D}(t))^2 + \frac{1}{6}\gamma_D(t)\hat{\theta}_{b,D}^3(t)\right\}}{\sqrt{1 + \gamma_D(t)\hat{\theta}_{b,D}(t)}},$$

$$S_{W,r}(t) = \frac{\exp\left\{\frac{1}{2}(b - \hat{\theta}_{b,W,r}(t))^2 + \frac{1}{6}\gamma_{W,r}(t)\hat{\theta}_{b,W,r}^3(t)\right\}}{\sqrt{1 + \gamma_{W,r}(t)\hat{\theta}_{b,W,r}(t)}},$$

with

$$\hat{\theta}_{b,D}(t) = \frac{\sqrt{1+2\gamma_D(t)b}-1}{\gamma_D(t)}, \quad \gamma_D(t) = \mathsf{E}\left(Z_D^3(t)\right),$$
$$\hat{\theta}_{b,W,r}(t) = \frac{\sqrt{1+2\gamma_{W,r}(t)b}-1}{\gamma_{W,r}(t)}, \quad \gamma_{W,r}(t) = \mathsf{E}\left(Z_{W,r}^3(t)\right).$$

To obtain $\gamma_D(t)$ and $\gamma_{W,r}(t)$, we need to figure out $\mathsf{E}\left(D^3(t)\right)$ and $\mathsf{E}\left(W_r^3(t)\right)$. The exact analytic expressions of $\mathsf{E}\left(D^3(t)\right)$ and $\mathsf{E}\left(W_r^3(t)\right)$ are complicated and they are provided in Appendix B.3.

REMARK 3.3.3. *When the marginal distribution is highly left-skewed, the skewness is so small that* $1+2\gamma(t)b$ *could be negative. Since this problem usually happens when* t/n *is close to 0 or 1, we apply a heuristic fix discussed in [**11**], the extrapolation for* $\theta_b(t)$ *using its values outside the problematic region.*

3.3.4. Checking p-value approximations under finite n. In this section, we check how the analytical p-value approximations work for finite samples. To this end, we compare the critical values for 0.05 p-value threshold obtained from doing 10,000 permutations and the critical values obtained in Section 3.3.2 and 3.3.3 under various simulation settings. Here, we focus on the single-change-point alternative, and the results for the changed-interval alternative are provided in Appendix B.5.

We consider three distributions (multivariate Gaussian (C1), multivariate t_5 (C2), multivariate log-normal (C3)) under various dimensions ($d = 100, 500, 1000$). In each simulation, two randomly simulated $n = 1,000$ sequences are generated. The analytic approximations depend on constraints n_0 and n_1. To make things simple, we set $n_1 = n - n_0$.

Since the asymptotic p-value approximation of $Z_D(t)$ without skewness correction does not depend on kernel values, the critical value is determined by n, n_0, and n_1 only. On the other hand, the asymptotic p-value approximation of $Z_{W,r}(t)$ without skewness correction and the skewness corrected p-value approximations of $Z_D(t)$ and $Z_{W,r}(t)$ depend on certain kenel values.

Table 3.2 shows the results of the scan statistic of $Z_D(t)$. The critical values without skewness corrections are presented labeld by ‘A1’. ‘A2’ denotes the skewness-corrected analytical critical values and ‘Per’ presents the critical values obtained from 10,000 permutation. We see that the asymptotic p-value approximation performs well in all cases, even without the skewness correction.

Table 3.3 shows the results of the scan statistic of $Z_{W,1.2}(t)$. Here, since the convergence of $Z_{W,r}(t)$ to the Gaussian process becomes slow as r is close to 1, r is set to 1.2 (see Section 3.3.5). We see that the

TABLE 3.2. Critical values for the single change-point scan statistic $\max_{n_0 \le t \le n_1} Z_D(t)$ at 0.05 significance level. $n = 1000$

	$n_0 = 100$	$n_0 = 75$	$n_0 = 50$	$n_0 = 25$
A1	3.00	3.05	3.10	3.16

	Critical Values							
	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$	
	A2	Per	A2	Per	A2	Per	A2	Per
Gaussian	3.00	3.01	3.05	3.04	3.10	3.09	3.16	3.14
$d = 100$	3.00	3.01	3.05	3.03	3.10	3.11	3.16	3.15
Gaussian	3.00	3.01	3.05	3.04	3.10	3.10	3.16	3.16
$d = 500$	3.00	3.01	3.05	3.05	3.10	3.10	3.16	3.16
Gaussian	3.00	3.01	3.05	3.04	3.10	3.10	3.16	3.14
$d = 1000$	3.00	2.99	3.05	3.06	3.10	3.10	3.16	3.15

	Critical Values							
	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$	
	A2	Per	A2	Per	A2	Per	A2	Per
MV-t_5	3.00	3.02	3.05	3.03	3.10	3.10	3.16	3.16
$d = 100$	3.00	3.00	3.05	3.04	3.10	3.10	3.16	3.16
MV-t_5	3.00	2.99	3.04	3.04	3.10	3.09	3.16	3.16
$d = 500$	3.00	2.99	3.04	3.03	3.10	3.09	3.16	3.16
MV-t_5	3.00	2.99	3.05	3.04	3.10	3.08	3.17	3.18
$d = 1000$	3.00	2.99	3.05	3.05	3.10	3.09	3.17	3.16

	Critical Values							
	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$	
	A2	Per	A2	Per	A2	Per	A2	Per
Log-normal	3.00	2.98	3.05	3.02	3.10	3.08	3.16	3.16
$d = 100$	3.00	2.99	3.05	3.04	3.10	3.04	3.16	3.15
Log-normal	3.00	3.00	3.04	3.04	3.10	3.09	3.16	3.16
$d = 500$	3.00	2.99	3.04	3.03	3.10	3.09	3.16	3.16
Log-normal	3.00	2.99	3.05	3.06	3.10	3.07	3.17	3.15
$d = 1000$	3.00	2.99	3.05	3.02	3.10	3.09	3.17	3.14

asymptotic p-value approximation performs reasonably well when $n_0 \geq 50$. However, the asymptotic p-value approximation with skewness correction is doing much better than the critical values without skewness correction, even when $n_0 \geq 25$. In particular, we see that the asymptotic p-value approximation works better as the dimension increases.

TABLE 3.3. Critical values for the single change-point scan statistic $\max_{n_0 \le t \le n_1} Z_{W,1.2}(t)$ at 0.05 significance level. $n = 1000$

	Critical Values											
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$		
	A1	A2	Per	A1	A2	Per	A1	A2	Per	A1	A2	Per
Gaussian	2.79	2.87	2.88	2.84	2.93	2.95	2.90	3.00	3.02	2.99	3.11	3.12
$d = 100$	2.79	2.86	2.86	2.85	2.93	2.94	2.91	3.00	3.03	2.99	3.10	3.08
Gaussian	2.79	2.82	2.83	2.85	2.88	2.89	2.91	2.94	2.93	2.99	3.04	3.02
$d = 500$	2.79	2.82	2.78	2.85	2.88	2.88	2.91	2.94	2.94	2.99	3.04	3.04
Gaussian	2.79	2.81	2.79	2.84	2.87	2.84	2.91	2.94	2.93	2.99	3.04	3.04
$d = 1000$	2.79	2.81	2.78	2.84	2.87	2.87	2.91	2.94	2.93	2.99	3.03	3.00

	Critical Values											
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$		
	A1	A2	Per	A1	A2	Per	A1	A2	Per	A1	A2	Per
MV-t_5	2.76	2.88	2.91	2.81	2.94	2.93	2.87	3.02	3.05	2.95	3.13	3.14
$d = 100$	2.76	2.88	2.92	2.81	2.94	2.97	2.87	3.02	3.03	2.96	3.13	3.13
MV-t_5	2.76	2.81	2.82	2.81	2.86	2.86	2.87	2.94	2.93	2.95	3.04	3.04
$d = 500$	2.75	2.81	2.80	2.81	2.87	2.86	2.87	2.93	2.92	2.95	3.03	3.02
MV-t_5	2.75	2.79	2.79	2.81	2.86	2.86	2.87	2.92	2.90	2.95	3.01	3.00
$d = 1000$	2.75	2.79	2.79	2.81	2.85	2.85	2.87	2.91	2.91	2.95	3.01	3.01

	Critical Values											
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$		
	A1	A2	Per	A1	A2	Per	A1	A2	Per	A1	A2	Per
Log-normal	2.74	3.01	3.14	2.80	3.08	3.22	2.85	3.18	3.29	2.93	3.31	3.48
$d = 100$	2.74	3.01	3.12	2.80	3.08	3.23	2.85	3.18	3.30	2.93	3.32	3.49
Log-normal	2.74	2.91	2.97	2.79	2.98	3.05	2.85	3.06	3.11	2.93	3.18	3.24
$d = 500$	2.74	2.90	2.96	2.79	2.97	3.04	2.85	3.05	3.09	2.93	3.17	3.24
Log-normal	2.74	2.88	2.90	2.79	2.94	2.97	2.85	3.02	3.04	2.92	3.13	3.17
$d = 1000$	2.74	2.88	2.92	2.79	2.94	2.98	2.85	3.02	3.07	2.92	3.13	3.17

We also check for $Z_{W,0.8}(t)$ and the results are presented in Table 3.4. The pattern is similar to that for $Z_{W,1.2}(t)$: the skewness corrected asymptotic p-value approximation shows better accuracy than that without skewness correction and the approximation becomes more accurate as the dimension increases.

3.3.5. Fast tests. Although $Z_{W,r}(t)$ ($r \neq 1$) converges to the Gaussian process under mild conditions, the performance of the test decreases as r goes away from 1. Table 3.5 shows the estimated power of $Z_{W,r}(t)$ under various r for Gaussian data where the first 100 observations are generated from from $N_d(\mu_1, I_d)$ and the second 100 observations are generated from $N_d(\mu_2, I_d)$, where $\Delta = \|\mu_1 - \mu_2\|_2$. The p-values of each

TABLE 3.4. Critical values for the single change-point scan statistic $\max_{n_0 \le t \le n_1} Z_{W,0.8}(t)$ at 0.05 significance level. $n = 1000$

	Critical Values											
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$		
	A1	A2	Per	A1	A2	Per	A1	A2	Per	A1	A2	Per
Gaussian	2.77	2.84	2.84	2.82	2.89	2.88	2.89	2.97	2.98	2.97	3.07	3.04
$d = 100$	2.78	2.84	2.84	2.83	2.90	2.90	2.89	2.97	2.99	2.98	3.07	3.06
Gaussian	2.78	2.80	2.80	2.82	2.85	2.84	2.89	2.92	2.91	2.97	3.01	2.99
$d = 500$	2.78	2.80	2.80	2.83	2.85	2.87	2.89	2.92	2.92	2.98	3.01	3.00
Gaussian	2.77	2.79	2.79	2.83	2.85	2.85	2.89	2.92	2.89	2.97	3.01	2.96
$d = 1000$	2.77	2.80	2.80	2.82	2.85	2.82	2.89	2.91	2.91	2.97	3.00	2.99

	Critical Values											
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$		
	A1	A2	Per	A1	A2	Per	A1	A2	Per	A1	A2	Per
MV-t_5	2.77	2.85	2.88	2.81	2.91	2.94	2.86	2.99	2.99	2.94	3.09	3.11
$d = 100$	2.77	2.85	2.87	2.80	2.91	2.95	2.87	2.99	3.00	2.95	3.09	3.06
MV-t_5	2.75	2.79	2.79	2.80	2.85	2.84	2.86	2.92	2.92	2.94	3.01	3.01
$d = 500$	2.75	2.80	2.80	2.80	2.85	2.83	2.86	2.92	2.92	2.94	3.01	3.00
MV-t_5	2.75	2.79	2.79	2.80	2.85	2.84	2.86	2.92	2.92	2.94	3.00	3.01
$d = 1000$	2.75	2.80	2.80	2.80	2.85	2.83	2.86	2.92	2.92	2.94	2.99	2.99

	Critical Values											
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$		
	A1	A2	Per	A1	A2	Per	A1	A2	Per	A1	A2	Per
Log-normal	2.74	2.96	3.04	2.79	3.04	3.11	2.85	3.13	3.24	2.93	3.25	3.39
$d = 100$	2.74	2.96	3.05	2.79	3.04	3.12	2.85	3.13	3.21	2.93	3.26	3.40
Log-normal	2.74	2.88	2.92	2.79	2.94	2.96	2.85	3.02	3.04	2.92	3.13	3.19
$d = 500$	2.74	2.88	2.93	2.79	2.94	2.97	2.85	3.02	3.06	2.92	3.13	3.19
Log-normal	2.73	2.84	2.84	2.79	2.92	2.97	2.85	2.99	3.01	2.92	3.10	3.12
$d = 1000$	2.73	2.85	2.86	2.79	2.92	2.94	2.85	2.99	3.01	2.92	3.10	3.13

test are approximated by 10,000 permutations for fair comparison. We see that the performance of the test decreases as r goes away from 1.

To make use of the asymptotic result of $Z_{W,r}(t)$ and maximize the power of the test, we propose to use $Z_{W,1.2}(t)$ and $Z_{W,0.8}(t)$ together. The power of the test is compromised by using both test statistics together as they cover different regions of alternatives and their r's are far enough away from 1 so that the Gaussian approximation is reasonable while maintaing a good power.

TABLE 3.5. Estimated power (by 100 simulation runs) of $Z_{W,r}(t)$ at 0.05 significance level. The first 100 observations are generated from $N_d(\mu_1, I_d)$ and the second 100 observations are generated from $N_d(\mu_2, I_d)$, where $\Delta = \|\mu_1 - \mu_2\|_2$

Location Alternatives					
d	10	30	50	70	100
Δ	0.31	0.60	0.77	0.96	1.13
$r = 1.3$	0.13	0.16	0.21	0.24	0.31
$r = 1.2$	0.13	0.22	0.34	0.40	0.47
$r = 1.1$	0.11	0.33	0.46	0.67	0.72
$r = 1.0$	0.12	**0.43**	**0.56**	**0.80**	**0.88**
$r = 0.9$	0.11	0.34	0.45	0.66	0.80
$r = 0.8$	0.12	0.25	0.27	0.38	0.49
$r = 0.7$	0.11	0.14	0.16	0.20	0.26

We now define two fast tests based on the asymptotic results. Let p_D , $p_{W,1.2}$, and $p_{W,0.8}$ be the approximated p-values of the test that reject for large values of $|Z_D(t)|$, $Z_{W,1.2}(t)$, and $Z_{W,0.8}(t)$, respectively.

- fGKCP$_1$: rejects the null hypothesis of homogeneity if $3\min(p_D, p_{W,1.2}, p_{W,0.8})$ is less than the significance level.
- fGKCP$_2$: rejects the null hypothesis of homogeneity if $2\min(p_{W,1.2}, p_{W,0.8})$ is less than the significance level.

It is expected that fGKCP$_1$ performs well for a wide range of alternatives, especially for scale alternatives due to $Z_D(t)$. Since $Z_W(t)$ is sensitive to location alternatives, we expect fGKCP$_2$ to be powerful for location alternatives. Furthermore, according to the simulation results, it turns out that fGKCP$_2$ can also detect variance changes to some extent as $r = 1.2, 0.8$ cover more types of alternatives than $r = 1$. When the null hypothesis is rejected, the location of change-point can be estimated by $\max_{n_0 \leq t \leq n_1}$ GKCP(t) so that the effects of $Z_D(t)$ and $Z_W(t)$ are fully utilized. Hence, as long as $\max_{n_0 \leq t \leq n_1} Z_D(t)$, $\max_{n_0 \leq t \leq n_1} Z_{W,1.2}(t)$, and $\max_{n_0 \leq t \leq n_1} Z_{W,0.8}(t)$ are computed, the tests can be conducted instantly.

REMARK 3.3.4. *We adopt the Bonferroni procedure for the fast tests to combine the advantages of each test statistic. To improve the power of the tests, the Simes procedure can be used and this also controls type I error well empirically (see Section 3.6).*

3.4. Performance of the new tests

We examine the performance of the new tests under various simulation settings. Each data sequence in the simulation is of length $n = 200$ with various dimensions d, where $y_1, \ldots, y_\tau \overset{iid}{\sim} F_0$ and $y_{\tau+1}, \ldots, y_n \overset{iid}{\sim} F_1$. Here, τ is the change-point. We consider the following setting:

- Multivariate Gaussian data Type I: $F_0 \sim N_d(\mathbf{0}_d, \Sigma)$ vs. $F_1 \sim N_d(a\mathbf{1}_d, \sigma^2\Sigma)$, where $\Delta = \|a\mathbf{1}_d\|_2$ and $\Sigma_{i,j} = 0.4^{|i-j|}$.
- Multivariate Gaussian data Type II: $F_0 \sim N_d(\mathbf{0}_d, \Sigma)$ vs. $F_1 \sim N_d(a\nu_d, \sigma^2\Sigma)$, where $\Delta = \|a\nu_d\|_2$, d-dimensional vector ν_d with half of it being zeros and half of it being 1's, and $\Sigma_{i,j} = 0.4^{|i-j|}$.
- Multivariate log-normal data: $F_0 \sim \exp(N_d(\mathbf{0}_d, \Sigma))$ vs. $F_1 \sim \exp(N_d(a\mathbf{1}_d, \Sigma))$, where $\Delta = \|a\mathbf{1}_d\|_2$ and $\Sigma_{i,j} = 0.4^{|i-j|}$.

In addition, we consider the multivariate Gaussian distribution with different structures and results are in Appendix B.4.

We simulate 100 datasets to estimate the power of the tests and the significance level is set to be 0.05 for all tests. To examine the empirical size of the test, we simulate 1,000 datasets. We also examine the accuracy of the estimated change-point location and the count where the location of estimated change-point is within 20 from the true change-point is provided in parentheses when the null is rejected.

It is usually hard to offer false positive controls as well as the estimation of the location of change-points. We compare the results for the new tests to the recent feasible kernel-based method, KCP [**2**], which can be implemented by `R` package `ecp` [**36**]. We also compare the new tests with other feasible nonparametric methods using interpoint distances (ECP) [**43**] and similarity graphs (GCP) [**13**], which can be implemented by `R` packages `ecp` and `gSeg`, respectively. Here, we approximate the p-value by 1,000 permutation for ECP and use the max-type method with 5-MST for GCP, following the suggestion in [**11**]. Lastly, we include the method using Fréchet means and variances (FCP) [**15**].

Table 3.6 and 3.7 show results for the multivariate Gaussian data with different means and/or variances. We see that KCP and ECP perform well for location alternatives, while they have considerable low or no power for scale alternatives. On the other hand, the new test GKCP performs very well for both location and scale alternatives and the fast tests, fGKCP$_1$ and fGKCP$_2$, also perform well. Other tests, GCP and FCP, do not work well for Gaussian settings.

TABLE 3.6. Estimated power of the tests for multivariate Gaussian data Type I. $n = 200$. Top 1 method and those higher than 95% of the top 1 are in bold

	Mean Change (τ at center)				Mean Change (τ at three quarter)			
d	100	500	1000	2000	100	500	1000	2000
Δ	1.20	1.90	2.40	3.13	1.30	2.12	2.68	3.39
fGKCP_1	50 (43)	68 (62)	78 (76)	**96 (95)**	45 (38)	67 (63)	82 (78)	93 (92)
fGKCP_2	58 (49)	73 (67)	84 (80)	**97 (96)**	51 (44)	73 (68)	87 (84)	**94 (94)**
GKCP	**75 (63)**	**88 (82)**	**95 (91)**	**99 (98)**	**64 (55)**	**84 (79)**	**95 (91)**	**97 (96)**
KCP	71 (61)	**85 (79)**	**93 (90)**	**98 (97)**	59 (50)	**81 (77)**	**92 (90)**	**97 (96)**
ECP	**76 (65)**	**89 (79)**	**96 (90)**	**99 (95)**	**61 (54)**	**84 (77)**	**93 (87)**	**98 (94)**
GCP	22 (9)	27 (14)	34 (20)	46 (32)	20 (10)	26 (15)	28 (20)	38 (28)
FCP	6 (1)	1 (0)	0 (0)	0 (0)	14 (4)	2 (0)	1 (0)	0 (0)

	Variance Change (τ at center)				Variance Change (τ at three quarter)			
d	100	500	1000	2000	100	500	1000	2000
σ^2	1.07	1.04	1.03	1.03	1.08	1.05	1.03	1.03
fGKCP_1	**46 (30)**	**68 (52)**	**79 (64)**	**93 (81)**	**46 (35)**	**69 (56)**	**75 (62)**	**89 (78)**
fGKCP_2	40 (25)	58 (43)	68 (54)	85 (73)	37 (28)	58 (47)	62 (51)	80 (71)
GKCP	41 (27)	**67 (51)**	**79 (63)**	**93 (80)**	44 (33)	**69 (55)**	**74 (61)**	**88 (78)**
KCP	18 (2)	15 (3)	12 (2)	7 (1)	18 (6)	13 (3)	8 (3)	12 (2)
ECP	5 (2)	6 (2)	6 (2)	6 (2)	5 (1)	5 (1)	4 (1)	6 (2)
GCP	27 (11)	40 (21)	49 (27)	64 (41)	13 (3)	19 (6)	20 (6)	26 (10)
FCP	13 (5)	0 (0)	0 (0)	0 (0)	9 (5)	0 (0)	0 (0)	0 (0)

	Both Change (τ at center)				Both Change (τ at three quarter)			
d	100	500	1000	2000	100	500	1000	2000
Δ	0.65	0.69	0.70	0.71	1.00	1.23	1.42	2.01
σ^2	1.06	1.04	1.03	1.03	1.06	1.04	1.03	1.03
fGKCP_1	**52 (36)**	**66 (49)**	**80 (63)**	**98 (91)**	**59 (49)**	**76 (67)**	**80 (70)**	**99 (95)**
fGKCP_2	**53 (37)**	63 (46)	75 (61)	**96 (89)**	**61 (51)**	**76 (68)**	**80 (69)**	**99 (95)**
GKCP	50 (35)	63 (47)	**78 (64)**	**98 (91)**	**62 (51)**	**76 (67)**	**80 (70)**	**99 (95)**
KCP	11 (7)	3 (1)	2 (1)	1 (0)	24 (20)	9 (7)	9 (6)	10 (9)
ECP	21 (13)	8 (5)	9 (4)	9 (4)	29 (24)	19 (15)	18 (14)	29 (24)
GCP	24 (10)	34 (16)	43 (22)	78 (56)	14 (4)	16 (4)	19 (5)	35 (16)
FCP	16 (7)	1 (0)	0 (0)	0 (0)	10 (4)	0 (0)	0 (0)	0 (0)

Table 3.8 shows results for the multivariate log normal data. Here, alternatives yield the changes in both the mean and variance of distributions. We see that the new tests exhibit high power not only for symmetric distributions but also for asymmetric distributions under moderate to high dimensions. However, KCP and GCP lose power in this case, while ECP still performs well. Compared with Gaussian settings, FCP exhibits high power, but it is outperformed by the new tests.

TABLE 3.7. Estimated power of the tests for multivariate Gaussian data Type II. $n = 200$

	Mean Change (τ at center)				Mean Change (τ at three quarter)			
d	100	500	1000	2000	100	500	1000	2000
Δ	0.99	2.37	2.46	3.16	0.99	2.05	2.90	3.64
$fGKCP_1$	17 (10)	39 (31)	57 (51)	84 (81)	22 (15)	49 (44)	82 (79)	**97 (96)**
$fGKCP_2$	21 (13)	46 (36)	64 (57)	89 (87)	27 (19)	55 (50)	87 (84)	**98 (97)**
GKCP	**34 (24)**	**64 (52)**	**81 (72)**	**97 (94)**	**33 (24)**	**66 (61)**	**94 (90)**	**99 (99)**
KCP	31 (22)	59 (49)	78 (70)	**94 (92)**	**32 (24)**	**65 (61)**	90 (87)	**98 (98)**
ECP	32 (21)	**63 (52)**	**85 (75)**	**98 (90)**	27 (21)	**67 (60)**	**98 (92)**	**99 (92)**
GCP	12 (3)	18 (6)	24 (11)	31 (20)	12 (3)	19 (10)	41 (30)	45 (35)
FCP	4 (0)	0 (0)	0 (0)	0 (0)	6 (1)	1 (0)	0 (0)	0 (0)

	Both Change (τ at center)				Both Change (τ at three quarter)			
d	100	500	1000	2000	100	500	1000	2000
Δ	0.65	0.69	0.70	0.71	0.70	0.87	1.00	1.42
σ^2	1.06	1.04	1.03	1.03	1.06	1.04	1.03	1.03
$fGKCP_1$	**46 (30)**	**63 (46)**	**79 (63)**	**99 (90)**	**39 (29)**	**65 (52)**	**75 (61)**	**98 (93)**
$fGKCP_2$	43 (27)	58 (43)	72 (56)	**96 (88)**	**40 (30)**	61 (49)	71 (59)	**96 (91)**
GKCP	42 (27)	**61 (45)**	**78 (61)**	90 (90)	**41 (30)**	**64 (51)**	**73 (59)**	**98 (93)**
KCP	5 (2)	2 (1)	1 (0)	1 (0)	11 (7)	5 (2)	4 (2)	4 (3)
ECP	11 (5)	7 (3)	5 (2)	7 (2)	14 (8)	9 (4)	10 (5)	10 (6)
GCP	23 (9)	34 (17)	47 (27)	77 (54)	12 (2)	15 (4)	17 (4)	33 (15)
FCP	12 (6)	0 (0)	0 (0)	0 (0)	6 (2)	0 (0)	0 (0)	0 (0)

TABLE 3.8. Estimated power of the tests for multivariate log-normal data. $n = 200$

	Mean Change (τ at center)				Mean Change (τ at three quarter)			
d	100	500	1000	2000	100	500	1000	2000
Δ	1.20	1.90	2.30	3.04	1.35	2.12	2.65	3.42
$fGKCP_1$	47 (35)	70 (57)	81 (71)	**96 (90)**	47 (38)	67 (58)	83 (77)	**95 (92)**
$fGKCP_2$	55 (41)	76 (63)	85 (75)	**97 (91)**	53 (43)	73 (63)	86 (80)	**97 (94)**
GKCP	63 (48)	**83 (68)**	**91 (80)**	**99 (93)**	**65 (53)**	**86 (75)**	**95 (88)**	**99 (96)**
KCP	20 (16)	6 (5)	10 (9)	5 (4)	20 (16)	5 (4)	12 (9)	7 (5)
ECP	**69 (52)**	**85 (72)**	**91 (80)**	**98 (91)**	**62 (54)**	81 (70)	90 (82)	**97 (90)**
GCP	32 (12)	33 (7)	32 (6)	36 (8)	17 (3)	12 (0)	10 (0)	12 (0)
FCP	32 (18)	57 (40)	69 (53)	83 (70)	36 (24)	55 (41)	66 (55)	77 (67)

The empirical size of the tests at 0.05 significance level for the multivariate Gaussian and log-normal data is presented in Table 3.9. We see that the new tests control the type I error rate well. However, KCP relies on a cumbersome method, such as the line search, to find the suitable penalty constant and this step is very sensitive, so it is difficult to control the type I error well.

TABLE 3.9. Empirical size of the tests at 0.05 significance level. $n = 200$

	Multivariate Gaussian				Multivariate log-normal			
d	100	500	1000	2000	100	500	1000	2000
fGKCP$_1$	0.032	0.047	0.047	0.037	0.038	0.039	0.041	0.036
fGKCP$_2$	0.043	0.057	0.055	0.052	0.051	0.050	0.050	0.055
GKCP	0.052	0.049	0.053	0.049	0.049	0.051	0.038	0.056
KCP	0.067	0.045	0.060	0.040	0.093	0.040	0.081	0.067
ECP	0.054	0.043	0.056	0.045	0.054	0.057	0.051	0.042
GCP	0.072	0.073	0.069	0.077	0.090	0.132	0.098	0.113
FCP	0.018	0.001	0.000	0.000	0.053	0.051	0.036	0.027

TABLE 3.10. Average runtimes in seconds from 10 simulations for each length n. All experiments were run by R on 2.2 GHz Intel Core i7

n	200	400	600	800	1000	2000
fGKCP$_1$	0.034	0.177	0.642	1.507	2.997	24.07
KCP	0.218	3.282	17.27	53.32	132.0	2161.83
ECP	1.440	5.053	12.00	19.22	30.25	144.38
GCP	0.009	0.041	0.095	0.185	0.342	2.008
FCP	26.57	94.37	209.1	369.5	544.0	2251.7

We also compare the computational cost of the tests and check runtimes of the tests for Gaussian data under various n. Table 3.10 shows average runtimes for each length n when $d = 100$. Although KCP utilizes a dynamic programming, we see that the fast test based on the new scan statistics is faster than KCP. Notice that the average runtimes of KCP in Table 3.10 only present the actual testing runtimes. If we consider the runtime for the parameter optimization procedure simultaneously, KCP is computationally infeasible to run. FCP is as slows as KCP. ECP relies on the permutation approach, so it is slower than the new tests. Though GCP is the fastest, the new test is comparably faster than other tests with great performance.

The overall pattern of the simulation results shows that the new tests exhibit high power for a wide range of alternatives. Unlike the existing kernel change-point detection method, the new tests are effective and easy to implement without any time-consuming procedures, such as parameter tuning, as long as the kernel matrix is computed. In practice, fGKCP$_1$ and fGKCP$_2$ would be preferred as the fast tests. If the test result is ambiguous and further investigation is needed, the permuation test of GKCP would also be useful.

3.5. A real data example

We apply the new tests to the phone-call network dataset. The MIT Media Laboratory studied with 87 subjects who used mobile phones with a pre-installed device that can record call logs. The study lasted for 330 days from July 2004 to June 2005 [**16**]. We use it to illustrate the new tests by detecting any change in the phone-call pattern among subjects over time. This can be viewed as the change of friendship along time.

We bin the phone-calls by day and we construct $n = 330$ of networks in total with 87 subjects as nodes. We encode each network by the adjacency matrix with value 1 for element (i, j) if subject i called j on day t and 0 otherwise. We then construct the Gaussian kernel matrix with the median heuristic.

We apply the single change-point detection method to the phone-call network dataset recursively in order to detect all possible change-points. Since this dataset has a lot of noise, we focus on the estimated change-points with p-value less than 0.001.

TABLE 3.11. Estimated change-points

	Days (t)				
Estimated change-points	53	90	141	251	293

Table 3.11 shows the estimated change-points until the new tests do not reject the null. In this analysis, all new tests yield the same results about whether to reject the null or not in each iteration and the estimated locations of change-point. Since the underlying distribution of the dataset is unknown, we perform a more sanity check with the kernel matrix of the whole period (Figure 3.4). It is evident that there are some changes occuring in the period and they match the results of the new test fairly well.

We also compare the results of the new tests with their nearby academic events (Table 3.12). We see that the new tests detect change-points at around the beginning of the Fall term, family weekend, and the end of the Fall term that could cause phone-call pattern changes among subjects. The new tests also detect the Spring break and the end of the Spring term. These are all reasonable times when there are some significant changes in phone-call pattern.

3.6. Discussion and conclusion

We proposed the new kernel-based scan statistic, GKCP, for the testing and estimation of change-points. The new tests are versatile and effective for a wide range of alternatives. Analytical p-value approximations

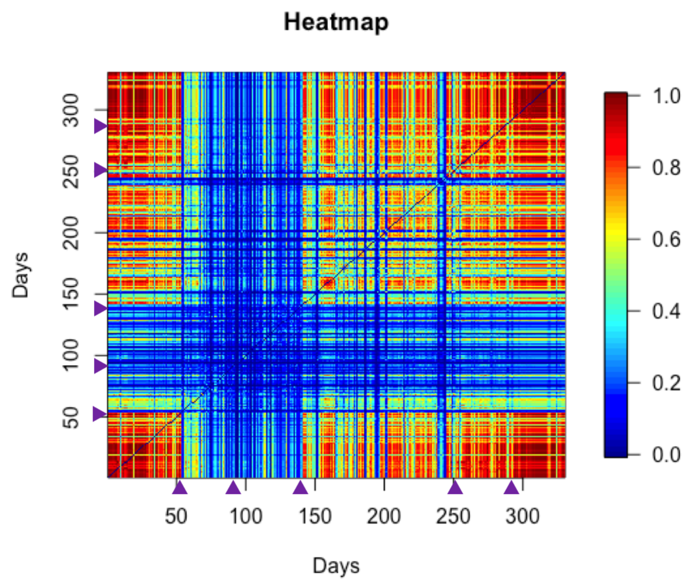


FIGURE 3.4. The heatmap of the kernel matrix corresponding to 330 networks. Purple triangles in the heatmap indicate estimated change-points.

TABLE 3.12. Estimated change-points and nearby academic events. The dates of the academic events are from the 2011-2012 academic calendar of MIT that is the closest academic calendar of MIT to 2004-2005 available online

Estimated change-points	Nearby academic events
$n = 53$: 2004/09/10	2004/09/07: Fall classes begin
$n = 90$: 2004/10/17	2004/10/14: Family weekend
$n = 141$: 2004/12/07	2004/12/14: Last day of Fall classes
$n = 251$: 2005/03/27	2005/03/26: Spring break begins
$n = 293$: 2005/05/08	2005/05/17: Last day of Spring classes

based on the limiting distributions were derived and the skewness-corrected versions were proposed. We also proposed two fast tests, fGKCP$_1$ and fGKCP$_2$, based on asymptotic results. The new tests exhibit superior power and work well particularly for high-dimensional settings. In practice, we recommend to use fGKCP$_1$ and fGKCP$_2$ as they are fast to implement. When the results are ambiguous, the permutation test based on GKCP could be run for the final conclusion.

Since the Bonferroni procedure is a bit conservative, the Simes procedure may be used to improve the power of the fast tests (fGKCP$_1$, fGKCP$_2$). Let $p_{(1)} \leq p_{(2)} \leq p_{(3)}$ be the ordered p-values of p_D, $p_{W,1.2}$, and $p_{W,0.8}$ and $p'_{(1)} \leq p'_{(2)}$ be the ordered p-values of $p_{W,1.2}$ and $p_{W,0.8}$. Then, the fast tests are defined that

- fGKCP$_1$-Simes: rejects the null hypothesis of homogeneity if $\min(3p_{(1)}, 1.5p_{(2)}, p_{(3)})$ is less than the significance level.
- fGKCP$_2$-Simes: rejects the null hypothesis of homogeneity if $\min(2p'_{(1)}, p'_{(2)})$ is less than the significance level.

It has been shown that the Simes procedure is exact under independent distributions, while it becomes conservative under positively dependent distributions and slightly liberal under negatively dependency. There have been a lot of works to prove the validity of the Simes test under dependency [**6**, **7**, **19**, **24**, **25**, **34**, **52**, **53**, **54**], but they are restricted to special cases. Nevertheless, the Simes test is widely used in many applications. [**48**] proved that the overall relative deviation of the Simes p-value from the true p-value is strongly bounded and showed that, although the Simes procedure may be liberal, it cannot be consistently. It is therefore reasonably expected that the Simes p-value will be asymptotically valid in most practical cases. Table 3.13 shows the empirical size of the tests for the multivariate Gaussian and log-normal data used in

TABLE 3.13. Empirical size of the tests at 0.05 significance level. $n = 200$

	Multivariate Gaussian				Multivariate log-normal			
d	100	500	1000	2000	100	500	1000	2000
fGKCP$_1$	0.032	0.047	0.047	0.037	0.030	0.039	0.036	0.035
fGKCP$_1$-Simes	0.036	0.048	0.048	0.038	0.031	0.040	0.036	0.036
fGKCP$_2$	0.043	0.057	0.055	0.052	0.051	0.052	0.055	0.052
fGKCP$_2$-Simes	0.044	0.057	0.057	0.052	0.051	0.052	0.055	0.052

Section 3.4. We see that the Simes procedure also controls type I error well. Hence, if we want to focus on the performance of the test and improve the power of the fast tests, the Simes procedure would be useful for the fast tests.

The new methods detect the most significant single change-point or changed-interval in the sequence. If two or more changes are presented in the sequence, the new methods can be applied recursively with their own advantages using techniques, such as binary segmentation, circular binary segmentation, or wild binary segmentation [**23**, **45**, **62**].

APPENDIX A

Appendix for Chapter 2

A.1. Proof to Theorem 2.2.1

Under the permutation null distribution, we have

$$
\begin{aligned}
\mathsf{E}(\alpha) &= \frac{1}{m(m-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}\mathsf{P}(g_i = g_j = 1) \\
&= \frac{1}{m(m-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}\frac{m(m-1)}{N(N-1)} \\
&= \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}, \\
\mathsf{E}(\alpha^2) &= \frac{1}{m^2(m-1)^2}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}\sum_{u=1}^{N}\sum_{v=1,v\neq u}^{N} k_{uv}\mathsf{P}(g_i = 1, g_j = 1, g_u = 1, g_v = 1) \\
&= \frac{2}{m^2(m-1)^2}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}^2\mathsf{P}(g_i = 1, g_j = 1) \\
&\quad + \frac{4}{m^2(m-1)^2}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\sum_{u=1,u\neq i,u\neq j}^{N} k_{ij}k_{iu}\mathsf{P}(g_i = 1, g_j = 1, g_u = 1) \\
&\quad + \frac{1}{m^2(m-1)^2}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\sum_{u=1,u\neq i,u\neq j}^{N}\sum_{v=1,v\neq i,v\neq j,v\neq u}^{N} k_{ij}k_{uv}\mathsf{P}(g_i = 1, g_j = 1, g_u = 1, g_v = 1) \\
&= \frac{2}{m^2(m-1)^2}A\frac{m(m-1)}{N(N-1)} + \frac{4}{m^2(m-1)^2}B\frac{m(m-1)(m-2)}{N(N-1)(N-2)} \\
&\quad + \frac{1}{m^2(m-1)^2}C\frac{m(m-1)(m-2)(m-3)}{N(N-1)(N-2)(N-3)} \\
&= \frac{1}{m^2(m-1)^2}\left(2A\frac{m(m-1)}{N(N-1)} + 4B\frac{m(m-1)(m-2)}{N(N-1)(N-2)} + C\frac{m(m-1)(m-2)(m-3)}{N(N-1)(N-2)(N-3)}\right).
\end{aligned}
$$

Then, $\Sigma_{\alpha,\beta(1,1)} = \mathsf{E}(\alpha^2) - \mathsf{E}(\alpha)^2$ follows readily. The expectation and variance of β can be done in a similar manner. For the covariance between α and β, we have

$$\begin{aligned}
\mathsf{E}(\alpha\beta) &= \frac{1}{mn(m-1)(n-1)} \sum_{i=1}^{N} \sum_{j=1,j\neq i}^{N} \sum_{u=1}^{N} \sum_{v=1,v\neq u}^{N} k_{ij}k_{uv}\mathsf{P}(g_i = g_j = 1,\ g_u = g_v = 2) \\
&= \frac{1}{mn(m-1)(n-1)} \sum_{i=1}^{N} \sum_{j=1,j\neq i}^{N} \sum_{u=1,u\neq i,u\neq j}^{N} \sum_{v=1,v\neq i,v\neq j,v\neq u}^{N} k_{ij}k_{uv} \frac{m(m-1)n(n-1)}{N(N-1)(N-2)(N-3)} \\
&= \frac{C}{N(N-1)(N-2)(N-3)}.
\end{aligned}$$

Then, $\Sigma_{\alpha,\beta(1,2)} = \mathsf{E}(\alpha\beta) - \mathsf{E}(\alpha)\mathsf{E}(\beta)$ follows readily.

A.2. Proof to Theorem 2.2.2

The components in $\Sigma_{\alpha,\beta}$ can be rewritten as

$$\begin{aligned}
\Sigma_{\alpha,\beta(1,1)} &= \frac{1}{m^2(m-1)^2} \cdot \frac{mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)} \\
&\quad \times \left(2A + \frac{m-2}{n-1}\left((4A+4B) - \frac{4(2A+4B+C)}{N}\right) - \frac{2}{N(N-1)}(2A+4B+C)\right), \\
\Sigma_{\alpha,\beta(2,2)} &= \frac{1}{n^2(n-1)^2} \cdot \frac{mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)} \\
&\quad \times \left(2A + \frac{n-2}{m-1}\left((4A+4B) - \frac{4(2A+4B+C)}{N}\right) - \frac{2}{N(N-1)}(2A+4B+C)\right), \\
\Sigma_{\alpha,\beta(1,2)} &= \frac{1}{mn(m-1)(n-1)} \cdot \frac{mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)} \\
&\quad \times \left(2A - \left((4A+4B) - \frac{4(2A+4B+C)}{N}\right) - \frac{2}{N(N-1)}(2A+4B+C)\right).
\end{aligned}$$

Hence, the determinant of $\Sigma_{\alpha,\beta}$ can be computed as

$$\begin{aligned}
|\Sigma_{\alpha,\beta}| &= \frac{1}{mn(m-1)^2(n-1)^2N(N-1)(N-2)}\left((4A+4B) - \frac{4(2A+4B+C)}{N}\right) \\
&\quad \times \left((N-2)2A + \frac{2}{N-1}(2A+4B+C) - (4A+4B)\right).
\end{aligned}$$

Notice that $A = tr(\tilde{K}^2)$, $A + B = \sum_{i=1}^{N}\sum_{j=1}^{N}(\tilde{K}^2)_{ij}$, and $2A + 4B + C = \left(\sum_{i=1}^{N}\sum_{j=1}^{N}\tilde{K}_{ij}\right)^2$, where $\tilde{K} = K - I$ with K being the kernel matrix.

We first figure out the term

$$(A+B) - \frac{(2A+4B+C)}{N} = \sum_{i=1}^{N}\sum_{j=1}^{N}(\tilde{K}^2)_{ij} - \frac{\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\tilde{K}_{ij}\right)^2}{N}.$$

It is not hard to show that

$$\sum_{i=1}^{N}\sum_{j=1}^{N}(\tilde{K}^2)_{ij} - \frac{\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\tilde{K}_{ij}\right)^2}{N} = \sum_{i=1}^{N}\sum_{j=1}^{N}(K^2)_{ij} - \frac{\left(\sum_{i=1}^{N}\sum_{j=1}^{N}K_{ij}\right)^2}{N}.$$

Let $M = K - h\mathbf{1}_N\mathbf{1}_N^t$, where h is the mean of all elements in K. Then, $\sum_{i=1}^{N}\sum_{j=1}^{N} M_{ij} = 0$. Since

$$K^2 = \left(M + h\mathbf{1}_N\mathbf{1}_N^t\right)^2 = M^2 + 2hNV\mathbf{1}_N^t + h^2N\mathbf{1}_N\mathbf{1}_N^t,$$

where $V = (\bar{M}_{1\cdot}, \ldots, \bar{M}_{N\cdot})^t$ and $\bar{M}_{i\cdot} = \frac{1}{N}\sum_{j=1}^{N} M_{ij}$ for $i = 1, \ldots, N$. We have

$$\sum_{i=1}^{N}\sum_{j=1}^{N}(K^2)_{ij} = \sum_{i=1}^{N}\sum_{j=1}^{N}(M^2)_{ij} + h^2N^3,$$
$$\sum_{i=1}^{N}\sum_{j=1}^{N}(K)_{ij} = hN^2.$$

Hence,

$$(A+B) - \frac{(2A+4B+C)}{N} = \sum_{i=1}^{N}\sum_{j=1}^{N}(M^2)_{ij}.$$

From the Cauchy$-$Schwarz inequality, we have

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N}(M^2)_{ij} &= \sum_{i=1}^{N}\left(\sum_{j=1, j\neq i}^{N} M_{ij}\right)^2 - N(1-h)^2 \\ &\geq \frac{1}{N}\left(\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} M_{ij}\right)^2 - N(1-h)^2 = 0.\end{aligned}$$

Here, the equality holds only when $\sum_{j\neq 1}^{N} M_{1j} = \ldots = \sum_{j\neq N}^{N} M_{Nj}$, which leads to the first condition that $\sum_{j=1, j\neq i}^{N} k_{ij}$ is constant for all $i = 1, \ldots, n$.

We next figure out the term

$$(N-2)2A + \frac{2}{N-1}(2A+4B+C) - (4A+4B)$$

$$= (N-2)2tr(\tilde{K}^2) + \frac{2}{N-1}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\tilde{K}_{ij}\right)^2 - 4\sum_{i=1}^{N}\sum_{j=1}^{N}(\tilde{K}^2)_{ij}$$

$$= 2(N-2)\left(tr(M^2) - \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(M^2)_{ij}}{N-2} - \frac{N^2}{N-1}(1-h)^2\right).$$

It is not hard to show that

$$tr(M^2) - \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(M^2)_{ij}}{N-2} - \frac{N^2}{N-1}(1-h)^2$$

$$= \sum_{i=1}^{N}\sum_{j\neq i}^{N}(M_{ij} - \bar{M}_{i\cdot})^2 - \frac{N(N-1)}{N-2}\sum_{i=1}^{N}(\bar{M}_{i\cdot} - \bar{M}_{\cdot\cdot})^2, \tag{A.1}$$

where $\bar{M}_{i\cdot} = \frac{1}{N-1}\sum_{j\neq i}^{N} M_{ij}$, $\bar{M}_{\cdot i} = \frac{1}{N-1}\sum_{j\neq i}^{N} M_{ji}$, and $\bar{M}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{N} M_{i\cdot}$. Notice that

$$\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}(M_{ij} - \bar{M}_{i\cdot})^2 = \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}(M_{ij} - \bar{M}_{\cdot j})^2,$$

$$\sum_{j=1,j\neq i}^{N}(M_{ij} - \bar{M}_{\cdot j}) = (N-1)\bar{M}_{i\cdot} - (N\bar{M}_{\cdot\cdot} - \bar{M}_{i\cdot}) = N(\bar{M}_{i\cdot} - \bar{M}_{\cdot\cdot}),$$

since $\bar{M}_{i\cdot} = \bar{M}_{\cdot i}$. Hence,

$$\sum_{j=1,j\neq i}^{N}(M_{ij} - \bar{M}_{\cdot j})^2 = \sum_{j=1,j\neq i}^{N}\left(M_{ij} - \bar{M}_{\cdot j} - \frac{N(\bar{M}_{i\cdot} - \bar{M}_{\cdot\cdot})}{N-1}\right)^2 + \frac{N^2(\bar{M}_{i\cdot} - \bar{M}_{\cdot\cdot})^2}{N-1}$$

$$= \sum_{j=1,j\neq i}^{N}(M_{ij} - \bar{M}_{\cdot j} - \bar{M}_{i\cdot} + \bar{M}_{\cdot\cdot})^2 + (N+1)(\bar{M}_{i\cdot} - \bar{M}_{\cdot\cdot})^2.$$

Then (A.1) is equivalent to

$$\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}(M_{ij} - \bar{M}_{\cdot j} - \bar{M}_{i\cdot} + \bar{M}_{\cdot\cdot})^2 - \frac{2}{N-2}\sum_{i=1}^{N}(\bar{M}_{i\cdot} - \bar{M}_{\cdot\cdot})^2$$

$$\triangleq \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} Q_{ij}^2 - \frac{2}{N-2}\sum_{i=1}^{N} R_i^2, \tag{A.2}$$

where $Q_{ij} = Q_{ji}$ and $\sum_{i=1}^{N} R_i = \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} Q_{ij} = 0$.

Now, we use the method of induction to show

$$\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} Q_{ij}^2 - \frac{2}{N-2}\sum_{i=1}^{N} R_i^2 \geq 0. \tag{A.3}$$

It is easy to check that (A.3) holds for $N = 3$. Assume that (A.3) holds for N and consider the $N + 1$ case. Define $e_i = Q_{i,N+1}$ and $H = \frac{1}{N}\sum_{i=1}^{N} e_i$. Let $\tilde{Q}_{ij} = Q_{ij} + \frac{2H}{N-1}$ and $\tilde{R}_i = \sum_{j\neq i}^{N} \tilde{Q}_{ij}$. Since

$$\sum_{i=1}^{N} \tilde{R}_i = \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} \tilde{Q}_{ij} = \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} Q_{ij} + 2NH = 0 - 2\sum_{i=1}^{N} e_i + 2NH = 0$$

and $\tilde{Q}_{ij} = \tilde{Q}_{ji}$, we have

$$\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} \tilde{Q}_{ij}^2 - \frac{2}{N-2}\sum_{i=1}^{N} \tilde{R}_i^2 \geq 0. \tag{A.4}$$

as $\tilde{Q}_{ij}$ and $\tilde{R}_i$ satisfy the conditions for (A.3).

For the $N + 1$ case, we have

$$\begin{aligned}\sum_{i=1}^{N+1}\sum_{j=1,j\neq i}^{N+1} Q_{ij}^2 &= \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\left(\tilde{Q}_{ij} - \frac{2H}{N-1}\right)^2 + 2\sum_{i=1}^{N} e_i^2 \\ &= \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} \tilde{Q}_{ij}^2 + \sum_{i=1}^{N} e_i^2 + \frac{4NH^2}{N-1}\end{aligned}$$

since $\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} \tilde{Q}_{ij} = 0$. We also have

$$\sum_{i=1}^{N+1} R_i^2 = \sum_{i=1}^{N} R_i^2 + \left(\sum_{i=1}^{N} e_i\right)^2.$$

Since

$$\sum_{i=1}^{N} R_i^2 = \sum_{i=1}^{N}\left(\sum_{j=1,j\neq i}^{N+1} Q_{ij}\right)^2 = \sum_{i=1}^{N}\left(\sum_{j=1,j\neq i}^{N}\left(\tilde{Q}_{ij} - \frac{2H}{N-1}\right) + e_i\right)^2 = \sum_{i=1}^{N}(\tilde{R}_i - 2H + e_i)^2,$$

we have

$$\sum_{i=1}^{N+1} R_i^2 = \sum_{i=1}^{N}(\tilde{R}_i + e_i - 2H)^2 + \left(\sum_{i=1}^{N} e_i\right)^2 = \sum_{i=1}^{N} \tilde{R}_i^2 + 2\sum_{i=1}^{N} e_i\tilde{R}_i + \sum_{i=1}^{N} e_i^2 + N^2H^2.$$

Hence, for $N+1$ case, we have

$$
\begin{aligned}
&\sum_{i=1}^{N+1}\sum_{j=1,j\neq i}^{N+1} Q_{ij}^2 - \frac{2}{N-1}\sum_{i=1}^{N+1} R_i^2 \\
&= \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} \tilde{Q}_{ij}^2 + \sum_{i=1}^{N} e_i^2 + \frac{4NH^2}{N-1} - \frac{2}{N-1}\left(\sum_{i=1}^{N}\tilde{R}_i^2 + 2\sum_{i=1}^{N} e_i\tilde{R}_i + \sum_{i=1}^{N} e_i^2 + N^2H^2\right) \\
&\geq \left(\frac{2}{N-1} - \frac{2}{N-2}\right)\sum_{i=1}^{N}\tilde{R}_i^2 - \frac{4}{N-1}\sum_{i=1}^{N} e_i\tilde{R}_i + \left(1 - \frac{2}{N-1}\right)\sum_{i=1}^{N} e_i^2 + \frac{4NH^2}{N-1} - \frac{2N^2H^2}{N-1} \\
&= \frac{2}{(N-1)(N-2)}\sum_{i=1}^{N}\left(\tilde{R}_i - (N-2)e_i\right)^2 - \frac{2N(N-2)}{N-1}H^2 \\
&\geq \frac{2}{(N-1)(N-2)}\frac{1}{N}\left(\sum_{i=1}^{N}\left(\tilde{R}_i - (N-2)e_i\right)\right)^2 - \frac{2N(N-2)}{N-1}H^2 \\
&= \frac{2}{(N-1)(N-2)}\frac{1}{N}\Big((N-2)NH\Big)^2 - \frac{2N(N-2)}{N-1}H^2 \\
&= 0,
\end{aligned}
$$

where the first inequality comes from (A.4) and the second inquality comes from the Cauchy−Schwarz inequality.

The condition that makes the second inequality an equal sign holds only when $\tilde{R}_i - (N-2)e_i$ are all the same for $i = 1, \ldots, N$. This is equivalent to that

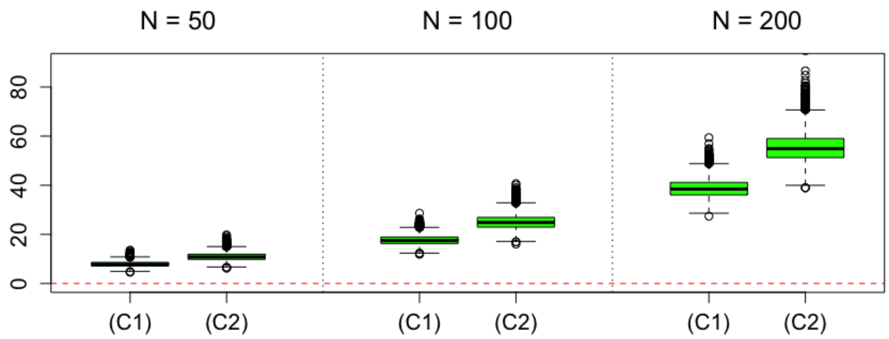
$$
\sum_{j=1,j\neq i}^{N} k_{ij} - (N-2)k_{iN} \tag{A.5}
$$

are the same for all $i = 1, \ldots, N-1$.

A.3. Illustration on conditions in Theorem 2.2.2

It is difficult to simplify the descriptions of (C1) and (C2) in Theorem 2.2.2 further. However, these two corner cases are rare to happen. To illustrate this, we simulate data from the standard multivariate Gaussian distribuion for different N's and dimension d's. For each combination of N and d, we randomly generate 10,000 datasets; and for each datasets, we compute the range of $\{\sum_{j=1,j\neq i}^{N} k_{ij}\}_{i=1,\ldots,N}$ and the range of $\{\sum_{j=1,j\neq i}^{N} k_{ij} - (N-2)k_{iN}\}_{i=1,\ldots,N-1}$. The corner cases happen when this range is 0. The boxplots of the ranges over 10,000 randomly simulated datasets are shown in Figure A.1. We see that none of the ranges reaches 0, and the range becomes larger as N increases. In practice, when applying the method, it is easy to check whether any of these two corner cases happen by plugging in the values of k_{ij}'s directly.

$d = 50$:



$d = 1000$:

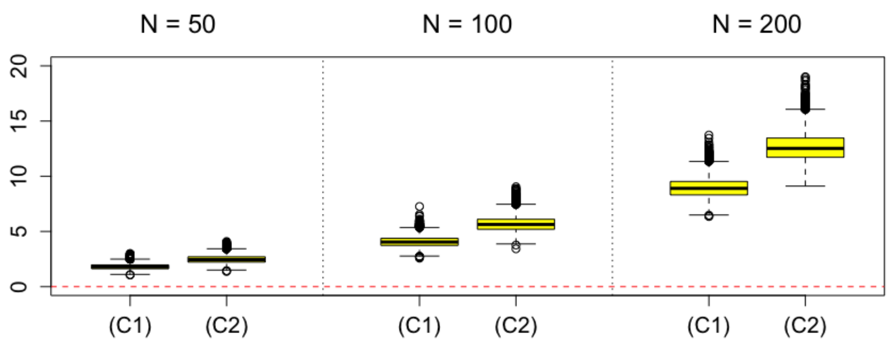


FIGURE A.1. Boxplots of the ranges of $\{\sum_{j=1,j\neq i}^{N} k_{ij}\}_{i=1,\ldots,N}$ (C1) and $\{\sum_{j=1,j\neq i}^{N} k_{ij} - (N-2)k_{iN}\}_{i=1,\ldots,N-1}$ (C2) in 10,000 simulation runs under different N's and d's.

A.4. Proof to Theorem 2.3.1

Let $U = (\alpha, \beta)^T$ and $V = \begin{pmatrix} m/N & n/N \\ m(m-1) & -n(n-1) \end{pmatrix}$. It is easy to show that V is invertible when $N > 2$. Then,

$$\text{GPK} = (U - \mathsf{E}(U))^T \, \Sigma_{\alpha,\beta}^{-1} (U - \mathsf{E}(U)) = (V(U - \mathsf{E}(U)))^T \left(V\Sigma_{\alpha,\beta}V^T\right)^{-1} (V(U - \mathsf{E}(U)))\,.$$

It is not hard to show that $V\Sigma_{\alpha,\beta}V^T$ is a 2×2 diagonal matrix with the first and the second diagonal elements $\frac{m^2}{N^2}\Sigma_{\alpha,\beta(1,1)} + \frac{2mn}{N^2}\Sigma_{\alpha,\beta(1,2)} + \frac{n^2}{N^2}\Sigma_{\alpha,\beta(2,2)}$ and $m^2(m-1)^2\Sigma_{\alpha,\beta(1,1)} - 2mn(m-1)(n-1)\Sigma_{\alpha,\beta(1,2)} +$

$n^2(n-1)^2\Sigma_{\alpha,\beta(2,2)}$, respectively. Then,

$$
\begin{aligned}
\text{GPK} = &\left(\frac{\frac{m}{N}\alpha + \frac{n}{N}\beta - \left(\mathsf{E}(\frac{m}{N}\alpha) + \mathsf{E}(\frac{n}{N}\beta)\right)}{\sqrt{\mathsf{Var}(\frac{m}{N}\alpha + \frac{n}{N}\beta)}}\right)^2 \\
&+ \left(\frac{m(m-1)\alpha - n(n-1)\beta - \left(\mathsf{E}(m(m-1)\alpha) - \mathsf{E}(n(n-1)\beta)\right)}{\sqrt{\mathsf{Var}(m(m-1)\alpha - n(n-1)\beta)}}\right)^2 \\
=&Z_W^2 + Z_D^2.
\end{aligned}
$$

A.5. Proof to Theorem 2.3.2

First, we observe that D can be expressed as a double-indexed permutation statistic as follows:

$$D = m(m-1)\alpha - n(n-1)\beta = \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}b_{ij}, \tag{A.6}$$

where

$$b_{ij} = \begin{cases} 1 & \text{if observations } i \text{ and } j \text{ are from Sample } X, \\ \text{-1} & \text{if observations } i \text{ and } j \text{ are from Sample } Y, \\ 0 & \text{otherwise.} \end{cases}$$

We define the mean-centered k'_{ij} and b'_{ij} for $i \neq j$,

$$k'_{ij} = k_{ij} - \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}, \quad b'_{ij} = b_{ij} - \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} b_{ij},$$

and let $k'_{ii} = b'_{ii} = 0$ for $i = 1, \ldots, N$. Then, $\sum_{i,j=1}^{N} k'_{ij} = \sum_{i,j=1}^{N} b'_{ij} = 0$. Also,

$$
\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}b_{ij} &= \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\left(k'_{ij} + \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k_{ij}\right)\left(b'_{ij} + \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} b_{ij}\right) \\
&= \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k'_{ij}b'_{ij} + E,
\end{aligned}
$$

where E is a constant that does not change under permutation. Hence, it is sufficient to show the asymptotic normality of $\sum_{i,j=1}^{N} k'_{ij}b'_{ij}$.

The asymptotic distribution of the double-indexed permutation statistic of the form (A.6) has been studied in [**14**]. Later [**21**] claimed that Daniels' conditions for the asymptotic normality can be weakened and proposed relexed conditions. However, Zhu and Chen (2021) recently study the conditions for the asymptotic normality and they claim the conditions presented in [**21**] are not sufficient. Instead, they show the following conditions are sufficient for the asymptotic normality of $\sum_{i,j=1}^{N} k'_{ij} b'_{ij}$:

$$\sum_{i,j=1}^{N} k'_{ij} = \sum_{i,j=1}^{N} b'_{ij} = 0, \tag{A.7}$$

$$\sum_{i,j,k=1}^{N} k'_{ij} k'_{ik} = O(N^3 k'_{\max}) \tag{A.8}$$

$$\sum_{i,j,k=1}^{N} b'_{ij} b'_{ik} = O(N^3 b'_{\max}), \tag{A.9}$$

where $k'_{\max}$ and $b'_{\max}$ are the largest order of k'_{ij} and b'_{ij}, respectively.

Equation (B.9) directly holds by the definiton. Let

$$k_{i\cdot} = \sum_{j=1, j\neq i}^{N} k_{ij}, \;\; \dot{k} = \frac{1}{N}\sum_{i=1}^{N} k_{i\cdot}, \;\; k'_{i\cdot} = \sum_{j=1}^{N} k'_{ij},$$
$$b_{i\cdot} = \sum_{j=1, j\neq i}^{N} b_{ij}, \;\; \dot{b} = \frac{1}{N}\sum_{i=1}^{N} b_{i\cdot}, \;\; b'_{i\cdot} = \sum_{j=1}^{N} b'_{ij}.$$

Then,

$$k'_{i\cdot} = k_{i\cdot} - \dot{k},$$
$$b'_{i\cdot} = b_{i\cdot} - \dot{b}.$$

We have

$$\sum_{i,j,k=1}^{N} k'_{ij} k'_{ik} = \sum_{i=1}^{N} k_{i\cdot}^2 - N\dot{k}^2 - \sum_{i,j=1}^{N} k_{ij}^2 + \frac{N}{N-1}\dot{k}^2.$$

Note that $k'_{\max} = k_{\max}$. When $k_{ij} = O(1)\ \forall i, j$, we have $k_{i.} = O(N)$, $\dot{k} = O(N)$, and $k_{\max} = O(1)$. Since the term $\sum_{i=1}^{N} k_{i\cdot}^2 - N\dot{k}^2$ dominates others, equation (A.8) holds when

$$\sum_{i=1}^{N} k_{i\cdot}^2 - N\dot{k}^2 = O\left(\sum_{i=1}^{N} k_{i\cdot}^2\right). \tag{A.10}$$

Let $m/N \to p$ and $n/N \to q$ with p, q constants and $0 < p, q < 1$, $p + q = 1$ as $N \to \infty$. Similarly, since $b_{\max} = O(1)$, we have

$$\lim_{N\to\infty} \frac{\sum_{i,j,k=1}^{N} b'_{ij} b'_{ik}}{N^3} = p^3 + q^3 - (p^2 - q^2)^2 = pq.$$

Hence, equation (A.9) also holds.

A.6. Proof to Theorem 2.3.3

We start from $W' = u\frac{m}{N}\alpha + v\frac{n}{N}\beta$, where u and v are constants. Similar to the proof in Supplement A.5, since

$$b_{ij} = \begin{cases} \frac{u}{N(m-1)} & \text{if observations } i \text{ and } j \text{ are from Sample } X, \\ \frac{v}{N(n-1)} & \text{if observations } i \text{ and } j \text{ are from Sample } Y, \\ 0 & \text{otherwise,} \end{cases}$$

Since $b_{\max} = O(1/N^2)$, we only need to show equation (A.9):

$$\sum_{i,j,k=1}^{N} b'_{ij} b'_{ik} = O(1/N).$$

We have

$$\lim_{N\to\infty} \frac{\sum_{i,j,k=1}^{N} b'_{ij} b'_{ik}}{1/N} = (up + vq)(1 - up - vq).$$

Hence, equation (A.9) holds unless $u = v$, that is, $r = 1$.

A.7. Numerical studies on $fGPK_M$

To check the effectiveness of fGPK_M, we compare fGPK_M with tests based on MMD. We use the same simulation setup in [**27**], and include in the comparison the Pearson approximation test (MMD-Pearson) and the bootstrap test (MMD-Bootstrap), which exhibited the best performance in [**27**]. In all the following experiments, the significance level is set to be 0.05 and 10,000 bootstrap replicates are used for approximating the p-value in MMD-Bootstrap.

Following the simulation setup in [**27**], for each dimension $d \in \{e^3, e^4, e^5, e^6, e^7\}$, we randomly draw 250 observations from $N_d(0, I_d)$ and 250 observations from $N_d(\mu\mathbf{1}_d/\sqrt{d}, I_d)$ for location alternatives, where 10 different values of μ's equally spaced from 0.05 to 40 are used. Each simulation setup is repeated for 100 trials, that is, a total of 1,000 trials for each dimension d. For scale alternatives, we randomly draw 250 observations from $N_d(0, I_d)$ and 250 observations from $N_d(0, \sigma^2 I_d)$, where 10 different σ^2 equally spaced from 1.05 to 100 are used (a total of 1,000 trials for each dimension d).

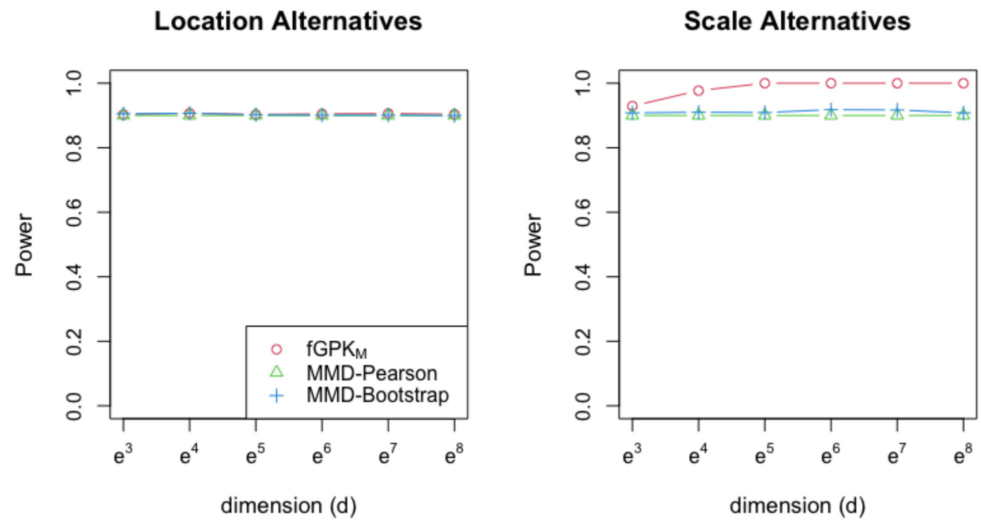


FIGURE A.2. Estimated power of fGPK (o), MMD-Pearson (△), and MMD-Bootstrap (+) at 0.05 significance level. $m = n = 250$.

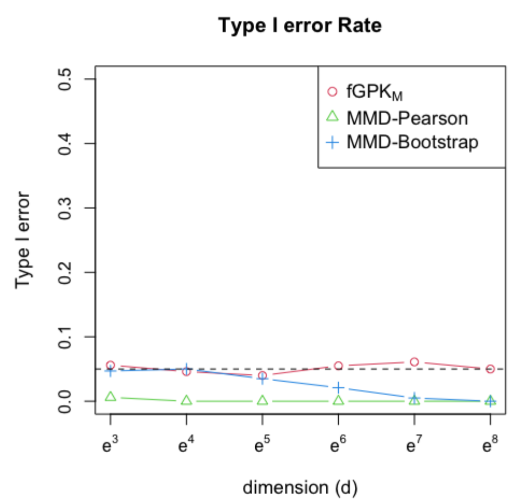


FIGURE A.3. Empirical size of fGPK$_M$ (o), MMD-Pearson (△), and MMD-Bootstrap (+), at 0.05 significance level. $m = n = 250$.

The estimated power is shown in Figure A.2. We see that, since the signal is quite strong, these three tests are doing very well in this simulation setups, while $\mathrm{fGPK_M}$ exhibits better performance than the other two tests for scale alternatives in high dimensions.

We also check the empirical size of the test and the results are plotted in Figure A.3. We see that all tests control the type I error well. However, MMD-Pearson is very conservative and MMD-Bootstrap becomes fairly conservative as dimension increases, which may lead to loss of efficiency of the test.

A.8. Runtimes implemented in R

Similar to Table 2.4 in Section 2.3.4, we also compare the computational cost of $\mathrm{fGPK_M}$, MMD-Pearson, and MMD-Boostrap with their R implementations and the results are presented in Table A.1. We see that all methods are slower than their corresponding Matlab equivalents since Matlab is a programming language especially developed for numerical linear algebra (matrix manipulations). Unlike the results under Matlab, MMD-Pearson is very slow since it is of $O(N^3 d)$ time complexity and the matrix calculation in R is less effective than Matlab. However, for both Matlab and R, $\mathrm{fGPK_M}$ is the fastest to run and fGPK is faster than the existing methods.

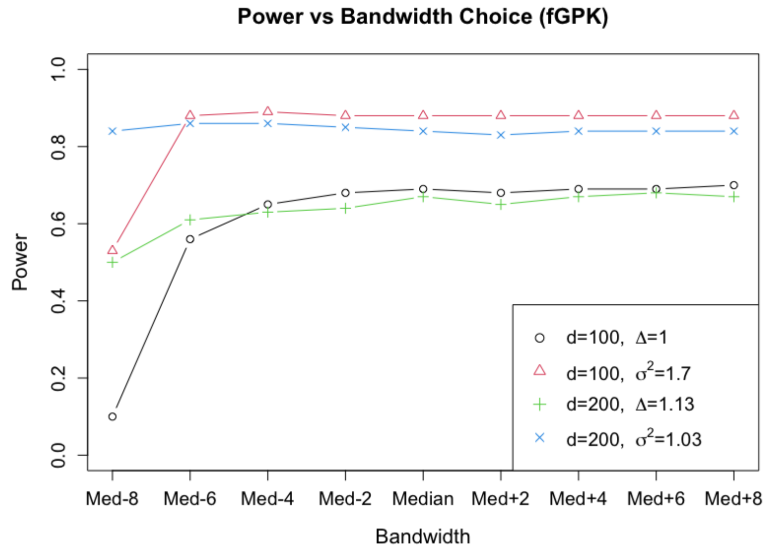
TABLE A.1. Average computation time in seconds from 10 simulations for each m. All experiments were run by R on 2.2 GHz Intel Core i7

m	50	100	250	500	1000
$\mathrm{fGPK_M}$	0.002	0.005	0.033	0.132	0.605
fGPK	0.001	0.005	0.036	0.149	0.688
MMD-Pearson	0.225	1.882	29.207	235.243	2023.701
MMD-Bootstrap	0.993	4.209	27.595	80.226	443.024

A.9. Additional Results for Bandwidth Discussion

Here, we present the performance of fGPK and $\mathrm{fGPK_M}$ for each bandwidth choice for four different settings used in Section 2.6.1.

The results are presented in Figure A.4. We see the results are consistent to the results of GPK and there is no significant difference in the performance of fGPK and $\mathrm{fGPK_M}$ unless the bandwidth is too small.



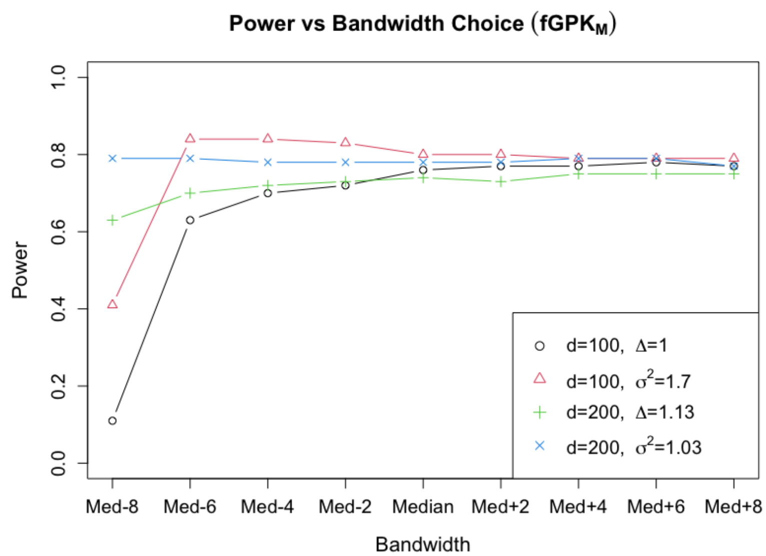


FIGURE A.4. Estimated power based on 100 trials of fGPK and fGPK$_{\mathrm{M}}$ vs bandwidth when 100 samples are generated from $N_d(\mathbf{0}, I_d)$ and 100 samples are drawn from $N_d(\mu, \sigma^2 I_d)$ with $\Delta = \|\mu\|_2$, $\alpha = 0.05$. 'Median' on the X axis indicates the averaged median heruistic in each simulation run.

APPENDIX B

Appendix for Chapter 3

B.1. Test statistics for the changed-interval alternative

Here, we propose the new test statistic for testing the null hypothesis H_0 (3.1) against the changed-interval alternative H_2 (3.3). The test statistics can be derived in a similar manner to the single change-point case. Since each possible interval $(t_1, t_2]$ divides the data sequence into two groups, we define

$$\alpha(t_1,t_2) = \frac{1}{(t_2-t_1)(t_2-t_1-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k_{ij} I_{g_i(t_1,t_2)=g_j(t_1,t_2)=1}, \tag{B.1}$$

$$\beta(t_1,t_2) = \frac{1}{(n-(t_2-t_1))(n-(t_2-t_1)-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k_{ij} I_{g_i(t_1,t_2)=g_j(t_1,t_2)=0}, \tag{B.2}$$

where $g_i(t_1, t_2) = I_{t_1 < i \leq t_2}$. Then, two basic quantities are defined as

$$\begin{aligned} D(t_1,t_2) &= (t_2-t_1)(t_2-t_1-1)\alpha(t_1,t_2) \\ &\quad - (n-(t_2-t_1))(n-(t_2-t_1)-1)\beta(t_1,t_2), \end{aligned} \tag{B.3}$$

$$\begin{aligned} W(t_1,t_2) &= \frac{n-(t_2-t_1)}{n}(t_2-t_1)(t_2-t_1-1)\alpha(t_1,t_2) \\ &\quad + \frac{(t_2-t_1)}{n}(n-(t_2-t_1))(n-(t_2-t_1)-1)\beta(t_1,t_2). \end{aligned} \tag{B.4}$$

Similarly, $D(t_1, t_2)$ and $W(t_1, t_2)$ are standardized:

$$Z_D(t_1,t_2) = \frac{D(t_1,t_2) - \mathsf{E}\left(D(t_1,t_2)\right)}{\sqrt{\mathsf{Var}\left(D(t_1,t_2)\right)}}, \quad Z_W(t_1,t_2) = \frac{W(t_1,t_2) - \mathsf{E}\left(W(t_1,t_2)\right)}{\sqrt{\mathsf{Var}\left(W(t_1,t_2)\right)}}. \tag{B.5}$$

Under the permutation null, the analytic expressions for the expectation and the variance of $D(t_1, t_2)$ and $W(t_1, t_2)$ can be obtained similarly as in the single change-point setting.

$$\mathsf{E}(\alpha(t_1,t_2)) = \mathsf{E}(\beta(t_1,t_2)) = \frac{1}{n(n-1)} R_0,$$

$$\mathsf{Var}(\alpha(t_1,t_2)) = \frac{1}{(t_2-t_1)^2(t_2-t_1-1)^2}\left(2R_1p_1(t_1,t_2)+4R_2p_2(t_1,t_2)+R_3p_3(t_1,t_2)\right),$$
$$\mathsf{Var}(\beta(t_1,t_2)) = \frac{1}{(n-(t_2-t_1))^2(n-(t_2-t_1)-1)^2}\left(2R_1q_1(t_1,t_2)+4R_2q_2(t_1,t_2)+R_3q_3(t_1,t_2)\right).$$

Then, we define the following test statistic for detecting and estimating the change-interval in the sequence as

(B.6)
$$GKCP(t_1,t_2) = Z_D^2(t_1,t_2) + Z_W^2(t_1,t_2).$$

The scan statistic involves a maximization over t_1 and t_2,

(B.7)
$$\max_{\substack{1<t_1<t_2\leq n\\ n_0\leq t_2-t_1\leq n_1}} GKCP(t_1,t_2),$$

where n_0 and n_1 are constraints on the window size. For example, we can set $n_1 = n - n_0$.

B.2. Proofs for Lemmas and Theorems

B.2.1. Proof of Theorem 3.2.1. Under the permutation null distribution, we have

$$\mathsf{E}\left(\alpha(t)\right) = \frac{1}{t(t-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} k_{ij}\mathsf{P}(g_i(t)=g_j(t)=1) = \frac{1}{t(t-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} k_{ij}p_1(t) = \frac{1}{n(n-1)}R_0,$$

$$\begin{aligned}
\mathsf{E}\left(\alpha^2(t)\right) &= \frac{1}{t^2(t-1)^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} k_{ij}\sum_{u=1}^{n}\sum_{v=1,v\neq u}^{n} k_{uv}\mathsf{P}(g_i(t)=1,g_j(t)=1,g_u(t)=1,g_v(t)=1)\\
&= \frac{2}{t^2(t-1)^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} k_{ij}^2\mathsf{P}(g_i(t)=g_j(t)=1)\\
&\quad + \frac{4}{t^2(t-1)^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\sum_{u=1,u\neq i,u\neq j}^{n} k_{ij}k_{iu}\mathsf{P}(g_i(t)=g_j(t)=g_u(t)=1)\\
&\quad + \frac{1}{t^2(t-1)^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\sum_{u=1,u\neq i,u\neq j}^{n}\sum_{v=1,v\neq i,v\neq j,v\neq u}^{n} k_{ij}k_{uv}\mathsf{P}(g_i(t)=g_j(t)=g_u(t)=g_v(t)=1)\\
&= \frac{2}{t^2(t-1)^2}R_1p_1(t) + \frac{4}{t^2(t-1)^2}R_2p_2(t) + \frac{1}{t^2(t-1)^2}R_3p_3(t)\\
&= \frac{1}{t^2(t-1)^2}\left(2R_1p_1(t)+4R_2p_2(t)+R_3p_3(t)\right).
\end{aligned}$$

Then, $\mathsf{Var}(\alpha(t)) = \mathsf{E}\left(\alpha^2(t)\right) - \mathsf{E}\left(\alpha(t)\right)^2$ follows readily. The expectation and the variance of $\beta(t)$ can be done in a similar manner. For the covariance between α and β, we have

$$\begin{aligned}\mathsf{E}\left(\alpha(t)\beta(t)\right) &= \frac{1}{t(t-1)(n-t)(n-t-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \sum_{u=1}^{n} \sum_{v=1, v\neq u}^{n} k_{ij}k_{uv}\mathsf{P}(g_i(t) = g_j(t) = 1,\ g_u(t) = g_v(t) = 2) \\ &= \frac{1}{t(t-1)(n-t)(n-t-1)} R_3 \frac{t(t-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)} \\ &= \frac{R_3}{n(n-1)(n-2)(n-3)}.\end{aligned}$$

Then, $\mathsf{Cov}\left(\alpha(t), \beta(t)\right) = \mathsf{E}(\alpha\beta) - \mathsf{E}(\alpha)\mathsf{E}(\beta)$ follows readily.

B.2.2. Proof of Theorem 3.3.1. Here, we prove $\{Z_D([nu]) : 0 < u < 1\}$ converges to a Gaussian process in finite dimensional distributions. The proof for the convergence of $\{Z_D([nu], [nv]) : 0 < u < v < 1\}$ to two-dimensional Gaussian random fields can be done in a similar manner.

To prove $\{Z_D([nu]) : 0 < u < 1\}$ converges to a Gaussian process, we only need to show that $(Z_D([nu_1]), Z_D([nu_2]), \ldots, Z_D([nu_Q]))$ coverges to a multivariate Gaussian distribution as $n \to \infty$ for any $0 < u_1 < \cdots < u_Q < 1$ and fixed Q under the permutation distribution. For notation simplicity, let $t_q = [nu_q]$, $q = 1, \ldots, Q$.

To prove $(Z_D(t_1), Z_D(t_2), \ldots, Z_D(t_Q))$ coverges to a multivariate Gaussian distribution, by Cramér-Wold device, it suffices to show that $\sum_{q=1}^{Q} a'_q Z_D(t_q)$ is asymptotically Gaussian distributed for any fixed a'_q with non-degenerating case that $\mathsf{Var}\left(\sum_{q=1}^{Q} a'_q Z_D(t_q)\right) > 0$.

First, based on the definition of $D(t)$, we observe that $\sum_{q=1}^{Q} a_q D(t_q)$ can be expressed as a double-indexed permutation statistic as follows:

$$\sum_{q=1}^{Q} a_q D(t_q) = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k_{ij} b_{ij}, \tag{B.8}$$

where $b_{ij} = \sum_{q=1}^{Q} b_{ij}(t_q)$ and

$$b_{ij}(t_q) = \begin{cases} a_q & \text{if } g_i(t_q) = g_j(t_q) = 0, \\ -a_q & \text{if } g_i(t_q) = g_j(t_q) = 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $g_i(t) = I_{i>t}$ and I_x is the indicator function. We define the mean-centered k'_{ij} and b'_{ij} for $i \neq j$,

$$k'_{ij} = k_{ij} - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} k_{ij}, \quad b'_{ij} = b_{ij} - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} b_{ij},$$

and let $k'_{ii} = b'_{ii} = 0$ for $i = 1, \ldots, n$. Then, $\sum_{i,j=1}^{n} k'_{ij} = \sum_{i,j=1}^{n} b'_{ij} = 0$. Also,

$$\begin{aligned}\sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} k_{ij} b_{ij} &= \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} \left(k'_{ij} + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} k_{ij}\right)\left(b'_{ij} + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} b_{ij}\right) \\ &= \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} k'_{ij} b'_{ij} + E,\end{aligned}$$

where E is a constant that does not change under permutation. Hence, it is sufficient to show the asymptotic normality of $\sum_{i,j=1}^{n} k'_{ij} b'_{ij}$.

The asymptotic distribution of the double-indexed permutation statistic of the form (B.8) has been studied in [**14**]. Later [**21**] claimed that Daniels' conditions for the asymptotic normality can be weakened and proposed relexed conditions. However, Zhu and Chen (2021) recently study the conditions for the asymptotic normality and they claim the conditions presented in [**21**] are not sufficient. Instead, they show the following conditions are sufficient for the asymptotic normality of $\sum_{i,j=1}^{n} k'_{ij} b'_{ij}$:

(B.9)
$$\sum_{i,j=1}^{n} k'_{ij} = \sum_{i,j=1}^{n} b'_{ij} = 0,$$

(B.10)
$$\sum_{i,j,k=1}^{n} k'_{ij} k'_{ik} = O(n^3 k'_{\max})$$

(B.11)
$$\sum_{i,j,k=1}^{n} b'_{ij} b'_{ik} = O(n^3 b'_{\max}),$$

where $k'_{\max}$ and $b'_{\max}$ are the largest order of k'_{ij} and b'_{ij}, respectively.

Equation (B.9) directly holds by the definiton. Let

$$k_{i\cdot} = \sum_{j=1,j\neq i}^{n} k_{ij}, \quad \dot{k} = \frac{1}{n} \sum_{i=1}^{n} k_{i\cdot}, \quad k'_{i\cdot} = \sum_{j=1}^{n} k'_{ij},$$

$$b_{i\cdot} = \sum_{j=1,j\neq i}^{n} b_{ij}, \quad \dot{b} = \frac{1}{n} \sum_{i=1}^{n} b_{i\cdot}, \quad b'_{i\cdot} = \sum_{j=1}^{n} b'_{ij}.$$

Then,

$$k'_{i\cdot} = k_{i\cdot} - \dot{k},$$

$$b'_{i\cdot} = b_{i\cdot} - \dot{b}.$$

We have

$$\sum_{i,j,k=1}^{n} k'_{ij}k'_{ik} = \sum_{i=1}^{n} k_{i\cdot}^2 - n\dot{k}^2 - \sum_{i,j=1}^{n} k_{ij}^2 + \frac{n}{n-1}\dot{k}^2.$$

Note that $k'_{\max} = k_{\max}$. When $k_{ij} = O(1)\ \forall i,j$, we have $k_{i.} = O(n)$, $\dot{k} = O(n)$, and $k_{\max} = O(1)$. Since the term $\sum_{i=1}^{n} k_{i\cdot}^2 - n\dot{k}^2$ dominates others, equation (B.10) holds when

(B.12)
$$\sum_{i=1}^{n} k_{i\cdot}^2 - n\dot{k}^2 = O\left(\sum_{i=1}^{n} k_{i\cdot}^2\right).$$

Let $\lim_{n\to\infty} t_q/n = u_q$. Since $b_{\max} = O(1)$ and

$$\sum_{i=1}^{n} b_{i\cdot}^2 = \sum_{q=0}^{Q}\left[(t_{q+1}-t_q)\left(\sum_{r=1}^{Q} a_r(t_r-1) + (2-n)\sum_{s=0}^{q} a_s\right)^2\right],$$

$$\dot{b} = \frac{1}{n}\sum_{q=0}^{Q}\left[(t_{q+1}-t_q)\left(\sum_{r=1}^{Q} a_r(t_r-1) + (2-n)\sum_{s=0}^{q} a_s\right)\right],$$

where $a_0 = t_0 = 0$ and $t_{Q+1} = n$, it is easy to show that $\sum_{i=1}^{n} b_{i\cdot}^2 - n\dot{b}^2$ is of order n^3 unless $a_1 = \cdots = a_Q = 0$.

Since

$$\begin{aligned}\lim_{n\to\infty} \mathsf{Var}\,(D(t_q)) = {} & \left(2R_1u_q^2 + 4R_2u_q^3 + R_3u_q^4\right) - u_q^4R_0^2 \\ & + \left(2R_1(1-u_q)^2 + 4R_2(1-u_q)^3 + R_3(1-u_q)^4\right) - (1-u_q)^4R_0^2 \\ & - u_q^2(1-u_q)^2R_3 - u_q^2(1-u_q)^2R_0^2,\end{aligned}$$

for all $i = 1,\ldots,Q$, all $\mathsf{Var}\,(D(t_q))$ are of the same order and this leads to the asymptotic normality of $\sum_{q=1}^{Q} a_q Z_D(t_q)$.

B.2.3. Proof of Theorem 3.3.2. Similar to the proof in B.2.2,

$$\sum_{q=1}^{Q} a_q W_r(t_q) = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k_{ij} b_{ij}, \tag{B.13}$$

where $b_{ij} = \sum_{q=1}^{Q} b_{ij}(t_q)$ and

$$b_{ij}(t_q) = \begin{cases} a_q r(n-t_q)/n & \text{if } g_i(t_q) = g_j(t_q) = 0, \\ a_q t_q/n & \text{if } g_i(t_q) = g_j(t_q) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Since $b_{\max} = O(1)$, we only need to show equation (B.11):

$$\sum_{i,j,k=1}^{n} b'_{ij} b'_{ik} = O(n^3).$$

Since,

$$\sum_{i=1}^{n} b_{i\cdot}^2 = \sum_{q=0}^{Q} \left[(t_{q+1} - t_q) \left(\sum_{h=1}^{Q} a_h r \frac{(n-t_h)(t_h-1)}{n} + \sum_{s=0}^{q} a_s \frac{t_s(n-t_s-1) - r(n-t_s)(t_s-1)}{n} \right)^2 \right],$$

$$\dot{b} = \frac{1}{n} \sum_{q=0}^{Q} \left[(t_{q+1} - t_q) \left(\sum_{h=1}^{Q} a_h r \frac{(n-t_h)(t_h-1)}{n} + \sum_{s=0}^{q} a_s \frac{t_s(n-t_s-1) - r(n-t_s)(t_s-1)}{n} \right) \right],$$

it is easy to show that $\sum_{i=1}^{n} b_{i\cdot}^2 - n\dot{b}^2$ is of order n^3 unless $\sum_{s=0}^{q} a_s \left(u_s(1-u_s) - r(1-u_s)u_s\right)$ are all the same for $q = 0, \ldots, Q$, that is, $r = 1$.

B.2.4. Proof of Theorem 3.3.3. Here, we show the derivation of $\rho_D^*(u,v)$, and that for $\rho_{W,r}^*(u,v)$ can be done in similar way.

Let $\rho_D(u,v) = \mathsf{Cov}\left(Z_D([nu]), Z_D([nv])\right)$. Then, $\rho_D^*(u,v) = \lim_{n\to\infty} \mathsf{Cov}\left(Z_D([nu]), Z_D([nv])\right)$.

We first show for $u \leq v$. Let $s = [nu]$ and $t = [nv]$. Then, $s < t$ and $\lim_{n\to\infty} s/n = u$, $\lim_{n\to\infty} t/n = v$.

Since $\mathsf{Cov}\left(Z_D(s), Z_D(t)\right) = \text{numerator}/\text{denominator}$, where

$$\begin{aligned} \text{numerator} = {} & \mathsf{E}\left(\big(s(s-1)\alpha(s) - (n-s)(n-s-1)\beta(s)\big)\big(t(t-1)\alpha(t) - (n-t)(n-t-1)\beta(t)\big)\right) \\ & - \mathsf{E}\left(s(s-1)\alpha(s) - (n-s)(n-s-1)\beta(s)\right) \mathsf{E}\left(t(t-1)\alpha(t) - (n-t)(n-t-1)\beta(t)\right) \end{aligned}$$

denominator $= \sqrt{\mathsf{Var}\left(s(s-1)\alpha(s)-(n-s)(n-s-1)\beta(s)\right)\mathsf{Var}\left(t(t-1)\alpha(t)-(n-t)(n-t-1)\beta(t)\right)}.$

Other terms follow easily from $\mathsf{E}\left(\alpha(t)\right)$, $\mathsf{E}\left(\beta(t)\right)$, and $\mathsf{Cov}\left(\alpha(t),\beta(t)\right)$ provided in Section 3.2. Hence, we only need to figure out

$$\begin{aligned}
&\mathsf{E}\left(\big(s(s-1)\alpha(s)-(n-s)(n-s-1)\beta(s)\big)\big(t(t-1)\alpha(t)-(n-t)(n-t-1)\beta(t)\big)\right)\\
&= s(s-1)t(t-1)\mathsf{E}\left(\alpha(s)\alpha(t)\right)-s(s-1)(n-t)(n-t-1)\mathsf{E}\left(\alpha(s)\beta(t)\right)\\
&\quad-(n-s)(n-s-1)t(t-1)\mathsf{E}\left(\beta(s)\alpha(t)\right)+(n-s)(n-s-1)(n-t)(n-t-1)\mathsf{E}\left(\beta(s)\beta(t)\right).
\end{aligned}$$

(i) $s(s-1)t(t-1)\mathsf{E}\left(\alpha(s)\alpha(t)\right)$

$$=\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}k_{ij}\sum_{u=1}^{n}\sum_{v=1,v\neq u}^{n}k_{uv}\mathsf{P}(g_i(s)=g_j(s)=0,\ g_u(t)=g_v(t)=0).$$

Since

$$P\big(g_i(s)=g_j(s)=0,\ g_u(t)=g_v(t)=0\big)
=\begin{cases}
\frac{s(s-1)}{n(n-1)}:=a_1(s) & if\ \begin{cases} i=u,j=v\\ i=v,j=u\end{cases}\\[2ex]
\frac{s(s-1)(t-1)}{n(n-1)(n-2)}:=a_2(s,t) & if\ \begin{cases} i=u,j\neq v\\ i=v,j\neq u\\ j=u,i\neq v\\ j=v,i\neq u\end{cases}\\[2ex]
\frac{s(s-1)(t-2)(t-3)}{n(n-1)(n-2)(n-3)}:=a_3(s,t) & if\ \ i\neq j\neq u\neq v
\end{cases}$$

we have

$$s(s-1)t(t-1)\mathsf{E}\left(\alpha(s)\alpha(t)\right)=2R_1(t)a_1(s)+4R_2a_2(s,t)+R_3a_3(s,t).$$

(ii) $s(s-1)(n-t)(n-t-1)\mathsf{E}\left(\alpha(s)\beta(t)\right)$

$$= \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} k_{ij} \sum_{u=1}^{n} \sum_{v=1, v \neq u}^{n} k_{uv} \mathsf{P}(g_i(s) = g_j(s) = 0,\ g_u(t) = g_v(t) = 1).$$

Since

$$P\big(g_i(s) = g_j(s) = 0,\ g_u(t) = g_v(t) = 1\big) = \frac{s(s-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)} := b_1(s,t)$$

we have

$$s(s-1)t(t-1)\mathsf{E}\left(\alpha(s)\alpha(t)\right) = R_3 b_1(s,t).$$

(iii) $(n-s)(n-s-1)t(t-1)\mathsf{E}\left(\beta(s)\alpha(t)\right)$

$$= \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} k_{ij} \sum_{u=1}^{n} \sum_{v=1, v \neq u}^{n} k_{uv} \mathsf{P}(g_i(s) = g_j(s) = 1,\ g_u(t) = g_v(t) = 0).$$

Since

$$P\big(g_i(s) = g_j(s) = 1,\ g_u(t) = g_v(t) = 0\big)$$

$$= \begin{cases} \frac{(t-s)(t-s-1)}{n(n-1)} := c_1(s,t) & if \begin{cases} i = u, j = v \\ i = v, j = u \end{cases} \\ \\ \frac{(t-s)\big((t-s-1)(t-2)+(t-1)(n-t)\big)}{n(n-1)(n-2)} := c_2(s,t)) & if \begin{cases} i = u, j \neq v \\ i = v, j \neq u \\ j = u, i \neq v \\ j = v, i \neq u \end{cases} \\ \\ \frac{(t-s)(n-s-2)\big((t-s-1)(n-s-3)+2s(n-s-1)\big)}{n(n-1)(n-2)(n-3)} & \\ + \frac{s(s-1)(n-s)(n-s-1)}{n(n-1)(n-2)(n-3)} := c_3(s,t) & if\ \ i \neq j \neq u \neq v \end{cases}$$

we have

$$(n-s)(n-s-1)t(t-1)\mathsf{E}\left(\beta(s)\alpha(t)\right) = 2R_1c_1(s,t) + 4R_2c_2(s,t) + R_3c_3(s,t).$$

(iv) $(n-s)(n-s-1)(n-t)(n-t-1)\mathsf{E}\left(\beta(s)\beta(t)\right)$

$$= \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k_{ij} \sum_{u=1}^{n} \sum_{v=1, v\neq u}^{n} k_{uv} \mathsf{P}(g_i(s) = g_j(s) = 1,\ g_u(t) = g_v(t) = 1).$$

Since

$$P\big(g_i(s) = g_j(s) = 1,\ g_u(t) = g_v(t) = 1\big)$$
$$= \begin{cases} \frac{(n-t)(n-t-1)}{n(n-1)} := d_1(t) & if \ \begin{cases} i = u, j = v \\ i = v, j = u \end{cases} \\ \\ \frac{(n-t)(n-t-1)(n-t-2)}{n(n-1)(n-2)} := d_2(s,t) & if \ \begin{cases} i = u, j \neq v \\ i = v, j \neq u \\ j = u, i \neq v \\ j = v, i \neq u \end{cases} \\ \\ \frac{(n-t)(n-t-1)(n-s-2)(n-s-3)}{n(n-1)(n-2)(n-3)} := d_3(s,t) & if \ \ i \neq j \neq u \neq v \end{cases}$$

we have

$$(n-s)(n-s-1)(n-t)(n-t-1)\mathsf{E}\left(\beta(s)\beta(t)\right) = 2R_1 d_1(t) + 4R_2 d_2(s,t) + R_3 d_3(s,t).$$

To sum, after tedious calculation, we have

$$\mathsf{E}\left(\big(s(s-1)\alpha(s) - (n-s)(n-s-1)\beta(s)\big)\big(t(t-1)\alpha(t) - (n-t)(n-t-1)\beta(t)\big)\right)$$
$$= 2R_1\left(a_1(s) - c_1(s,t) + d_1(t)\right) + 4R_2\left(a_2(s,t) - c_2(s,t) + d_2(s,t)\right)$$
$$+ R_3\left(a_3(s,t) - b_1(s,t) - c_3(s,t) + d_3(s,t)\right).$$

Since

$$\mathsf{E}\left(s(s-1)\alpha(s) - (n-s)(n-s-1)\beta(s)\right)\mathsf{E}\left(t(t-1)\alpha(t) - (n-t)(n-t-1)\beta(t)\right)$$
$$= R_0^2\left(p_1(s)p_1(t) - p_1(s)q_1(t) - q_1(s)p_1(t) + q_1(s)q_1(t)\right),$$

by plugging in and simplifying the expressions, we have

$$\begin{aligned}&\mathsf{E}\left(\big(s(s-1)\alpha(s)-(n-s)(n-s-1)\beta(s)\big)\big(t(t-1)\alpha(t)-(n-t)(n-t-1)\beta(t)\big)\right)\\&\qquad-\mathsf{E}\left(s(s-1)\alpha(s)-(n-s)(n-s-1)\beta(s)\right)\mathsf{E}\left(t(t-1)\alpha(t)-(n-t)(n-t-1)\beta(t)\right)\\&\quad=\frac{4s(n-t)}{n(n-1)}\left(R_1+R_2-\frac{1}{n}R_0^2\right).\end{aligned}$$

Since

$$\mathsf{Var}\left(s(s-1)\alpha(s)-(n-s)(n-s-1)\beta(s)\right)=\frac{4s(n-s)}{n(n-1)}\left(R_1+R_2-\frac{1}{n}R_0^2\right),$$

we have

$$\begin{aligned}&\mathsf{Cov}\left(Z_D(s),Z_D(t)\right)\\&=\frac{\frac{4s(n-t)}{n(n-1)}\left(R_1+R_2-\frac{1}{n}R_0^2\right)}{\sqrt{\frac{4s(n-s)}{n(n-1)}\left(R_1+R_2-\frac{1}{n}R_0^2\right)\frac{4t(n-t)}{n(n-1)}\left(R_1+R_2-\frac{1}{n}R_0^2\right)}}\\&=\sqrt{\frac{s(n-t)}{t(n-s)}}.\end{aligned}$$

Hence, for $u\le v$,

$$\rho_D^*(u,v)=\lim_{n\to\infty}\mathsf{Cov}\left(Z_D(s),Z_D(t)\right)=\sqrt{\frac{u(1-v)}{v(1-u)}}.$$

Similarly, for $v\le u$,

$$\rho_D^*(u,v)=\sqrt{\frac{v(1-u)}{u(1-v)}},$$

and the result in the proposition follows.

B.3. Analytic expressions for the third moments

Since

$$\mathsf{E}\left(Z_D^3(t)\right)=\frac{\mathsf{E}\left(D^3(t)\right)-3\mathsf{E}\left(D(t)\right)\mathsf{Var}\left(D(t)\right)-\mathsf{E}^3\left(D(t)\right)}{\mathsf{Var}^{3/2}\left(D(t)\right)},$$

$$\mathsf{E}\left(Z_{W,r}^3(t)\right) = \frac{\mathsf{E}\left(W_r^3(t)\right) - 3\mathsf{E}\left(W_r(t)\right)\mathsf{Var}\left(W_r(t)\right) - \mathsf{E}^3\left(W_r(t)\right)}{\mathsf{Var}^{3/2}\left(W_r(t)\right)},$$

and the analytic expressions for the expectations and variances of $Z_D(t)$ and $Z_{W,r}(t)$ can be found in Section 3.2, we only need to figure out the analytic expressions for the expectations of $\mathsf{E}\left(D^3(t)\right)$ and $\mathsf{E}\left(W_r^3(t)\right)$ and they can be obtained based on the following theorem.

THEOREM B.3.1. *Let $k_{ii} = 0$ for all $i = 1, \ldots, n$. We have*

$$\begin{aligned}
\mathsf{E}\left(\alpha^3(t)\right) &= \frac{1}{t^3(t-1)^3} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k_{ij} \sum_{u=1}^{n} \sum_{v=1, v\neq u}^{n} k_{uv} \sum_{r=1}^{n} \sum_{s=1, s\neq r}^{n} k_{rs} \\
&\quad \times \mathsf{P}\left(g_i(t) = g_j(t) = g_u(t) = g_v(t) = g_r(t) = g_s(t) = 0\right) \\
&= \frac{1}{t^3(t-1)^3}\Bigg(4\sum_{i,j=1}^{n} k_{ij}^3 p_1(t) + 24 \sum_{i,j,u=1}^{n} k_{ij}^2 k_{iu} p_2(t) + 8 \sum_{i,j,u=1}^{n} k_{ij}k_{ju}k_{ui}p_2(t) \\
&\quad + 6 \sum_{i,j,u,v=1}^{n} k_{ij}^2 k_{uv} p_3(t) + 8 \sum_{i,j,u,v=1}^{n} k_{ij}k_{iu}k_{iv}p_3(t) + +24 \sum_{i,j,u,v=1}^{n} k_{ij}k_{ju}k_{uv}p_3(t) \\
&\quad + 12 \sum_{i,j,u,v,r=1}^{n} k_{ij}k_{ju}k_{vr}p_4(t) + \sum_{i,j,u,v,r,s=1}^{n} k_{ij}k_{uv}k_{rs}p_5(t)\Bigg), \\
\mathsf{E}\left(\alpha^2(t)\beta(t)\right) &= \frac{1}{t^2(t-1)^2(n-t)(n-t-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k_{ij} \sum_{u=1}^{n} \sum_{v=1, v\neq u}^{n} k_{uv} \sum_{r=1}^{n} \sum_{s=1, s\neq r}^{n} k_{rs} \\
&\quad \times \mathsf{P}\left(g_i(t) = g_j(t) = g_u(t) = g_v(t) = 0,\ g_r(t) = g_s(t) = 1\right) \\
&= \frac{1}{t^2(t-1)^2(n-t)(n-t-1)}\Bigg(2 \sum_{i,j,u,v=1}^{n} k_{ij}^2 k_{uv} p_1'(t) + 4 \sum_{i,j,u,v,r=1}^{n} k_{ij}k_{ju}k_{vr} f_2'(t) \\
&\quad + \sum_{i,j,u,v,r,s=1}^{n} k_{ij}k_{uv}k_{rs} f_3'(t)\Bigg), \\
\mathsf{E}\left(\alpha(t)\beta^2(t)\right) &= \frac{1}{t(t-1)(n-t)^2(n-t-1)^2} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k_{ij} \sum_{u=1}^{n} \sum_{v=1, v\neq u}^{n} k_{uv} \sum_{r=1}^{n} \sum_{s=1, s\neq r}^{n} k_{rs} \\
&\quad \times \mathsf{P}\left(g_i(t) = g_j(t) = 0,\ g_u(t) = g_v(t) = g_r(t) = g_s(t) = 1\right) \\
&= \frac{1}{t(t-1)(n-t)^2(n-t-1)^2}\Bigg(2 \sum_{i,j,u,v=1}^{n} k_{ij}^2 k_{uv} q_1'(t) + 4 \sum_{i,j,u,v,r=1}^{n} k_{ij}k_{ju}k_{vr} q_2'(t) \\
&\quad + \sum_{i,j,u,v,r,s=1}^{n} k_{ij}k_{uv}k_{rs} q_3'(t)\Bigg),
\end{aligned}$$

$$
\begin{aligned}
\mathsf{E}\left(\beta^3(t)\right) = {} & \frac{1}{t^3(t-1)^3}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} k_{ij}\sum_{u=1}^{n}\sum_{v=1,v\neq u}^{n} k_{uv}\sum_{r=1}^{n}\sum_{s=1,s\neq r}^{n} k_{rs} \\
& \times \mathsf{P}(g_i(t) = g_j(t) = g_u(t) = g_v(t) = g_r(t) = g_s(t) = 1) \\
= {} & \frac{1}{t^3(t-1)^3}\Bigg(4\sum_{i,j=1}^{n} k_{ij}^3 q_1(t) + 24\sum_{i,j,u=1}^{n} k_{ij}^2 k_{iu} q_2(t) + 8\sum_{i,j,u=1}^{n} k_{ij}k_{ju}k_{ui} q_2(t) \\
& + 6\sum_{i,j,u,v=1}^{n} k_{ij}^2 k_{uv} q_3(t) + 8\sum_{i,j,u,v=1}^{n} k_{ij}k_{iu}k_{iv} q_3(t) + +24\sum_{i,j,u,v=1}^{n} k_{ij}k_{ju}k_{uv} q_3(t) \\
& + 12\sum_{i,j,u,v,r=1}^{n} k_{ij}k_{ju}k_{vr} q_4(t) + \sum_{i,j,u,v,r,s=1}^{n} k_{ij}k_{uv}k_{rs} q_5(t)\Bigg),
\end{aligned}
$$

where

$$
\begin{aligned}
& p_4(t) = p_3(t)\frac{t-4}{n-4},\ \ p_5(t) = p_4(t)\frac{t-5}{n-5},\ \ q_4(t) = q_3(t)\frac{n-t-4}{n-4},\ \ q_5(t) = q_4(t)\frac{n-t-5}{n-5}, \\
& p_1'(t) = \frac{t(t-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)},\ \ p_2'(t) = p_1'(t)\frac{t-2}{n-4},\ \ p_3'(t) = p_2'(t)\frac{t-3}{n-5}, \\
& q_1'(t) = p_1'(t),\ \ q_2'(t) = q_1'(t)\frac{n-t-2}{n-4},\ \ q_3'(t) = q_2'(t)\frac{n-t-3}{n-5}.
\end{aligned}
$$

B.4. More Experiment Results

Here, we consider the multivariate Gaussian data with the following structures: We simulate 100 datasets

TABLE B.1. Three toy problems

DATA	F_0	F_1
LOCATION ALTERNATIVE	$N_d(\mathbf{0}_d, I_d)$	$N_d((1.6, 0, \ldots, 0)^T, I_d)$
SCALE ALTERNATIVE	$N_d(\mathbf{0}_d, I_d)$	$N_d(\mathbf{0}_d, \text{DIAG}(15, 1, \ldots, 1))$

to estimate the power of the tests and the significance level is set to be 0.05 for all tests.

The results are shown in Table B.2. Here, it becomes harder to detect changes as d increases since the signal gets weaker. We first see that FCP has lower or no power in all settings. KCP and ECP exhibit high power for the location alternatives, but they lose power for the scale alternatives. However, the new tests perform well for both location and scale alternatives. GCP also works well for both location and scale alternatives, but it is outperformed by the new tests.

TABLE B.2. Estimated power of the tests for multivariate Gaussian data. $n = 200$

	Mean Change (τ at center)				Variance Change (τ at center)			
d	100	500	1000	2000	100	500	1000	2000
fGKCP_1	98 (98)	47 (44)	37 (23)	16 (9)	98 (77)	45 (28)	22 (13)	12 (5)
fGKCP_2	99 (99)	55 (50)	44 (29)	24 (13)	96 (76)	38 (20)	23 (12)	16 (7)
GKCP	100 (100)	73 (67)	61 (43)	25 (17)	95 (75)	45 (28)	17 (10)	15 (4)
KCP	100 (100)	78 (71)	54 (42)	22 (14)	33 (22)	9 (0)	4 (1)	7 (1)
ECP	100 (98)	85 (77)	57 (45)	38 (23)	15 (9)	10 (2)	9 (5)	10 (4)
GCP	72 (61)	24 (9)	14 (4)	7 (1)	94 (82)	33 (15)	19 (6)	17 (4)
FCP	18 (8)	0 (0)	0 (0)	0 (0)	87 (74)	0 (0)	0 (0)	0 (0)

B.5. Checking analytic p-value approximations for the changed-interval

Here, we examine the performance of the analytical p-value approximations of the new tests for the changed-interval alternative. The simulation setting and notation are identical to the single change-point alternative in Section 3.3.4.

The results are shown in Table B.3, B.4, and B.5. According to the tables, conclusions similar to the single change-point alternative can be drawn. The analytical p-value approximation with skewness correction performs better than the p-value apprpoximation without skewness correction, especially when the window size increases. In general, the skewness-corrected p-value approximation works well when $n_0 \geq 50$.

TABLE B.3. Critical values for the single change-point scan statistic $\max_{n_0 \le t_2 - t_1 \le n_1} Z_D(t_1, t_2)$ at 0.05 significance level. $n = 1000$

	$n_0 = 100$	$n_0 = 75$	$n_0 = 50$	$n_0 = 25$
A1	3.91	3.98	4.07	4.20

	Critical Values							
	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$	
	A2	Per	A2	Per	A2	Per	A2	Per
Gaussian	3.90	3.85	3.97	3.95	4.05	4.01	4.17	4.18
$d = 100$	3.91	3.87	3.97	3.95	4.05	4.01	4.20	4.18
Gaussian	3.91	3.87	3.98	3.93	4.07	4.02	4.19	4.15
$d = 500$	3.91	3.87	3.98	3.93	4.07	4.02	4.20	4.18
Gaussian	3.91	3.87	3.97	3.94	4.06	4.03	4.19	4.17
$d = 1000$	3.91	3.87	3.97	3.94	4.06	4.04	4.23	4.18

	Critical Values							
	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$	
	A2	Per	A2	Per	A2	Per	A2	Per
MV-t_5	3.86	3.87	3.92	3.93	3.97	4.06	3.99	4.26
$d = 100$	3.85	3.86	3.90	3.93	3.95	4.06	4.02	4.32
MV-t_5	3.88	3.87	3.94	3.94	3.97	4.04	4.02	4.21
$d = 500$	3.87	3.88	3.91	3.94	4.01	4.06	4.08	4.22
MV-t_5	3.87	3.86	3.90	3.92	3.95	4.03	4.00	4.21
$d = 1000$	3.86	3.87	3.93	3.96	3.99	4.05	4.01	4.28

	Critical Values							
	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$	
	A2	Per	A2	Per	A2	Per	A2	Per
Log-normal	3.85	3.88	3.89	3.95	3.93	4.08	3.98	4.29
$d = 100$	3.85	3.86	3.89	3.95	3.94	4.09	3.98	4.30
Log-normal	3.82	3.85	3.86	3.95	3.90	4.10	3.94	4.37
$d = 500$	3.83	3.87	3.87	3.96	3.91	4.11	3.96	4.41
Log-normal	3.83	3.87	3.87	3.95	3.94	4.09	3.98	4.37
$d = 1000$	3.85	3.88	3.89	3.96	3.91	4.04	3.95	4.31

TABLE B.4. Critical values for the single change-point scan statistic $\max_{n_0 \le t_2 - t_1 \le n_1} Z_{W,1.2}(t_1, t_2)$ at 0.05 significance level. $n = 1000$

	Critical Values											
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$		
	A1	A2	Per	A1	A2	Per	A1	A2	Per	A1	A2	Per
Gaussian	3.79	3.97	3.97	3.88	4.10	4.10	3.99	4.26	4.27	4.16	4.53	4.55
$d = 100$	3.78	4.00	3.98	3.89	4.11	4.10	3.99	4.27	4.27	4.17	4.55	4.59
Gaussian	3.80	3.88	3.86	3.88	3.98	3.96	4.00	4.12	4.10	4.17	4.34	4.31
$d = 500$	3.80	3.88	3.86	3.88	3.98	3.96	4.00	4.12	4.09	4.17	4.34	4.34
Gaussian	3.79	3.85	3.84	3.88	3.95	3.92	3.99	4.09	4.05	4.16	4.30	4.30
$d = 1000$	3.79	3.86	3.86	3.88	3.95	3.94	3.99	4.07	4.05	4.16	4.28	4.27

	Critical Values											
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$		
	A1	A2	Per	A1	A2	Per	A1	A2	Per	A1	A2	Per
MV-t_5	3.74	4.02	4.02	3.82	4.12	4.12	3.93	4.27	4.27	4.10	4.53	4.58
$d = 100$	3.74	3.98	3.96	3.82	4.10	4.07	3.93	4.25	4.25	4.10	4.50	4.53
MV-t_5	3.73	3.86	3.84	3.82	3.94	3.92	3.93	4.07	4.06	4.09	4.29	4.27
$d = 500$	3.73	3.85	3.84	3.82	3.95	3.93	3.92	4.06	4.05	4.09	4.25	4.25
MV-t_5	3.73	3.81	3.78	3.81	3.90	3.87	3.92	4.01	3.99	4.08	4.16	4.16
$d = 1000$	3.73	3.80	3.77	3.81	3.88	3.85	3.92	3.98	3.97	4.08	4.20	4.19

	Critical Values											
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$		
	A1	A2	Per	A1	A2	Per	A1	A2	Per	A1	A2	Per
Log-normal	3.71	4.38	4.46	3.79	4.52	4.64	3.89	4.73	4.92	4.05	5.12	5.49
$d = 100$	3.71	4.39	4.45	3.79	4.52	4.64	3.89	4.73	4.92	4.05	5.11	5.44
Log-normal	3.70	4.07	4.06	3.78	4.18	4.19	3.88	4.35	4.37	4.03	4.61	4.65
$d = 500$	3.70	4.08	4.08	3.78	4.20	4.21	3.87	4.33	4.36	4.03	4.63	4.65
Log-normal	3.70	3.97	3.96	3.78	4.09	4.08	3.88	4.23	4.23	4.03	4.48	4.49
$d = 1000$	3.70	4.02	4.03	3.78	4.12	4.13	3.88	4.26	4.27	4.03	4.54	4.57

TABLE B.5. Critical values for the single change-point scan statistic $\max_{n_0 \le t_2 - t_1 \le n_1} Z_{W,0.8}(t_1, t_2)$ at 0.05 significance level. $n = 1000$

	Critical Values											
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$		
	A1	A2	Per	A1	A2	Per	A1	A2	Per	A1	A2	Per
Gaussian	2.77	2.84	3.91	2.82	2.89	4.01	2.89	2.97	4.15	2.97	3.07	4.37
$d = 100$	2.78	2.84	3.93	2.83	2.90	4.00	2.89	2.97	4.15	2.98	3.07	4.40
Gaussian	2.78	2.80	2.80	2.82	2.85	2.84	2.89	2.92	2.91	2.97	3.01	2.99
$d = 500$	2.78	2.80	2.80	2.83	2.85	2.87	2.89	2.92	2.92	2.98	3.01	3.00
Gaussian	2.77	2.79	2.79	2.83	2.85	2.85	2.89	2.92	2.89	2.97	3.01	2.96
$d = 1000$	2.77	2.80	2.80	2.82	2.85	2.82	2.89	2.91	2.91	2.97	3.00	2.99

	Critical Values											
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$		
	A1	A2	Per	A1	A2	Per	A1	A2	Per	A1	A2	Per
MV-t_5	3.72	4.00	4.01	3.80	4.12	4.12	3.90	4.28	4.36	4.05	4.55	4.68
$d = 100$	3.72	4.01	4.02	3.80	4.13	4.13	3.90	4.29	4.36	4.05	4.57	4.72
MV-t_5	3.72	4.00	4.00	3.80	4.12	4.12	3.90	4.28	4.36	4.05	4.33	4.34
$d = 500$	3.72	4.01	4.02	3.80	4.13	4.13	3.90	4.29	4.36	4.05	4.39	4.45
MV-t_5	3.72	3.84	3.84	3.79	3.94	3.92	3.89	4.07	4.06	4.04	4.31	4.32
$d = 1000$	3.72	3.85	3.84	3.79	3.96	3.95	3.89	4.11	4.09	4.04	4.37	4.41

	Critical Values											
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$		
	A1	A2	Per	A1	A2	Per	A1	A2	Per	A1	A2	Per
Log-normal	3.70	4.31	4.42	3.78	4.45	4.61	3.88	4.66	4.97	4.03	5.02	5.62
$d = 100$	3.70	4.31	4.42	3.78	4.45	4.61	3.88	4.66	4.98	4.03	5.02	5.63
Log-normal	3.70	4.13	4.15	3.78	4.27	4.32	3.87	4.48	4.59	4.02	4.76	5.13
$d = 500$	3.70	4.13	4.14	3.78	4.27	4.32	3.87	4.48	4.58	4.02	4.83	5.26
Log-normal	3.70	4.05	4.08	3.78	4.18	4.22	3.87	4.37	4.39	4.02	4.70	5.01
$d = 1000$	3.70	4.05	4.08	3.78	4.16	4.20	3.87	4.34	4.38	4.02	4.65	4.89

Bibliography

[1] A. Abou-Elailah, V. Gouet-Brunet, and I. Bloch, *Detection of abrupt changes in spatial relationships in video sequences*, in International Conference on Pattern Recognition Applications and Methods, Springer, 2015, pp. 89–106.

[2] S. Arlot, A. Celisse, and Z. Harchaoui, *A kernel multiple change-point algorithm via model selection.*, Journal of Machine Learning Research, 20 (2019), pp. 1–56.

[3] W. Baumgartner, P. Weiss, and H. Schindler, *A nonparametric test for the general two-sample problem*, Biometrics, (1998), pp. 1129–1135.

[4] M. Biswas and A. K. Ghosh, *A nonparametric two-sample test applicable to high dimensional data*, Journal of Multivariate Analysis, 123 (2014), pp. 160–171.

[5] C. Blake, *Uci repository of machine learning databases*, http://www. ics. uci. edu/˜ mlearn/MLRepository. html, (1998).

[6] H. W. Block, T. H. Savits, and M. Shaked, *Some concepts of negative dependence*, The Annals of Probability, 10 (1982), pp. 765–772.

[7] H. W. Block, T. H. Savits, and J. Wang, *Negative dependence and the simes inequality*, Journal of statistical planning and inference, 138 (2008), pp. 4107–4110.

[8] M. Bosc, F. Heitz, J.-P. Armspach, I. Namer, D. Gounot, and L. Rumbach, *Automatic change detection in multimodal serial mri: application to multiple sclerosis lesion evolution*, NeuroImage, 20 (2003), pp. 643–656.

[9] W.-C. Chang, C.-L. Li, Y. Yang, and B. Póczos, *Kernel change-point detection with auxiliary deep generative models*, arXiv preprint arXiv:1901.06077, (2019).

[10] H. Chen and J. H. Friedman, *A new graph-based two-sample test for multivariate and object data*, Journal of the American statistical association, 112 (2017), pp. 397–409.

[11] H. Chen and N. Zhang, *Graph-based change-point detection*, The Annals of Statistics, 43 (2015), pp. 139–176.

[12] J. Chen and A. K. Gupta, *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*, Springer Science & Business Media, 2011.

[13] L. Chu and H. Chen, *Asymptotic distribution-free change-point detection for multivariate and non-euclidean data*, The Annals of Statistics, 47 (2019), pp. 382–414.

[14] H. E. Daniels, *The relation between measures of correlation in the universe of sample permutations*, Biometrika, 33 (1944), pp. 129–135.

[15] P. Dubey and H.-G. Müller, *Fréchet change-point detection*, Annals of Statistics, 48 (2020), pp. 3312–3335.

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

[17] M. Eric, F. R. Bach, and Z. Harchaoui, *Testing for homogeneity with kernel fisher discriminant analysis*, in Advances in Neural Information Processing Systems, 2008, pp. 609–616.

[18] C.-Y. Fang, S.-W. Chen, and C.-S. Fuh, *Automatic change detection of driving environments in a vision-based driver assistance system*, IEEE Transactions on Neural Networks, 14 (2003), pp. 646–657.

[19] H. Finner, M. Roters, and K. Strassburger, *On the simes test under dependence*, Statistical Papers, 58 (2017), pp. 775–789.

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

[21] ———, *Graph-theoretic measures of multivariate association and prediction*, The Annals of Statistics, 11 (1983), pp. 377–391.

[22] M. Frisén, *Statistical surveillance. optimality and methods*, International Statistical Review, 71 (2003), pp. 403–434.

[23] P. Fryzlewicz et al., *Wild binary segmentation for multiple change-point detection*, The Annals of Statistics, 42 (2014), pp. 2243–2281.

[24] J. Gou, *Least conservative critical boundaries of multiple hypothesis testing in a range of correlation values*, Statistics in Biopharmaceutical Research, (2021), pp. 1–9.

[25] J. Gou and A. C. Tamhane, *Hochberg procedure under negative dependence*, Statistica Sinica, (2018), pp. 339–362.

[26] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, *A kernel method for the two-sample-problem*, in Advances in neural information processing systems, 2007, pp. 513–520.

[27] A. Gretton et al., *A kernel two-sample test*, Journal of Machine Learning Research, 13 (2012), pp. 723–773.

[28] ———, *Optimal kernel choice for large-scale two-sample tests*, in Advances in neural information processing systems, 2012, pp. 1205–1213.

[29] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur, *A fast, consistent kernel two-sample test*, in Advances in neural information processing systems, 2009, pp. 673–681.

[30] Z. Harchaoui and O. Cappé, *Retrospective mutiple change-point estimation with kernels*, in 2007 IEEE/SP 14th Workshop on Statistical Signal Processing, IEEE, 2007, pp. 768–772.

[31] Z. Harchaoui, E. Moulines, and F. R. Bach, *Kernel change-point analysis*, in Advances in Neural Information Processing Systems, 2009, pp. 609–616.

[32] N. Henze, *A multivariate two-sample test based on the number of nearest neighbor type coincidences*, The Annals of Statistics, (1988), pp. 772–783.

[33] T. P. Hettmansperger, J. Möttönen, and H. Oja, *Affine invariant multivariate rank tests for several samples*, Statistica Sinica, (1998), pp. 785–800.

[34] Y. Hochberg and D. Rom, *Extensions of multiple testing procedures based on simes' test*, Journal of Statistical Planning and Inference, 48 (1995), pp. 141–152.

[35] S. Huang, Z. Kong, and W. Huang, *High-dimensional process monitoring and change point detection using embedding distributions in reproducing kernel hilbert space*, IIE Transactions, 46 (2014), pp. 999–1016.

[36] N. A. James and D. S. Matteson, *ecp: An r package for nonparametric multiple change point analysis of multivariate data*, arXiv preprint arXiv:1309.3295, (2013).

[37] D. Ji, J. Lu, Y. Zhang, S. Gao, and H. Zhao, *Inference of dynamic graph changes for functional connectome*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 3230–3240.

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

[39] ———, *Asymptotic distribution-free change-point detection based on interpoint distances for high-dimensional data*, Journal of Nonparametric Statistics, 32 (2020), pp. 157–184.

[40] S. Li, Y. Xie, H. Dai, and L. Song, *M-statistic for kernel change-point detection*, in Advances in Neural Information Processing Systems, 2015, pp. 3366–3374.

[41] D. Lopez-Paz and M. Oquab, *Revisiting classifier two-sample tests*, arXiv preprint arXiv:1610.06545, (2016).

[42] A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé, *Homogeneity and change-point detection tests for multivariate data using rank statistics*, arXiv preprint arXiv:1107.1971, (2011).

[43] D. S. Matteson and N. A. James, *A nonparametric approach for multiple change point analysis of multivariate data*, Journal of the American Statistical Association, 109 (2014), pp. 334–345.

[44] H. Oja, *Multivariate nonparametric methods with R: an approach based on spatial signs and ranks*, Springer Science & Business Media, 2010.

[45] A. B. Olshen, E. Venkatraman, R. Lucito, and M. Wigler, *Circular binary segmentation for the analysis of array-based dna copy number data*, Biostatistics, 5 (2004), pp. 557–572.

[46] W. Pan, Y. Tian, X. Wang, and H. Zhang, *Ball divergence: nonparametric two sample test*, Annals of statistics, 46 (2018), p. 1109.

[47] A. Ramdas, S. J. Reddi, B. Poczos, A. Singh, and L. Wasserman, *Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing*, arXiv preprint arXiv:1508.00655, (2015).

[48] E. A. Rødland, *Simes' procedure is valid on average*, Biometrika, 93 (2006), pp. 742–746.

[49] P. R. Rosenbaum, *An exact distribution-free test comparing two multivariate distributions based on adjacency*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 515–530.

[50] V. Rousson, *On distribution-free tests for the multivariate two-sample location-scale model*, Journal of multivariate analysis, 80 (2002), pp. 43–57.

[51] D. Rybach, C. Gollan, R. Schluter, and H. Ney, *Audio segmentation for speech recognition using segment features*, in 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2009, pp. 4197–4200.

[52] E. Samuel-Cahn, *Is the simes improved bonferroni procedure conservative?*, Biometrika, 83 (1996), pp. 928–933.

[53] S. K. Sarkar, *Some probability inequalities for ordered mtp2 random variables: a proof of the simes conjecture*, Annals of Statistics, (1998), pp. 494–504.

[54] S. K. Sarkar and C.-K. Chang, *The simes method for multiple hypothesis testing with positively dependent test statistics*, Journal of the American Statistical Association, 92 (1997), pp. 1601–1608.

[55] M. F. Schilling, *Multivariate two-sample tests based on nearest neighbors*, Journal of the American Statistical Association, 81 (1986), pp. 799–806.

[56] B. Schölkopf, A. J. Smola, F. Bach, et al., *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press, 2002.

[57] D. Siegmund and B. Yakir, *The statistics of gene mapping*, Springer Science & Business Media, 2007.

[58] H. Song and H. Chen, *Generalized kernel two-sample tests*, arXiv preprint arXiv:2011.06127, (2020).

[59] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet, *Hilbert space embeddings and metrics on probability measures*, Journal of Machine Learning Research, 11 (2010), pp. 1517–1561.

[60] G. J. Székely and M. L. Rizzo, *Energy statistics: A class of statistics based on distances*, Journal of statistical planning and inference, 143 (2013), pp. 1249–1272.

[61] K. M. Tyssowski, N. R. DeStefino, J.-H. Cho, C. J. Dunn, R. G. Poston, C. E. Carty, R. D. Jones, S. M. Chang, P. Romeo, M. K. Wurzelmann, et al., *Different neuronal activity patterns induce different gene expression programs*, Neuron, 98 (2018), pp. 530–546.

[62] L. Y. Vostrikova, *Detecting disorder in multidimensional random processes*, in Doklady Akademii Nauk, vol. 259, Russian Academy of Sciences, 1981, pp. 270–274.

[63] G. Wang, C. Zou, and G. Yin, *Change-point detection in multinomial data with a large number of categories*, Annals of Statistics, (2017).

[64] T. Wang and R. J. Samworth, *High dimensional change point estimation via sparse projection*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80 (2018), pp. 57–83.

[65] Y. Wang, C. Wu, Z. Ji, B. Wang, and Y. Liang, *Non-parametric change-point method for differential gene expression detection*, PloS one, 6 (2011), p. e20060.

[66] S. Zahid, F. Hussain, M. Rashid, M. H. Yousaf, and H. A. Habib, *Optimized audio classification and segmentation algorithm by using ensemble methods*, Mathematical Problems in Engineering, 2015 (2015).

[67] W. Zaremba, A. Gretton, and M. Blaschko, *B-test: A non-parametric, low variance kernel two-sample test*, in Advances in neural information processing systems, 2013, pp. 755–763.

[68] N. R. Zhang, D. O. Siegmund, H. Ji, and J. Z. Li, *Detecting simultaneous changepoints in multiple sequences*, Biometrika, (2010).