DISTRIBUTED INFERENCE FOR DEGENERATE U-STATISTICS WITH APPLICATION TO ONE AND TWO SAMPLE TEST

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Ernest Atta-Asiamah

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Statistics

April 2020

Fargo, North Dakota

NORTH DAKOTA STATE UNIVERSITY

Graduate School

Title

DISTRIBUTED INFERENCE FOR DEGENERATE U-STATISTICS WITH APPLICATION TO ONE AND TWO SAMPLE TEST

By

Ernest Atta-Asiamah

The supervisory committee certifies that this dissertation complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Mingao Yuan
Chair

Dr. Gang Shen

Dr. Megan Orr

Dr. Edward Deckard

Approved:

26 May 2020
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

In many hypothesis testing problems such as one-sample and two-sample test problems, the test statistics are degenerate U-statistics. One of the challenges in practice is the computation of U-statistics for a large sample size. Besides, for degenerate U-statistics, the limiting distribution is a mixture of weighted chi-squares, involving the eigenvalues of the kernel of the U-statistics. As a result, it's not straightforward to construct the rejection region based on this asymptotic distribution. In this research, we aim to reduce the computation complexity of degenerate U-statistics and propose an easy-to-calibrate test statistic by using the divide-and-conquer method. Specifically, we randomly partition the full n data points into k_n even disjoint groups, and compute U-statistics on each group and combine them by averaging to get a statistic T_n. We proved that the statistic T_n has the standard normal distribution as the limiting distribution. In this way, the running time is reduced from $O(n^m)$ to $O(\frac{n^m}{k_n^{m-1}})$, where m is the order of the one sample U-statistics. Besides, for a given significance level α, it's easy to construct the rejection region. We apply our method to the goodness of fit test and two-sample test. The simulation and real data analysis show that the proposed test can achieve high power and fast running time for both one and two-sample tests.

ACKNOWLEDGEMENTS

I want to thank Dr. Mingao Yuan for his tirelessness support and readily guidance towards this dissertation. I am delighted to work with you and very appreciate his kindness. I would also like to express my great appreciation and thanks to Dr. Gang Shen, Dr. Megan Orr, and Dr. Edward L. Deckard for their time, efforts and contribution toward the completion of this research.

My special thanks and gratitude goes to my wife, Salomey Owusu, for their advice and support towards my education. She was always there stood by me. She has accompanied me through good times and bad.

I also want to express my gratitude to my cousin and co-teaching assistant, Adu Boampong Asare, for his contribution toward the completion of this dissertation. He is always there for me in terms of advice and motivation. It was through him that I joined NDSU.

I also want to thank my siblings for their constant support and encouragement toward this academic journey. Lastly, my appreciation and thanks to Yaa Sefah, and Dominic Bosomtwe for their incredible support. God Bless you all.

DEDICATION

This dissertation is dedicated to my late great grandmother, Madam Yaa Abayewa

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

The U-statistic is one of the most commonly used non-linear and non-parametric statistics, introduced by Hoeffding (1948). The class arises as a generalization of the notion of an unbiased sample means, sample variance, Kendall's rank correlation coefficient, Gini's mean difference. Usually, U-statistics produce minimum-variance unbiased estimators.

U-statistics have been widely investigated in theoretical and applied statistics. For example, Bickel and Freedman(1981) studied the bootstrap of non-degenerate U-statistics and proved the bootstrap consistency. Arcones and Gine(1992) proved the bootstrap CLT for U statistics under minimal integrability conditions; Peng and Tan(2018) proved the Wilks theorems for jackknife empirical likelihood for vector U-statistics; Cheng et al. (2018) investigated the two-sample U-statistics via jackknife empirical likelihood. Huang, W et al. (2006) and Dewan et al. (2001) proposed a central limit theorem for degenerate and non-degenerate U-statistics when the sequence is negatively related to random variables.

U-statistics have full application in many estimations and machine learning problems. For instance, the MeanNN approach estimation for differential entropy introduced by Faivishevsky and Goldberger (2008) is a U statistic. Using U-statistics, Liu et al. (2016) proposed a new test statistic for goodness-of-fit tests; Clemencon(2011) defined a measure by U-statistics to quantify the clustering quality of a partition.

U-statistics applied in statistical inference and estimation, including the simultaneous testing of different hypotheses, the estimation of high dimensional graphical models. For high dimensional hypothesis testing, the new methods based on U-statistics have been proposed and studied in Chen, Zhang, and Zhong (2010) and Zhong and Chen (2011). Further, the degenerate of order-1 U-statistics aroused in the context of testing for independence in paired circular data, for instance, the tests studied in Fisher and Lee (1982). Other degenerate U-statistics proposed for testing goodness-of-fit, for example, Watson's U^2 for two-sample, goodness-of-fit on a circle Persson (1979), the Cramer-von Mises type statistics for one-sample goodness-of-fit in Anderson and Darling (1952).

U-statistic can be degenerate or non-degenerate. Even though most of the literature on U-statistics focuses on the non-degenerate case, degenerate U-statistics are very useful in many

hypothesis testing problems. For example, the energy test statistic or maximum mean discrepancy (MMD)(Gretton et al. 2012) for two sample problems are degenerate U-statistics under the null hypothesis; in the goodness of fit test, the kernelized stein discrepancy(KSD) test is also a degenerate U-statistics(Liu et al. 2016) and Atta-Asiamah and Yuan (2019). Many other examples of degenerate U-statistics could be found in testing for independence and model misspecification, in the field of physical and social science such as geophysics Stephens (1979), econometrics Bierens and Ploberger (1997) and ecology (Fisher and Lee, 1982).

The asymptotic behavior of U-statistics are well studied. A non-degenerate U-statistic has the normal distribution as the limiting distribution. In the degenerate case, the limiting distribution is a mixture of independent chi-square distributions, weighted by the eigenvalues of the kernel of the U statistic. Concentration inequalities are also available in the literature.

In the application of (non-degenerate or degenerate) U-statistics, one challenge is the calculation when the sample size is large. For U-statistics of order m, the computational complexity is $O(n^m)$. Another challenge in using degenerate U-statistics for hypothesis testing is: under the null hypothesis H_0, the limiting distribution of a degenerate U-statistic is a quadratic form of independent standard Gaussian random variables, weighted by the eigenvalues of the kernel of the U-statistics. To solve these challenges, Atta-Asiamah, E and Yuan, M (2019) proposed a divide-and-conquer method to deal with the eigenvalues challenges and computational cost. In practice, it is hard to get the closed-form expression of the eigenvalues. Other techniques like bootstrapping or permutation test should be employed, which may increase the computation burden for large sample size.

2. METHODOLOGY

2.1. Divide and conquer for degenerate U-statistics

Let $(X_1, X_2, \ldots, X_n)$ be i.i.d. data from some distribution. For some symmetric function $h : \mathbb{R}^m \to \mathbb{R}$, the U-statistics of order m is defined as

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < \cdots < i_m \le n} h(X_{i_1}, \ldots, X_{i_m}),$$

which is an unbiased estimator of $\theta = \mathbb{E}h(X_1, \ldots, X_m)$. For $c = 0, 1, \ldots, m$, let $h_c(x_1, \ldots, x_c) = \mathbb{E}h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m)$ and $\sigma_c^2 = Var(h_c(X_1, \ldots, X_c))$. A U-statistics is said to be k-degenerate of order m if $\sigma_1^2 = \cdots = \sigma_k^2 = 0$ and $\sigma_{k+1}^2 \neq 0$. When $k = 0$, the U-statistics is non-degenerate. For convenience and simplicity, in this research we consider 1-degenerate U-statistics of order 2, that is, $m = 2$ and $k = 1$. In the following, we provide several examples of U-statistics.

Example 2.1.1. *Suppose $X \sim F$ with variance $\theta = Var(X)$, defined as*

$$\theta = \int (x - \mu)^2 dF(x).$$

An unbiased estimator of θ is the sample variance, which is a U-statistics. Specifically, define kernel h as

$$h(x_1, x_2) = \frac{{x_1}^2 + {x_2}^2 - 2x_1x_2}{2} = \frac{1}{2}(x_1 - x_2)^2$$

and the corresponding U-statistic is given as

$$\begin{aligned} U_n(X_1, \ldots, X_n) &= \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} h(X_i, X_j) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{n} {x_1}^2 - n\bar{x} \right) \\ &= S^2 \end{aligned}$$

is the sample variance.

Example 2.1.2. *Let X_1 and X_2 be independent samples from a distribution F. Suppose $\theta_F = E_F|X_1 - X_2|$ is measure of the concentration, called Gini's difference. Define the kernel h as $h(x_1, x_2) = |x_1 - x_2|$, the estimator of θ is a U-statistics of order 2:*

$$U_n = U(X_1, \ldots, X_n) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} |X_i - X_j|.$$

Example 2.1.3. *We consider a kernel, $h(x_1, x_2) = x_1 x_2$. Then $h_1(x_1) = \mathbb{E}(x_1 X_2) = x_1 \mathbb{E}(X_2) = x_1 \mu$ and $\sigma_1^2 = Var(h_1(X_1)) = Var(X_1 \mu) = \mu^2 Var(X_1) = \mu^2 \sigma^2$. Suppose that $\mu = E(X_1) = 0$, then $\sigma_1^2 = 0$. But assuming $\sigma^2 > 0$, then $\sigma_2^2 = Var(X_1 X_2) = Var(X_1) Var(X_2) = \sigma^2 \sigma^2 = \sigma^4 > 0$. Therefore it is degenerate of order 1.*

In this thesis, we are mainly interested in testing the following hypothesis

$$H_0 : \theta = 0, \ v.s. \ H_1 : \theta \neq 0. \tag{2.1}$$

In many cases, the U-statistics U_n serve as a test statistic and frequently the U_n is degenerate under H_0, see (Liu et al (2016)), (Gretton, et al(2012)) and Atta-Asiamah, E and Yuan, M (2019) for example. The classical asymptotic distribution of degenerate U_n is given in Lemma 2.1.1 as follows.

Lemma 2.1.1. *Let U_n be the 1-degenerate U-statistics of order 2 and $\mathbb{E}U_n = \theta$. Then we have*

$$n(U_n - \theta) \to \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1),$$

where Z_j, $j = 1, 2, \ldots$ follow independent standard normal distribution, and λ_j, $j = 1, 2, \ldots$ are eigenvalues of the function $h(x_1, x_2) - \theta$.

In practice, there are two challenges in using Lemma 2.1.1 to construct the rejection region: it is computationally hard to compute U_n for large sample size and it is pretty challenging to calculate the eigenvalues λ_i. We address these two issues simultaneously by the divide-and-conquer method as follows.

Firstly, we partition the i.i.d. sample $(X_1, X_2, \ldots, X_n)$ into k_n disjoint blocks, each block with $m_n = \frac{n}{k_n}$ samples. Then construct U-statistics U_i for i-th block$(1 \leq i \leq k_n)$ and combine them to get the following quantity $\mathcal{T}_n$

$$\mathcal{T}_n = \frac{\sum_{i=1}^{k_n} m_n U_i}{s_n},$$

where $s_n^2 = \sum_{i=1}^{k_n} Var(m_n U_i) = k_n \sigma_{m_n}^2$. Note that $\sigma_{m_n}^2$ converges to σ^2, the variance of $G = \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1)$, as m_n goes to infinity. Under H_0, we derive the limiting distribution of $\mathcal{T}_n$ as in the following Theorem 2.1.1.

Theorem 2.1.1. *Suppose $h(x_1, x_2)$ is a bounded symmetric function. If $k_n \to \infty$ and $m_n = \frac{n}{k_n} \to \infty$ as n goes to infinity, then under H_0 we have*

$$\mathcal{T}_n = \frac{\sum_{i=1}^{k_n} m_n U_i}{s_n} \to N(0, 1).$$

Note that $\mathcal{T}_n$ involves unknown variance $\sigma_{m_n}^2$. We propose to estimate it by $\hat{\sigma}_{m_n}^2 = \frac{1}{k_n} \sum_{i=1}^{k_n} \left(m_n U_i\right)^2$. By the proof of Theorem 2.1.1, $\{(m_n U_i)^2\}$ is uniformly integrable, which implies that $\hat{\sigma}_{m_n}^2$ converges to σ^2 in probability. Then the test statistic for (2.1) is defined as

$$\hat{\mathcal{T}}_n = \frac{\sum_{i=1}^{k_n} m_n U_i}{\sqrt{k_n} \hat{\sigma}_{m_n}}. \tag{2.2}$$

For given type I error α, we reject H_0 if $|\hat{\mathcal{T}}_n| > Z_{\frac{\alpha}{2}}$, where $Z_{\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})100\%$ quantile of the standard normal distribution.

We point out two advantages of the test statistic $\hat{\mathcal{T}}_n$. Firstly, the running time of $\hat{\mathcal{T}}_n$ is $O(\frac{n^2}{k_n})$, which is significantly faster than $O(n^2)$, the running time of U_n. For example, if we take $k_n = \frac{n}{\log n}$, then the running time of $\hat{\mathcal{T}}_n$ is $O(n \log n)$, almost linear. Besides, the limiting distribution of $\hat{\mathcal{T}}_n$ is standard normal distribution, which doesn't require computing the eigenvalues of the kernel function h.

To study the power of the proposed test, we consider the following hypothesis.

$$H_0 : \theta = 0, \ v.s. \ H_a : \theta = \frac{c}{\sqrt{n}}, \tag{2.3}$$

where $c \neq 0$ is some constant. Under H_a, we have

Theorem 2.1.2. *If $k_n = o(n)$ and the U-statistic is non-degenerate under H_a, $\mathcal{T}_n$ converges in distribution to $N(\frac{c}{\sigma}, 1)$ under H_a.*

By Theorem 2.1.2, the proposed test statistic $\hat{\mathcal{T}}_n$ can achieve high power as n tends to infinity. Moreover, the optimal test rate of our test statistic is $\sqrt{n}$, the same as the rate for the non-degenerate case (see Theorem 2.2.2). In many cases, under H_a, the U-statistic is non-degenerate, see (Liu et al (2016)), (Gretton, et al(2012)) for example. If under H_a, the U-statistic is still 1-degenerate, the optimal test rate would be $\frac{\sqrt{k_n}}{n}$, which is faster than $\sqrt{n}$(see the Remark below the proof of Theorem 2.1.2).

2.2. Divide and conquer for non-degenerate U-statistics

The divide-and-conquer method for non-degenerate U-statistics was studied in Lin, N. and Xi, R.(2010) and Atta-Asiamah, E and Yuan, M (2019). For completeness and to compare with the degenerate case, we adjust their results to hypothesis testing problem. Consider the following hypothesis

$$H_0 : \theta = \theta_0, \ v.s. \ H_1 : \theta \neq \theta_0. \tag{2.4}$$

Following a similar procedure as in the degenerate case, we define the test statistic for (2.4) as

$$\mathcal{T}_n = \frac{\sqrt{k_n}(T_n - \theta_0)}{\sqrt{\frac{1}{k_n}\sum_{i=1}^{k_n}(U_i - \overline{U})^2}}. \tag{2.5}$$

where $T_n = \frac{1}{k_n}\sum_{i=1}^{k_n} U_i$ and $\overline{U} = \frac{1}{k_n}\sum_{i=1}^{k_n} U_i$. Under H_0, the limiting distribution of $\mathcal{T}_n$ is also standard normal distribution (Lin, N. and Xi, R.(2010)).

Theorem 2.2.1. *If $k_n = o(n)$, then under H_0, $\mathcal{T}_n$ converges in distribution to $N(0, 1)$.*

Given type I error α, we reject H_0 if $|\mathcal{T}_n| > Z_{\frac{\alpha}{2}}$, where $Z_{\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})100\%$ quantile of the standard normal distribution. To study the power of the proposed test, we consider the following hypothesis test,

$$H_0 : \theta = \theta_0, \ v.s. \ H_a : \theta = \theta_0 + \frac{c}{\sqrt{n}}, \tag{2.6}$$

where $c \neq 0$ is some constant. Under H_a, we can easily derive the asymptotic distribution.

Theorem 2.2.2. *If $k_n = o(n)$ and the U-statistic is non-degenerate under H_a, then $\mathcal{T}_n$ converges in distribution to $N(\frac{c}{\sigma}, 1)$ under H_a, where $\frac{1}{k_n}\sum_{i=1}^{k_n}(\sqrt{m_n}U_i - \sqrt{m_n}\overline{U})^2$ converges to σ^2 in probability.*

By Theorem 2.2.2, the optimal rate of the test statistic $\mathcal{T}_n$ is $\sqrt{n}$, equal to the optimal test rate of degenerate case. The computation complexity of the test is $O(\frac{n^m}{k_n^{m-1}})$, much smaller than that of the full sample $O(n^m)$. In terms of computation time and optimal test rate, there is no difference between the non-degenerate and the degenerate case of the divide-and-conquer method.

2.3. Divide and conquer for two-sample U-statistics

In two sample testing problems, the test statistic are usually degenerate U-statistics under the null hypothesis. For instances, the maximum mean discrepancy for comparing two distributions is a degenerate two-sample U-statistic (Gretton et al.(2012)). We propose the divide and conquer methods for two sample U-statistics in this subsection.

Let G and H be independent continuous distributions and $\theta = \theta(G, H)$ is a parameter defined as follows: for a measurable function $h(x_1, x_2, x_3, \ldots, x_{k_1}; y_1, y_2, y_3, \ldots, y_{k_2})$,

$$\theta = \int_{-\infty}^{+\infty} h(x_1, x_2, x_3, \ldots, x_{k_1}; y_1, y_2, y_3, \ldots, y_{k_2}) \prod_{i=1}^{k_1} dG(x_i) \prod_{j=1}^{k_2} dH(x_j)$$

Suppose $h(x_1, x_2, x_3, \ldots, x_{k_1}; y_1, y_2, y_3, \ldots, y_{k_2})$ is symmetric with respect to $x_1, x_2, x_3, \ldots, x_{k_1}$ and $y_1, y_2, y_3, \ldots, y_{k_2}$, respectively. Suppose $X_1, X_2, X_3, \ldots, X_{n_1}$ and $Y_1, Y_2, Y_3, \ldots, Y_{n_2}$ be the two independent samples from the distributions G and H respectively. Then an unbiased estimator of θ is a U-statistic of two-sample U_{n_1,n_2} with degree (k_1, k_2) given as

$$U_{n_1,n_2} = \frac{1}{\binom{n_1}{k_1}\binom{n_2}{k_2}} \sum_{(n_1,k_1)} \sum_{(n_2,k_2)} h(X_{i_1}, \ldots, X_{i_{k_1}}; Y_1, \ldots, Y_{k_2}),$$

where the summation $\sum_{(n_1,k_1)}$ is taken all possible values of $x_1, x_2, x_3, \ldots, x_{k_1}$ which satisfying $1 \le i_1 < \cdots < i_{k_1} \le n_1$ and $\sum_{(n_2,k_2)}$ is similarly defined. Some examples of two sample U-statistic is given below.

Example 2.3.1. *Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be independent samples from continuous distributions F and G, respectively. Let θ be a parameter defined as*

$$\theta(F, G) = \int F dG = P(X \leq Y).$$

Therefore, an unbiased estimator of θ is

$$U = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(X_i \leq Y_j).$$

This is the Wilcoxon 2-sample U-statistics.

In this thesis, we are specially interested in the two sample U statistics in two sample test problems. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independently and identically distributed random variables from distribution p and q respectively. The two sample test problem is to test the following hypotheses

$$H_0 : p = q, \qquad H_1 : p \neq q. \tag{2.7}$$

Under H_0, the two samples are from the same distribution, while under H_1, they are from different distributions.

Many test procedures are available in the literature. Among them, the Maximum Mean Discrepancy(MMD) test statistic is one of the most popular and has good performance. Let $K(x, y)$ be a positive definite symmetric kernel function for RKHS. The Maximum Mean Discrepancy(MMD) test statistic is defined as (Gretton, et al (2012)).

$$T_{mn} = \frac{1}{\binom{m}{2}} \sum_{1 \leq i < j \leq m} K(X_i, X_j) + \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} K(Y_i, Y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} K(X_i, Y_j).$$

Under H_0, the limiting distribution of T_{mn} is given by the following result.

Proposition 2.3.1 (Gretton, et al (2012))**.** *Let $\frac{m}{m+n} \to \rho_x$ and $\frac{n}{m+n} \to \rho_y$ with $0 < \rho_x < 1$, $0 < \rho_y < 1$. Then under H_0, we have*

$$(m+n) T_{mn} \to \sum_{l=1}^{\infty} \lambda_l \left[(\rho_x^{-\frac{1}{2}} a_l - \rho_y^{-\frac{1}{2}} b_l)^2 - (\rho_x \rho_y)^{-1} \right],$$

where a_l, b_l are independent standard normal random variables, λ_l are the eigenvalues of the centered kernel $\widetilde{K}(x,y)$ of $K(x,y)$.

The computational cost of MMD is $O(m+n)^2$. To alleviate the intensive computation for large sample sizes m and n, a linear time test statistic was proposed if $m = n$ Gretton, et al (2012), but it performs poorly based on our simulation. The limiting distribution contains the eigenvalues of the kernel for the RKHS, which is usually very difficult to estimate in the application. We need to either estimate them or use other techniques such as bootstrap to get the critical value. However, all these procedures will definitely increase the computation burden.

In this research, we propose a test statistic by using the divide-and-conquer method to overcome these issues simultaneously. Specifically, we divide randomly and evenly $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ into k groups respectively. Based on the i-th group samples, calculate the MMD test statistic, denoted as $T_{(i)}$, and then average them as

$$\widehat{T}_{mn} = \frac{1}{k}\sum_{i=1}^{k} T_{(i)}.$$

Let $m_1 = \frac{m}{k}$ and $n_1 = \frac{n}{k}$. Then the divide-and-conquer test statistic for (2.7) is defined as

$$T_k = \frac{\sqrt{k}(m_1+n_1)\widehat{T}_{mn}}{s_k},$$

where $s_k^2 = \frac{1}{k}\sum_{i=1}^{k}(m_1+n_1)^2 T_{(i)}^2$.

The asymptotic distribution of T_k under the null hypothesis is given in the following theorem.

Theorem 2.3.1. *Suppose the kernel function $K(x,y)$ is bounded, $m = cn$ for some constant $c > 0$, $k \to \infty$ and $k = o(m)$ as m goes to infinity. Then under H_0, T_k converges to $N(0,1)$ in distribution.*

According to Theorem 2.3.1, the limiting distribution is the standard normal distribution. It's easy to calibrate the test statistic, without calculating the eigenvalues of the kernel function. Given significance level α, reject H_0 if $|T_k| > Z_{\frac{\alpha}{2}}$, where $Z_{\frac{\alpha}{2}}$ is the $100(1-\frac{\alpha}{2})\%$ quantile of the standard normal distribution. The computation complexity of T_k is at most $O\Big(\frac{\max\{m^2,n^2\}}{k}\Big)$, which can be almost linear if $k = \frac{n}{\log n}$. However, larger k will decrease the power based on our simulation. There is a trade-off between running time and power.The power of our test statistic is evaluated by simulation study

3. PROOF OF THE MAIN THEOREM

In this section, we provide the proof of the main results.

3.1. Proof of main theorem in one sample case

Proof of Theorem 2.1.1: Let c_1, c_2, c_3 be universal constants. For 1-degenerate U-statistics of order 2, we have the following tail probability Arcones and Gine (1993) and Atta-Asiamah and Yuan (2019).

$$\mathbb{P}\Big(|U_i(h)| \geq c_1||h||_\infty \frac{\log \frac{c_2}{\delta}}{m_n}\Big) \leq \delta,$$

which implies that for $t > 0$

$$\mathbb{P}\Big(|U_i(h)| \geq t) \leq c_2 \exp\big(-\frac{m_n t}{c_1||h||_\infty}\big).$$

Hence, we have

$$\mathbb{P}\Big(|m_n U_i(h)|^2 \geq t) \leq c_2 \exp\big(-\frac{\sqrt{t}}{c_1||h||_\infty}\big).$$

Direct computation yields

$$\int_t^{+\infty} \mathbb{P}\Big(|m_n U_i(h)|^2 \geq x)dx \leq \int_t^{+\infty} c_2 \exp\big(-\frac{\sqrt{x}}{c_1||h||_\infty}\big)dx \leq c_3 \int_t^{+\infty} \frac{1}{x^2}dx = \frac{c_3}{t} \to 0,$$

as $t \to \infty$. Here, in the second inequality we used the fact that h is bounded. Then one has

$$\begin{aligned}
\mathbb{E}(|m_n U_i|^2 1[|m_n U_i| > t]) &= \int_0^{+\infty} \mathbb{P}(|m_n U_i|^2 1[|m_n U_i| > t] > x)dx \\
&\leq \int_0^{t^2} \mathbb{P}(|m_n U_i|^2 > t^2)dx + \int_{t^2}^{+\infty} \mathbb{P}(|m_n U_i|^2 > x)dx \\
&\leq c_2 t^2 \exp\big(-\frac{t}{c_1||h||_\infty}\big) + \frac{c_3}{t^2}.
\end{aligned}$$

Next we can verify the Lindeberg condition. For any $\epsilon > 0$, let $t = \epsilon s_n$, it follows that

$$\frac{1}{s_n^2}\sum_{i=1}^{k_n} \mathbb{E}(|m_n U_i|^2 1[|m_n U_i| > t]) \leq \frac{1}{\sigma_{m_n}^2}\Big(c_2 \epsilon^2 s_n^2 \exp\big(-\frac{\epsilon s_n}{c_1||h||_\infty}\big) + \frac{c_3}{\epsilon^2 s_n^2}\Big) \to 0,$$

if $m_n \to \infty$ and $k_n \to \infty$. Hence, by the central limit theorem, the desired result follows. □

Proof of Theorem 2.1.2: Under H_a, we have

$$\frac{\frac{1}{\sqrt{k_n}}\sum_{i=1}^{k_n} m_n U_i}{\sqrt{Var(m_n U_i)}} = \frac{\frac{1}{\sqrt{k_n}}\sum_{i=1}^{k_n} m_n (U_i - \frac{c}{\sqrt{n}})}{\sqrt{m_n Var(\sqrt{m_n} U_i)}} + \frac{\sqrt{k_n m_n}\frac{c}{\sqrt{n}}}{\sqrt{Var(\sqrt{m_n} U_i)}}. \tag{3.1}$$

For the first term in the right-hand side of (3.1), if U_i is non-degenerate H_a, it converges to $N(0,1)$ in distribution by a similar proof of Theorem 2 in Lin and Xi (2010) and the second term converges to $\frac{c}{\sigma}$ in probability. Then the limiting distribution of $\mathcal{T}_n$ is $N(\frac{c}{\sigma}, 1)$ under H_a. □

Remark If U_i is still degenerate under H_a, the first term converges to $N(0,1)$ in distribution by a similar proof of Theorem 2.1.1, while the second term is $O_p\Big(\sqrt{\frac{n}{k_n}}c\Big)$. In this case, the optimal test rate is $\frac{\sqrt{k_n}}{n}$.

Proof of Theorem 2.2.1: The proof follows directly from that of Theorem 2 in Lin and Xi (2010). □

Proof of Theorem 2.2.2: The proof is similar to that of Theorem 2.1.2. □

3.2. Proof of main theorem in two-sample test

Suppose $m = cn$ for some constant $c \geq 0$. In this case, $m_1 = cn_1$. We center the kernel to get $\widetilde{K}(X_i, X_j)$ as follows

$$\widetilde{K}(X_i, X_j) = K(X_i, X_j) - \mathbb{E}_x K(X_i, x) - \mathbb{E}_x K(x, X_j) + \mathbb{E}_{x,y} K(x, y).$$

Note that $T_{(i)}$ is a sum of three U-statistics, that is,

$$\begin{aligned} T_{(i)} &= \frac{1}{\binom{m_i}{2}} \sum_{1\leq i<j\leq m_i} \widetilde{K}(X_i, X_j) + \frac{1}{\binom{n_i}{2}} \sum_{1\leq i<j\leq n_i} \widetilde{K}(Y_i, Y_j) - \frac{2}{m_i n_i} \sum_{i=1}^{m_i}\sum_{j=1}^{n_i} \widetilde{K}(X_i, Y_j) \\ &= U_{i1} + U_{i2} + U_{i3}, \end{aligned}$$

where U_{i1} and U_{i2} are degenerate one-sample U-statistics, and U_{i3} is a two-sample U-statistic.

By Arcones and Gine (1993), if the kernel K is bounded, for some generic positive constants c_1 and c_2, we get for $t > 0$,

$$\mathbb{P}\big(|U_{i1}| > t\big) \leq c_1 \exp\{-c_3 m_1 t\}, \tag{3.2}$$

$$\mathbb{P}\big(|U_{i2}| > t\big) \leq c_1 \exp\{-c_3 n_1 t\}. \tag{3.3}$$

Next we show a similar concentration inequality holds for U_{i3}. Let $Z_i = X_i$ for $i = 1, 2, \ldots, m$ and $Z_i = Y_i$ for $i = m+1, \ldots, m+n$. Then we have

$$\begin{aligned}
\sum_{i,j=1}^{m+n} \widetilde{K}(Z_i, Z_j) &= \sum_{i,j=1}^{m} \widetilde{K}(Z_i, Z_j) + \sum_{i,j=m+1}^{m+n} \widetilde{K}(Z_i, Z_j) + \sum_{i=1}^{m}\sum_{j=m+1}^{m+n} \widetilde{K}(Z_i, Z_j) \\
&\quad + \sum_{j=1}^{m}\sum_{i=m+1}^{m+n} \widetilde{K}(Z_i, Z_j) \\
&= \sum_{i,j=1}^{m} \widetilde{K}(X_i, X_j) + \sum_{i,j=m+1}^{m+n} \widetilde{K}(Y_i, Y_j) + \sum_{i=1}^{m}\sum_{j=1}^{n} \widetilde{K}(X_i, Y_j) \\
&\quad + \sum_{j=1}^{m}\sum_{i=1}^{n} \widetilde{K}(Y_i, X_j) \\
&= \sum_{i,j=1}^{m} \widetilde{K}(X_i, X_j) + \sum_{i,j=1}^{n} \widetilde{K}(Y_i, Y_j) + 2\sum_{i=1}^{m}\sum_{j=1}^{n} \widetilde{K}(X_i, Y_j).
\end{aligned}$$

Then it follows that

$$\begin{aligned}
U_{i3} &= \frac{1}{mn} \sum_{i,j=1}^{m+n} \widetilde{K}(Z_i, Z_j) - \frac{1}{mn} \sum_{i,j=1}^{m} \widetilde{K}(X_i, X_j) - \frac{1}{mn} \sum_{i,j=m+1}^{m+n} \widetilde{K}(Y_i, Y_j) \\
&= \frac{2}{mn} \sum_{i<j}^{m+n} \widetilde{K}(Z_i, Z_j) - \frac{2}{mn} \sum_{i<j}^{m} \widetilde{K}(X_i, X_j) - \frac{2}{mn} \sum_{i<j}^{n} \widetilde{K}(Y_i, Y_j) \\
&= U_{i3}^{(1)} + U_{i3}^{(2)} + U_{i3}^{(3)}.
\end{aligned}$$

Under Null hypothesis, H_0, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are from the same distribution. Hence each term in U_{i3} is a degenerate U-statistics. If $m = cn$ for some constant $c > 0$, by Arcones and Gine

(1993), for bounded K and some generic positive constants c_1 and c_3, we get for $t > 0$,

$$\begin{aligned}
\mathbb{P}\Big(|U_{i3}| > t\Big) &\leq \mathbb{P}\Big(|U_{i3}^{(1)}| > \frac{t}{3}\Big) + \mathbb{P}\Big(|U_{i3}^{(2)}| > \frac{t}{3}\Big) + \mathbb{P}\Big(|U_{i3}^{(3)}| > \frac{t}{3}\Big) \\
&\leq c_1 \exp\{-c_3(n_1+m_1)t\} + c_1 \exp\{-c_3 m_1 t\} + c_1 \exp\{-c_3 n_1 t\} \\
&\leq c_1 \exp\{-c_3 n_1 t\} \qquad (3.4)
\end{aligned}$$

By (3.2), (3.3) and (3.4), we have

$$\mathbb{P}\big(|(m_1+n_1)U_{i1}|^2 > t\big) \leq c_1 \exp\{-c_3\sqrt{t}\},$$

$$\mathbb{P}\big(|(m_1+n_1)U_{i2}|^2 > t\big) \leq c_1 \exp\{-c_3\sqrt{t}\},$$

$$\mathbb{P}\big(|(m_1+n_1)U_{i3}|^2 > t\big) \leq c_1 \exp\{-c_3\sqrt{t}\}.$$

Note that for any $x > 0$,

$$\begin{aligned}
\mathbb{P}(|(m_1+n_1)T_{(i)}|^2 \geq x) &\leq \mathbb{P}\Big(3\big(|(m_1+n_1)U_{i1}|^2 + |(m_1+n_1)U_{i1}|^2 + |(m_1+n_1)U_{i1}|^2\big) \geq x\Big) \\
&\leq \mathbb{P}\Big(|(m_1+n_1)U_{i1}|^2 \geq \frac{x}{9}\Big) + \mathbb{P}\Big(|(m_1+n_1)U_{i2}|^2 \geq \frac{x}{9}\Big) \\
&\quad + \mathbb{P}\Big(|(m_1+n_1)U_{i3}|^2 \geq \frac{x}{9}\Big).
\end{aligned}$$

Hence it's easy to get

$$\begin{aligned}
&\int_t^\infty \mathbb{P}(|(m_1+n_1)T_{(i)}|^2 \geq x)dx \leq \int_t^\infty \mathbb{P}(|(m_1+n_1)U_{i1}|^2 \geq c_1 x)dx \\
&+ \int_t^\infty \mathbb{P}(|(m_1+n_1)U_{i2}|^2 \geq c_1 x)dx + \int_t^\infty \mathbb{P}(|(m_1+n_1)U_{i3}|^2 \geq c_1 x)dx \\
\leq\ & c_1 \int_t^\infty \exp\{-c_3\sqrt{x}\}dx \leq c_1 \int_t^\infty \frac{1}{x^2}dx = \frac{c_1}{t}.
\end{aligned}$$

Then it follows that

$$\begin{aligned}\mathbb{E}(|(m_1+n_1)T_{(i)}|^2 1[|(m_1+n_1)T_{(i)}|>t]) &\leq \int_0^{t^2} P(|(m_1+n_1)T_{(i)}|^2 \geq t^2)dx \\ &\quad + \int_{t^2}^{\infty} P(|(m_1+n_1)T_{(i)}|^2 \geq x)dx \\ &\leq c_1 t^2 \exp\{-c_2 t\} + \frac{c_1}{t^2}.\end{aligned}$$

Let $s_n^2 = \sum_{i=1}^k Var((m_1+n_1)T_{(i)}) = k\sigma_{m_1}^2$. For any $\epsilon > 0$, it follows that

$$\frac{1}{s_n^2}\sum_{i=1}^{k}\mathbb{E}\big(|(m_1+n_1)T_{(i)}|^2 1[|(m_1+n_1)T_{(i)}|>\epsilon s_n]\big) \leq \frac{1}{\sigma_{m_1}^2}\Big(c_1\epsilon^2 s_n^2 \exp\big(-c_2\epsilon s_n\big) + \frac{c_2}{\epsilon^2 s_n^2}\Big) \to 0,$$

if $k, m_1 \to \infty$. By the Lindeberg Central Limit Theorem, the proof is complete.

4. THE SIMULATION AND REAL DATA

In this section, we apply the proposed divide and conquer methods for U-statistic to the goodness of fit test and two-sample test.

4.1. The simulation of goodness-of-fit test

Throughout this simulation, we run the experiments 500 times to get the size and power of the test. For the degenerate case, we consider the U-statistics for goodness-of-fit test proposed in Liu et al (2016) and Atta-Asiamah, E and Yuan, M (2019). Specifically, given $X_1, X_2, \ldots, X_n \sim q(x)$ (unknown), we consider the following hypothesis test

$$H_0 : q(x) = p(x), \ v.s. \ H_1 : q(x) = p_1(x).$$

Liu et al.(2016) proposed the *Kernel Stein Discrepancy*(KSD) test statistics in terms of U-statistics:

$$\hat{S}(p, q) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} u_p(X_i, X_j),$$

where

$$\begin{aligned} u_p(x, y) &= \nabla_x^T \log p(x) \nabla_y \log p(y) k(x, y) + \nabla_x^T \log p(x) \nabla_y k(x, y) \\ &\quad + \nabla_y^T \log p(y) \nabla_x k(x, y) + \sum_{i=1}^{d} \frac{\partial^2 k(x, y)}{\partial x_i \partial y_i}, \end{aligned}$$

and $k(x, y)$ is a symmetric kernel, which is usually taken to be the Gaussian kernel, $k(x, y) = e^{-\frac{(x-y)^2}{2}}$. Under H_0, the degenerate U-statistics $\hat{S}(p, q)$ has the limiting distribution in Lemma 2.1.1, which is not convenient to use in practice. They proposed to use bootstrap to get the rejection region. However, in the big data region, bootstrapping is quiet time-consuming. Instead, we use (2.2) as the test statistic and evaluate its size and power.

4.1.1. Degenerate U-statistics of 1-dimension with one sample case

Firstly, we consider the univariate case. Let $p(x) = N(0, 1)$ under H_0 and $p_1(x) = Laplace(0, \frac{1}{\sqrt{2}}), N(0, 1.1), N(0, 1.15)$ under H_1. Here, $\frac{1}{\sqrt{2}}$ in the Laplace distribution is the

same variance as the standard normal distribution as proposed in Atta-Asiamah, E and Yuan, M (2019). In this case,

$$\begin{aligned}
\frac{\partial k(x,y)}{\partial x} &= -e^{-\frac{(x-y)^2}{2}}(x-y) \\
\frac{\partial k(x,y)}{\partial y} &= e^{-\frac{(x-y)^2}{2}}(x-y) \\
\frac{\partial^2 k(x,y)}{\partial x \partial y} &= e^{-\frac{(x-y)^2}{2}} - e^{-\frac{(x-y)^2}{2}}(x-y)^2 \\
\frac{\partial \log p(x)}{\partial x} &= -x \\
\frac{\partial \log p(y)}{\partial y} &= -y.
\end{aligned}$$

Then the kernel of $\hat{S}(p,q)$ is

$$u_p(x,y) = e^{-\frac{(x-y)^2}{2}}\Big((xy+1) - 2(x-y)^2\Big),$$

which is bounded and symmetric.

Table 4.1. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$ and $k_n \in \{30, 50, 80, 100\}$ and $N(0,1)$, $N(0,1.1)$, $N(0,1.15)$ and $L(0, \frac{1}{\sqrt{2}})$

(n, k_n)	α	$N(0,1)$	$N(0,1.1)$	$N(0,1.15)$	$L(0, \frac{1}{\sqrt{2}})$
(2400, 30)	0.05	0.058	0.370	0.890	1.000
(2400, 50)	0.05	0.060	0.360	0.726	1.000
(2400, 80)	0.05	0.056	0.162	0.620	1.000
(2400, 100)	0.05	0.052	0.112	0.446	1.000
(4800, 30)	0.05	0.068	0.850	1.000	1.000
(4800, 50)	0.05	0.058	0.710	1.000	1.000
(4800, 80)	0.05	0.054	0.544	0.982	1.000
(4800, 100)	0.05	0.052	0.450	0.974	1.000

Table 4.1 summarizes the simulated size and power for various n and k_n. For fixed (n, k_n), the power increases and can approach 1 as σ increases. When n is fixed, the power get larger when m_n increases. For instance, when $n = 2400$, the power of 2nd row has larger power than the 5th row. Moreso, for fixed k_n, the powers of $n = 4800$ are larger than the powers of $n = 2400$ which

implies that large n has larger power. When $p_1(x)$ is the Laplace distribution $Laplace(0, \frac{1}{\sqrt{2}})$, all the powers are 1. This indicates that when σ is far lower than 1, the power approaches 1. Likewise, when the σ is far larger than 1, the power approaches to 1.

Table 4.2. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$ and $k_n \in \{30, 50, 80, 100\}$, and $N(0,1)$, $N(0, 0.95)$, $N(0, 0.90)$ and $N(0, 0.85)$

(n, k_n)	α	$N(0,1)$	$N(0,0.95)$	$N(0,0.90)$	$N(0,0.85)$
(2400,30)	0.05	0.045	0.084	0.57	0.990
(2400,50)	0.05	0.054	0.070	0.342	0.978
(2400,80)	0.05	0.046	0.064	0.266	0.912
(2400,100)	0.05	0.050	0.078	0.242	0.848
(4800,30)	0.05	0.044	0.138	0.964	1.000
(4800,50)	0.05	0.050	0.110	0.912	1.000
(4800,80)	0.05	0.052	0.088	0.758	1.000
(4800,100)	0.05	0.050	0.072	0.676	1.000

In reference to Table 4.2, it summarizes the simulated size and power for various n and k_n where the $\sigma = 0.95$, 0.90 and 0.85. The size of the distribution is close to nominal level $\alpha = 0.05$ justifying that the limiting distribution is valid. For fixed (n, k_n), the power increases as σ increases. When $n = 4800$ and $\sigma = 0.85$, the power approaches 1. This implies that for any number $0 < \sigma \leq 0.85$, the power will be 1 for the same n and k_n. For fixed k_n and σ, the power gets large when n is large. For example, the powers of $n = 4800$ are larger than that of $n = 2400$.

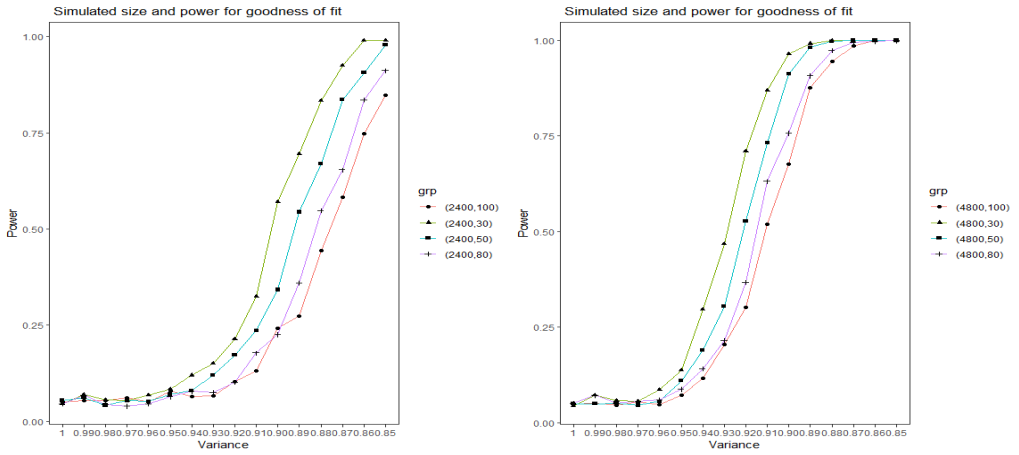

Figure 4.1. Plot of the simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, $\mu = 0$, and $\sigma \in \{1, 0.99, 0.98, \ldots, 0.87, 0.86, 0.85\}$

Figure 4.1 represents the simulated size and power of goodness of fit test of sample size n of 2400 and 4800. The null and alternative hypothesis are $N(0,1)$, and $N(0,\sigma)$, respectively where σ varies. The graph on the left and right hand sides are $n = 2400$ and $n = 4800$ respectively. Both plots show upward sloping with starting point almost 0.05, which therefore implies that the σ decreases but positive as the power increases. The curves of $n = 4800$ is steeper than curves of $n = 2400$ which implies $n = 4800$ has higher powers than that of $n = 2400$. The graphs also show that the alternative hypothesis has high powers. The plot shift outward as k_n increase, indicating that the power of the test decrease as k_n increases.

Table 4.3. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$ and $k_n \in \{30, 50, 80, 100\}$ and $N(0,1)$, $N(0.05,1)$, $N(0.10,1)$ and $N(0.15,1)$

(n, k_n)	α	$N(0,1)$	$N(0.05,1)$	$N(0.10,1)$	$N(0.15,1)$
(2400,30)	0.05	0.045	0.062	0.230	0.792
(2400,50)	0.05	0.054	0.070	0.170	0.608
(2400,80)	0.05	0.046	0.062	0.120	0.506
(2400,100)	0.05	0.050	0.068	0.136	0.394
(4800,30)	0.05	0.044	0.120	0.692	1.000
(4800,50)	0.05	0.050	0.096	0.530	0.996
(4800,80)	0.05	0.052	0.092	0.418	0.936
(4800,100)	0.05	0.050	0.056	0.324	0.922

Table 4.3 represents the simulated size and power of sample of $n = 2400$ and $n = 4800$ and the blocks of $k_n = 30$, 50, and 100 where the $\sigma = 0.95$, 0.90 and 0.85. The size of the distribution is close to $\alpha = 0.05$ justifying that the limiting distribution is valid. The power increases as σ increases for fixed n and k_n which implies the power depends on the σ. For fixed k_n and σ, the power is large when n is large. For example, the powers of $n = 4800$ are larger than that of $n = 2400$.

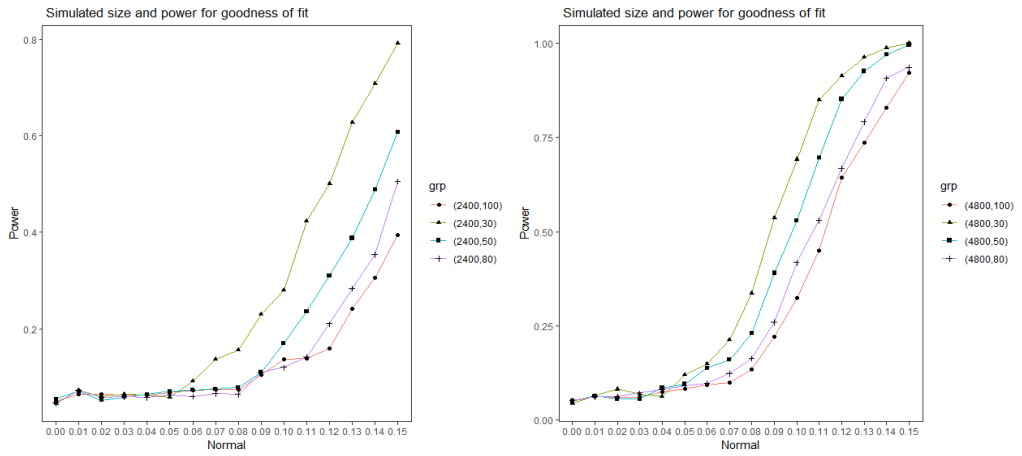

Figure 4.2. Plot of the simulated size and power for goodness of fit with $n \in (2400, 4800)$ and $k_n \in (30, 50, 80, 100)$, $\sigma = 0.05$ and $\mu \in \{0.00, 0.01, 0.02, \ldots, 0.14, 0.15\}$

Figure 4.2 represents the plot of simulated size and power of goodness of fit test with $n =$ 2400 and 4800. The curves of $n = 4800$ is steeper than curves of $n = 2400$ which indicating higher powers of $n = 4800$. The graph confirms that the power increases as the μ gets larger.

Figure 4.3 represents the plot of the simulated size and power of goodness of fit with the μ from -0.05 to -0.15 with an increment of 0.01. Figure 4.3 has similar characteristics, and same pattern to Figure 4.2 since with all other variables held constant under standard normal distribution, the negative of the μ gives the equal values of that positive of the μ.

Table 4.4. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$ and $k_n \in \{30, 50, 80, 100\}$ and $N(0, 1)$, $N(-0.05, 1)$, $N(-0.10, 1)$ and $N(-0.15, 1)$

(n, k_n)	α	$N(0,1)$	$N(-0.05,1)$	$N(-0.10,1)$	$N(-0.15,1)$
(2400,30)	0.05	0.045	0.074	0.254	0.780
(2400,50)	0.05	0.054	0.068	0.190	0.638
(2400,80)	0.05	0.046	0.084	0.130	0.510
(2400,100)	0.05	0.050	0.062	0.103	0.420
(4800,30)	0.05	0.044	0.088	0.696	1.000
(4800,50)	0.05	0.050	0.084	0.546	0.982
(4800,80)	0.05	0.052	0.078	0.364	0.946
(4800,100)	0.05	0.050	0.082	0.342	0.920

In Table 4.4, the mean of the distribution varies, and a fixed variance. In Table 4.5, Table 4.6, Table 4.7 and Table 4.8 have fixed mean and change of variance. They summarize the simulated sizes and powers for various n and k_n. As explained in previous tables, Table 4.5, Table 4.6, Table 4.7 and Table 4.8 have similar characteristics. From our findings, for fixed (n, k_n), the power increases as σ increases. Similarly, the power increases as μ increases in Table 4.4. For fixed n, when m_n becomes larger, the power gets larger. For instance, the second row has a larger power than the fifth row. For fixed k_n, large n has larger power. For example, the powers of the $n = 4800$ are larger than $n = 2400$ in the Tables we have discussed. It confirms that the proposed method gives high power.

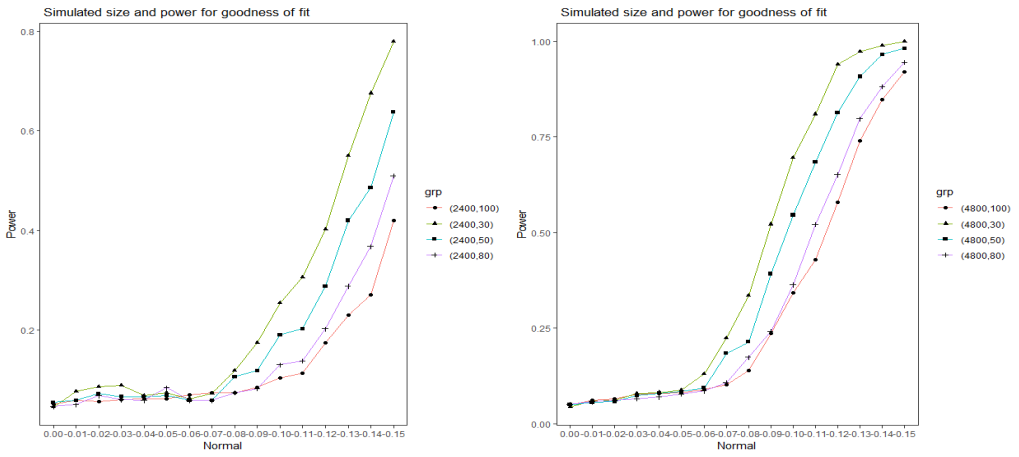

Figure 4.3. Plot of the simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, $\sigma = 0.05$ and $\mu \in \{0.00, -0.01, -0.02, \ldots, -0.13, -0.14, -0.15\}$

Table 4.5. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, and $N(0, 1)$, $N(0.05, 1.05)$, $N(0.05, 1.1)$ and $N(0.05, 1.15)$

(n, k_n)	α	$N(0, 1)$	$N(0.05, 1.05)$	$N(0.05, 1.1)$	$N(0.05, 1.15)$
(2400,30)	0.05	0.045	0.086	0.456	0.908
(2400,50)	0.05	0.054	0.072	0.300	0.752
(2400,80)	0.05	0.046	0.076	0.206	0.624
(2400,100)	0.05	0.050	0.070	0.182	0.568
(4800,30)	0.05	0.044	0.270	0.926	1.000
(4800,50)	0.05	0.050	0.212	0.820	0.998
(4800,80)	0.05	0.052	0.150	0.676	0.994
(4800,100)	0.05	0.050	0.126	0.616	0.982

Table 4.6. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, and $N(0,1)$, $N(0.05, 0.95)$, $N(0.05, 0.90)$ and $N(0.05, 0.85)$

(n, k_n)	α	$N(0,1)$	$N(0.05, 0.95)$	$N(0.05, 0.90)$	$N(0.05, 0.85)$
(2400,30)	0.05	0.045	0.120	0.696	0.998
(2400,50)	0.05	0.054	0.082	0.518	0.986
(2400,80)	0.05	0.046	0.084	0.356	0.918
(2400,100)	0.05	0.050	0.084	0.308	0.862
(4800,30)	0.05	0.044	0.348	0.994	1.000
(4800,50)	0.05	0.050	0.274	0.958	1.000
(4800,80)	0.05	0.052	0.172	0.866	1.000
(4800,100)	0.05	0.050	0.150	0.844	1.000

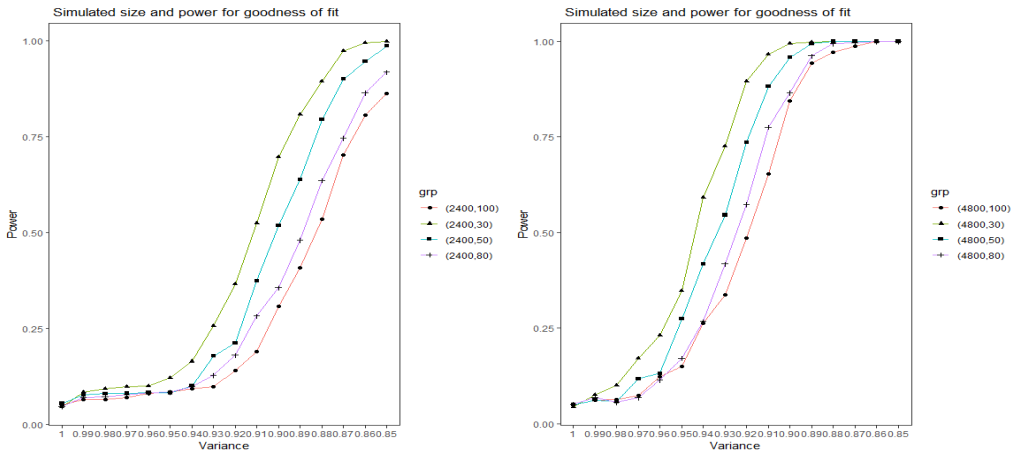

Figure 4.4. Plot of the simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, $\mu = 0.05$, and $\sigma \in \{1, 0.99, 0.98, \ldots, 0.87, 0.86, 0.85\}$

Table 4.7. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, and $N(0,1)$, $N(0.1, 1.05)$, $N(0.1, 1.1)$ and $N(0.1, 1.15)$

(n, k_n)	α	$N(0,1)$	$N(0.1, 1.05)$	$N(0.1, 1.1)$	$N(0.1, 1.15)$
(2400,30)	0.05	0.045	0.326	0.716	0.982
(2400,50)	0.05	0.054	0.230	0.552	0.930
(2400,80)	0.05	0.046	0.166	0.424	0.804
(2400,100)	0.05	0.050	0.150	0.324	0.692
(4800,30)	0.05	0.044	0.852	0.996	1.000
(4800,50)	0.05	0.050	0.722	0.966	1.000
(4800,80)	0.05	0.052	0.524	0.934	1.000
(4800,100)	0.05	0.050	0.498	0.868	0.998

Table 4.8. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, and $N(0,1)$, $N(0.1, 0.95)$, $N(0.1, 0.90)$ and $N(0.1, 0.85)$

(n, k_n)	α	$N(0,1)$	$N(0.1, 0.95)$	$N(0.1, 0.90)$	$N(0.1, 0.85)$
(2400,30)	0.05	0.045	0.402	0.920	1.000
(2400,50)	0.05	0.054	0.294	0.802	0.996
(2400,80)	0.05	0.046	0.208	0.614	0.982
(2400,100)	0.05	0.050	0.184	0.564	0.966
(4800,30)	0.05	0.044	0.922	1.000	1.000
(4800,50)	0.05	0.050	0.812	0.998	1.000
(4800,80)	0.05	0.052	0.690	0.988	1.000
(4800,100)	0.05	0.050	0.610	0.984	1.000

Table 4.9. Simulated size and power for goodness of fit of t-distribution with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, and $N(0,1)$, $v = 1$, $v = 2$ and $v = 3$

(n, k_n)	α	N(0,1)	$v = 1$	$v = 2$	$v = 3$
(2400,30)	0.05	0.045	0.996	1.000	1.000
(2400,50)	0.05	0.054	0.994	1.000	1.000
(2400,80)	0.05	0.046	0.988	0.998	0.990
(2400,100)	0.05	0.050	0.988	0.970	0.952
(4800,30)	0.05	0.044	0.996	1.000	1.000
(4800,50)	0.05	0.050	0.996	1.000	1.000
(4800,80)	0.05	0.052	0.994	1.000	1.000
(4800,100)	0.05	0.050	0.992	1.000	1.000

Table 4.9 provides the simulated size and power of various n and k_n with the t-distribution. Our size of the model is close to the nominal $\alpha = 0.05$, which indicates that the limiting distribution is valid. The power of both samples for fixed n and k_n are large, with a small degree of freedom and

approach to 1. The $n = 4800$ is performed better than the $n = 2400$ which concluded that larger n has higher power.

Similarly, from Table 4.10, with the degree of freedom $v = 5$, 10 and 20 and fixed n and k_n, the power decreases. This shows that power reduces as the degree of freedom increases. For instance, the power of column $v = 5$ is larger than $v = 10$ and $v = 20$. The power of $n = 4800$ still performs better than $n = 2400$ with fixed k_n. The results in both Table 4.9 and Table 4.10 is consistent with the fact that the t-distribution approaches a standard normal distribution as the degree of freedom increases. In case of column $v = 20$, powers are closer to size $\alpha = 0.05$.

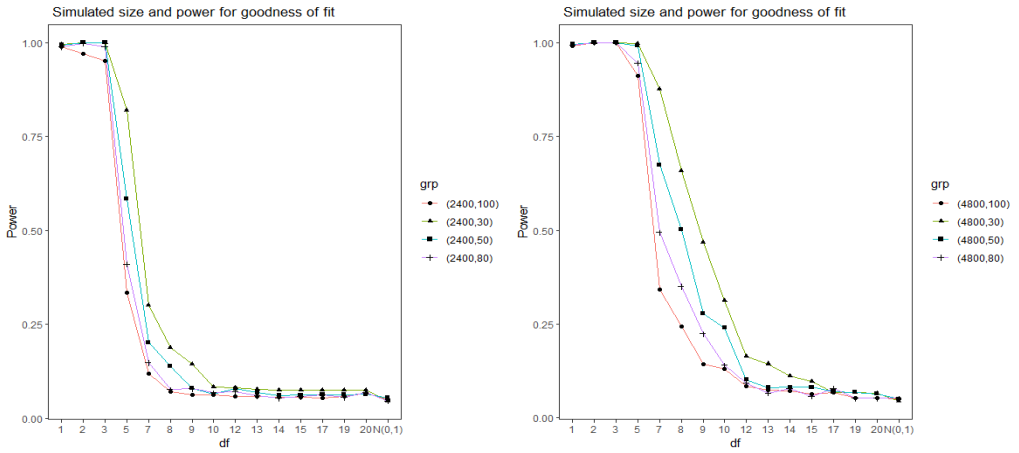

Figure 4.5. Plot of the simulated size and power for goodness of fit of t-distribution with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, and $N(0, 1)$, $v \in \{1, 2, 3, \ldots, 17, 19, 20)\}$

Figure 4.5 represents the simulated size and power of t-distribution with the degree of freedom, v from 1 to 20. The left and right of the plot are a sample of $n = 2400$ and $n = 4800$, respectively. Both plots show sharp downward sloping until they get to $v = 8$ on sample size n =2400 and $v = 13$ on sample size $n = 4200$ and afterward they become constant. The curves of $n = 4800$ have the same pattern showing that the powers approach 1 with a small degree of freedom and a larger n and decreases for a high degree of freedom. At point $v = 20$, the curves are equal as the size of the test confirming that t-distribution gets to a standard normal distribution with a larger degree of freedom.

Table 4.10. Simulated size and power for goodness of fit of t-distribution with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, and $N(0,1)$, $v = 5$, $v = 10$ and $v = 20$

(n, k_n)	α	N(0,1)	$v = 5$	$v = 10$	$v = 20$
(2400,30)	0.05	0.045	0.820	0.084	0.074
(2400,50)	0.05	0.054	0.584	0.064	0.064
(2400,80)	0.05	0.046	0.410	0.068	0.068
(2400,100)	0.05	0.050	0.334	0.063	0.066
(4800,30)	0.05	0.044	0.998	0.312	0.066
(4800,50)	0.05	0.050	0.992	0.240	0.064
(4800,80)	0.05	0.052	0.946	0.140	0.052
(4800,100)	0.05	0.050	0.912	0.130	0.052

4.1.2. Degenerate U-statistics of d-dimension with one sample case

In this section, we consider multivariate distributions of null hypothesis $p(x) = N(0, I_d)$ and alternative hypothesis $q(x) = N(0, \sigma I_d)$,

$\prod_{i=1}^{d} L(x_i, 0, \frac{1}{\sqrt{2}})$ where σ varies from each simulation. The kernel in this case is $k(x,y) = e^{-\frac{||x-y||^2}{2}}$. Let $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$. Direct computation yields the following:

$$\begin{aligned}
\frac{\partial \log p(x)}{\partial x_i} &= -x_i, \\
\frac{\partial k(x,y)}{\partial x_i} &= -e^{-\frac{||x-y||^2}{2}}(x_i - y_i), \\
\frac{\partial k(x,y)}{\partial y_i} &= e^{-\frac{||x-y||^2}{2}}(x_i - y_i),
\end{aligned}$$

$$\frac{\partial^2 k(x,y)}{\partial x_i \partial y_i} = e^{-\frac{||x-y||^2}{2}}\Big(1 - (x_i - y_i)^2\Big),$$

$$u_p(x,y) = e^{-\frac{||x-y||^2}{2}}\Big(\sum_{i=1}^{d}(1 + x_i y_i) - 2\sum_{i=1}^{d}(x_i - y_i)^2\Big).$$

Clearly, the $u_p(x,y)$ is also bounded and symmetric.

We evaluate the performance of our method with $n = 2400$ and $n = 4800$ and the trend of the powers. The two samples are randomly divided into $k = 30, 50, 80$ groups, respectively, and calculate the size and power using the divide-and-conquer test statistic T_k. The size and power

of the test detect the location difference from various dimensions. The $\mu = 0$ and σ varies from locations and different dimensions of d From Table 4.11 to Table 4.17, the limiting distribution is valid indicating that the sizes are close to nominal value $\alpha = 0.05$.

The power decreases as the dimension,d, increases for fixed n, k_n and σ, which is the variation of crime locations. Moreover, for fixed n, d and σ, the power of the simulation declines as the k_n increases from 30 to 80. It implies the trade-off between computational cost and power of the test statistic. Increasing k_n reduces the running time, but in effects, it reduces the power; therefore, k_n should be carefully chosen. An Increase of σ far from 1, resulting in bigger power, for instance, in Table 4.17. Similarly, when the value of σ is close to 1, then the power also gets closer to the size. For all other variables hold constant, the power of $n = 4800$ is larger than $n = 2400$ as k_n increases. Therefore, the sample n needs to be increased in order to raise power.

Table 4.11. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0, 1.3I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400, 30)	0.05	(0.056) 0.906	(0.048) 0.810	(0.044) 0.340
(2400, 50)	0.05	(0.046) 0.782	(0.050) 0.628	(0.042) 0.120
(2400, 80)	0.05	(0.058) 0.570	(0.030) 0.430	(0.052) 0.106
(4800, 30)	0.05	(0.054) 1.000	(0.060) 1.000	(0.050) 0.800
(4800, 50)	0.05	(0.056) 1.000	(0.060) 0.992	(0.054) 0.590
(4800, 80)	0.05	(0.052) 0.986	(0.040) 0.952	(0.040) 0.396

Table 4.12. Simulated size and power for goodness of fit with, $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$,$q(x) = \prod_{i=1}^{d} L(x_i, 0, \frac{1}{\sqrt{2}})$, and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400, 30)	0.05	(0.056) 1.000	(0.048) 1.000	(0.044) 1.000
(2400, 50)	0.05	(0.046) 1.000	(0.050) 1.000	(0.042) 1.000
(2400, 80)	0.05	(0.058) 1.000	(0.030) 1.000	(0.052) 1.000

Table 4.12 reports the power of the Multivariate Laplace distribution as an alternative hypothesis. The powers of the test are all 1. The high values of the power irrespective of the dimension show the optimism of the power. This behavior explains the characteristics of the Multivariate Laplace distribution. It has heavy tails compare to normal distribution since Laplace distribution expressed in terms of absolute difference from the mean.

Table 4.13. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0, 1.2I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)0.404	(0.040)0.352	(0.050)0.366
(2400,50)	0.05	(0.044)0.256	(0.054)0.238	(0.048)0.242
(2400,80)	0.05	(0.042)0.174	(0.048)0.164	(0.052)0.156
(4800,30)	0.05	(0.058)0.916	(0.050)0.818	(0.042)0.342
(4800,50)	0.05	(0.058)0.786	(0.054)0.656	(0.048)0.226
(4800,80)	0.05	(0.056)0.618	(0.054)0.486	(0.044)0.158

Table 4.14. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0, 1.4I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)1.000	(0.040)0.992	(0.050)0.520
(2400,50)	0.05	(0.044)0.998	(0.054)0.926	(0.048)0.316
(2400,80)	0.05	(0.042)0.930	(0.048)0.763	(0.052)0.202
(4800,30)	0.05	(0.058)1.000	(0.050)1.000	(0.042)0.978
(4800,50)	0.05	(0.058)1.000	(0.054)1.000	(0.048)0.902
(4800,80)	0.05	(0.056)1.000	(0.054)1.000	(0.044)0.664

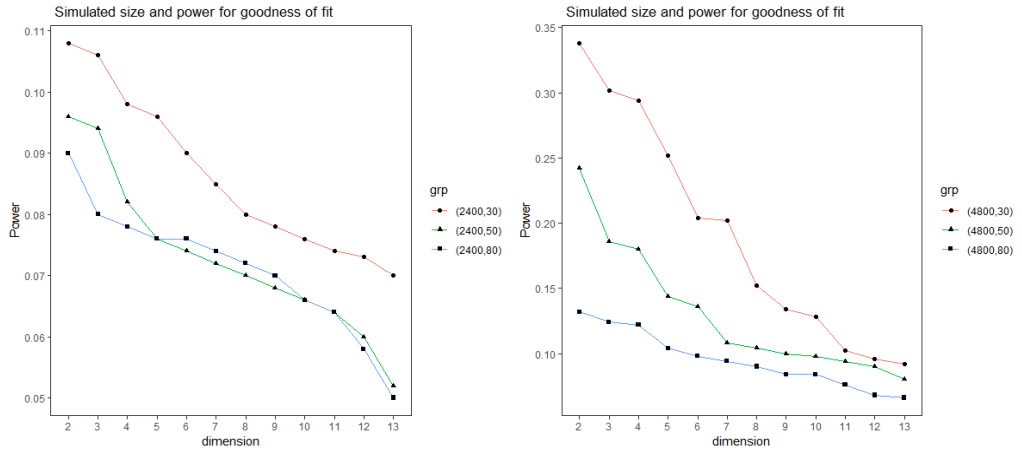

Figure 4.6. Plot of the simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, $\sigma = 1.4$, and $d \in \{2, 3, 4, \ldots, 12, 13\}$

The Figure 4.6 represents the plot of sample size of $n = 2400$ and 4800, $sigma = 1.4$ and dimension,d, from 2 to 13. The curves are downward sloping, which implies the power of the simulation decreases as the d increases. The curves of $n = 4800$ is steeper than curves of $n = 2400$ which indicates that higher powers when $n = 4800$. The graph confirms that in order to increase power, the sample size should increase.

Table 4.15. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0, 0.9I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)0.106	(0.040)0.108	(0.050)0.070
(2400,50)	0.05	(0.044)0.072	(0.054)0.076	(0.048)0.064
(2400,80)	0.05	(0.042)0.068	(0.048)0.074	(0.052)0.064
(4800,30)	0.05	(0.058)0.238	(0.050)0.252	(0.042)0.138
(4800,50)	0.05	(0.058)0.142	(0.054)0.144	(0.048)0.108
(4800,80)	0.05	(0.056)0.132	(0.054)0.104	(0.044)0.084

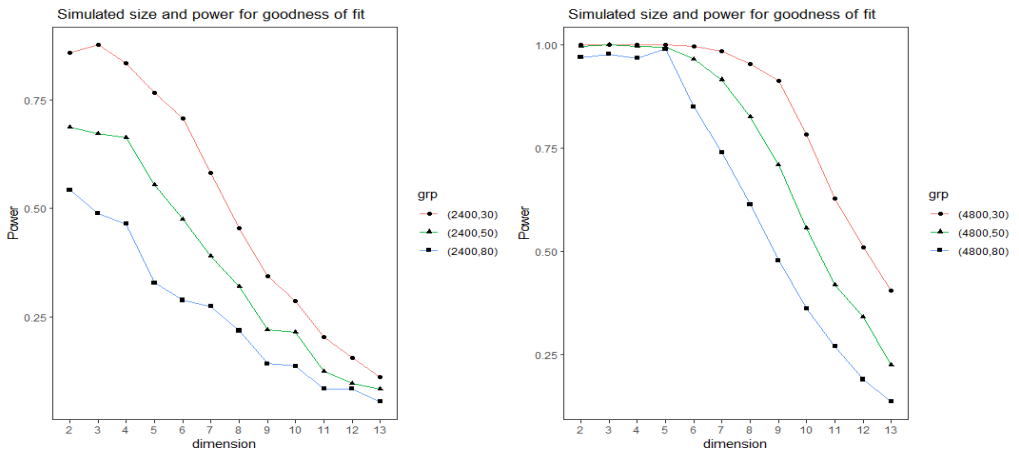

Figure 4.7. The plot of simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, $\sigma = 0.9$) and $d \in \{2, 3, 4, \ldots, 12, 13\}$

Figure 4.7 shows the plot of simulated size and power of goodness of fit test with $\sigma = 0.9$. The curves show negative sloping, which implies the power of the simulation decreases as the dimension increases. Initially, the curves of $n = 4800$ are linear, and after $d = 4$, they start declining. The plot of the (2400, 30) and (4800, 30) are higher than that of (2400, 80) and (4800, 80) because the power decreases as the dimension increases.

Table 4.16. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0, 0.85I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)0.310	(0.040)0.290	(0.050)0.184
(2400,50)	0.05	(0.044)0.228	(0.054)0.206	(0.048)0.130
(2400,80)	0.05	(0.042)0.156	(0.048)0.160	(0.052)0.094
(4800,30)	0.05	(0.058)0.840	(0.050)0.830	(0.042)0.546
(4800,50)	0.05	(0.058)0.696	(0.054)0.648	(0.048)0.332
(4800,80)	0.05	(0.056)0.496	(0.054)0.506	(0.044)0.224

Table 4.17. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0, 0.70I_d)$, $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)1.000	(0.040)1.000	(0.050)1.000
(2400,50)	0.05	(0.044)1.000	(0.054)1.000	(0.048)0.998
(2400,80)	0.05	(0.042)0.998	(0.048)0.998	(0.052)0.972
(4800,30)	0.05	(0.058)1.000	(0.050)1.000	(0.042)1.000
(4800,50)	0.05	(0.058)1.000	(0.054)1.000	(0.048)1.000
(4800,80)	0.05	(0.056)1.000	(0.054)1.000	(0.044)1.000

From Table 4.18 to Table 4.28, the validity of limiting distribution confirms since sizes are close nominal values, $\alpha = 0.05$. The power of the test decreases as the dimension,d, increases for fixed n, k_n, μ and σ. Furthermore, for fixed n, d, μ and σ, the power declines as the k_n increases from 30 to 80. It implies the trade-off between the power of the test statistic and computational cost. Though increasing k_n reduces the running time but in effects, it reduces the power. The powers of $n = 4800$ are larger than $n = 2400$ as k_n increases. Therefore, the sample size n needs to be increased in order to raise power.

Table 4.18. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0.1, 1.2I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)0.858	(0.040)0.766	(0.050)0.286
(2400,50)	0.05	(0.044)0.686	(0.054)0.554	(0.048)0.214
(2400,80)	0.05	(0.042)0.542	(0.048)0.328	(0.052)0.136
(4800,30)	0.05	(0.058)1.000	(0.050)1.000	(0.042)0.784
(4800,50)	0.05	(0.058)0.996	(0.054)0.994	(0.048)0.556
(4800,80)	0.05	(0.056)0.970	(0.054)0.900	(0.044)0.362

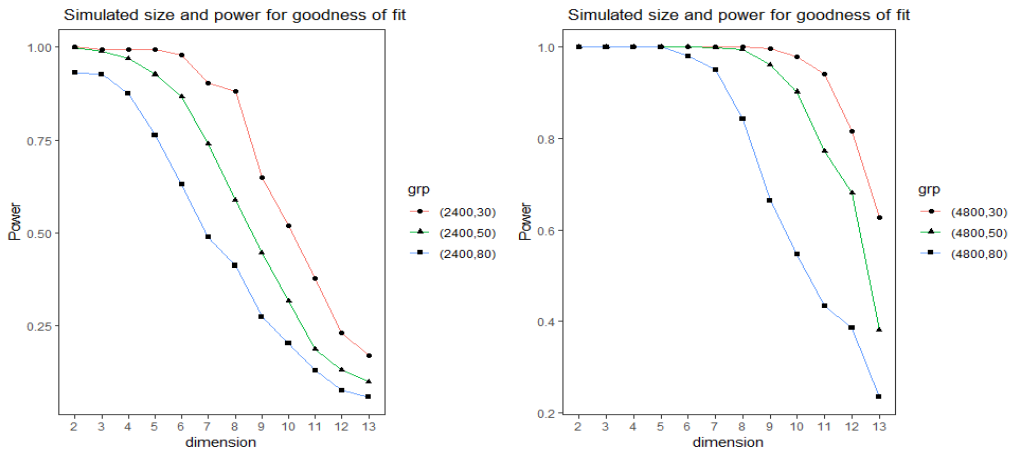

Figure 4.8. Plot of simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, $\mu = 0.1$, $\sigma = 1.2$ and $d \in \{2, 3, 4, \ldots, 12, 13\}$

The Figure 4.8 shows the plot of simulated size and power of goodness of fit test with n = (2400, 4800), $\mu = 0.1$ and $\sigma = 1.2$. The plot of $n = 2400$ declines faster than $n = 4800$ with different dimension, d and it is as a result of smaller sample size. The curves are downward sloping since power decreases as the dimension increases. The curves of $n = 4800$ are steeper than curves of $n = 2400$, which indicates that a large sample size increases power. The plot of the (2400, 30) and (4800, 30) are higher than that of (2400, 80) and (4800, 80) because the power decreases with increasing dimensions.

Table 4.19. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0.15, 1.2I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)0.998	(0.040)0.984	(0.050)0.608
(2400,50)	0.05	(0.044)0.972	(0.054)0.906	(0.048)0.380
(2400,80)	0.05	(0.042)0.870	(0.048)0.740	(0.052)0.250
(4800,30)	0.05	(0.058)1.000	(0.050)1.000	(0.042)0.988
(4800,50)	0.05	(0.058)1.000	(0.054)1.000	(0.048)0.940
(4800,80)	0.05	(0.056)1.000	(0.054)1.000	(0.044)0.756

Table 4.20. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, $p(x) = N(0, I_d)$, $q(x) = N(0.1, 1.3I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)0.998	(0.040)0.968	(0.050)0.500
(2400,50)	0.05	(0.044)0.930	(0.054)0.860	(0.048)0.278
(2400,80)	0.05	(0.042)0.874	(0.048)0.710	(0.052)0.170
(4800,30)	0.05	(0.058)1.000	(0.050)1.000	(0.042)0.966
(4800,50)	0.05	(0.058)1.000	(0.054)1.000	(0.048)0.856
(4800,80)	0.05	(0.056)1.000	(0.054)0.996	(0.044)0.668

Table 4.21. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0.15, 1.3I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)1.000	(0.040)1.000	(0.050)0.758
(2400,50)	0.05	(0.044)1.000	(0.054)0.988	(0.048)0.544
(2400,80)	0.05	(0.042)0.976	(0.048)0.910	(0.052)0.316
(4800,30)	0.05	(0.058)1.000	(0.050)1.000	(0.042)1.000
(4800,50)	0.05	(0.058)1.000	(0.054)1.000	(0.048)0.980
(4800,80)	0.05	(0.056)1.000	(0.054)1.000	(0.044)0.902

Table 4.22. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) \in (N(0, I_d)$, $q(x) = N(0.1, 1.4I_d))$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)1.000	(0.040)1.000	(0.050)0.660
(2400,50)	0.05	(0.044)0.998	(0.054)0.984	(0.048)0.468
(2400,80)	0.05	(0.042)0.990	(0.048)0.904	(0.052)0.296
(4800,30)	0.05	(0.058)1.000	(0.050)1.000	(0.042)1.000
(4800,50)	0.05	(0.058)1.000	(0.054)1.000	(0.048)0.970
(4800,80)	0.05	(0.056)1.000	(0.054)1.000	(0.044)0.874

Table 4.23. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0.1, 0.9I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)0.616	(0.040)0.590	(0.050)0.286
(2400,50)	0.05	(0.044)0.450	(0.054)0.410	(0.048)0.184
(2400,80)	0.05	(0.042)0.342	(0.048)0.332	(0.052)0.126
(4800,30)	0.05	(0.058)0.996	(0.050)0.990	(0.042)0.800
(4800,50)	0.05	(0.058)0.960	(0.054)0.944	(0.048)0.618
(4800,80)	0.05	(0.056)0.884	(0.054)0.834	(0.044)0.426

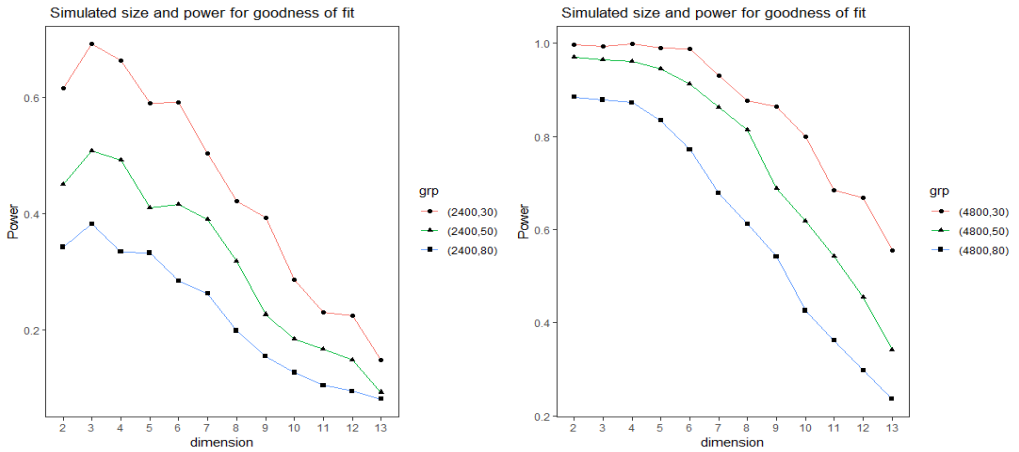

Figure 4.9. Plot of simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, $\mu = 0.1$, $\sigma = 0.9$ and $d \in \{2, 3, 4, \ldots, 12, 13\}$

The Figure 4.6 represents the plot of simulated size and power of goodness of fit test with $n = 2400$ and 4800, the dimension, d and the parameters of $\mu = 0.1$ and $\sigma = 0.9$. Similar to Figure 4.8, the plot of $n = 2400$ declines faster than 4800, and this is a result of the smaller sample with lower power. The curves are downward sloping indicates the power decreases as the dimension increases. The plot of the (2400, 30) and (4800, 30) are higher than that of (2400, 80) and (4800, 80) because the power decreases with increasing dimensions.

Table 4.24. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0.15, 0.9I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)0.982	(0.040)0.970	(0.050)0.744
(2400,50)	0.05	(0.044)0.934	(0.054)0.926	(0.048)0.0.544
(2400,80)	0.05	(0.042)0.818	(0.048)0.790	(0.052)0.398
(4800,30)	0.05	(0.058)1.000	(0.050)1.000	(0.042)0.998
(4800,50)	0.05	(0.058)1.000	(0.054)1.000	(0.048)0.990
(4800,80)	0.05	(0.056)1.000	(0.054)0.998	(0.044)0.932

Table 4.25. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0.1, 0.85I_d)$, and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)0.884	(0.040)0.880	(0.050)0.574
(2400,50)	0.05	(0.044)766	(0.054)0.726	(0.048)0.382
(2400,80)	0.05	(0.042)0.576	(0.048)0.570	(0.052)0.200
(4800,30)	0.05	(0.058)1.000	(0.050)1.000	(0.042)0.994
(4800,50)	0.05	(0.058)0.1.000	(0.054)0.996	(0.048)0.916
(4800,80)	0.05	(0.056)0.982	(0.054)0.980	(0.044)0.774

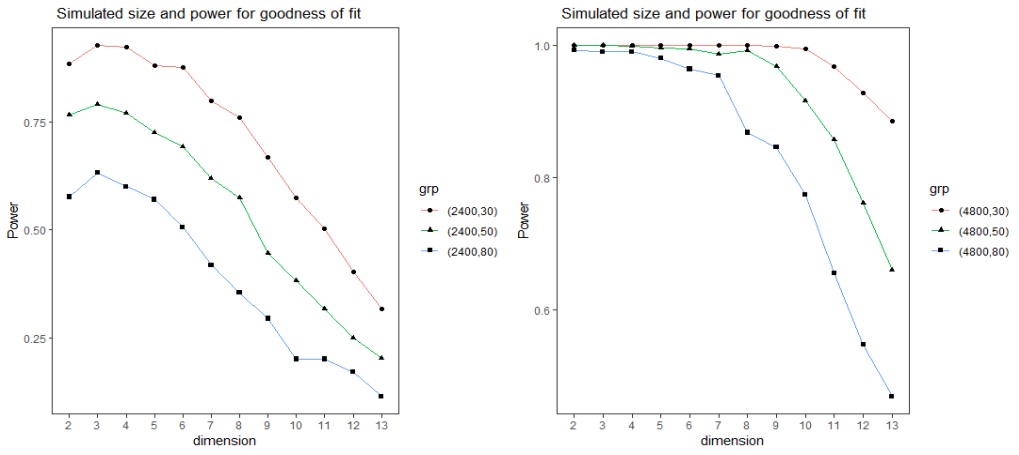

Figure 4.10. Plot of simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, $\mu = 0.1$, $\sigma = 0.85$ and $d \in \{2, 3, 4, \ldots, 12, 13\}$

The Figure 4.10 is similar to Figure 4.8 which represents the plot of simulated size and power of goodness of fit test with $n = 2400$ and 4800, dimension, d, $\mu = 0.1$ and $\sigma = 0.85$. The curves are downward sloping, which implies the power decreases as the dimension increases. The curves of $n = 4800$ is steeper than curves of $n = 2400$ which indicates higher powers of $n = 4800$. The plot of the (2400, 30) and (4800, 30) are higher than that of (2400, 80) and (4800, 80) because the power decreases with increasing dimensions.

Table 4.26. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0.15, 0.85I_d)$, and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)1.000	(0.040)1.000	(0.050)0.946
(2400,50)	0.05	(0.044)1.000	(0.054)0.994	(0.048)0.768
(2400,80)	0.05	(0.042)0.926	(0.048)0.924	(0.052)0.590
(4800,30)	0.05	(0.058)1.000	(0.050)1.000	(0.042)1.000
(4800,50)	0.05	(0.058)1.000	(0.054)1.000	(0.048)1.000
(4800,80)	0.05	(0.056)1.000	(0.054)1.000	(0.044)0.996

Table 4.27. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = N(0.1, 0.70I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)1.000	(0.040)1.000	(0.050)1.000
(2400,50)	0.05	(0.044)1.000	(0.054)1.000	(0.048)0.998
(2400,80)	0.05	(0.042)0.998	(0.048)0.998	(0.052)0.998
(4800,30)	0.05	(0.058)1.000	(0.050)1.000	(0.042)1.000
(4800,50)	0.05	(0.058)1.000	(0.054)1.000	(0.048)1.000
(4800,80)	0.05	(0.056)1.000	(0.054)1.000	(0.044)1.000

Table 4.28. Simulated size and power for goodness of fit with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = (N(0, I_d)$, $q(x) = N(0.15, 0.70I_d)$ and $d \in \{2, 5, 10\}$

(n, k_n)	α	$d = 2$	$d = 5$	$d = 10$
(2400,30)	0.05	(0.054)1.000	(0.040)1.000	(0.050)1.000
(2400,50)	0.05	(0.044)1.000	(0.054)1.000	(0.048)1.000
(2400,80)	0.05	(0.042)1.000	(0.048)1.000	(0.052)1.000
(4800,30)	0.05	(0.058)1.000	(0.050)1.000	(0.042)1.000
(4800,50)	0.05	(0.058)1.000	(0.054)1.000	(0.048)1.000
(4800,80)	0.05	(0.056)1.000	(0.054)1.000	(0.044)1.000

This section analyses the performance of the power of our method from t-distribution with sample sizes, $n = 2400$ and $n = 4800$, dimensions, $d = (2, 5, 10)$ and degree of freedom, $v =$. From Table 4.29 to Table 4.31, the simulated sizes are close to nominal value $\alpha = 0.05$. If all the variables are held constant, the power of $d = 2$ is higher than $d = 5$. It concludes that power declines as the d increases. Comparing Table 4.29 and Table 4.30, Table 4.29 are stronger power than Table 4.30 due to change of degree of freedom. For fixed n and k_n, the power decreases as the degree of freedom increases.

Moreso, the power of the test decreases as the groups,k_n, increases from 30 to 80 for fixed n and v. In order to increase power, the sample size should be increased (see Table 4.29). It implies the trade-off between computational cost and power of the test statistic. Though increasing k reduces the running time, but in effects, it reduces power.

Table 4.29. Simulated size and power for goodness of fit of t- distribution with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, $p(x) = (N(0, I_d)$, $q(x) = \prod_{i=1}^{d} t(x_i, v)$, and $d = 2$

(n, k_n)	α	N(0,1)	$v = 1$	$v = 2$	$v = 3$
(2400,30)	0.05	0.054	1.000	1.000	0.998
(2400,50)	0.05	0.044	1.000	1.000	0.936
(2400,80)	0.05	0.042	1.000	1.000	0.792
(4800,30)	0.05	0.058	1.000	1.000	1.000
(4800,50)	0.05	0.058	1.000	1.000	1.000
(4800,80)	0.05	0.056	1.000	1.000	1.000

Table 4.30. Simulated size and power for goodness of fit of t-distribution with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = (N(0, I_d)$, $q(x) = \prod_{i=1}^{d} t(x_i, v)$ and $d = 5$

(n, k_n)	α	N(0,1)	$v = 1$	$v = 2$	$v = 3$
(2400,30)	0.05	0.040	1.000	0.996	0.898
(2400,50)	0.05	0.054	1.000	0.974	0.744
(2400,80)	0.05	0.048	1.000	0.900	0.564
(4800,30)	0.05	0.050	1.000	1.000	1.000
(4800,50)	0.05	0.054	1.000	1.000	0.998
(4800,80)	0.05	0.054	1.000	1.000	0.968

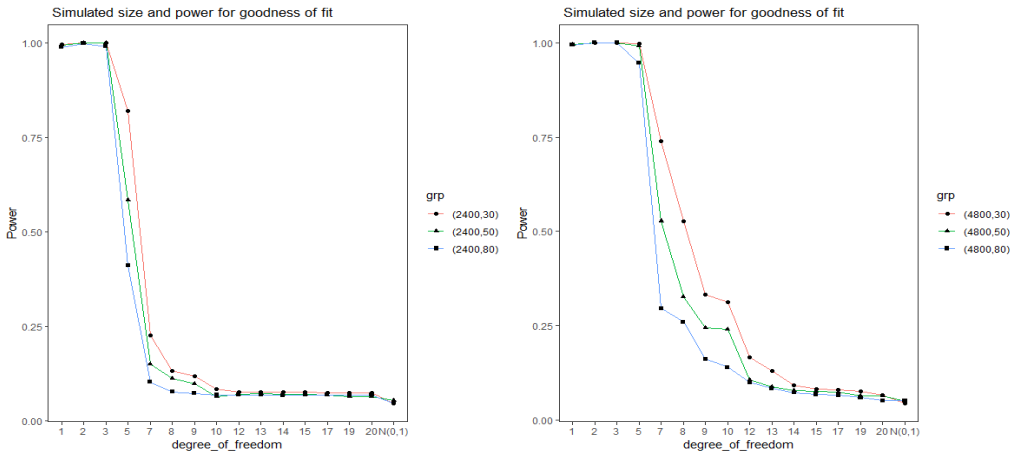

Figure 4.11. Plot of simulated size and power for goodness of fit of t-distribution with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, $q(x) = \prod_{i=1}^{d} t(x_i, v)$, $V \in \{1, 2, 3, \ldots, 17, 19, 20\}$, and $d = 5$

In Figure 4.11, the left side plot is linear until it reaches $v = 3$, and afterward, it starts to decline and approaches to standard normal. The right side plot also linear at the initial stage and start falling until it gets close to standard normal. The curves of $n = 4800$ are partly linear from $v = 1$ to $v = 5$ indicating that power is approximate to 1. There is not much difference in the powers of the degree of freedom $v = 1$, 2, 3, and 5. The plots exhibit higher power when the degree of freedom is small.

Table 4.31. Simulated size and power for goodness of fit of t-distribution with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, $p(x) = (N(0, I_d)$, $q(x) = \prod_{i=1}^{d} t(x_i, v)$, and $d = 10$

(n, k_n)	α	N(0,1)	$v = 1$	$v = 2$	$v = 3$
(2400,30)	0.05	0.050	1.000	1.000	1.000
(2400,50)	0.05	0.048	1.000	1.000	1.000
(2400,80)	0.05	0.052	1.000	1.000	0.983
(4800,30)	0.05	0.042	1.000	1.000	1.000
(4800,50)	0.05	0.048	1.000	1.000	1.000
(4800,80)	0.05	0.044	1.000	1.000	1.000

From Table 4.32 to Table 4.34, the simulated sizes are close to nominal level $\alpha = 0.05$. Comparatively, the power of $n = 2400$ is smaller than $n = 4800$. In order to increase the simulated powers, the sample size should increase. The power of $d = 2$ is higher than $d = 5$, which concludes that the power declines as the d increases. For example, comparing Table 4.32 and Table 4.33, the power of Table 4.32 are stronger than Table 4.33.

For fixed n and k_n, the power decreases as the degree of freedom and dimensions increase. Further, the power of the test decreases as the block, k_n, increases from 30 to 80 for fixed n and v. From all Tables, the powers reduce approximately to the size when the degree of freedom is $v = 20$. It indicates that the t-distribution approximates to a standard normal distribution as the degree of freedom increases.

Table 4.32. Simulated size and power for goodness of fit of t-distribution with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = \prod_{i=1}^{d} t(x_i, v)$, $d = 2$

(n, k_n)	α	N(0,1)	$v = 5$	$v = 10$	$v = 20$
(2400,30)	0.05	0.054	0.594	0.092	0.074
(2400,50)	0.05	0.044	0.384	0.076	0.067
(2400,80)	0.05	0.042	0.286	0.066	0.064
(4800,30)	0.05	0.058	0.988	0.264	0.078
(4800,50)	0.05	0.058	0.928	0.146	0.074
(4800,80)	0.05	0.056	0.770	0.114	0.064

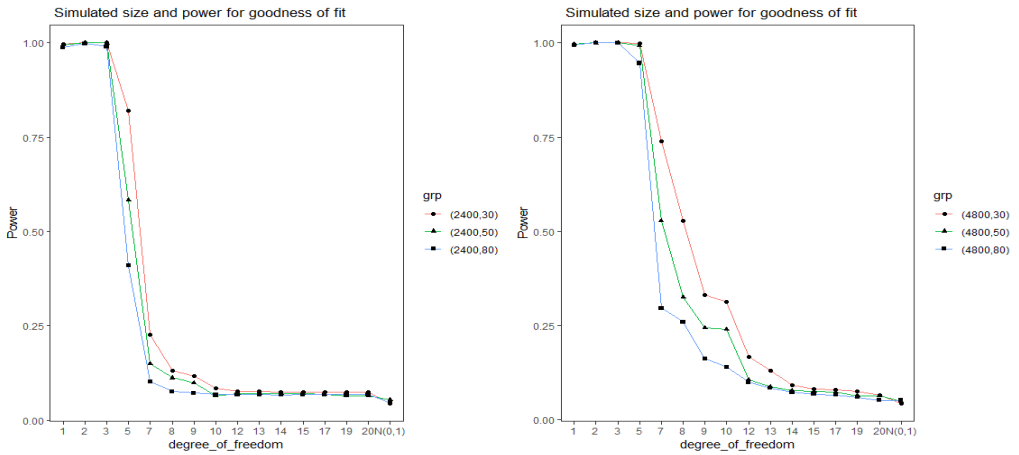

Figure 4.12. Plot of simulated size and power for goodness of fit of t-distribution with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80, 100\}$, $q(x) = \prod_{i=1}^{d} t(x_i, v)$, $V \in \{1, 2, 3, \ldots, 17, 19, 20\}$, and $d = 2$

Figure 4.12, the left and right of the plot are sample size of $n = 2400$ and $n = 4800$ respectively with degree of freedom, v, and dimension,$d = 2$. Between the degree of freedom $v = 3$ to 10 and $v = 7$ to 14, the gap between the graphs is wider for $n = 2400$ and $n = 4800$ respectively and after that, it gets tighter. It implies for a large degree of freedom and fixed sample size, t-distribution approaches a standard normal distribution. At point $v = 20$, the power is equally as the size confirming that t-distribution gets to standard normal with a large degree of freedom.

Table 4.33. Simulated size and power for goodness of fit of t-distribution with $n \in \{2400, 4800\}$, $k_n \in \{30, 50, 80\}$, and $p(x) = N(0, I_d)$, $q(x) = \prod_{i=1}^{d} t(x_i, v)$, $d = 5$

(n, k_n)	α	N(0,1)	$v = 5$	$v = 10$	$v = 20$
(2400,30)	0.05	0.040	0.404	0.102	0.064
(2400,50)	0.05	0.054	0.266	0.066	0.054
(2400,80)	0.05	0.048	0.192	0.082	0.064
(4800,30)	0.05	0.050	0.912	0.210	0.064
(4800,50)	0.05	0.054	0.762	0.134	0.068
(4800,80)	0.05	0.054	0.608	0.120	0.066

Table 4.34. Simulated size and power for goodness of fit of t-distribution with $n \in \{2400, 4800\}$, $k_n \in \{30, 50 and 80\}$, and $p(x) = N(0, I_d)$, $q(x) = \prod_{i=1}^{d} t(x_i, v)$ and $d = 10$

(n, k_n)	α	N(0,1)	$v = 5$	$v = 10$	$v = 20$
(2400,30)	0.05	0.050	0.904	0.242	0.076
(2400,50)	0.05	0.048	0.726	0.140	0.058
(2400,80)	0.05	0.052	0.502	0.084	0.052
(4800,30)	0.05	0.042	1.000	0.666	0.142
(4800,50)	0.05	0.048	1.000	0.488	0.094
(4800,80)	0.05	0.044	0.984	0.322	0.082

Table 4.35. Running time for goodness of fit with $d = 10$

(n, k_n)	(4800, 30)	(4800, 50)	(4800, 80)	(4800,1)	(10000, 1)
Time	1.58	1.12	1.00	38.34	169.17

In Table 4.35, we recorded the running time(in seconds) of the divide-and-conquer method and the full sample when $d = 10$. For $n = 4800$ and groups given as k_n= 1, 30, 50 and 80, when $k_n = 1$ implies full sample. The full sample U-statistic takes 38.34 seconds, while the divide-and-conquer method only requires less than 2 seconds for all the groups. As the k_n increases, the running time in seconds decreases, thus larger k_n saves more time. For $n = 10,000$ the running time for full sample is 169.17 sec. while 38.34 sec. for full sample of $n = 4800$. The computation becomes costly and time-consuming as a result of an increase in sample size. Then, our method proved to be more efficient, especially when the running simulation of the large sample where it demands much time.

4.1.3. Non-degenerate (Gini's Difference) U-statistics of 1-dimension with one sample

For non-degenerate U-statistics, we consider the Gini difference as an example. For independent X_1 and X_2 from the same distribution, the Gini difference is defined as

$$\theta = E[|X_1 - X_2|].$$

Given i.i.d. data $X_1, X_2, \ldots, X_n$, the corresponding U-statistics is

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} |X_i - X_j|.$$

Suppose we want to test the following hypothesis

$$H_0 : \theta = \theta_0, \ v.s. \ H_1 : \theta = \theta_1,$$

and use (2.5) to construct the test statistic.

For any independent variables $X, Y \sim N(\mu, \sigma^2)$, the Gini difference is $\theta = \frac{2\sigma}{\sqrt{\pi}}$. In Table 4.36, under H_0, we generate the data from $N(1,1)$, while under H_1, the data are generated from $N(1, 1.02)$, $N(1, 1.03)$, $N(1, 1.04)$ respectively. Two different sample sizes 2400 and 4800 with same blocks k_n=10, 30, 50 and 80 were simulated. For fixed n, the larger m_n has larger power. It also shows that the power increases as m_n increases. Moreover, for the fixed n and m_n, the power increases as the standard deviation increases too. Comparatively, the power of the sample size 4800 is higher than that 2400 which implies the high power depends on the sample size.

Table 4.36. Simulated size and power for Gini difference with $n \in \{2400, 4800\}$, $k_n \in \{10, 30, 50, 80\}$, and $N(1,1)$, $N(1, 1.02)$, $N(1, 1.03)$ and $N(1, 1.04)$

(n, k_n)	α	$N(1,1)$	$N(1, 1.02)$	$N(1, 1.03)$	$N(1, 1.04)$
(2400, 10)	0.05	0.051	0.294	0.560	0.750
(2400, 30)	0.05	0.050	0.274	0.562	0.758
(2400, 50)	0.05	0.056	0.264	0.494	0.754
(2400, 80)	0.05	0.054	0.270	0.524	0.728
(4800, 10)	0.05	0.050	0.640	0.810	0.970
(4800, 30)	0.05	0.054	0.474	0.786	0.974
(4800, 50)	0.05	0.050	0.530	0.800	0.952
(4800, 80)	0.05	0.050	0.414	0.772	0.958

Table 4.37. Simulated size and power for Gini difference with $n \in \{2400, 4800\}$, $k_n \in \{10, 30, 50, 80\}$, and $N(1,1)$, $N(1, 0.98)$, $N(1, 0.95)$ and $N(1, 0.90)$

(n, k_n)	α	N(1,1)	N(1,0.98)	N(1,0.95)	N(1,0.90)
(2400,10)	0.05	0.045	0.350	0.958	1.000
(2400,30)	0.05	0.046	0.314	0.942	0.996
(2400,50)	0.05	0.052	0.278	0.944	0.982
(2400,80)	0.05	0.054	0.288	0.942	0.966
(4800,10)	0.05	0.056	0.506	0.996	1.000
(4800,30)	0.05	0.054	0.486	1.000	1.000
(4800,50)	0.05	0.050	0.510	1.000	1.000
(4800,80)	0.05	0.050	0.538	0.998	1.000

In the Table 4.37, it follows the same pattern as Table 4.36. The power of the hypothesis increases as the m_n increases. The large sample size, n, performs better than the small size. Therefore, the bottom table has higher powers. The power approximates to 1 as the standard deviation is getting far from 1. When the $\sigma \leq 0.90$, the power approaches 1, and therefore it is insignificant to simulate the power with the same n.

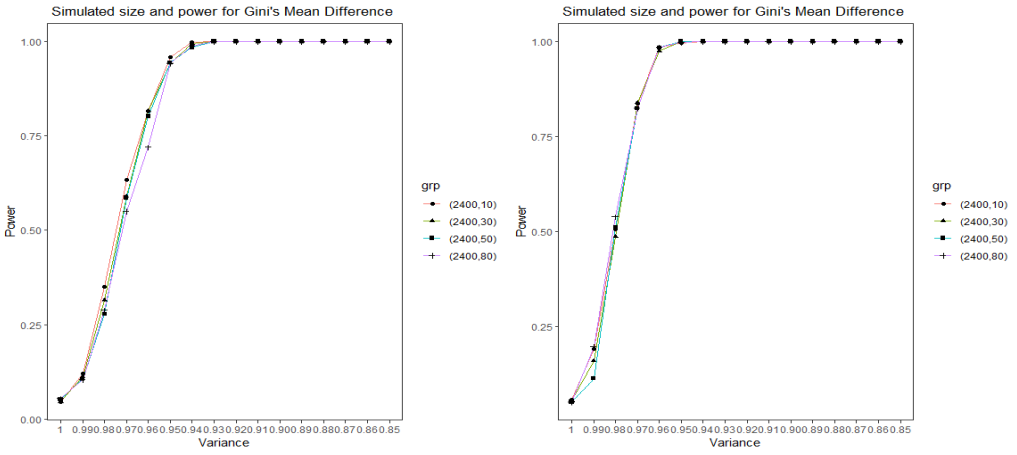

Figure 4.13. Plot of simulated size and power for Gini difference with $n \in \{2400, 4800\}$, $k_n \in \{10, 30, 50, 80\}$, and $\sigma \in \{1.00, 0.99, 0.98, \ldots, 0.87, 0.86, 0.85\}$

Figure 4.14 shows the visual representation of the simulated size and power of Gini difference with a sample size of $n = (2400, 4800)$. The left plot is the plot of the powers with $n = 2400$, and the right plot is $n = 4800$. The plots show upward sloping with all starting point of almost 0.05. The curves of $n = 4800$ are steeper than curves of $n = 2400$, which implies they have high powers.

The graph visualization confirms that the power increases as the σ get far larger. For any $\sigma \leq 0.90$, the powers are 1 which implies the graphs are linear after $\sigma = 1.04$.

Table 4.38. Simulated size and power for Gini difference with $n \in \{2400, 4800\}$, $k_n \in \{10, 30, 50, 80\}$, and $N(1,1)$, $N(1,1.01)$, $N(1,1.02)$ and $N(1,1.03)$

(n, k_n)	α	N(1,1)	N(1,1.01)	N(1,1.02)	N(1,1.03)
(2400,10)	0.05	0.045	0.128	0.284	0.556
(2400,30)	0.05	0.046	0.106	0.258	0.530
(2400,50)	0.05	0.052	0.096	0.228	0.500
(2400,80)	0.05	0.054	0.110	0.280	0.492
(4800,10)	0.05	0.056	0.204	0.486	0.772
(4800,30)	0.05	0.054	0.162	0.496	0.806
(4800,50)	0.05	0.050	0.142	0.434	0.808
(4800,80)	0.05	0.050	0.144	0.460	0.798

From Table 4.38, the tests exhibit increasing higher powers but decreasing rate as the k_n increases. The powers are rising significantly as the σ increases. As usual, the power of a large sample size higher than which implies that one needs to increase the sample size in order to increase power.

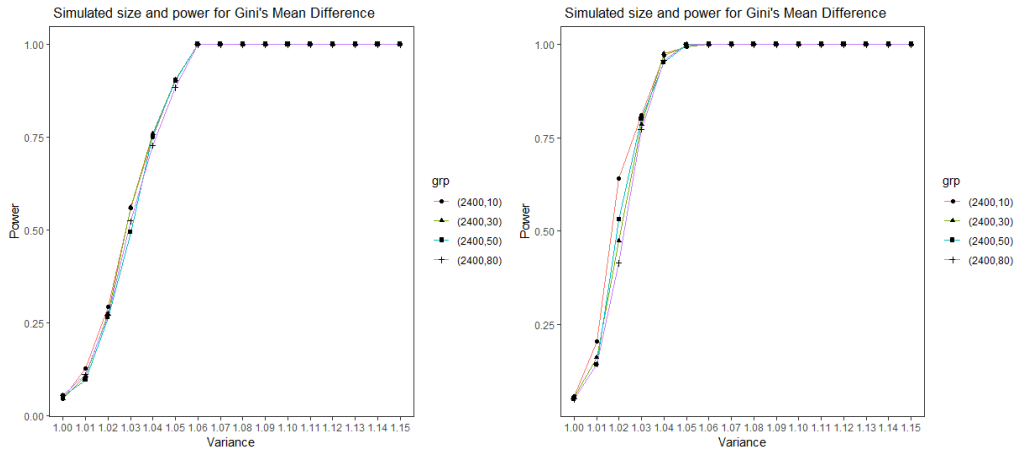

Figure 4.14. Plot of simulated size and power for Gini difference with $n \in \{2400, 4800\}$, $k_n \in \{10, 30, 50, 80\}$, and $\sigma \in \{1.00, 1.01, 1.03, \ldots, 1.12, 1.13, 1.15\}$

The Figure 4.13 represents the simulated size and power of Gini difference with sample size $n = (2400, 4800)$ and $\sigma \in \{1.00, 1.01, 1.03, \ldots, 1.12, 1.13, 1.15\}$. Both plots show upward sloping

with a starting point of almost 0.05. The curves of $n = 4800$ are steeper than curves of $n = 2400$, which implies they have high powers. The graph confirms that the power increases as the σ gets larger and larger until approaches 1.

Table 4.39. Simulated size and power for Gini difference with $n \in \{2400, 4800\}$, $k_n \in \{10, 30, 50, 80\}$, and $N(1, 1)$, $N(1, 1.05)$, $N(1, 1.10)$ and $N(1, 1.15)$

(n, k_n)	α	N(1,1)	N(1,1.05)	N(1,1.10)	N(1,1.15)
(2400,10)	0.05	0.045	0.904	1.000	1.000
(2400,30)	0.05	0.046	0.904	1.000	1.000
(2400,50)	0.05	0.052	0.902	1.000	1.000
(2400,80)	0.05	0.054	0.884	1.000	1.000
(4800,10)	0.05	0.056	0.992	1.000	1.000
(4800,30)	0.05	0.054	0.992	1.000	1.000
(4800,50)	0.05	0.050	0.998	1.000	1.000
(4800,80)	0.05	0.050	1.000	1.000	1.000

In Table 4.39, it follows the same pattern as Table 4.37. The power of the hypothesis increases as the m_n and σ increase. The larger sample size n is performed better than the small size. Then, the bottom table has higher powers. The power of the hypothesis approximate to 1 as the standard deviation is getting far larger than 1. When the $\sigma = 1.10$, the power approaches 1 and therefore it is irrelevant to simulate the power of $\sigma > 1.10$ with the same n and k_n. It supports that the divide-and-conquer method has high power.

Table 4.40. Running time for Gini difference with $n = 4800$ and $k_n \in (1, 30, 50, 80)$

(n, k_n)	(4800, 30)	(4800, 50)	(4800, 80)	(4800,1)
Time	0.44	0.32	0.25	11.36

In Table 4.40, we record the running time for a fixed sample size of n and different groups. With the n of 4800 and groups,k_n= 30, 50, 80 and 1, the time recorded in seconds are 0.44, 0.32, 0.25 and 11.36 respectively. We realized that time decreases as the group increases. It justifies that the divide-and-conquer method significantly reduces the running time. Moreover, our method is more time-efficient than the old method in terms of running time and strong power.

4.1.4. The real data for one sample: goodness-of-fit test

We use the crime data from the city of Chicago in 2016, 2017, 2018, and 2019 to evaluate the performance of the proposed method. The sample size of the data is $n = 20000$ locations of crime events expressed as latitude and longitude coordinates were selected. It is publicly available at https://data.cityofchicago.org/browse?category=
Public%20Safety. We test whether the location follows the bivariate normal distribution $p(x) = N(\mu_1, \sigma_1^2)N(\mu_2, \sigma_2^2)$. We randomly selected 5000 dataset of the locations to estimate the mean values and variances, which yields $\hat{\mu}_1 = 41.844$, $\hat{\mu}_2 = -87.673$, $\hat{\sigma}_1 = 0.085$, $\hat{\sigma}_2 = 0.058$. Then the rest 15,000 location is used to do the test. Let $x = (x_1, x_2)$, $y = (y_1, y_2)$. The direct computation yields the following kernel for U-statistics

$$\nabla_x \log p(x) \quad = \quad \Big(- \frac{x_1 - \mu_1}{\sigma_1^2}, - \frac{x_2 - \mu_2}{\sigma_2^2} \Big),$$

$$\begin{aligned} u_p(x, y) \quad = \quad & \exp\Big\{ - \frac{(x_1 - y_1)^2 + (x_2 - y_2)^2}{2} \Big\} \Big\{ \frac{(x_1 - \mu_1)(y_1 - \mu_1)}{\sigma_1^4} + \frac{(x_2 - \mu_2)(y_2 - \mu_2)}{\sigma_2^4} \\ & - \frac{(x_1 - \mu_1)(x_1 - y_1)}{\sigma_1^2} - \frac{(x_2 - \mu_2)(x_2 - y_2)}{\sigma_2^2} + \frac{(y_1 - \mu_1)(x_1 - y_1)}{\sigma_1^2} \\ & + \frac{(y_2 - \mu_2)(x_2 - y_2)}{\sigma_2^2} + 2 - (x_1 - y_1)^2 - (x_2 - y_2)^2 \Big\}, \end{aligned}$$

which is symmetric and bounded.

We verify if our data follow the bivariate distribution, 5000 samples randomly selected from the data of 20,000 observations of the location of crimes. A visual representation is made by plotting the data using the geographical of Latitude and Longitude. By the scatter plots of the crime location of 5000 data in Figure 4.15, 4.16, 4.17 and 4.18, they show that plots have two clusters or concentration. Therefore, they have two peaks, but for the bivariate distribution case, it should have one cluster and peak. Further, it is, therefore, that the data does not follow the bivariate normal distribution. Our method correctly rejects the null hypothesis using visualization that it is the bivariate normal distribution.

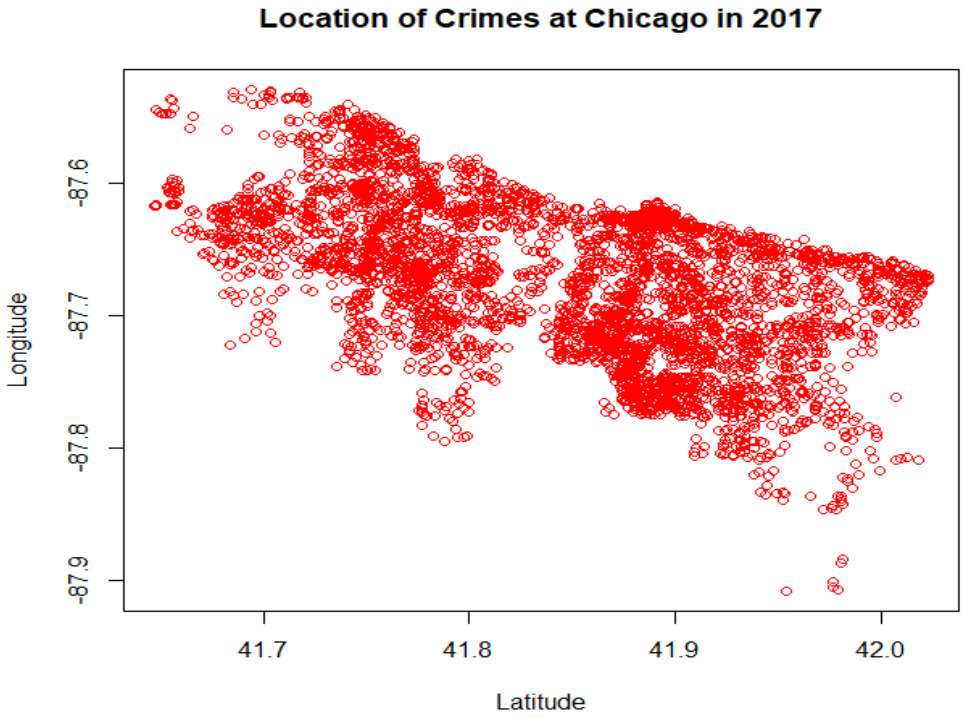

Figure 4.15. Location of crimes at Chicago in 2017 of 5000 observations

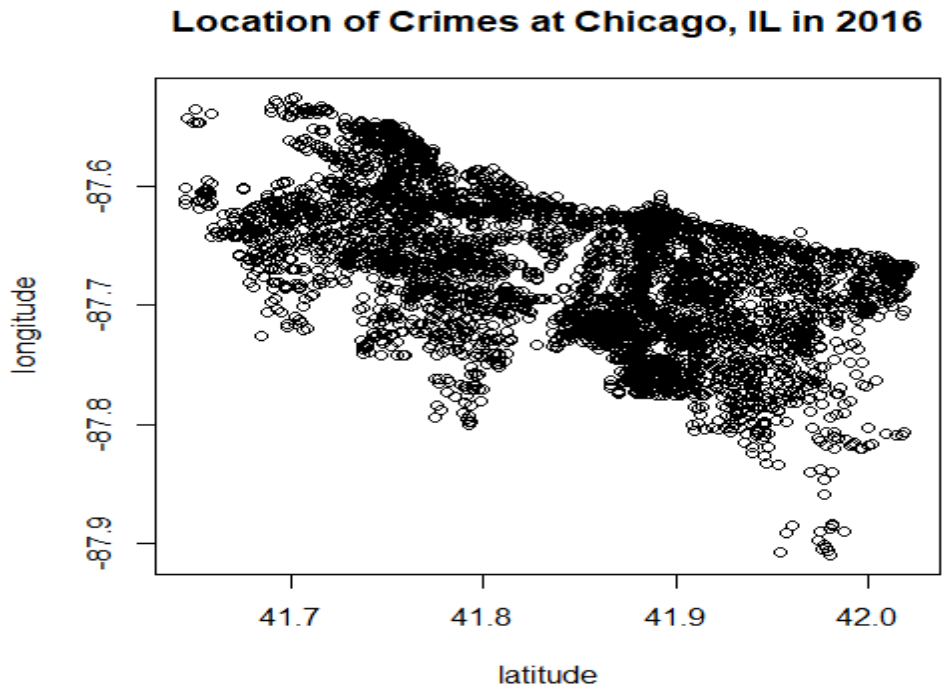

Figure 4.16. Location of crimes at Chicago in 2016 of 5000 observations

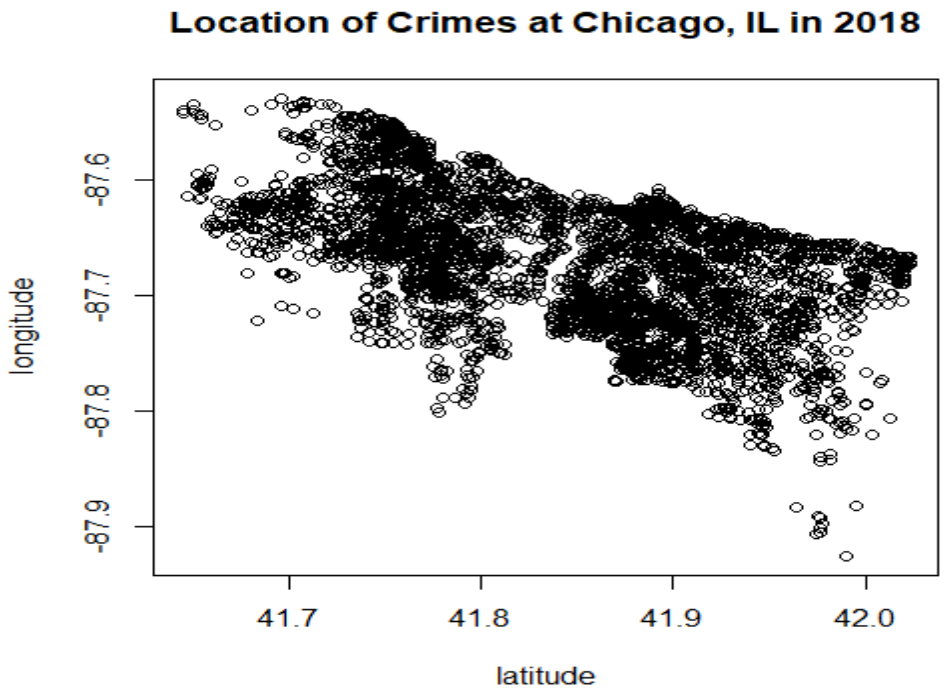

Figure 4.17. Location of crimes at Chicago in 2018 of 5000 observations

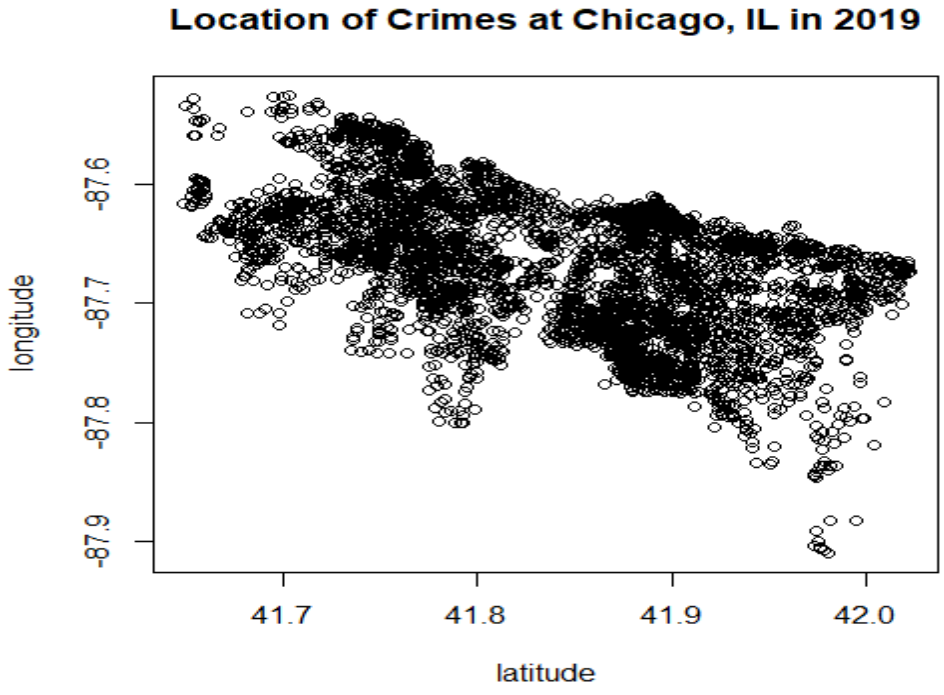

Figure 4.18. Location of crimes at Chicago in 2019 of 5000 observations

We prove that our proposed method is more efficient and fastest in running time. Then, 15,000 samples randomly selected from 20,000 from data of the location of crimes at the city of Chicago, IL, from 2016 to 2019 for validation to support our claim. We divide the 15000 data into $k_n = 30, 50, 80$ groups and calculate the test statistics (2.2) respectively. The groups of our proposed test is $k_n = 30$, 50 and 80 and $k_n = 1$ is the full sample size. From Table 4.41, 4.42, 4.43 and 4.44, the proposed divide-and-conquer method running time is less than 12 seconds for group 30 and less than 5 seconds for group 80, while it takes approximately 6 minutes to calculate the test statistic for the full sample. It indicates that the running time reduces as the k_n increases. This result further confirms that our method saves time and can have high power. We also find out whether our data does follow the bivariate normal distribution. From Table 4.41, 4.42, 4.43 and 4.44, all the p-values of corresponding (n, k_n) are less than the nominal level, $\alpha = 0.05$, therefore we reject the null hypothesis and conclude that the locations of crimes does not follows the bivariate distribution under the null hypothesis.

Table 4.41. The p-values and running time of crimes at Chicago, IL for 2016

(n, k_n)	(15,000, 30)	(15,000, 50)	(15,000, 80)	(15,000,1)
p-value	0.0000	0.0000	0.0000	0.0000
Time	11.3100	6.7200	4.2900	334.2200

Table 4.42. The p-values and running time of crimes at Chicago, IL for 2017

(n, k_n)	(15,000, 30)	(15,000, 50)	(15,000, 80)	(15,000,1)
p-value	0.0000	0.0024	0.0078	0.0000
Time	10.4700	6.3500	4.0300	316.7200

Table 4.43. The p-values and running time of crimes at Chicago, IL for 2018

(n, k_n)	(15,000, 30)	(15,000, 50)	(15,000, 80)	(15,000,1)
p-value	0.0000	0.0000	0.0000	0.0000
Time	11.5700	7.2200	4.4400	341.8600

Table 4.44. The p-values and running time of crimes at Chicago, IL for 2019

(n, k_n)	(15,000, 30)	(15,000, 50)	(15,000, 80)	(15,000,1)
p-value	0.0000	0.0000	0.0000	0.0000
Time	11.1200	6.6100	4.2500	341.3100

4.2. The numerical experiment of two-sample test

Base on the proposed method, the divide-and-conquer method, we perform the simulation of equal and unequal sample sizes. We run a simulation and real data to evaluate the performance of samples of our method. The simulation study applied to evaluate the efficiency of the powers of the test with different dimensions with both equal and unequal samples.

4.2.1. Simulation

Generate i.i.d. data $X_1, \ldots, X_m$ from distribution p and i.i.d. data $Y_1, \ldots, Y_n$ from distribution q. Consider the following hypotheses

$$H_0 : p = q, \qquad H_1 : p \neq q.$$

In this simulation, we use the standard Gaussian kernel function $K(x, y) = e^{\frac{(x-y)^2}{2}}$, where $p = N(0, I_d)$ and $q = N(\mu, I_d)$ or $q = N(0, \sigma^2 I_d)$ are null and alternative hypothesis respectively, where d is the dimension of the normal distribution, the mean μ and variance σ^2 will be varied to assess the power. Example of the distribution includes the following, the Null hypothesis is $p = N(0, I_2)$ and Alternative hypothesis can either be $q = N(0.20, I_2)$ or $q = N(0, 1.2 * I_2)$ where $d = 2$, $\mu = 0.20$ and $\sigma^2 = 1.20$. Let $\alpha = 0.05$, and we repeat the experiment 500 times to calculate the empirical size and power.

We run the simulation for equal sample sizes $m = n$ and unequal sample sizes $m \neq n$. The simulation results for $m = n$ and $m \neq n$ are presented in two separate sections. For the same sample size case, we also compare the power of our method with the linear-time approximation method proposed in (Gretton et al. (2012)), while for the unequal sample size case, the linear-time approximation method is not available.

4.2.1.1. Simulation result for equal sample size

In this section, we empirically evaluate the performance of our method when the sample sizes are equal($m = n$) and compare it with the linear time test statistic. Two sample sizes are

taken, $m = n = 2,000$ and $m = n = 4,000$. We evenly and randomly divide the two samples into $k = 10, 20, 40$ groups, respectively, and calculate the divide-and-conquer test statistic T_k.

Firstly, we assess the power of our test to detect the location difference for various dimensions, that is, $q = N(\mu, I_d)$, where $\mu = 0.00, 0.10, 0.20$ and $d = 2, 5, 10$. When $\mu = 0.00$, $p = q$, it yields the empirical size. The results are summarized in Table 4.45-4.47. All the simulated sizes of our method are close to the nominal level $\alpha = 0.05$. For fixed (n, m, k), the power increases as μ get further away from 0.00. Especially, when $n = 4,000$, $\mu = 0.20$ and $d = 2$, all the powers are greater than 0.90 (see Table 4.45). For fixed μ, n, and d, it is clear that the powers decline as k increases from 10 to 40. It shows the trade-off between computational cost and power of the test statistic. Increasing k reduces the running time but results in loss of power. The power increases for fixed μ, d and k, as the sample size doubles. Due to the curse of dimensionality, as the dimension d increases, the power drops significantly. In order to achieve high power, we need a large sample size.

In this setting, the liner time test statistic performs poorly. The sizes vary a lot from 0.02 to 0.06. Besides, it has no power to detect the location difference between q and p. When $n = 4000, d = 2$, $\mu = 0.60$, the power of linear method is 0.964, but our test has power 1 in this case.

Table 4.45. Simulated size and power for goodness of fit with $n = m \in \{2000, 4000\}$, $k_n \in \{10, 20, 40\}$, $\mu \in \{0.00, 0.10, 0.20\}$ and $d = 2$.

μ	$(n, k_n) = (2,000, 10)$	$(2,000, 20)$	$(2,000, 40)$	Linear
0.00	0.060	0.053	0.052	0.030
0.10	0.080	0.072	0.064	0.050
0.20	0.674	0.600	0.374	0.042
μ	$(n, k_n) = (4,000, 10)$	$(4,000, 20)$	$(4,000, 40)$	Linear
0.00	0.064	0.050	0.052	0.060
0.10	0.254	0.210	0.124	0.066
0.20	0.998	0.990	0.930	0.038

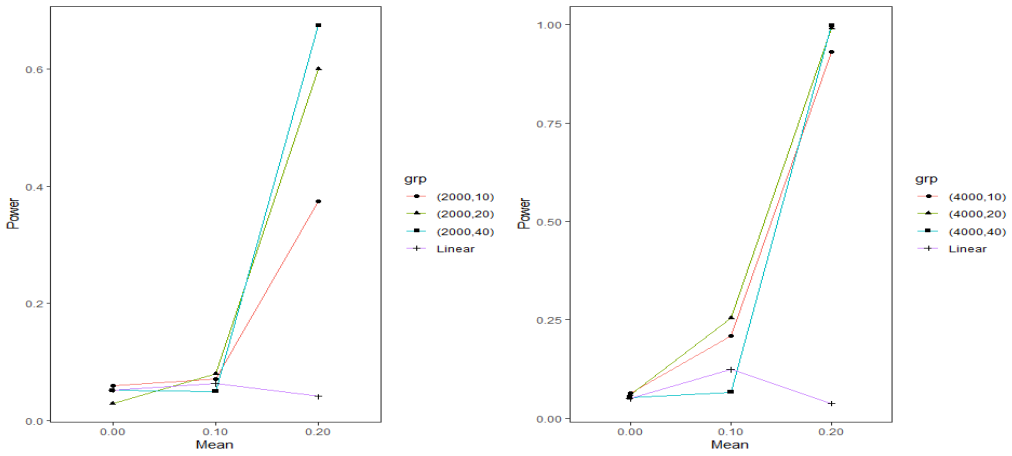

Figure 4.19. The plot of the simulated size and power for goodness of fit with $n = m \in \{2000, 4000\}$, $k_n \in \{10, 20, 40\}$, $\mu \in \{0.00, 0.10, 0.20\}$ and $d = 2$.

From Figure 4.19, the left and right sides of the plot are sample sizes of 2,000 and 4,000, respectively. For varied mean and fixed variance, it observed that except for the linear time test statistic plot, all plot of our proposed method has positively sloped. In this case, the power increases as a result of the changes in the mean. The plots (2,000, 10) are higher than that of (2,000, 20) and (2,000, 40), which indicates that the plot moves outward. It is due to the significant increase in the group.

It observed that our proposed method reduces the steepness of the graph as the group increases. The linear time test statistic plot is approximate to linear in that there is no significant difference from the size when the mean varies. Besides, the plots of the $n = 4000$ are steeper than that of $n = 2000$. It is as a result of the doubling of sample size.

Table 4.46. Simulated size and power for goodness of fit with $n = m \in \{2000, 4000\}$, $k_n \in \{10, 20, 40\}$, $\mu \in \{0.00, 0.10, 0.20\}$ and $d = 5$

μ	$(n, k_n) = (2,000, 10)$	$(2,000, 20)$	$(2,000, 40)$	Linear
0.00	0.044	0.060	0.040	0.056
0.10	0.094	0.067	0.044	0.056
0.20	0.528	0.406	0.270	0.054
μ	$(n, k_n) = (4,000, 10)$	$(4,000, 20)$	$(4,000, 40)$	Linear
0.00	0.046	0.038	0.052	0.054
0.10	0.194	0.154	0.110	0.042
0.20	0.986	0.926	0.708	0.050

Table 4.47. Simulated size and power for goodness of fit with $n = m \in \{2000, 4000\}$, $k_n \in \{10, 20, 40\}$, $\mu \in \{0.00, 0.10, 0.20\}$ and $d = 10$

μ	$(n, k_n) = (2,000, 10)$	$(2,000, 20)$	$(2,000, 40)$	Linear
0.00	0.036	0.058	0.056	0.028
0.10	0.052	0.050	0.058	0.020
0.20	0.088	0.064	0.064	0.020
μ	$(n, k_n) = (4,000, 10)$	$(4,000, 20)$	$(4,000, 40)$	Linear
0.00	0.030	0.048	0.048	0.020
0.10	0.066	0.054	0.070	0.010
0.20	0.296	0.166	0.120	0.008

Secondly, we assess the power of our test to detect the scale difference for various dimensions, that is, $q = N(0, \sigma^2 I_d)$, where $\sigma^2 = 1.00, 1.20, 1.30, 1.40$ and $d = 2, 5, 10$. When $\sigma^2 = 1.00$, and null hypothesis $p = q$, it yields the empirical size. The results are summarized in Table 4.48-4.50. All the simulated sizes of our method are close to the nominal level of $\alpha = 0.05$. The power pattern is similar to the previous case. In this setting, the liner time test statistic still performs poorly. It has no power to detect the scale difference between q and p.

Table 4.48. Simulated size and power for goodness of fit with $n = m \in \{2000, 4000\}$, $k_n \in \{10, 20, 40\}$, $\sigma^2 \in \{1.00, 1.20, 1.30, 1.40\}$ and $d = 2$

σ^2	$(n, k_n) = (2,000, 10)$	$(2,000, 20)$	$(2,000, 40)$	Linear
1.00	0.040	0.058	0.058	0.040
1.20	0.154	0.090	0.102	0.062
1.30	0.500	0.360	0.286	0.078
1.40	0.852	0.774	0.522	0.066
σ^2	$(n, k_n) = (4,000, 10)$	$(4,000, 20)$	$(4,000, 40)$	Linear
1.00	0.042	0.050	0.032	0.052
1.20	0.430	0.396	0.248	0.042
1.30	0.950	0.878	0.724	0.054
1.40	0.998	0.998	0.992	0.084

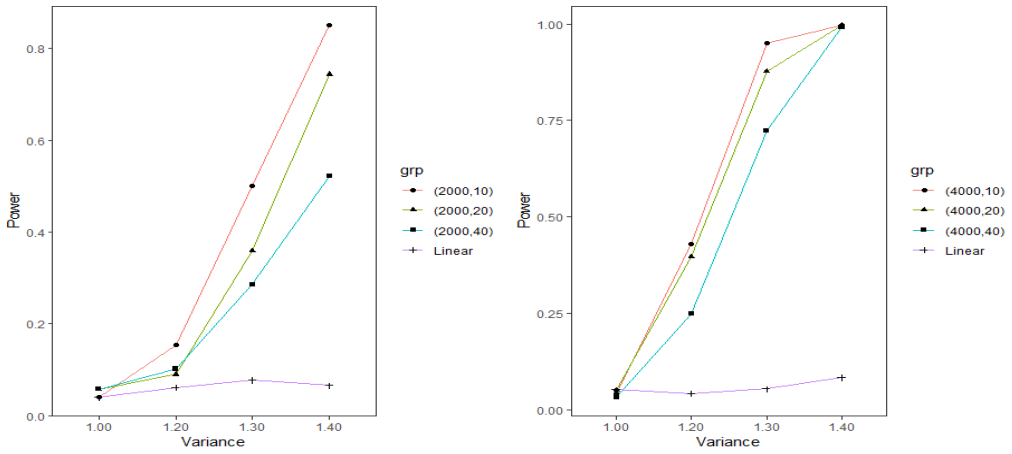

Figure 4.20. The plot of the simulated size and power for goodness of fit with $n = m \in \{2000, 4000\}$, $k_n \in \{10, 20, 40\}$, $\sigma^2 \in \{1.00, 1.20, 1.30, 1.40\}$, and $d = 2$

In Figure 4.20 shows the plot with dimension, $d = 2$, sample sizes of n = 2,000 and 4,000 respectively where mean is fixed and variance differs. The left and right sides of the plot are sample sizes of 2,000 and 4,000, respectively. The plots have a similar pattern to the previous plot. It showed that our proposed method has steep graphs, and the steepness reduces as the group increases. The linear time test statistic plot is approximate to linear, which implies that it is no significant difference from the size when the variance varies.

Also, the plots of the $n = 2000$ are more gentle than that of $n = 4000$. It is as a result of the doubling of sample size. It can also observe that except for the linear time test statistic plot, all plot of our proposed method has positively sloped. In that case, the power increases as a result of the change of variance. The plots (2,000, 10) are higher than that of (2,000, 20) and (2,000, 40). It indicates that the plot moves outward as a result of a significant increase in the group.

Table 4.49. Simulated size and power for goodness of fit with $n = m \in \{2000, 4000\}$, $k_n \in \{10, 20, 40\}$, $\sigma^2 \in \{1.00, 1.20, 1.30, 1.40\}$ and $d = 5$

σ^2	$(n, k_n) = (2,000, 10)$	$(2,000, 20)$	$(2,000, 40)$	Linear
1.00	0.022	0.030	0.044	0.042
1.20	0.252	0.168	0.112	0.040
1.30	0.628	0.500	0.312	0.048
1.40	0.950	0.856	0.688	0.068
σ^2	$(n, k_n) = (4,000, 10)$	$(4,000, 20)$	$(4,000, 40)$	Linear
1.00	0.040	0.048	0.032	0.042
1.20	0.682	0.522	0.326	0.058
1.30	0.992	0.966	0.832	0.046
1.40	1.000	1.000	0.990	0.106

Table 4.50. Simulated size and power for goodness of fit with $n = m \in \{2000, 4000\}$, $k_n \in \{10, 20, 40\}$, $\sigma^2 \in \{1.00, 1.20, 1.30, 1.40\}$ and $d = 10$

σ^2	$(n, k_n) = (2,000, 10)$	$(2,000, 20)$	$(2,000, 40)$	Linear
1.00	0.042	0.048	0.044	0.008
1.20	0.082	0.052	0.062	0.028
1.30	0.258	0.150	0.114	0.012
1.40	0.436	0.318	0.182	0.018
σ^2	$(n, k_n) = (4,000, 10)$	$(4,000, 20)$	$(4,000, 40)$	Linear
1.00	0.040	0.044	0.060	0.012
1.20	0.222	0.164	0.100	0.024
1.30	0.590	0.442	0.276	0.028
1.40	0.924	0.800	0.548	0.032

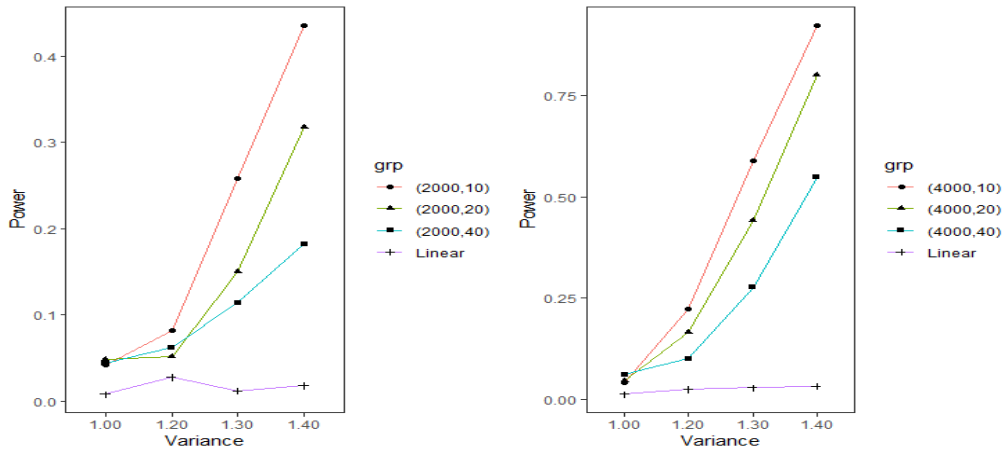

Figure 4.21. The plot of the simulated size and power for goodness of fit with $n = m \in \{2000, 4000\}$, $k_n \in \{10, 20, 40\}$, $\sigma^2 \in \{1.00, 1.20, 1.30, 1.40\}$ and $d = 10$.

Figure 4.21 indicates that the left and right sides of the plot are sample sizes of 2,000 and 4,000 respectively with dimension, $d = 10$ and fixed mean, and variance varies. It has similar in pattern to that of Figure 4.20. It observed that all plot of our proposed method has positively sloped apart from the linear time test statistic plot. Therefore, power increases as a result of the increase in variance. The plots (2,000, 10) are higher than that of (2,000, 20) and (2,000, 40), which indicates that the plot moves outward. It is due to the significant increase in the group.

Also, the plots of the $n = 4000$ are steeper than that of $n = 2000$. It is as a result of the doubling of sample size. Comparing the 4.21 and 4.20, the entire graphs in 4.20 is steeper than graphs in 4.21 and it is due curse of dimensionality, that is changes of the dimension.

4.2.1.2. Simulation result for unequal sample size

In this subsection, we empirically evaluate the performance of our method under the same setup as in subsection 5.2.1.1 except that the sample sizes are unequal, that is, $n = 4,000$, $m = 2,000$ and $n = 4,000$, $m = 3,000$. The linear time test statistic is not available. The simulated results are presented in Table 4.51-4.56. All the simulated sizes are close to the nominal level of $\alpha = 0.05$. Overall, the powers for unequal sample size are smaller than the equal sample size cases($m = n = 4000$), due to less sample size $m \leq 3,000$. The powers have a similar trend as in the equal sample sizes($m = n$) case.

Table 4.51. Simulated size and power for goodness of fit with n=4,000, $m \in \{2000, 3,000\}$, $k_n \in \{10, 20, 40\}$, $\mu \in \{0.00, 0.10, 0.20\}$ and $d = 2$.

μ	$(n, m, k_n) = (4,000, 2,000, 10)$	$(4,000, 2,000, 20)$	$(4,000, 2,000, 40)$
0.00	0.050	0.060	0.048
0.10	0.090	0.112	0.086
0.20	0.904	0.848	0.648
μ	$(n, m, k_n) = (4,000, 3,000, 10)$	$(4,000, 3,000, 20)$	$(4,000, 3,000, 40)$
0.00	0.048	0.050	0.044
0.10	0.166	0.136	0.098
0.20	0.960	0.956	0.850

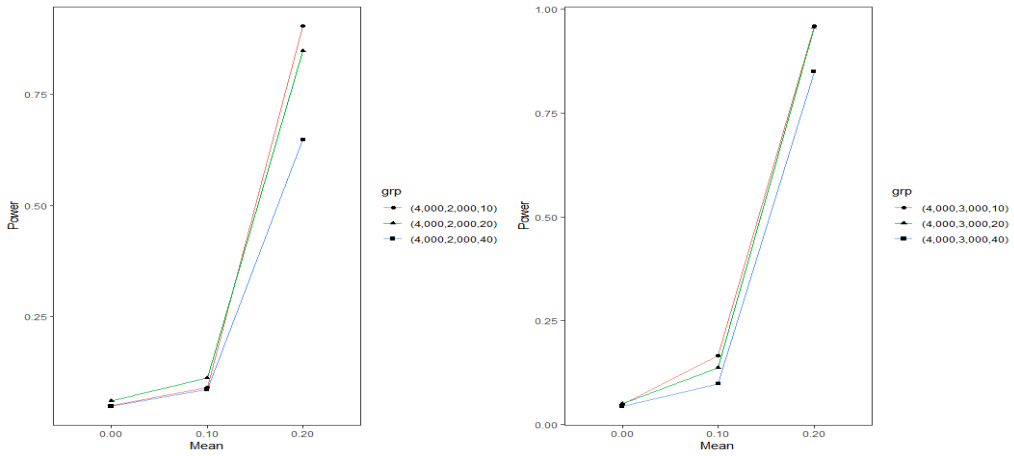

Figure 4.22. The plot of the simulated size and power for goodness of fit with n=4,000, $m \in \{2000, 3,000\}$, $k_n \in \{10, 20, 40\}$, $\mu \in \{0.00, 0.10, 0.20\}$, and $d = 2$

In Figure 4.22, the dimension, $d = 2$ and the left and right sides of the plot are unequal sample sizes of (4,000 and 2,000) and (4,000 and 3000) respectively. The mean is varied from 0.00 to 0.20, and fixed variance. It observed that the plots have positively sloped. In this case, the power increases as a result of the changes in mean. The plots with dimension, $d = 10$ is higher than that of $d = 20$ and 40, which implies that the plot moves outward as the group increases.

Likewise, it observed that the steepness of the plot reduces as the group increases. In addition, the plots of the $n = 4{,}000$ and $m = 3{,}000$ is steeper than that of $n = 4{,}000$ and $m = 2{,}000$ and it is as a result of the increase of sample size.

Table 4.52. Simulated size and power for goodness of fit with n=4,000, $m \in \{2000, 3,000\}$, $k_n \in \{10, 20, 40\}$, $\mu \in \{0.00, 0.10, 0.20\}$ and $d = 5$.

μ	$(n, m, k_n) = (4,000, 2,000, 10)$	$(4,000, 2,000, 20)$	$(4,000, 2,000, 40)$
0.00	0.036	0.030	0.056
0.10	0.078	0.064	0.064
0.20	0.784	0.600	0.366
μ	$(n, m, k_n) = (4,000, 3,000, 10)$	$(4,000, 3,000, 20)$	$(4,000, 3,000, 40)$
0.00	0.048	0.028	0.058
0.10	0.146	0.116	0.072
0.20	0.928	0.818	0.606

Table 4.53. Simulated size and power for goodness of fit with $n = 4,000$, $m \in \{2000, 3,000\}$, $k_n \in \{10, 20, 40\}$, $\mu \in \{0.00, 0.10, 0.20\}$ and $d = 10$.

μ	$(n, m, k_n) = (4,000, 2,000, 10)$	$(4,000, 2,000, 20)$	$(4,000, 2,000, 40)$
0.00	0.048	0.044	0.044
0.10	0.064	0.058	0.052
0.20	0.116	0.088	0.072
μ	$(n, m, k_n) = (4,000, 3,000, 10)$	$(4,000, 3,000, 20)$	$(4,000, 3,000, 40)$
0.00	0.046	0.044	0.046
0.10	0.074	0.060	0.053
0.20	0.178	0.122	0.090

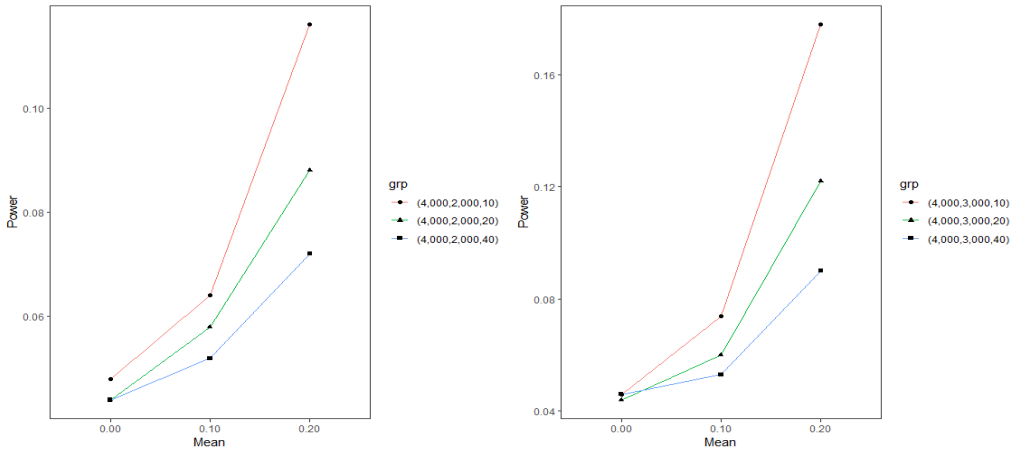

Figure 4.23. The plot of the simulated size and power for goodness of fit with $n = 4,000$, $m \in \{2000, 3,000\}$, $k_n \in \{10, 20, 40\}$, $\mu \in \{0.00, 0.10, 0.20\}$, $d = 10$

In Figure 4.23, the dimension, $d = 10$ and the left and right side of the plot are unequal sample sizes of (4,000 and 2,000) and (4,000 and 3000) respectively. Figure 4.23 has also has the same pattern as 4.22. The mean is varied from 0.00 to 0.20, and fixed variance. It showed that the plots have positively sloped. In this case, the power increases as the mean vary. The plots with dimension, $d = 10$ is higher than that of $d = 20$ and 40, which implies that the plot moves outward as the dimension increases. In addition, it is also showed that the steepness of the plot reduces as the group increases. The plots of the $n = 4,000$ and $m = 3,000$ is steeper than that of $n = 4,000$ and $m = 2,000$ and it is as a result of the increase of sample size.

Table 4.54. Simulated size and power for goodness of fit with n=4,000, $m \in \{2000, 3,000\}$, $k_n \in \{10, 20, 40\}$, $\sigma^2 \in \{1.00, 1.20, 1.30, 1.40\}$ and $d = 2$

σ^2	$(n, m, k_n) = (4,000, 2,000, 10)$	$(4,000, 2,000, 20)$	$(4,000, 2,000, 40)$
1.00	0.052	0.070	0.042
1.20	0.238	0.192	0.132
1.30	0.700	0.578	0.414
1.40	0.968	0.952	0.792
σ^2	$(n, m, k_n) = (4,000, 3,000, 10)$	$(4,000, 3,000, 20)$	$(4,000, 3,000, 40)$
1.00	0.056	0.052	0.044
1.20	0.364	0.312	0.208
1.30	0.888	0.800	0.624
1.40	0.998	0.992	0.950

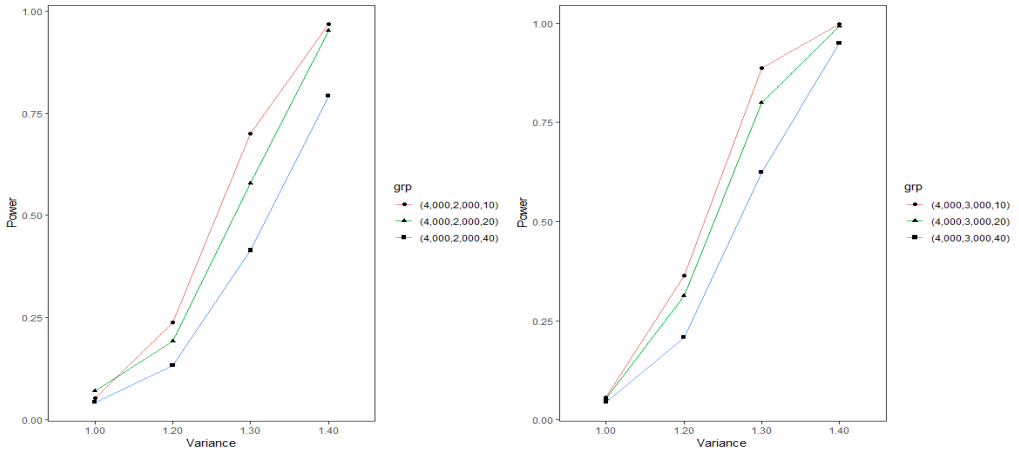

Figure 4.24. The plot of the simulated size and power for goodness of fit with n=4,000, $m \in \{2000, 3,000\}$, $k_n \in \{10, 20, 40\}$, $\sigma^2 \in \{1.00, 1.20, 1.30, 1.40\}$ and $d = 2$

In Figure 4.24 indicated the dimension, $d = 2$ and unequal sample sizes of (4,000 and 2,000) and (4,000 and 3000) respectively. The fixed mean and variance vary from 1.00 to 1.40. The plot is not different from previous plots. They followed upward sloping from left to right. In this case, the power increases as a result of the change of variance. The plots of the group, $k_n = 10$ is higher than that of $k_n = 20$ and 40. It implies the power of the test decrease as the group increases.

Comparatively, an increase of dimension and group reduces the steepness of the plot. Also, the plots of the $n = 4,000$ and 3,000 are steeper than that of $n = 4,000$ and 2,000. It is a result of the increase in sample size.

Table 4.55. Simulated size and power for goodness of fit with $n = 4,000$, $m \in \{2000, 3,000\}$, $k_n \in \{10, 20, 40\}$, $\sigma^2 \in \{1.00, 1.20, 1.30, 1.40\}$ and $d = 5$

σ^2	$(n, m, k_n) = (4,000, 2,000, 10)$	$(4,000, 2,000, 20)$	$(4,000, 2,000, 40)$
1.00	0.044	0.042	0.056
1.20	0.360	0.300	0.190
1.30	0.894	0.786	0.580
1.40	0.998	0.984	0.928
σ^2	$(n, m, k_n) = (4,000, 3,000, 10)$	$(4,000, 3,000, 20)$	$(4,000, 3,000, 40)$
1.00	0.048	0.062	0.058
1.20	0.552	0.424	0.224
1.30	0.968	0.908	0.768
1.40	1.000	1.000	0.984

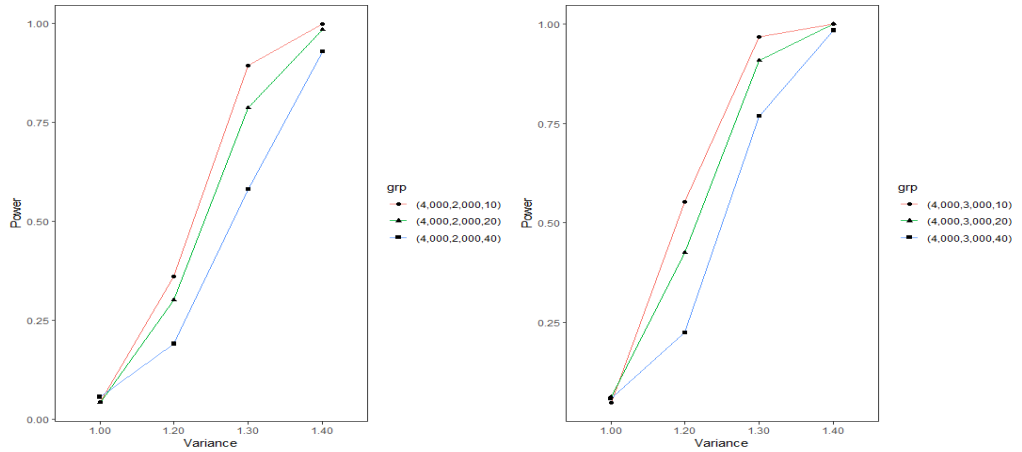

Figure 4.25. The plot of the simulated size and power for goodness of fit with $n = 4,000$, $m \in \{2000, 3,000\}$, $k_n \in \{10, 20, 40\}$, $\sigma^2 \in \{1.00, 1.20, 1.30, 1.40\}$ and $d = 5$

In Figure 4.25 has the similar characteristics as Figure 4.24. The only change is the dimension of the test, where $d = 5$. With unequal sample sizes of (4,000 and 2,000) and (4,000 and 3000), fixed mean and variance varies from 1.00 to 1.40, the plot still follows upward sloping from left to right. Therefore, power increases as a result of an increase in variance. The plots of the group, $k_n = 10$ is higher than that of $k_n = 20$ and 40. This implies the power of the test decrease as the group increases. The plots of the $n = 4,000$ and 3,000 are steeper than that of $n = 4,000$ and 2,000. It is a result of the increase in sample size.

Table 4.56. Simulated size and power for goodness of fit with $n = 4,000$, $m \in \{2000, 3,000\}$, $k_n \in \{10, 20, 40\}$, $\sigma^2 \in \{1.00, 1.20, 1.30, 1.40\}$ and $d = 10$

σ^2	$(n, m, k_n) = (4,000, 2,000, 10)$	$(4,000, 2,000, 20)$	$(4,000, 2,000, 40)$
1.00	0.042	0.038	0.038
1.20	0.146	0.106	0.094
1.30	0.392	0.288	0.186
1.40	0.716	0.606	0.434
σ^2	$(n, m, k_n) = (4,000, 3,000, 10)$	$(4,000, 3,000, 20)$	$(4,000, 3,000, 40)$
1.00	0.042	0.046	0.048
1.20	0.204	0.110	0.092
1.30	0.516	0.370	0.230
1.40	0.890	0.692	0.438

4.2.2. Running time comparison

We compare the running time of our divide-and-conquer test statistic, the full sample MMD test statistic, and the linear time test statistic with the dimension of $d = 10$ and equal sample size of $n = 4000$. The running time reported in Table 4.57. The full sample MMD test statistic (that is $k_n = 1$) takes significantly more time than the proposed divide-and-conquer test statistic. The linear time test statistic has the least running time. For our proposed test statistic, it takes less time to calculate for a large k_n.

Table 4.57. Running time in seconds with $n = 4000$, $k_n \in \{1, 10, 20, 40\}$ and $d = 10$

(n, k_n)	$(4,000, 1)$	$(4,000, 10)$	$(4,000, 20)$	$(4,000, 40)$	Linear
Time	39.4060	2.4657	1.2659	0.6778	0.0464

Based on our simulation, our test statistic can achieve high power and at the same time. It significantly reduces the running time of the full sample MMD test statistic. Besides, our test statistic overwhelmingly outperforms the linear time test statistic in terms of power. Also, our proposed test statistic is more effective than the full sample MMD test statistic in terms of running time.

4.2.3. The real data for two samples: Maximum Mean Discrepancy (MMD) test

We apply our method to the City of Chicago Crime data, where the latitude and longitude of each crime event recorded yearly. We would like to test whether the crime events from two years follow the same distribution. Firstly, we extract the crime data for 2001 and 2009 with an equal

sample size of 200,000, which plotted in Figure 4.26 and Figure 4.27, respectively. Graphically, the scatter plots look almost the same, which implies it follows the same distribution.

We want to justify this conclusion if it follows the same distribution theoretically. We therefore formally perform our divide-and-conquer method hypothesis test with a significant level of $\alpha = 0.05$. The test statistic(TS) value, p-value, running time(in hours) and number of groups k_n are recorded in Table 4.58. All the p-values for different k_n are greater than 0.05. It implies that we fail to reject the null hypothesis that the data follows the same distribution. It takes 65.2876 hours to calculate the test statistic for $k_n = 500$. It reduced to 3.1782 hours when $k_n = 8000$. By the running time change pattern, it is almost impossible to calculate the full sample test statistic.

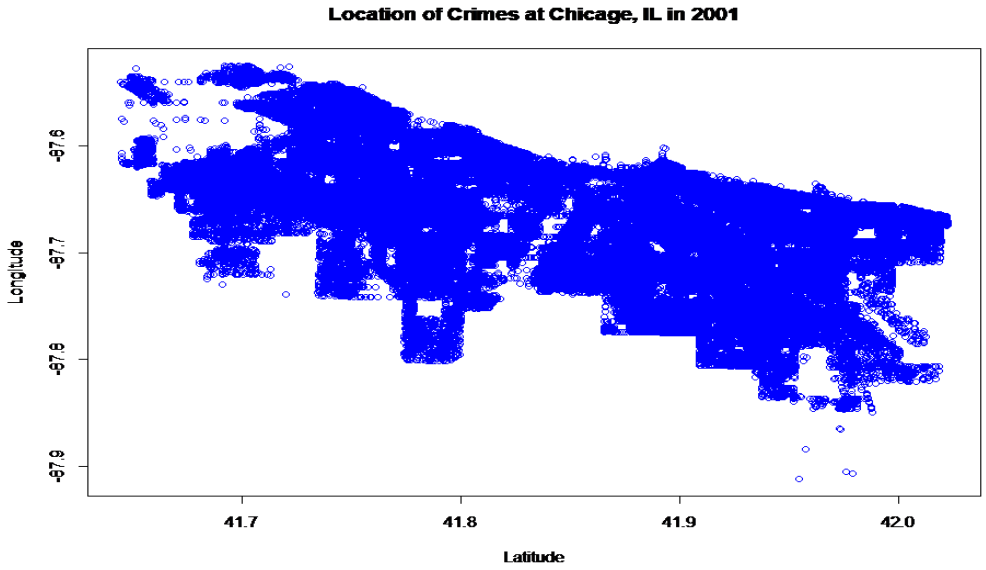

Figure 4.26. Location of crimes at Chicago in 2001 of 200,000 observations

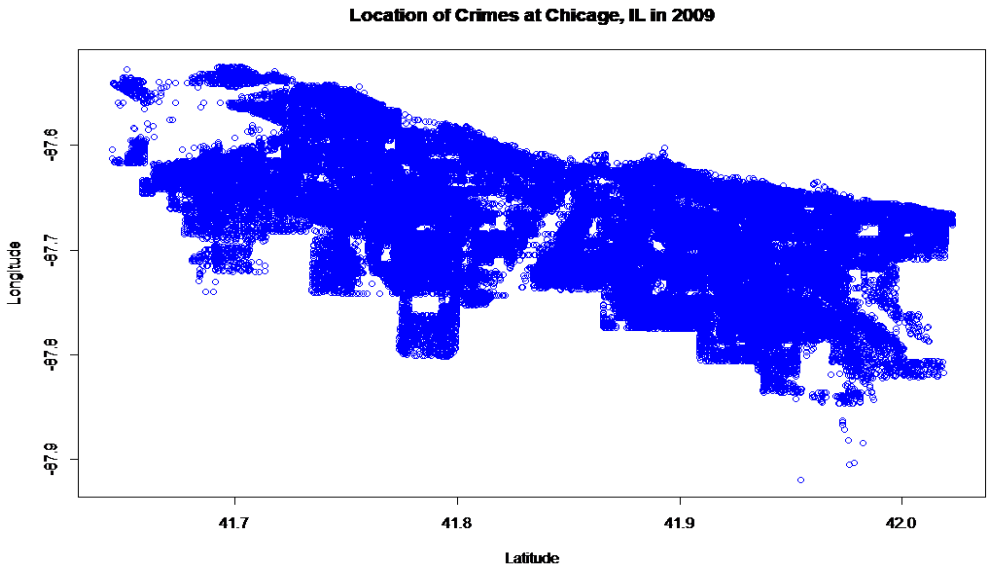

Figure 4.27. Location of crimes at Chicago in 2009 of 200,000 observations

Table 4.58. Chicago crime data analysis for year 2001 and 2009

k_n	500	1000	2,000	4,000	8,000
Time(h)	65.2876	27.8643	13.0803	7.8563	3.1782
TS	-1.1194	-0.7252	0.6858	0.4170	0.0694
p-value	0.2630	0.4683	0.4928	0.6767	0.9447

We again consider the crime events for 2002 and 2010, with an equal sample size of 400,000. In Figure 4.28 and Figure 4.29 below, the scatter plot indicates that there is a clear difference in the lower right corner. Figure 4.29 has a heavy tail at the lower right corner. It concludes that the two data are not of the same distribution.

In order to justify that the two data do not have the same distribution base on the plotting, we use our test statistics to test hypothetically. Table 4.59 reported the test statistic(TS) value, p-value, and time in hours. Our test statistic successfully detects the difference since all the p-values in Table 4.59 are significantly less than 0.05. Therefore, the null hypothesis rejected that the two data follow the same distribution. The running time of $k_n = 500$ is 148.8818 hours while the running of $k_n = 4,000$ is 18.6494 hours. Moreover, the running time decreased as the k_n increases.

The TS value has a relationship with the change of k_n. The TS values decrease as the k_n reduces. Comparatively, the running time for each k_n in Table 4.59 is longer than that in Table 4.58, due to larger sample sizes.

Table 4.59. Chicago crime data analysis of year 2002 and 2010

k_n	500	1,000	2,000	4,000	8,000
Time(h)	148.8818	78.2395	36.8120	18.6494	9.3309
TS	8.8532	6.3389	5.7188	4.7147	2.9334
p-value	0.0000	0.0000	0.0000	0.0000	0.0036

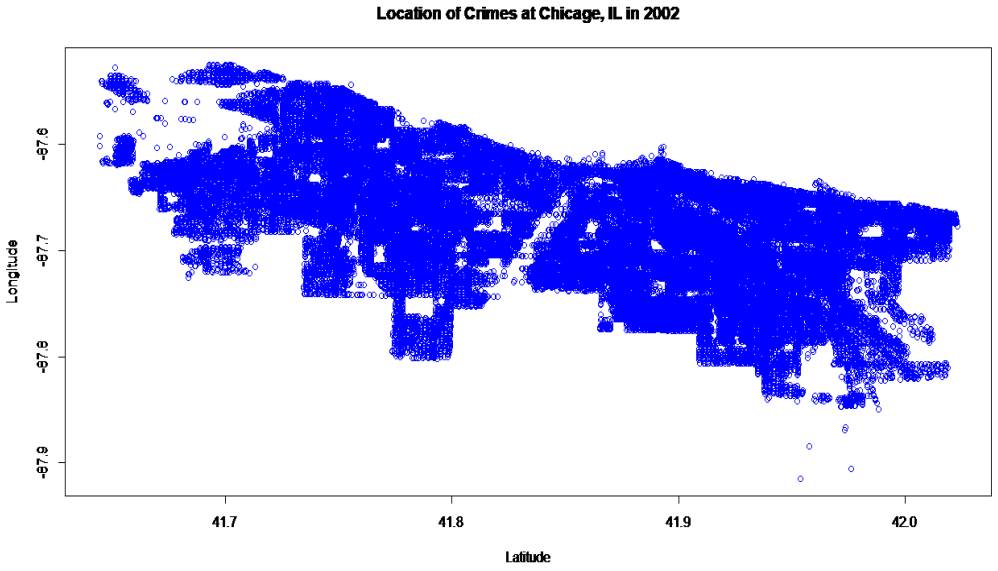

Figure 4.28. Location of crimes at Chicago in 2002 of 200,000 observations

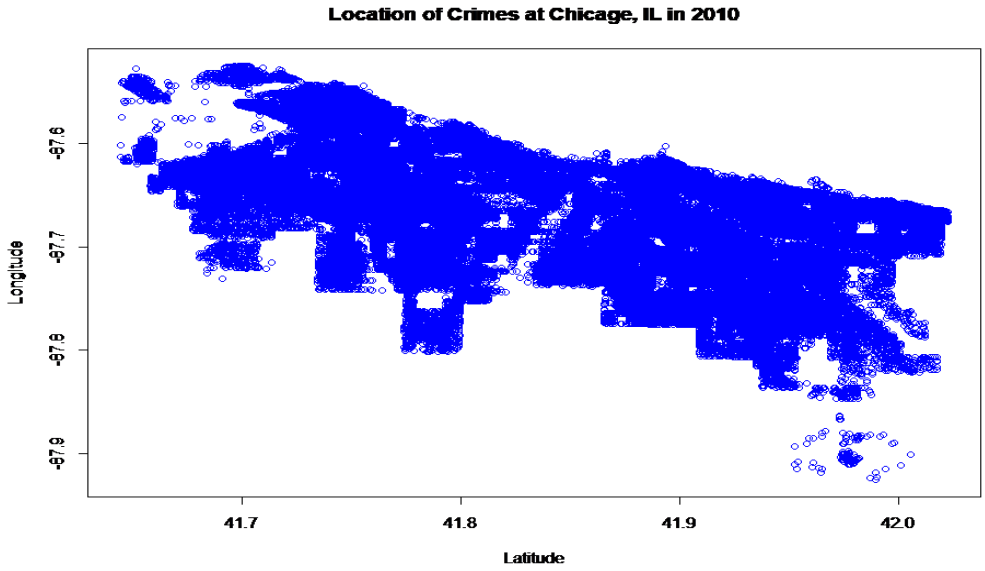

Figure 4.29. Location of crimes at Chicago in 2010 of 200,000 observations

5. DISCUSSION

U-statistics is a class of widely used unbiased nonparametric estimators. They play an essential role in many estimations or statistical inference problems. In many hypothesis testing problems, the test statistics are degenerate U-statistics under the null hypothesis, which motivates us to study the degenerate U-statistics in a big data context. Specifically, we propose a divide and conquer method to solve the computation challenge of degenerate U statistics, and at the same time, the proposed test has the standard normal distribution as the limiting distribution. We apply the methods to the goodness of fit test(KSD) and two-sample test(MMD) and evaluate the performance of the method by extensive simulation.

In the simulation studies, the power of the tests assessed under various block sizes, sample sizes, and multiple distributions(uniquely determined by mean and standard deviation) for both degenerate and non-degenerate test statistics. In the goodness of fit test, for one-dimensional data, for fixed k_n, mean, and standard deviation, the power of the test increases as the sample size increases. The power gets smaller for large k_n for fixed n, mean, and standard deviation. Moreover, for fixed n and k_n, the power is high for either a large discrepancy in the mean or standard deviation. For multivariate data, the power also depends on the dimension. For a higher dimension, power reduces when all other factors are constant.

In applying the method of MMD for a two-sample test, it confirms that with both equal and unequal sample sizes, the power of the hypothesis testing increases as either mean or variance increases for fixed n, m, k_n, and dimension. Likewise, as k_n increases, the power decreases. It constitutes the trade-off between time and power. Furthermore, power significantly affected by the dimension of data.

The k_n controls the balance between running time and test power. For fixed n, larger k_n saves more computation time but sacrifices some power by the simulation. In practice, we recommend choosing moderately large k_n to save time and achieve adequate power. Besides, one byproduct of our method is that we avoid calculating the eigenvalues of the kernel function of the degenerate U-statistics involved in the limiting distribution of degenerate U-statistics.

In summary, under the null hypothesis, our test statistic converges in law to the standard normal distribution. The running time reduced to almost linear. Moreover, the simulation and real data analysis show that our test can achieve high power.

REFERENCES

[1] Arcones, M. A. and Gine, E. (1993). Limit theorems for U-processes. *The Annals of Probability*, (21) 1494–1542.

[2] Arcones, M. and Gine, E. (1995). On the law of the iterated logarithm for canonical U-statistics and processes, *Stochastic processes and their applications*, 58(2): 217-245.

[3] Arcones, M. (1995). On the central limit theorem for U-statistics under absolute regularity, *Statistics and probability letters*, 24(3): 245-249.

[4] Arcones, M. (1992). Large deviations for U-statistics, *Journal of Multivariate Analysis*, 42(2): 299-301.

[5] Arcones, M. and Gine, E. (1992). On the bootstrap of U- and V -statistics. *Annals of Statistics*, 20(2): 655-674.

[6] Arcones, M. A. and Gine, E. (1993). Limit theorems for U-processes. *The Annals of Probability*, (21) 1494-1542.

[7] Anderson, N., Hall, P. and Titterington, D. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, (50) 41-54.

[8] Anderson, T. W and Darling, D. A. (1952). Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes, *Ann. Math. Statist.*, (23) 193-212.

[9] Atta-Asiamah, E and Yuan, M. (2019). Distributed inferences for degenerate U-statistics, *Stat*, (8) e234

[10] Biau, G. and Gyorfi, L. (2005). On the asymptotic properties of a nonparametric l1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11): 3965-3973.

[11] Bickel, P. and Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9(6): 1196-1217.

[12] Bickel, P. (1969). A distribution free version of the Smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, 40(1): 1-23.

[13] Bierens, H. J. and Ploberger, W. (1997). Asymptotic theory of integrated conditional moment tests, *Econometrica*, (65) 1129-1151.

[14] Borgwardt, K.M., Gretton,A., Rasch, M.J., Kriegel, H.-P., Scholkopf, B. and Smola.A. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14): e49-e57, 2006.

[15] Chen, S. X., Zhang, L. X. and Zhong, P. S. (2010), Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.*, (105) 810–819

[16] Cheng, C., Liu, Y., Liu, Z. and Zhou, W. (2018). Balanced augmented jackknife empirical likelihood for two sample U-statistics, *Science China Mathematics*, 61(6): 1129-1138.

[17] Chikkagoudar and Bhat (2014). Limiting Distribution of Two-Sample Degenerate U-Statistic under Contiguous Alternatives and Applications, *Journal of Applied Statistical Science*, (22) 127-139

[18] Clemencon, S. (2011). On U-processes and clustering performance. *NIPS*, 37-45.

[19] Colin, I., Bellet, A., Salmon, J. and Clemencon, S. (2015). Extending gossip algorithms to distributed estimation of U-statistics. *NIPS*, 271-279.

[20] Dehling, H. and Mikosch, T. (1994). Random quadratic forms and the bootstrap for U statistics. *Journal of Multivariate Analysis*, 51(2): 392-413.

[21] Dewan, I. and Prakasa-Rao, B.L.S. (2001). Asymptotic normality for U-statistics of associated random variables, *Journal of Statistical Planning and Inference*, (97) 201–225

[22] Dwass, M. (1956). "The large-sample power of rank tests in the two-sample problem," *Ann. Math. Statist.*, (27) 352-374.

[23] Faivishevsky, L. and Goldberger, J. (2008). ICA based on a smooth estimation of the differential entropy. *NIPS*, 433-440.

[24] Fisher, N. I. and Lee, A. J. (1982). Nonparametric measures of angular-angular association. *Biometrika*, (69) 315-321

[25] Friedman J. and Rafsky, L. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4): 697-717.

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

[27] Hall P. and Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2): 359-374.

[28] Halmos, P. R. (1946). "The theory of unbiased estimation," *Ann. Marh. Statisr.*, (11) 3443

[29] Ho, Hwai-Chung and Shieh, G. S. (2006). Two-stage U-statistics for Hypothesis Testing, *Scandinavian Journal of Statistics*, (33) 861-873

[30] Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19(3): 293-325.

[31] Huang, W. and Zhang, L. (2006). Asymptotic Normality for U-statistics of negatively associated random variables, *Statistics and Probability Letters*, (76) 1125-1131

[32] Huskova, M. and Janssen, P. (1993). Consistency of the generalized bootstrap for degenerate U-statistics. *Annals of Statistics*, 21(4): 1811-1823.

[33] Jing, B., Yuan, J. and Zhou, W. (2008). Empirical likelihood for non-degenerate U-statistics, *Statistics and probability letters*, 78(6): 599-607.

[34] Lee, A. (1990). *U-Statistics: Theory and Practice. Statistics: A Series of Textbooks and Monographs*. CRC Press.

[35] Lehmann, E. L. (1951). "Consistency and unbiasedness of certain nonparametric tests," *Ann. Math. Statist.*, (22) 165-179.

[36] Lin, N. and Xi, R. (2010). Fast surrogates of U-statistics, *Computational Statistics and Data Analysis*, (54) 16-24.

[37] Liu, Q., Lee, J. and Jordan, M. (2016). A Kernelized Stein Discrepancy for Goodness-of-fit Tests, *ICML*, 276-284.

[38] Lloyd J. R. and Ghahramani, Z. (2014). Statistical model criticism using kernel two sample tests. *Technical report.*

[39] Pelckmans, K. and Suykens, J. (2009). Gossip Algorithms for computing U-statistics. *Proceeding of the first IFAC workshop on estimation and control of networked systems.* 2009, Venice, Italy.

[40] Peng, H. and Tan, F. (2018). Jackknife empirical likelihood goodness-of-fit tests for U-statistics based general estimating equations, *Bernoulli*, 24(1): 449-464.

[41] Persson, T. (1979). A new way to obtain Watson's U^2, *Scand. J. Statist.*, (6) 119-122.

[42] Rosenbaum, P. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society B*, 67(4): 515-530.

[43] Serfling, R.J., (1980). Approximation Theorems of Mathematical Statistics, *Wiley*, New York

[44] Stephens, M. A. (1979). Vector correlation, *Biometrika* , (66) 41-48.

[45] Von Mises, R. (1947). "On the asymptotic distribution of differentiable statistical functions," *Ann. Math. Statist.*, (18) 309-348.

[46] Zaremba, W., Gretton A., and M. Blaschko (2013). B-test: A non-parametric, low variance kernel two-sample test. In Advances in Neural Information Processing Systems, *Curran Associates, Inc.*, (26) 755–763

[47] Zhong, P.S. and Chen, S. X. (2011). Test of high-dimensional regression coefficients with factorial designs, *J. Amer. Statist. Assoc.*, (106) 260-274.