SOME COMPUTATIONALLY EFFICIENT METHODS IN STATISTICS AND THEIR APPLICATIONS IN PARAMETER ESTIMATION AND HYPOTHESES TESTING

A Thesis
Presented to
The Academic Faculty

by

Cheng Huang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2017

SOME COMPUTATIONALLY EFFICIENT METHODS IN STATISTICS AND THEIR APPLICATIONS IN PARAMETER ESTIMATION AND HYPOTHESES TESTING

Approved by:

Dr. Xiaoming Huo, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Jeff Wu
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yajun Mei
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yao Xie
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Vladimir Kolthinskii
School of Mathematics
Georgia Institute of Technology

Date Approved: May 5, 2017

To my parents,

thanks for your support and encoragement.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my advisor, Professor Xiaoming Huo, for his guidance and support on academic research, without which this dissertation would not have been possible. I am truly grateful for his mentoring on professional and personal development.

Also, I would like to thank Professor Jeff Wu, Professor Yajun Mei, Professor Yao Xie, and Professor Vladimir Kolthinskii, for sitting in my committee and their insightful suggestions and comments on this work.

Thank all my friends for the joy that they brought to me during the last five years.

Again, I would like to thank everyone who gave me any kind of help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

In this dissertation, we will consider three fundamental problems under the setting of high data volume: statistical inference with distributed data, testing of independence, and two-sample testing.

The first part of this dissertation focuses on distributed statistical inference, which has recently attracted enormous attention. Many existing work focuses on the averaging estimator, e.g., [93] together with many others. We propose a one-step approach to enhance a simple-averaging-based distributed estimator. We derive the corresponding asymptotic properties of the newly proposed estimator. We find that the proposed one-step estimator enjoys the same asymptotic properties as the centralized estimator. The proposed one-step approach merely requires one additional round of communication in relative to the averaging estimator; so the extra communication burden is insignificant. In finite sample cases, numerical examples show that the proposed estimator outperforms the simple averaging estimator with a large margin in terms of the mean squared errors. A potential application of the one-step approach is that one can use multiple machines to speed up large scale statistical inference with little compromise in the quality of estimators. The proposed method becomes more valuable when data can only be available at distributed machines with limited communication bandwidth.

The second part is a statistically and computationally efficient test of independence based on distance covariance and random projections. As we know, test of independence plays a fundamental role in many statistical techniques. Among the nonparametric approaches, the distance-based methods (such as the distance correlation based hypotheses testing for independence) have numerous advantages, comparing with many other alternatives. A known limitation of the distance-based method is that its computational complexity

can be high. In general, when the sample size is n, the order of computational complexity of a distance-based method, which typically requires computing of all pairwise distances, can be $O(n^2)$. Recent advances have discovered that in the *univariate* cases, a fast method with $O(n \log n)$ computational complexity and $O(n)$ memory requirement exists. In this part, we show the potential of random projection in converting the *multivariate* problems into multiple *univariate* ones. As an immediate consequence, we develop a novel test of independence method based on random projection and distance covariance. We name our method a Randomly Projected Distance Covariance (RPDC), which achieves nearly the same power as the state-of-the-art distance-based approach, works in the *multivariate* cases, and enjoys the $O(nK \log n)$ computational complexity and $O(\max\{n, K\})$ memory requirement, where K is the number of random projections. The empirical results even suggest that fixed number of random projections suffice. The statistical theoretical analysis takes advantage of some techniques on random projections, which are rooted in contemporary machine learning. Numerical experiments demonstrate the efficiency of the proposed method, in relative to several competitors.

In the third part, we apply the technique of random projections on energy statistics to develop an efficient algorithm and derive a corresponding two-sample test. A common disadvantage in existing distribution-free two-sample testing approaches is that the computational complexity could be high. Specifically, if the sample size is N, the computational complexity of those two-sample tests is at least $O(N^2)$. In this part, we develop an efficient algorithm with complexity $O(N \log N)$ for computing energy statistics in *univariate* cases. For *multivariate* cases, we introduce a two-sample test based on energy statistics and random projections, which enjoys the $O(KN \log N)$ computational complexity, where K is the number of random projections. We name our method for *multivariate* cases as Randomly Projected Energy Statistics (RPES). We can show RPES achieves nearly the same test power with energy statistics both theoretically and empirically. Numerical experiments also demonstrate the efficiency of the proposed method over the competitors.

CHAPTER I

INTRODUCTION

Parameter estimation and hypotheses testing are fundamental problems in statistics. Many existing methods have been developed for the problems with moderate amount of data. Unfortunately, some of those methods could be computationally costly or even infeasible when the volume of data is high. This dissertation is an attempt to fulfill the needs for computationally efficient methods in statistics. Specifically, we focus on three main topics: the first one is distributed statistical estimation; the second one is a fast algorithm of distance covariance and corresponding test of independence; the third one is an efficient algorithm for energy statistics and its application in the two-sample test.

1.1 Distributed Statistical Inference

In many important contemporary applications, data are often partitioned across multiple servers. For example, a search engine company may have data coming from a large number of locations, and each location collects tera-bytes of data per day [20]. On a different setting, high volume of data (like videos) have to be stored distributively, instead of on a centralized server [55]. Given the modern "data deluge", it is often the case that centralized methods are no longer possible to implement. It has also been notified by various researchers (e.g., [35]) that the speed of local processors can be thousands time faster than the rate of data transmission in a modern network. Consequently it is evidently advantageous to develop communication-efficient method, instead of transmitting data to a central

location and then apply a global estimator.

In statistical inference, estimators are introduced to infer some important hidden quantities. In ultimate generality, a statistical estimator of a parameter $\theta \in \Theta$ is a measurable function of the data, taking values in the parameter space Θ. Many statistical inference problems could be solved by finding the maximum likelihood estimators (MLE), or more generally, M-estimators. In either case, the task is to maximize an objective function, which is the average of a criterion function over the entire data, which is typically denoted by $S = \{X_1, X_2, \ldots, X_N\}$, where N is called the sample size. Here we choose a capitalized N to distinguish from a lower n that will be used later. Traditional centralized setting requires access to entire data set S simultaneously. However, due to the explosion of data size, it may be infeasible to store all the data in a single machine like we did during past several decades. Distributed (sometimes, it is called *parallel*) statistical inference would be an indispensable approach for solving these large-scale problems.

At a high level, there are at least two types of distributed inference problems. In the first type, each sample X_i is completely observed at one location; at the same time, different samples (i.e., X_i and X_j for $i \neq j$) may be stored at different locations. We will focus to this type of problems. On the other hand, it is possible that for the same sample X_i, different parts are available at different locations, and they are *not* available in a centralized fashion. The latter has been studied in the literature (see [27] and references therein). We will not study the second type.

For distributed inference in the first type of the aforementioned setting, data are split into several subsets and each subset is assigned to a processor. This chapter will focus on

the M-estimator framework, in which an estimator is obtained by solving a distributed optimization problem. The objective in the distributed optimization problem may come from an M-estimator framework (or more particularly from the maximum likelihood principle), empirical risk minimization, and/or penalized version of the above. Due to the type 1 setting, we can see that the objective functions in the corresponding optimization problem are separable; in particular, the global objective function is a summation of functions such that each of them only depends on data reside on one machine. The exploration in this chapter will base on this fact. As mentioned earlier, a distributed inference algorithm should be communication-efficient because of high communication cost between different machines or privacy concerns (such as sensitive personal information or financial data). It is worth noting that even if the data could be handled by a single machine, distributed inference would still be beneficial for reducing computing time.

Our work has been inspired by recent progress in distributed optimization. We review some noticeable progress in numerical approaches and their associated theoretical analysis. Plenty of research work has been done in distributed algorithms for large scale optimization problems during recent years. [13] suggests to use Alternating Direction Method of Multipliers (ADMM) to solve distributed optimization problems in statistics and machine learning. Using a trick of *consistency* (or sometimes called *consensus*) constraints on local variables and a global variable, ADMM can be utilized to solve a distributed version of the Lasso problem [82, 18]. ADMM has also been adopted in solving distributed logistic regression problem, and many more. ADMM is feasible for a wide range of problems, but it requires iterative communication between local machines and the center. In comparison, we will propose a method that only requires two times iteration. [96] proposes a

parallelized stochastic gradient descent method for empirical risk minimization and proves its convergence. The established contractive mappings technique seems to be a powerful method to quantify the speed of convergence of the derived estimator to its limit. [71] presents the Distributed Approximate Newton-type Method (DANE) for distributed statistical optimization problems. Their method firstly averages the local gradients then follows by averaging all local estimators in each iteration until convergence. They prove that this method enjoys linear convergence rate for quadratic objectives. For non-quadratic objectives, it has been showed that the value of objective function has geometric convergence rate. [35] proposes a communication-efficient method for distributed optimization in machine learning, which uses local computation with randomized dual coordinate descent in a primal-dual setting. They also prove the geometric convergence rate of their method. The above works focused on the properties of numerical solutions to the corresponding optimization problems. Nearly all of them require more than two rounds of communication. Due to different emphasis, they did not study the statistical asymptotic properties (such as convergence in probability, asymptotic normality, Fisher information bound) of the resulting estimators.

Now we switch the gear to statistical inference. Distributed inference has been studied in many existing works, and various proposals have been made in different settings. To the best of our knowledge, the distributed one-step estimator has *not* been studied in any of these existing works. We review a couple of state-of-the-art approaches in the literature. Our method builds on a closely related recent line of work of [93], which presents a straight forward approach to solve large scale statistical optimization problem, where the

local empirical risk minimizers are simply averaged. They showed that this averaged estimator achieves mean squared error that decays as $O(N^{-1} + (N/k)^{-2})$, where N stands for the total number of samples and k stands for the total number of machines. They also showed that the mean squared error could be even reduced to $O(N^{-1} + (N/k)^{-3})$ with one more bootstrapping sub-sampling step. Obviously, there exists efficiency loss in their method since the centralized estimator could achieve means squared error $O(N^{-1})$. [47] proposes an inspiring two-step approach: firstly find local maximum likelihood estimators, then subsequently combine them by minimizing the total Kullback-Leibler divergence (KL-divergence). They proved the exactness of their estimator as the global MLE for the full exponential family. They also estimated the mean squared errors of the proposed estimator for a curved exponential family. Due to the adoption of the KL-divergence, the effectiveness of this approach heavily depends on the parametric form of the underlying model. [19] proposes a split-and-conquer approach for a penalized regression problem (in particular, a model with the canonical exponential distribution) and show that it enjoys the same oracle property as the method that uses the entire data set in a single machine. Their approach is based on a majority voting, followed by a weighted average of local estimators, which somewhat resembles a one-step estimator however is different. In addition, their theoretical results requires $k \leq O(N^{\frac{1}{5}})$, where k is the number of machines and N is the total number of samples; this is going to be different from our needed condition for theoretical guarantees. Their work considers a high-dimensional however sparse parameter vector, which is not considered in this chapter. [64] analyzes the error of averaging estimator in distributed statistical learning under two scenarios. The number of machines is fixed in the first one

and the number of machines grows in the same order with the number of samples per machine. They presented asymptotically exact expression for estimator error in both scenarios and showed that the error grows linearly with the number of machines in the latter case. Their work does not consider the one-step updating that will be studied in this chapter. Although it seems that their work proves the asymptotic optimality of the simple averaging, our simulations will demonstrate the additional one-step updating can improve over the simple averaging, at least in some interesting finite sample cases. [6] study the distributed parameter estimation method for penalized regression and establish the oracle asymptotic property of an averaging estimator. They also discussed hypotheses testing, which is not covered in this chapter. Precise upper bounds on the errors of their proposed estimator have been developed. We benefited from reading the technical proofs of their paper; however unlike our method, their method is restricted to linear regression problems with penalty and requires the number of machine $k = o(\sqrt{N})$. [41] devise a one-shot approach, which averages "debiased" lasso estimators, to distributed sparse regression in the high-dimensional setting. They show that their approach converges at the same order of rate as the Lasso when the data set is not split across too many machines.

It is worth noting that near all existing distributed estimator are *averaging* estimators. The idea of applying one additional updating, which correspondingly requires one additional round of communication, has not be explicitly proposed. We may notice some precursor of this strategy. For example, in [71], an approximate Newton direction was estimated at the central location, and then broadcasted to local machines. Another occurrence is that in [41], some intermediate quantities are estimated in a centralized fashion, and then distributed to local machines. None of them explicitly described what we will propose.

In the theory on maximum likelihood estimators (MLE) and M-estimators, there is a one-step method, which could make a consistent estimator as efficient as MLE or M-estimators with a single Newton-Raphson iteration. (Here, efficiency stands for the relative efficiency converges to 1.) See [84] for more details. There have been numerous papers utilizing this method. See [8], [25] and [97]. One-step estimator enjoys the same asymptotic properties as the MLE or M-estimators as long as the initial estimators are $\sqrt{n}$-consistent. A $\sqrt{n}$-consistent estimator is much easier to find than the MLE or an M-estimator. For instance, the simple averaging estimator (e.g., the one proposed by [93]) is good enough as a starting point for a one-step estimator.

In this dissertation, we propose a one-step estimator for distributed statistical inference. The proposed estimator is built on the well-analyzed simple averaging estimator. We show that the proposed one-step estimator enjoys the same asymptotic properties (including convergence and asymptotic normality) as the centralized estimator, which would utilize the entire data. Given the amount of knowledge we had on the distributed estimators, the above result may not be surprising. However, when we derive an upper bound for the error of the proposed one-step estimator, we found that we can achieve a slightly better one than those in the existing literature. We also perform a detailed evaluation of our one-step method, comparing with simple averaging method and centralized method using synthetic data. The numerical experiment is much more encouraging than the theory predicts: in nearly all cases, the one-step estimator outperformed the simple averaging one with a clear margin. We also observe that the one-step estimator achieves the comparable performance as the global estimator at a much faster rate than the simple averaging estimator. Our work may indicate that in practice, it is better to apply a one-step distributed estimator, than a

simple-average one. See [32] for a stand-alone paper on this topic.

1.2 Distance Covariance and Testing of Independence

Test of independence plays a fundamental role in many statistical techniques. Among the nonparametric approaches, the distance-based methods (such as the distance correlation based hypotheses testing for independence) have numerous advantages, comparing with many other alternatives. A known limitation of the distance-based method is that its computational complexity can be high. In general, when the sample size is n, the order of computational complexity of a distance-based method, which typically requires computing of all pairwise distances, can be $O(n^2)$. Recent advances have discovered that in the *univariate* cases, a fast method with $O(n \log n)$ computational complexity and $O(n)$ memory requirement exists. We will introduces a test of independence method based on random projection and distance correlation, which achieves nearly the same power as the state-of-the-art distance-based approach, works in the *multivariate* cases, and enjoys the $O(nK \log n)$ computational complexity and $O(\max\{n, K\})$ memory requirement, where K is the number of random projections. Note that saving is achieved when $K < n/\log n$. We name our method a Randomly Projected Distance Covariance (RPDC). The statistical theoretical analysis takes advantage of some techniques on random projection which are rooted in contemporary machine learning. Numerical experiments demonstrate the efficiency of the proposed method, in relative to several competitors.

Test of independence is a fundamental problem in statistics, with many existing work including the maximal information coefficient (MIC) [62], the copula based measures [68, 72], the kernel based criterion [29] and the distance correlation [80, 77], which motivated

our current work. Note that the above works as well as ours focus on the detection of the presence of the independence, which can be formulated as statistical hypotheses testing problems. On the other hand, interesting developments (e.g., [61]) aim at a more general framework for interpretable statistical dependence, which is not the goal of this dissertation.

Distance correlation proposed by [80] is an indispensable method in test of independence. The direct implementation of distance correlation takes $O(n^2)$ time, where n is the sample size. The time cost of distance correlation could be substantial when sample size is just a few thousands. When the random variables are univariate, there exist efficient numerical algorithms of time complexity $O(n \log n)$ [34]. However, for the multivariate random variables, we have not found any efficient algorithms in existing papers after an extensive literature survey.

Independence tests of multivariate random variables could have a wide range of applications. In many problem settings, as metioned in [81], each experimental unit will be measured multiple times, resulting in multivariate data. Researchers are often interested in exploring potential relationships among subsets of these measurements. For example, some measurements may represent attributes of physical characteristics while others represent attributes of psychological characteristics. It may be of interests to determine whether there exists a relationship between the physical and the psychological characteristics. A test of independence between pairs of vectors, where the vectors may have different dimensions and scales, becomes crucial. Moreover, the number of experimental units, or equivalently, sample size, could be massive, which requires the test to be computationally efficient. This work will meet the demands for numerically efficient independence tests of multivariate random variables.

The newly proposed test of independence between two (potentially multivariate) random variable X and Y works as follows. Firstly, both X and Y are randomly projected to one-dimensional spaces. Then the fast computing method for distance covariances between a pair of *univariate* random variables is adopted to compute for an surrogate distance covariance. The above two steps are repeated for numerous times. The final estimate of the distance covariance is the average of all aforementioned surrogate distance covariances.

For numerical efficiency, we will show (in Theorem 3.2.1) that the newly proposed algorithm enjoys the $O(Kn \log n)$ computational complexity and $O(\max\{n, K\})$ memory requirement, where K is the number of random projections and n is the sample size. On the statistical efficiency, we will show (in Theorem 3.3.18) that the asymptotic power of the test of independence by utilizing the newly proposed statistics is as efficient as its original multivariate counterpart, which achieves the stat-of-the-art rates.

Another contribution of this work is that we show potential of random projection in distance-based methods. Specifically, we can convert *multivariate* problems into *univariate* problems by projecting the data in some random directions. People have long conjectured that this random-projection approach may work, however it is not solved yet. Our work has the potential to significantly advance the frontier of this line of research. Moreover, in lemma 3.3.1, 3.3.2 and 3.3.3, we reveal the sufficiency and necessity of random projections for distance covariance. Lemma C.2.1, which is foundation of aforementioned three lemmas, even indicates that random projection should also work for other distance-based statistics.

1.3 Energy Statistics and Two-Sample Testing

Testing the equality of distributions is one of the most fundamental problems in statistics. Formally, let F and G denote two distribution function in $\mathbb{R}^p$. Given independent and identically distributed samples

$$\{X_1, \ldots, X_n\} \text{ and } \{Y_1, \ldots, Y_m\}$$

from two unknown distribution F and G, respectively, the two-sample testing problem is to test hypotheses

$$\mathcal{H}_0 : F = G \text{ v.s. } \mathcal{H}_1 : F \neq G.$$

There are a few recent advances in two sample testing that attract attentions in statistics and machine learning communities. [63] propose a test statistic based on the optimal non-bipartite matching, and, [9] develop a test based on shortest Hamiltonian path, both of which are distribution-free. [28] develop a kernel method based on maximum mean discrepancy. [75], [76] and [5] consider a test statistic based on pairwise distance within the same sample and across two different samples, which also motivates this work.

Computational complexity is a common limitation in the aforementioned methods. Let $N = n + m$ denote the size of the two-sample testing problem. The Cross Match (CM) test in [63] requires solving the non-bipartite matching problem, whose computational complexity is: (1) $O(N^3)$ with optimal solution, see [23]; (2) $O(N^2)$ with greedy heuristic. The two-sample test in [9] is based on shortest Hamilton path, which is an NP-complete problem, and its computational complexity is $O(N^2 \log N)$ with heuristic method based on Kruskal's algorithm ([39]). The Maximum Mean Discrepancy (MMD) proposed by [28] requires computing the kernel function values of all pairs of samples, whose complexity is

$O(N^2)$. Similarly, the energy statistics based methods in [75] and [5] typical requires the pairwise Euclidean distance, which also costs $O(N^2)$ complexity.

As a summary, the computational complexity of the aforementioned two-sample tests is at least $O(N^2)$, which leads to substantial computing time and prohibits their feasibility when the sample size N is too large. As a solution, we develop an efficient algorithm for computing the energy statistics in [75] with complexity $O(N \log N)$ for univariate random variables. For multivariate random variables, we propose an efficient algorithm of complexity $O(KN \log N)$ with the technique of random projection, where K is the number of random projections. The main idea of the multivariate algorithm is as follows: firstly, we project the data along some random direction; then, we use the univariate fast algorithm to compute the energy statistics with the univariate projected data; lastly, we repeat previous procedure for multiple times and take the average. As we will show in Theorem 4.3.12, the proposed test statistic based on random projections has nearly the same power with energy statistics.

The technique of random projection has been widely used in two-sample testing problems. [48] propose a new method, which firstly projects data along a few random direction; and then, applies the classical Hotelling T^2 statistic, for testing the equality of means in different samples. [74] develop a similar approach based on random projection and the Hotelling T^2 statistic, but the random projection is taken with respect to sample mean vectors and sample covariance matrices. These two papers focus on the problem under multivariate Gaussian settings while our work is more general and does not impose any assumptions in the distributions.

CHAPTER II

DISTRIBUTED STATISTICAL INFERENCE

This chapter is organized as follows. Section 2.1 describes details of our problem setting and two methods—the simple averaging method and the proposed one-step method. In Section 2.2, we study the asymptotic properties of the one-step estimator in the M-estimator framework and analyze the upper bound of its estimation error. Section 2.3 provides some numerical examples of distributed statistical inference with synthetic data. We conclude in Section 2.4. When appropriate, detailed proofs are relegated to the appendix.

2.1 Problem Formulation

2.1.1 Notations

In this subsection, we will introduce some notations that will be used in this chapter. Let $\{m(x;\theta) : \theta \in \Theta \subset \mathbb{R}^d\}$ denote a collection of criterion functions, which should have continuous second derivative. Consider a data set S consisting of $N = nk$ samples, which are drawn i.i.d. from $p(x)$ (for simplicity, we assume that the sample size N is a multiple of k). This data set is divided evenly at random and stored in k machines. Let S_i denote the subset of data assigned to machine i, $i = 1, \ldots, k$, which is a collection of n samples drawn i.i.d. from $p(x)$. Note that any two subsets in those S_i's are not overlapping.

For each $i \in \{1, \ldots, k\}$, let the local empirical criterion function that is based on the local

data set on machine i and the corresponding maximizer be denoted by

$$M_i(\theta) = \frac{1}{|S_i|} \sum_{x \in S_i} m(x;\theta) \text{ and } \theta_i = \arg\max_{\theta \in \Theta} M_i(\theta). \tag{2.1.1}$$

Let the global empirical criterion function be denoted by

$$M(\theta) = \frac{1}{k} \sum_{i=1}^{k} M_i(\theta). \tag{2.1.2}$$

And let the population criterion function and its maximizer be denoted by

$$M_0(\theta) = \int_{\mathcal{X}} m(x;\theta) p(x) dx \text{ and } \theta_0 = \arg\max_{\theta \in \Theta} M_0(\theta), \tag{2.1.3}$$

where $\mathcal{X}$ is the sample space. Note that θ_0 is the parameter of interest. The gradient and Hessian matrix of $m(x;\theta)$ with respect to θ are denoted by

$$\dot{m}(x;\theta) = \frac{\partial m(x;\theta)}{\partial \theta}, \ddot{m}(x;\theta) = \frac{\partial^2 m(x;\theta)}{\partial \theta \, \partial \theta^T}. \tag{2.1.4}$$

We also let the gradient and Hessian of local empirical criterion function be denoted by

$$\dot{M}_i(\theta) = \frac{\partial M_i(\theta)}{\partial \theta} = \frac{1}{|S_i|} \sum_{x \in S_i} \frac{\partial m(x;\theta)}{\partial \theta}, \ddot{M}_i(\theta) = \frac{\partial^2 M_i(x;\theta)}{\partial \theta \, \partial \theta^T} = \frac{1}{|S_i|} \sum_{x \in S_i} \frac{\partial^2 m(x;\theta)}{\partial \theta \, \partial \theta^T}, \tag{2.1.5}$$

where $i \in \{1, 2, \ldots, k\}$, and let the gradient and Hessian of global empirical criterion function be denoted by

$$\dot{M}(\theta) = \frac{\partial M(\theta)}{\partial \theta}, \ddot{M}(\theta) = \frac{\partial^2 M(\theta)}{\partial \theta \, \partial \theta^T}. \tag{2.1.6}$$

Similarly, let the gradient and Hessian of population criterion function be denoted by

$$\dot{M}_0(\theta) = \frac{\partial M_0(\theta)}{\partial \theta}, \ddot{M}_0(\theta) = \frac{\partial^2 M_0(\theta)}{\partial \theta \, \partial \theta^T}. \tag{2.1.7}$$

The vector norm $\|\cdot\|$ for $a \in \mathbb{R}^d$ that we use in this chapter is the usual Euclidean norm $\|a\| = (\sum_{j=1}^{d} a_j^2)^{\frac{1}{2}}$. And we also use $|||\cdot|||$ to denote a norm for matrix $A \in \mathbb{R}^{d\times d}$, which is defined as its maximal singular value, i.e., we have

$$|||A||| = \sup_{u:u\in R^d, \|u\|\leq 1} \|Au\|.$$

The aforementioned matrix norm will be the major matrix norm that is used throughout the chapter. The only exception is that we will also use Frobenius norm in Appendix B.1. And the Euclidean norm is the only vector norm that we use throughout this chapter.

2.1.2 Review on M-estimators

In this chapter, we will study the distributed scheme for large-scale statistical inference. To make our conclusions more general, we consider M-estimators, which could be regarded as a generalization of the Maximum Likelihood Estimators (MLE). The M-estimator $\hat{\theta}$ could be obtained by maximizing empirical criterion function, which means

$$\hat{\theta} = \arg\max_{\theta\in\Theta} M(\theta) = \arg\max_{\theta\in\Theta} \frac{1}{|S|} \sum_{x\in S} m(x;\theta).$$

Note that, when the criterion function is the log likelihood function, i.e., $m(x;\theta) = \log f(x;\theta)$, the M-estimator is exactly the MLE. Let us recall that $M_0(\theta) = \int_{\mathcal{X}} m(x;\theta)p(x)dx$ is the population criterion function and $\theta_0 = \arg\max_{\theta\in\Theta} M_0(\theta)$ is the maximizer of population criterion function. It is known that $\hat{\theta}$ is a consistent estimator for θ_0, i.e., $\hat{\theta} - \theta_0 \xrightarrow{P} 0$. See Chapter 5 of [84].

2.1.3 Simple Averaging Estimator

Let us recall that $M_i(\theta)$ is the local empirical criterion function on machine i,

$$M_i(\theta) = \frac{1}{|S_i|} \sum_{x\in S_i} m(x;\theta).$$

And, θ_i is the local M-estimator on machine i,

$$\theta_i = \arg\max_{\theta \in \Theta} M_i(\theta).$$

Then as mentioned in [93], the simplest and most intuitive method is to take average of all local M-estimators. Let $\theta^{(0)}$ denote the average of these local M-estimators, we have

$$\theta^{(0)} = \frac{1}{k} \sum_{i=1}^{k} \theta_i, \tag{2.1.8}$$

which is referred as the simple averaging estimator in the rest of this chapter.

2.1.4 One-step Estimator

Under the problem setting above, starting from the simple averaging estimator $\theta^{(0)}$, we can obtain the one-step estimator $\theta^{(1)}$ by performing a single Newton-Raphson update, i.e.,

$$\theta^{(1)} = \theta^{(0)} - [\ddot{M}(\theta^{(0)})]^{-1}[\dot{M}(\theta^{(0)})], \tag{2.1.9}$$

where $M(\theta) = \frac{1}{k}\sum_{i=1}^{k} M_i(\theta)$ is the global empirical criterion function, $\dot{M}(\theta)$ and $\ddot{M}(\theta)$ are the gradient and Hessian of $M(\theta)$, respectively. The whole process to compute one-step estimator can be summarized as follows.

(1) For each $i \in \{1, 2, \ldots, k\}$, machine i compute the local M-estimator with its local data set,

$$\theta_i = \arg\max_{\theta \in \Theta} M_i(\theta) = \arg\max_{\theta \in \Theta} \frac{1}{|S_i|} \sum_{x \in S_i} m(x; \theta).$$

(2) All local M-estimators are averaged to obtain simple averaging estimator,

$$\theta^{(0)} = \frac{1}{k} \sum_{i=1}^{k} \theta_i \,.$$

Then $\theta^{(0)}$ is sent back to each local machine.

(3) For each $i \in \{1, 2, \ldots, k\}$, machine i compute the gradient and Hessian matrix of its local empirical criterion function $M_i(\theta)$ at $\theta = \theta^{(0)}$. Then send $\dot{M}_i(\theta^{(0)})$ and $\ddot{M}_i(\theta^{(0)})$ to the central machine.

(4) Upon receiving all gradients and Hessian matrices, the central machine computes gradient and Hessian matrix of $M(\theta)$ by averaging all local information,

$$\dot{M}(\theta^{(0)}) = \frac{1}{k}\sum_{i=1}^{k} \dot{M}_i(\theta^{(0)}),\ \ddot{M}(\theta^{(0)}) = \frac{1}{k}\sum_{i=1}^{k} \ddot{M}_i(\theta^{(0)}).$$

Then the central machine would perform a Newton-Raphson iteration to obtain a one-step estimator,

$$\theta^{(1)} = \theta^{(0)} - [\ddot{M}(\theta^{(0)})]^{-1}[\dot{M}(\theta^{(0)})].$$

Note that $\theta^{(1)}$ is not necessarily the maximizer of empirical criterion function $M(\theta)$ but it shares the same asymptotic properties with the corresponding global maximizer (M-estimator) under some mild conditions, i.e., we will show

$$\theta^{(1)} \xrightarrow{P} \theta_0, \sqrt{N}(\theta^{(1)} - \theta_0) \xrightarrow{d} \mathbf{N}(0, \Sigma),\ \text{as } N \to \infty,$$

where the covariance matrix Σ will be specified later.

The one-step estimator has advantage over simple averaging estimator in terms of estimation error. In [93], it is showed both theoretically and empirically that the MSE of simple averaging estimator $\theta^{(0)}$ grows significantly with the number of machines k when the total number of samples N is fixed. More precisely, there exists some constant $C_1, C_2 > 0$ such that

$$\mathbb{E}[\|\theta^{(0)} - \theta_0\|^2] \leq \frac{C_1}{N} + \frac{C_2 k^2}{N^2} + O(kN^{-2}) + O(k^3 N^{-3}).$$

Fortunately, one-step method $\theta^{(1)}$ could achieve a lower upper bound of MSE with only one additional step. we will show the following in Section 2.2:

$$\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2] \leq \frac{C_1}{N} + O(N^{-2}) + O(k^4 N^{-4}).$$

2.2 Main Results of One-Step Estimator

At first, some assumptions will be introduced in Section 3.2.1. After that, we will study the asymptotic properties of one-step estimator in Section 3.2.2, i.e., convergence, asymptotic normality and mean squared error (MSE). In Section 3.2.3, we will consider the one-step estimator under the presence of information loss.

2.2.1 Assumptions

Throughout this chapter, we impose some regularity conditions on the criterion function $m(x;\theta)$, the local empirical criterion function $M_i(\theta)$ and population criterion function $M_0(\theta)$. We use the similar assumptions in [93]. Those conditions are also standard in classical statistical analysis of M-estimators (cf. [84]).

First assumption restricts the parameter space to be compact, which is reasonable and not rigid in practice. One reason is that the possible parameters lie in a finite scope for most cases. Another justification is that the largest number that computers could cope with is always limited.

Assumption 2.2.1 (parameter space)**.** *The parameter space $\Theta \in \mathbb{R}^d$ is a compact convex set. And let $D \triangleq \max_{\theta,\theta' \in \Theta} \|\theta - \theta'\|$ denote the diameter of Θ.*

We also assume that $m(x;\theta)$ is concave with respect to θ and $M_0(\theta)$ has some curvature around the unique optimal point θ_0, which is a standard assumption for any method requires

consistency.

Assumption 2.2.2 (invertibility)**.** *The Hessian of population criterion function $M_0(\theta)$ at θ_0 is a nonsingular matrix, which means $\ddot{M}_0(\theta_0)$ is negative definite and there exists some $\lambda > 0$ such that $\sup_{u\in\mathbb{R}^d:\|u\|<1} u^t \ddot{M}_0(\theta_0)u \leq -\lambda$.*

In addition, we require the criterion function $m(x;\theta)$ to be smooth enough, at least in the neighborhood of the optimal point θ_0, $B_\delta = \{\theta \in \Theta : \|\theta - \theta_0\| \leq \delta\}$. So, we impose some regularity conditions on the first and second derivative of $m(x;\theta)$. We assume the gradient of $m(x;\theta)$ is bounded in moment and the difference between $\ddot{m}(x;\theta)$ and $\ddot{M}_0(\theta)$ is also bounded in moment. Moreover, we assume that $\ddot{m}(x;\theta)$ has Lipschitz continuity in B_δ.

Assumption 2.2.3 (smoothness)**.** *There exist some constants G and H such that*

$$\mathbb{E}[\|\dot{m}(X;\theta)\|^8] \leq G^8 \ \text{ and } \ \mathbb{E}\left[\left|\left|\left|\ddot{m}(X;\theta) - \ddot{M}_0(\theta)\right|\right|\right|^8\right] \leq H^8, \forall\theta \in B_\delta.$$

For any $x \in \mathcal{X}$, the Hessian matrix $\ddot{m}(x;\theta)$ is $L(x)$-Lipschitz continuous,

$$|||\ddot{m}(x;\theta) - \ddot{m}(x;\theta')||| \leq L(x)\|\theta - \theta'\|, \forall\theta, \theta' \in B_\delta,$$

where $L(x)$ satisfies

$$\mathbb{E}[L(X)^8] \leq L^8 \ \text{ and } \ \mathbb{E}[(L(X) - \mathbb{E}[L(X)])^8] \leq L^8,$$

for some finite constant $L > 0$.

By Theorem 8.1 in Chapter XIII of [40], $m(x;\theta)$ enjoys interchangeability between differentiation on θ and integration on x, which means the following two equations hold:

$$\dot{M}_0(\theta) = \frac{\partial}{\partial\theta}\int_{\mathcal{X}} m(x;\theta)p(x)dx = \int_{\mathcal{X}} \frac{\partial m(x;\theta)}{\partial\theta}p(x)dx = \int_{\mathcal{X}} \dot{m}(x;\theta)p(x)dx,$$

and,

$$\ddot{M}_0(\theta) = \frac{\partial^2}{\partial\theta^t\partial\theta}\int_{\mathcal{X}} m(x;\theta)p(x)dx = \int_{\mathcal{X}} \frac{\partial^2 m(x;\theta)}{\partial\theta^t\partial\theta}p(x)dx = \int_{\mathcal{X}} \ddot{m}(x;\theta)p(x)dx.$$

2.2.2 Asymptotic Properties and Mean Squared Error (MSE) Bound

Our main result is that one-step estimator enjoys oracle asymptotic properties and has mean squared error of $O(N^{-1})$ under some mild conditions.

Theorem 2.2.4. *Let $\Sigma = \ddot{M}_0(\theta_0)^{-1}\mathbb{E}[\dot{m}(x;\theta_0)\dot{m}(x;\theta_0)^t]\ddot{M}_0(\theta_0)^{-1}$, where the expectation is taken with respect to $p(x)$. Under Assumption 2.2.1, 2.2.2, and 2.2.3, when the number of machines k satisfies $k = O(\sqrt{N})$, $\theta^{(1)}$ is consistent and asymptotically normal, i.e., we have*

$$\theta^{(1)} - \theta_0 \xrightarrow{P} 0 \ \text{ and } \ \sqrt{N}(\theta^{(1)} - \theta_0) \xrightarrow{d} \mathbf{N}(0,\Sigma) \text{ as } N \to \infty.$$

See Appendix B.3 for a proof. The above theorem indicates that the one-step estimator is asymptotically equivalent to the centralized M-estimator.

Remark. *It is worth noting that the condition $\|\sqrt{N}(\theta^{(0)} - \theta_0)\| = O_P(1)$ suffices for our proof to Theorem 2.2.4. Let $\tilde{\theta}^{(0)}$ denote another starting point for the one-step update, then the following estimator*

$$\tilde{\theta}^{(1)} = \tilde{\theta}^{(0)} - \ddot{M}(\tilde{\theta}^{(0)})^{-1}\dot{M}(\tilde{\theta}^{(0)})$$

also enjoys the same asymptotic properties with $\theta^{(1)}$ (and the centralized M-estimator $\hat{\theta}$) as long as $\sqrt{N}(\tilde{\theta}^{(0)} - \theta_0)$ is bounded in probability. Therefore, we can replace $\theta^{(0)}$ with any estimator $\tilde{\theta}^{(0)}$ that satisfies

$$\|\sqrt{N}(\tilde{\theta}^{(0)} - \theta_0)\| = O_P(1).$$

Theorem 2.2.5. *Under Assumption 2.2.1, 2.2.2, and 2.2.3, the mean squared error of the one-step estimator $\theta^{(1)}$ is bounded by*

$$\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2] \leq \frac{2Tr[\Sigma]}{N} + O(N^{-2}) + O(k^4 N^{-4}).$$

When the number of machines k satisfies $k = O(\sqrt{N})$, we have

$$\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2] \leq \frac{2Tr[\Sigma]}{N} + O(N^{-2}).$$

See Appendix B.4 for a proof.

In particular, when we choose the criterion function to be the log likelihood function, $m(x;\theta) = \log f(x;\theta)$, the one-step estimator has the same asymptotic properties with the maximum likelihood estimator (MLE), which is described below.

Corollary 2.2.6. *If $m(x;\theta) = \log f(x;\theta)$ and $k = O(\sqrt{N})$, one-step estimator $\theta^{(1)}$ is a consistent and asymptotic efficient estimator of θ_0,*

$$\theta^{(1)} - \theta_0 \xrightarrow{P} 0 \text{ and } \sqrt{N}(\theta^{(1)} - \theta_0) \xrightarrow{d} \mathbf{N}(0, I(\theta_0)^{-1}), \text{ as } N \to \infty,$$

where $I(\theta_0)$ is the Fisher's information at $\theta = \theta_0$. And the mean squared error of $\theta^{(1)}$ is bounded as follows:

$$\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2] \leq \frac{2Tr[I^{-1}(\theta_0)]}{N} + O(N^{-2}) + O(k^4 N^{-4}).$$

Proof. It follows immediately from Theorem 2.2.4, 2.2.5 and the definition of the Fisher's information. □

2.2.3 Under the Presence of Communication Failure

In practice, it is possible that the information (local estimator, local gradient and local Hessian) from a local machine *cannot* be received by the central machine due to various causes (for instance, network problem or hardware crash). We assume that the communication failure on each local machine occurs independently.

We now derive a distributed estimator under the scenario with possible information loss. We will also present the corresponding theoretical results. We use $a_i \in \{0, 1\}, i = 1, \ldots, k$, to denote the status of local machines: when machine i successfully sends all its local information to central machine, we have $a_i = 1$; when machine i fails, we have $a_i = 0$. The corresponding simple averaging estimator is computed as

$$\theta^{(0)} = \frac{\sum_{i=1}^{k} a_i \theta_i}{\sum_{i=1}^{k} a_i}.$$

And one-step estimator is as follows

$$\theta^{(1)} = \theta^{(0)} - \left[\sum_{i=1}^{k} a_i \ddot{M}_i(\theta^{(0)})\right]^{-1} \left[\sum_{i=1}^{k} a_i \dot{M}_i(\theta^{(0)})\right].$$

Corollary 2.2.7. *Suppose r is the probability (or rate) that a local machine fails to send its information to the central machine. When $n = N/k \to \infty$, $k \to \infty$ and $k = O(\sqrt{N})$, the one-step estimator is asymptotically normal:*

$$\sqrt{(1-r)N}(\theta^{(1)} - \theta_0) \xrightarrow{d} \mathbf{N}(0, \Sigma).$$

And more precisely, unless all machines fail, we have

$$\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2] \leq \frac{2\mathit{Tr}[\Sigma]}{N(1-r)} + \frac{6\mathit{Tr}[\Sigma]}{Nk(1-r)^2} + O(N^{-2}(1-r)^{-2}) + O(k^2 N^{-2}).$$

See Appendix B.5 for a proof. Note that the probability that all machines fail is r^k, which is negligible when r is small and k is large.

2.3 Numerical Examples for One-Step Estimators

In this section, we will discuss the results of simulation studies comparing the performance of the simple averaging estimator $\theta^{(0)}$ and the one-step estimator $\theta^{(1)}$, as well as the centralized M-estimator $\hat{\theta}$, which maximizes the global empirical criterion function $M(\theta)$ when the entire data are available centrally. Besides, we will also study the resampled averaging estimator, which is proposed by [93]. The main idea of a resampled averaging estimator is to resample $\lfloor sn \rfloor$ observations from each local machine to obtain another averaging estimator $\theta_1^{(0)}$. Then the resampled averaging estimator can be constructed as follows:

$$\theta_{re}^{(0)} = \frac{\theta^{(0)} - s\theta_1^{(0)}}{1-s}.$$

In our numerical examples, the resampling ratio s is chosen to be $s = 0.1$ based on past empirical studies. We shall implement these estimators for logistic regression, Beta distribution and Gaussian Distribution. We will also study the parameter estimation for Beta distribution with occurrence of communication failures, in which some local machines could fail to send their local information to the central machine.

2.3.1 Logistic Regression

In this example, we simulate the data from the following logistic regression model:

$$y \sim \text{Bernoulli}(p), \text{ where } p = \frac{\exp(x^t\theta)}{1+\exp(x^t\theta)} = \frac{\exp(\sum_{j=1}^d x_j\theta_j)}{1+\exp(\sum_{j=1}^d x_j\theta_j)}. \tag{2.3.10}$$

In this model, $y \in \{0,1\}$ is a binary response, $x \in R^d$ is a continuous predictor and $\theta \in R^d$ is the parameter of interest.

In each single experiment, we choose a fixed vector θ with each entry $\theta_j, j = 1, \ldots, d$, drawn from Unif$(-1,1)$ independently. Entry $x_j, j = 1, \ldots, d$ of $x \in R^d$ is sampled

from Unif$(-1, 1)$, independent from parameters θ_j's and other entries. After generating parameter θ and predictor x, we can compute the value of probability p and generate y according to (2.3.10). We fix the number of observed samples $N = 2^{17} = 131,072$ in each experiment, but vary the number of machines k. The target is to estimate θ with different number of parallel splits k of the data. The experiment is repeated for $K = 50$ times to obtain reliable average error. And the criterion function is the log-likelihood function,

$$m(x, y; \theta) = yx^t\theta - \log(1 + \exp(x^t\theta)).$$

The goal of each experiment is to estimate parameter θ_0 maximizing population criterion function

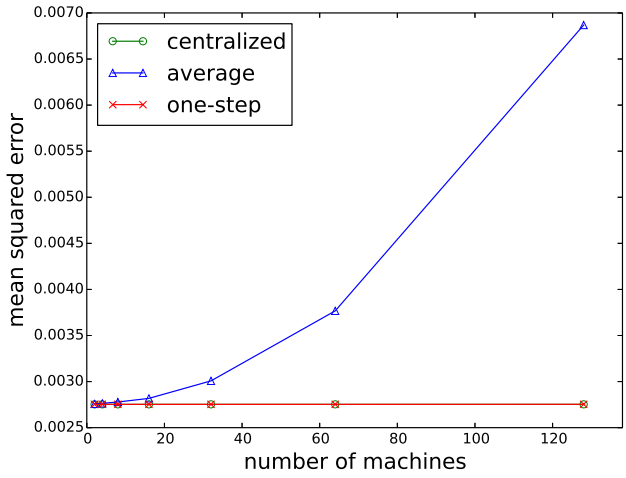
$$M_0(\theta) = \mathbb{E}_{x,y}[m(x, y; \theta)] = \mathbb{E}_{x,y}[yx^t\theta - \log(1 + \exp(x^t\theta))].$$

In this particular case, θ_0 is exactly the same with the true parameter.

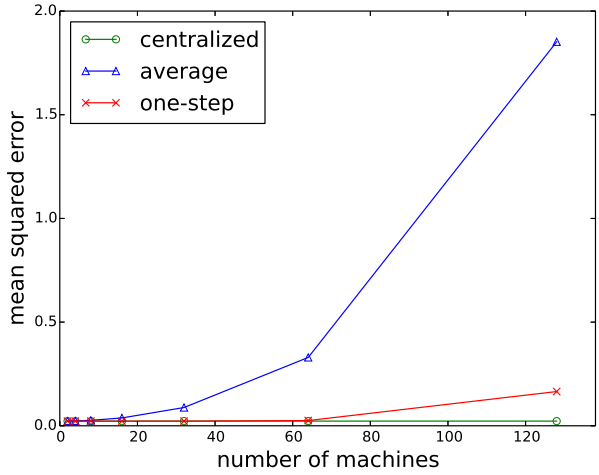
In each experiment, we split the data into $k = 2, 4, 8, 16, 32, 64, 128$ non-overlapping subsets of size $n = N/k$. We compute a local estimator θ_i from each subset. And simple averaging estimator is obtained by taking all local estimators, $\theta^{(0)} = \frac{1}{k}\sum_{i=1}^{k}\theta_i$. Then the one-step estimator $\theta^{(1)}$ could be computed by applying a Newton-Raphson update to $\theta^{(0)}$, i.e., equation (2.1.9).

The dimension is chosen to be $d = 20$ and $d = 100$, which could help us understand the performance of those estimators in both low and high dimensional cases. In Fig. 1, we plot the mean squared error of each estimator versus the number of machines k. As we expect, the mean squared error of simple averaging estimator grows rapidly with the number of machines. But, the mean squared error of one-step estimator remains the same with the mean squared error of oracle estimator when the number of machines k is not very

large. Even when the $k = 128$ and the dimension of predictors $d = 100$, the performance of one-step estimator is significantly better than simple averaging estimator. As we can easily find out from Fig. 1, the mean squared error of simple averaging estimator is about 10 times of that of one-step estimator when $k = 128$ and $d = 100$. Detailed values of



(a) $d = 20$



(b) $d = 100$

Figure 1: Logistic Regression: The mean squared error $\|\hat{\theta} - \theta_0\|^2$ versus number of machines, with fifty simulations. The "average" is $\theta^{(0)}$ and the "one-step" is $\theta^{(1)}$. The "centralized" denotes the oracle estimator with entire data.

mean squared error are listed in Table 1 and 2. From the tables, we can easily figure out

that the standard deviation of the error of one-step estimator is significantly smaller than that of simple averaging, especially when the number of machines k is large, which means one-step estimator is more stable.

Table 1: Logistic Regression ($d = 20$): Detailed values of squared error $\|\hat{\theta} - \theta_0\|^2$. In each cell, the first number is the mean of squared error in $K = 50$ experiments and the number in the brackets is the standard deviation of the squared error.

number of machines	2	4	8	16	32	64	128
simple avg ($\times 10^{-4}$)	28.036 (7.982)	28.066 (7.989)	28.247 (8.145)	28.865 (8.443)	30.587 (9.812)	38.478 (14.247)	69.898 (27.655)
one-step ($\times 10^{-4}$)	28.038 (7.996)	28.038 (7.996)	28.038 (7.996)	28.038 (7.996)	28.038 (7.996)	28.035 (7.998)	28.039 (8.017)
centralized ($\times 10^{-4}$)	28.038 (7.996)						

Table 2: Logistic Regression ($d = 100$): Detailed values of squared error $\|\hat{\theta} - \theta_0\|^2$. In each cell, the first number is the mean of squared error in $K = 50$ experiments and the number in the brackets is the standard deviation of squared error.

number of machines	2	4	8	16	32	64	128
simple avg ($\times 10^{-3}$)	23.066 (4.299)	23.818 (4.789)	26.907 (6.461)	38.484 (10.692)	87.896 (22.782)	322.274 (67.489)	1796.147 (324.274)
one-step ($\times 10^{-3}$)	22.787 (4.062)	22.784 (4.060)	22.772 (4.048)	22.725 (3.998)	22.612 (3.835)	24.589 (4.651)	151.440 (43.745)
centralized ($\times 10^{-3}$)	22.787 (4.063)						

2.3.2 Beta Distribution

In this example, we use data simulated from Beta distribution Beta(α, β), whose p.d.f. is as follows:

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

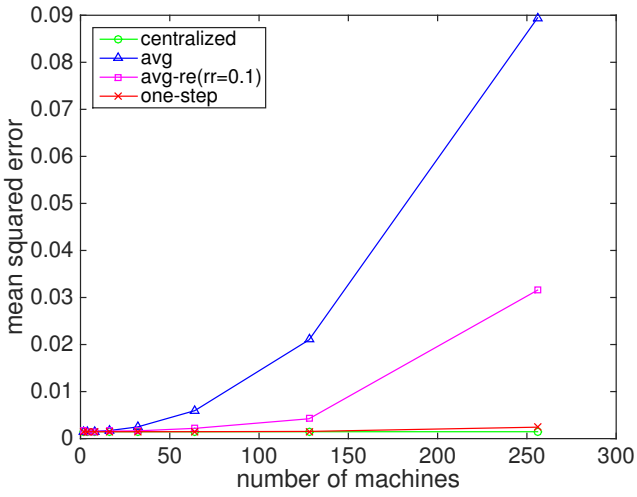
In each experiment, we generate the value of parameter as $\alpha \sim$ Unif$(1, 3)$ and $\beta \sim$ Unif$(1, 3)$, independently. Once (α, β) is determined, we can simulate samples from the

above density. In order to examine the performance of two distributed methods when k is extremely large, we choose to use a data set with relatively small size $N = 2^{13} = 8192$ and let number of machines vary in a larger range $k = 2, 4, 8, \ldots, 256$. And the objective is to estimate parameter (α, β) from the observed data. The experiment is again repeated for $K = 50$ times. The criterion function is $m(x; \theta) = \log f(x; \alpha, \beta)$, which implies that the centralized estimator is the MLE.

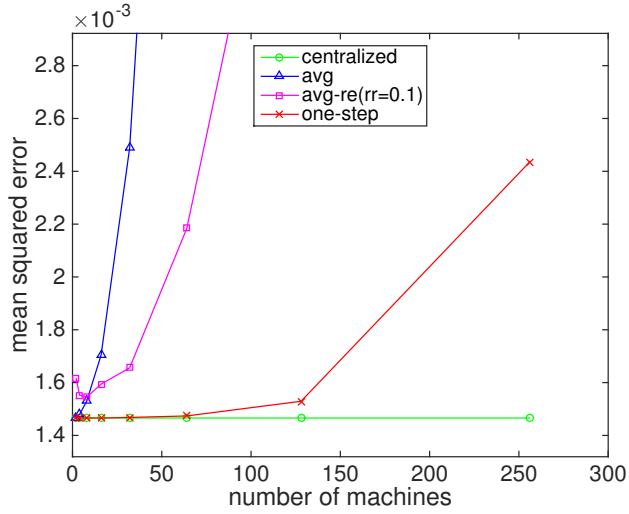
Figure 2 and Table 3 show that the one-step estimator has almost the same performance with centralized estimator in terms of MSE and standard deviation when the number of machines $k \leq \sqrt{N}$ (i.e., when $k \leq 64$). However, the one-step estimator performs worse than centralized estimator when $k > \sqrt{N}$ (i.e., when $k = 128$ or 256), which confirms the necessity of condition $k = O(\sqrt{N})$ in Theorem 2.2.4. In addition, we can easily find out that both simple averaging estimator and resampled averaging estimator are worse than the proposed one-step estimator regardless of the value of k.

Table 3: Beta Distribution: Detailed values of squared error $\|\hat{\theta} - \theta_0\|^2$. In each cell, the first number is the mean squared error with $K = 50$ experiments and the number in the brackets is the standard deviation of the squared error.

number of machines	simple avg ($\times 10^{-3}$)	resampled avg ($\times 10^{-3}$)	one-step ($\times 10^{-3}$)	centralized ($\times 10^{-3}$)
2	1.466 (1.936)	1.616 (2.150)	1.466 (1.943)	1.466 (1.943)
4	1.480 (1.907)	1.552 (2.272)	1.466 (1.943)	
8	1.530 (1.861)	1.545 (2.177)	1.466 (1.943)	
16	1.704 (1.876)	1.594 (2.239)	1.466 (1.946)	
32	2.488 (2.628)	1.656 (2.411)	1.468 (1.953)	
64	5.948 (5.019)	2.184 (3.529)	1.474 (1.994)	
128	21.002 (11.899)	4.221 (7.198)	1.529 (2.199)	
256	89.450 (35.928)	31.574 (36.518)	2.435 (3.384)	



(a) overview



(b) detailed view

Figure 2: Beta Distribution: The error $\|\theta - \theta_0\|^2$ versus the number of machines, with fifty simulations, where θ_0 is the true parameter. The "avg" is $\theta^{(0)}$, the "avg-re" is $\theta_{re}^{(0)}$ with resampling ratio $rr = 10\%$ and the "one-step" is $\theta^{(1)}$. The "centralized" denotes maximum likelihood estimator with the entire data.

2.3.3 Beta Distribution with Possibility of Losing Information

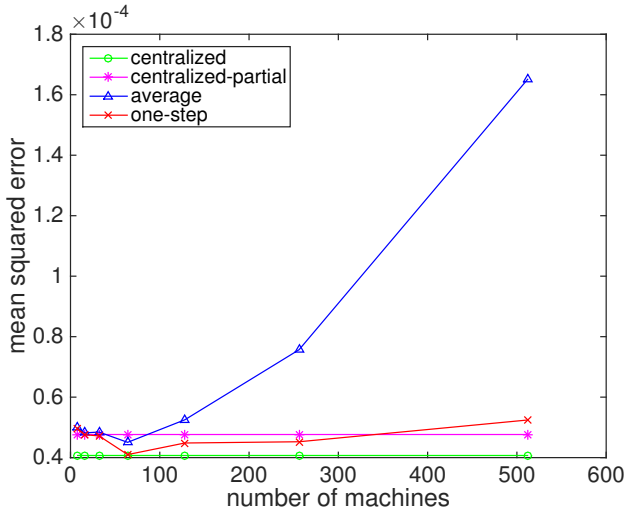
Now, we would like to compare the performance of simple averaging estimator and one-step estimator under a more practical scenario, in which each single local machine could

fail to send its information to central machine. We assume those failures would occur independently with probability $r = 0.05$. The simulation settings are similar to previous example in Section 4.2, however, we will generate $N = 409600$ samples from Beta distribution Beta(α, β), where α and β are chosen from Unif$(1, 3)$, independently. And the goal of experiment is to estimate parameter (α, β). In each experiment, we let the number of machines vary $k = 8, 16, 32, 64, 128, 256, 512$. We also compare the performance of the centralized estimator with entire data and centralized estimator with $(1-r) \times 100\% = 95\%$ of entire data. This experiment is repeated for $K = 50$ times.

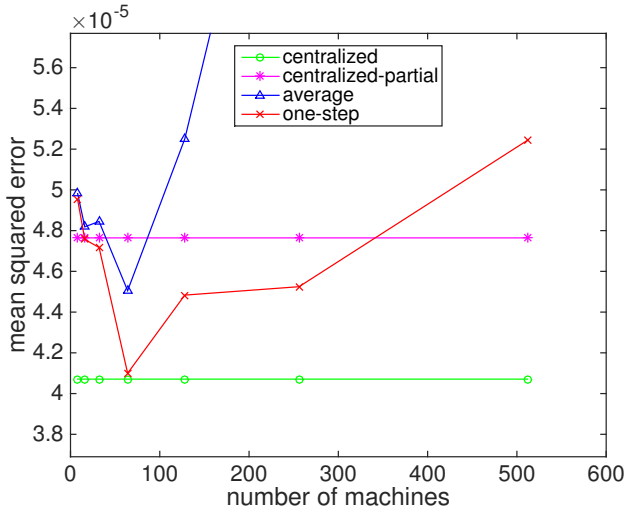
In Figure 3(a), we plot the MSE of each estimator against the number of machines. As expected, the MSE of simple averaging estimator grows significantly with the number of machines while the other three remains nearly the same. We can easily find out that performance of simple averaging estimator is far worse than others, especially when the number of machines is large (for instance, when $k = 256$ or 512). If we take a closer look at the other three estimators from Fig. 3(b), we will find that the performance of one-step estimator is volatile but always remains in a reasonable range. And as expected, the error of one-step estimator converges to the error of oracle estimator with partial data when number of machines k is large.

2.3.4 Gaussian Distribution with Unknown Mean and Variance

In this part, we will compare the performance of the simple averaging estimator, the resampled averaging estimator and the one-step estimator when fixing the number of machines $k = \sqrt{N}$ and letting the value of N increase. We draw N samples from $N(\mu, \sigma^2)$, where $\mu \sim$ Unif$(-2, 2)$ and $\sigma^2 \sim$ Unif$(0.25, 9)$, independently. We let N vary in $\{4^3, \ldots, 4^9\}$



(a) overview



(b) detailed view

Figure 3: Beta Distribution with Possibility of Losing Information: The error $\|\theta - \theta_0\|^2$ versus the number of machines, with fifty simulations, where θ_0 is the true parameter. The "average" is $\theta^{(0)}$ and the "one-step" is $\theta^{(1)}$. The "centralized" denotes maximum likelihood estimator with the entire data. And the "centralized-partial" denotes the maximum likelihood estimator with $(1-r) \times 100\% = 95\%$ of data.

and repeat the experiment for $K = 50$ times for each N. We choose the criterion function to be the log-likelihood function

$$m(x; \mu, \sigma^2) = -\frac{(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2.$$

Figure 4 and Table 5 show that one-step estimator is asymptotically efficient while

Table 4: Beta Distribution with Possibility of Losing Information: Detailed values of squared error $\|\hat{\theta} - \theta_0\|^2$. In each cell, the first number is the mean of squared error in $K = 50$ experiments and the number in the brackets is the standard deviation of squared error.

number of machines	simple avg ($\times 10^{-5}$)	one-step ($\times 10^{-5}$)	centralized ($\times 10^{-5}$)	centralized(95%) ($\times 10^{-5}$)
8	4.98 (10.76)	4.95 (10.62)	4.07 (6.91)	4.76 (9.81)
16	4.82 (7.61)	4.75 (7.40)		
32	4.85 (9.65)	4.72 (9.31)		
64	4.51 (7.89)	4.10 (7.04)		
128	5.25 (9.16)	4.48 (7.77)		
256	7.57 (12.26)	4.52 (7.70)		
512	16.51 (20.15)	5.24 (8.02)		

simple averaging estimator is absolutely not. It is worth noting that the resampled averaging estimator is not asymptotic efficient though it is better than simple averaging estimator. When the number of samples N is relatively small, the one-step estimator is worse than centralized estimator. When the number of samples N grows large, the differences between the one-step estimator and the centralized estimator become minimal in terms of both mean squared error and standard deviation. However, the error of the simple averaging estimator is significant larger than both the one-step estimator and the centralized estimator. When the sample size $N = 4^9 \approx 250,000$, the mean squared error of the simple averaging estimator is more than twice of that of the one-step and the centralized estimator.

2.4 Conclusions on One-Step Estimator

The M-estimator is a fundamental and high-impact methodology in statistics. The classic M-estimator theory is based on the assumption that the entire data are available at a central location, and can be processed/computed without considering communication issues. In many modern estimation problems arising in contemporary sciences and engineering, the

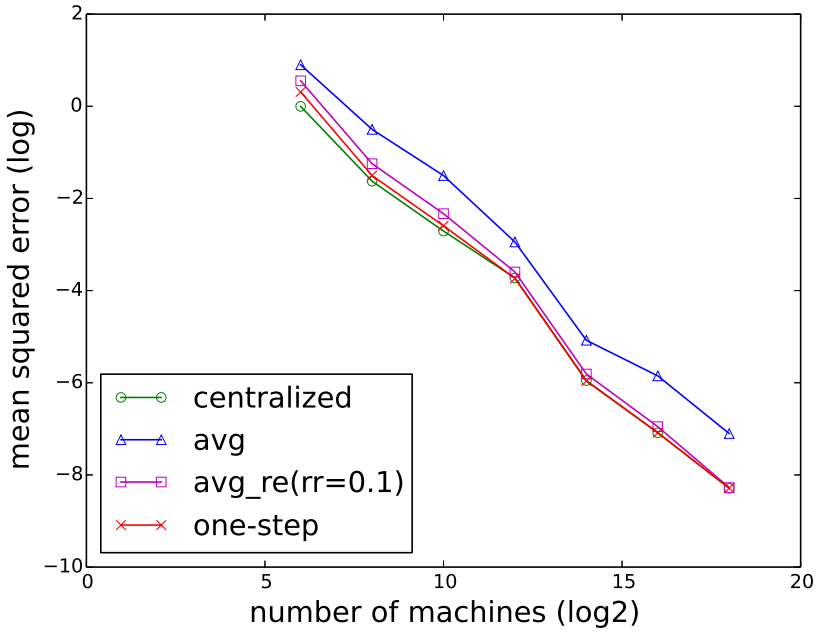


Figure 4: Gaussian Distribution with Unknown Mean and Variance: The log error $\log \|\theta - \theta_0\|^2$ versus the log number of machines ($\log_2 k$), with fifty repeated experiments for each N, where θ_0 is the true parameter. The "avg", "avg-re" and "one-step" denote $\theta^{(0)}$, $\theta^{(0)}_{re}$ with resampling ratio $rr = 10\%$ and $\theta^{(1)}$, respectively. The "centralized" denotes the maximum likelihood estimator with the entire data. The sample size is fixed to be $N = k^2$.

classical notion of asymptotic optimality suffers from a significant deficiency: it requires access to all data. The asymptotic property when the data has to be dealt with distributively is under-developed. In this chapter, we close this gap by considering a distributed one-step estimator.

Our one-step estimator builds on the existing *averaging* estimator. In a nutshell, after obtaining an averaging estimator, this initial estimate is broadcasted to local machines, to facilitate their computation of gradients and hessians of their objective functions. By doing so, the data do *not* need to be transmitted to the central machine. The central machine than collects the locally estimated gradients and hessians, to produce a global estimate of the overall gradient and overall hessian. Consequently, a one-step update of the initial estimator

Table 5: Gaussian Distribution with Unknown Mean and Variance: Detailed values of squared error $\|\hat{\theta} - \theta_0\|^2$. In each cell, the first number is the mean of squared error in $K = 50$ experiments and the number in the brackets is the standard deviation of squared error.

no. of machines	no. of samples	simple avg	resampled avg	one-step	centralized
8	64	3.022104 (4.385627)	2.153958 (3.458645)	1.694668 (2.882794)	1.388959 (2.424813)
16	256	0.739784 (1.209734)	0.392389 (0.739390)	0.318765 (0.621990)	0.286175 (0.566140)
32	1024	0.118766 (0.151695)	0.041050 (0.053808)	0.034494 (0.046586)	0.032563 (0.045779)
64	4096	0.026839 (0.046612)	0.016519 (0.030837)	0.014255 (0.029258)	0.014414 (0.030533)
128	16384	0.010996 (0.019823)	0.004542 (0.009089)	0.004329 (0.009453)	0.004357 (0.009315)
256	65536	0.002909 (0.005785)	0.001158 (0.002733)	0.001105 (0.002779)	0.001099 (0.002754)
512	262144	0.000843 (0.001426)	0.000461 (0.000744)	0.000376 (0.000596)	0.000376 (0.000595)

is implemented. Just like the one-step approach has improved the estimator in the classical (non-distributed) setting, we found that the one-step approach can improve the performance of an estimator under the distributed setting, both theoretically and numerically.

Besides the works that have been cited earlier, there are many other results that are in the relevant literature, however they may not be directly technically linked to what's been done here. We discuss their influence and insights in the next few paragraphs.

An interesting split-and-merge Bayesian approach for variable selection under linear models is proposed in [73]. The method firstly split the ultrahigh dimensional data set into a number of lower dimensional subsets and select relevant variables from each of the subsets, and then aggregate the variables selected from each subset and then select relevant variables from the aggregated data set. Under mild conditions, the authors show that the proposed

approach is consistent, i.e., the underlying true model will be selected in probability 1 as the sample size becomes large. This work differs from all the other approaches that we discussed in this chapter: it splits the variables, while all other approaches that we referenced (including ours) split the data according to observations. This paper certainly is in line with our research, however takes a very distinct angle.

An interesting piece of work that combines distributed statistical inference and information theory in communication is presented in [92]. Their current results need to rely on special model settings: uniform location family $\mathcal{U} = \{P_\theta, \theta \in [-1, 1]\}$, where P_θ denotes the uniform distribution on the interval $[\theta - 1, \theta + 1]$, or Gaussian location families $N_d([-1, 1]^d) = \{N(\theta, \sigma^2 I_{d\times d}) \mid \theta \in \Theta = [-1, 1]^d\}$. It will be interesting to see whether or not more general results are feasible.

[57] proposed a distributed expectation-maximization (EM) algorithm for density estimation and clustering in sensor networks. Though the studied problem is technically different from ours, it provides an inspiring historic perspective: distributed inference has been studied more than ten years ago.

[56] propose an asymptotically exact, embarrassingly parallel MCMC method by approximating each sub-posterior with Gaussian density, Gaussian kernel or weighted Gaussian kernel. They prove the asymptotic correctness of their estimators and bound rate of convergence. This dissertation does not consider the MCMC framework. The analytical tools that they used in proving their theorems are of interests.

[86] propose a distributed variable selection algorithm, which accepts a variable if more than half of machines select that variable. They give upper bounds for the success probability and Mean Squared Error (MSE) of estimator. This work bears similarity with [73]

and [19], however with somewhat different emphases.

[37] propose a scalable bootstrap (named 'bag of little bootstraps' (BLB)) for massive data to assess the quality of estimators. They also demonstrate its favorable statistical performance through both theoretical analysis and simulation studies. A comparison with this work will be interesting, however not included here.

[94] consider a partially linear framework for massive heterogeneous data and propose an aggregation type estimator for the commonality parameter that possesses the minimax optimal bound and asymptotic distribution when number of sub-populations does not grow too fast.

A recent work [2] shed interesting new light into the distributed inference problem. The authors studied the fundamental limits to communication-efficient distributed methods for convex learning and optimization, under different assumptions on the information available to individual machines, and the types of functions considered. The current problem formulation is more numerical than statistical properties. Their idea may lead to interesting counterparts in statistical inference.

Besides estimation, other distributed statistical technique may be of interests, such as the distributed principal component analysis [4]. We do not touch this line of research.

Various researchers have studied communication-efficient algorithms for statistical estimation (e.g., see the papers [22, 3, 85, 54] and references therein). They were not discussed in details here, because they are pretty much discussed/compared in other references of this dissertation.

There is now a rich and well-developed body of theory for bounding and/or computing the minimax risk for various statistical estimation problems, e.g., see [90] and references

therein. In several cited references, researchers have started to derive the optimal minimax rate for estimators under the distributed inference setting. This will be an exciting future research direction.

CHAPTER III

DISTANCE COVARIANCE AND TESTING OF INDEPENDENCE

This chapter is organized as follows. In Section 3.1, we review the definition of distance covariance, its fast algorithm in univariate cases and related distance-based independence tests. Section 3.2 gives the detailed algorithm for distance covariance of random vectors and corresponding independence tests. In Section 3.3, we present some theoretical properties on distance covariance and the asymptotic distribution of the proposed estimator. In Section 3.4, we conduct numerical examples to compare our method against others in existing literature. Some discussions are presented in Section 3.5. We conclude in Section 3.6. All technical proofs as well as formal presentation of algorithms are relegated to the appendix when appropriate.

Throughout this chapter, we adopt the following notations. We denote $c_p = \frac{\pi^{(p+1)/2}}{\Gamma((p+1)/2)}$ and $c_q = \frac{\pi^{(q+1)/2}}{\Gamma((q+1)/2)}$ as two constants, where $\Gamma(\cdot)$ denotes the Gamma function. We will also need the following constants: $C_p = \frac{c_1 c_{p-1}}{c_p} = \frac{\sqrt{\pi}\Gamma((p+1)/2)}{\Gamma(p/2)}$ and $C_q = \frac{c_1 c_{q-1}}{c_q} = \frac{\sqrt{\pi}\Gamma((q+1)/2)}{\Gamma(q/2)}$. For any vector v, let v^t denote its transpose.

3.1 Review of Distance Covariance: Definition, Fast Algorithm, and Related Independence Tests

In this section, we review some related existing works. In Section 3.1.1, we recall the concept of distance variances and correlations, as well as some of their properties. In Section 3.1.2, we discuss the estimators of distance covariances and correlations, as well

as their computation. We present their applications in testing of independence in Section 3.1.3.

3.1.1 Definition of Distance Covariances

Measuring and testing the dependency between two random variables is a fundamental problem in statistics. The classical Pearson's correlation coefficient can be inaccurate and even misleading when nonlinear dependency exists. [80] proposes the novel measure–distance correlation–which is exactly zero if and only if two random variables are independent. A limitation is that if the distance correlation is implemented based on its original definition, the corresponding computational complexity can be as high as $O(n^2)$, which is not desirable when n is large.

We review the definition of the distance correlation in [80]. Let us consider two random variables $X \in \mathbb{R}^p, Y \in \mathbb{R}^q, p \geq 1, q \geq 1$. Let the complex-valued functions $\phi_{X,Y}(\cdot)$, $\phi_X(\cdot)$, and $\phi_Y(\cdot)$ be the characteristic functions of the joint density of X and Y, the density of X, and the density of Y, respectively. For any function ϕ, we denote $|\phi|^2 = \phi\bar{\phi}$, where $\bar{\phi}$ is the conjugate of ϕ; in words, $|\phi|$ is the magnitude of ϕ at a particular point. For vectors, let us use $|\cdot|$ to denote the Euclidean norm. In [80], the definition of distance covariance between random variables X and Y is

$$\mathcal{V}^2(X,Y) = \int_{\mathbb{R}^{p+q}} \frac{|\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)|^2}{c_p c_q |t|^{p+1} |s|^{q+1}} dtds, \tag{3.1.11}$$

where two constants c_p and c_q have been defined at the beginning of this chapter. The distance correlation is defined as

$$\mathcal{R}^2(X,Y) = \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X,X)}\sqrt{\mathcal{V}^2(Y,Y)}}.$$

The following property has been established in the aforementioned paper.

Theorem 3.1.1. *Suppose $X \in \mathbb{R}^p, p \geq 1$ and $Y \in \mathbb{R}^q, q \geq 1$ are two random variables, the following statements are equivalent:*

(1) X is independent of Y;

(2) $\phi_{X,Y}(t,s) = \phi_X(t)\phi_Y(s)$, for any $t \in \mathbb{R}^p$ and $s \in \mathbb{R}^q$;

(3) $\mathcal{V}^2(X,Y) = 0$;

(4) $\mathcal{R}^2(X,Y) = 0$.

Given sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, we can estimate the distance covariance by replacing the population characteristic function with the sample characteristic function: for $i = \sqrt{-1}, t \in \mathbb{R}^p, s \in \mathbb{R}^q$, we define

$$\begin{aligned} \hat{\phi}_X(t) &= \frac{1}{n}\sum_{j=1}^{n} e^{iX_j^t t}, \\ \hat{\phi}_Y(s) &= \frac{1}{n}\sum_{j=1}^{n} e^{iY_j^t s}, \text{ and} \\ \hat{\phi}_{X,Y}(t,s) &= \frac{1}{n}\sum_{j=1}^{n} e^{iX_j^t t + iY_j^t s}. \end{aligned}$$

Consequently one can have the following estimator for $\mathcal{V}^2(X,Y)$:

$$\mathcal{V}_n^2(X,Y) = \int_{\mathbb{R}^{p+q}} \frac{|\hat{\phi}_{X,Y}(t,s) - \hat{\phi}_X(t)\hat{\phi}_Y(s)|^2}{c_p c_q |t|^{p+1} |s|^{q+1}} dt \cdot ds. \tag{3.1.12}$$

Note that the above formula is convenient to define a quantity, however is *not* convenient for computation, due to the integration on the right hand side. In the literature, other estimates have been introduced and will be presented in the following.

3.1.2 Fast Algorithm in the Univariate Cases

The paper [50] gives an equivalent definition for the distance covariance between random variables X and Y:

$$\mathcal{V}^2(X,Y) = \mathbb{E}[d(X,X')d(Y,Y')] = \mathbb{E}[|X-X'||Y-Y'|] \\ - 2\mathbb{E}[|X-X'||Y-Y''|] + \mathbb{E}[|X-X'|]\mathbb{E}[|Y-Y'|], \quad (3.1.13)$$

where the double centered distance $d(\cdot,\cdot)$ is defined as

$$d(X,X') = |X-X'| - \mathbb{E}_X[|X-X'|] - \mathbb{E}_{X'}[|X-X'|] + \mathbb{E}[|X-X'|],$$

where $\mathbb{E}_X$, $\mathbb{E}_{X'}$ and $\mathbb{E}$ are expectations over X, X' and (X,X'), respectively.

Motivated by the above definition, one can give an unbiased estimator for $\mathcal{V}^2(X,Y)$. The following notations will be utilized: for $1 \le i,j \le n$,

$$\begin{aligned} a_{ij} &= |X_i - X_j|, \quad b_{ij} = |Y_i - Y_j|, \\ a_{i\cdot} &= \sum_{l=1}^{n} a_{il}, \quad b_{i\cdot} = \sum_{l=1}^{n} b_{il}, \\ a_{\cdot\cdot} &= \sum_{k,l=1}^{n} a_{kl}, \quad \text{and} \quad b_{\cdot\cdot} = \sum_{k,l=1}^{n} b_{kl}. \end{aligned} \quad (3.1.14)$$

It has been proven [79, 34] that

$$\Omega_n(X,Y) = \frac{1}{n(n-3)} \sum_{i \ne j} a_{ij} b_{ij} \\ - \frac{2}{n(n-2)(n-3)} \sum_{i=1}^{n} a_{i\cdot} b_{i\cdot} + \frac{a_{\cdot\cdot} b_{\cdot\cdot}}{n(n-1)(n-2)(n-3)} \quad (3.1.15)$$

is an unbiased estimator of $\mathcal{V}^2(X,Y)$. In addition, a fast algorithm has been propose [34] for the aforementioned sample distance covariance in the univariate cases with complexity order $O(n \log n)$ and storage $O(n)$. We list the result below for reference purpose.

Theorem 3.1.2 (Theorem 3.2 & Corollary 4.1 in [34]). *Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n \in \mathbb{R}$. The unbiased estimator Ω_n defined in (3.1.15) can be computed by an $O(n \log n)$ algorithm.*

In addition, as a byproduct, the following result is established in the same paper.

Corollary 3.1.3. *The quantity*

$$\frac{a_{..}b_{..}}{n(n-1)(n-2)(n-3)} = \frac{\sum_{k,l=1}^{n} a_{kl} \sum_{k,l=1}^{n} b_{kl}}{n(n-1)(n-2)(n-3)}$$

can be computed by an $O(n \log n)$ algorithm.

We will use the above result in our test of independence. However, as far as we know, in the multivariate cases, there does not exist any work on fast algorithm of the order of complexity $O(n \log n)$. This chapter will fill in this gap by introducing an order $O(nK \log n)$ complexity algorithm in the multivariate cases.

3.1.3 Distance Based Independence Tests

In [80] an independence test is proposed using the distance covariance. We summarizes it below as a theorem, which serves as a benchmark. Our test will be aligned with the following one, except that we introduced a new test statistic, which can be more efficiently computed, and it has comparable asymptotic properties with the test statistic that is used below.

Theorem 3.1.4 ([80], Theorem 6). *For potentially multivariate random variables X and Y, a prescribed level α_s, and sample size n, one rejects the independence if and only if*

$$\frac{n\mathcal{V}_n^2(X,Y)}{S_2} > (\Phi^{-1}(1-\alpha_s/2))^2,$$

where $\mathcal{V}_n^2(X,Y)$ *has been defined in* (3.1.12)*,* $\Phi(\cdot)$ *denote the cumulative distribution function of the standard normal distribution and*

$$S_2 = \frac{1}{n^4}\sum_{i,j=1}^{n}|X_i - X_j|\sum_{i,j=1}^{n}|Y_i - Y_j|.$$

Moreover, let $\alpha(X,Y,n)$ *denote the achieved significance level of the above test. If* $\mathbb{E}[|X|+|Y|] < \infty$*, then for all* $0 < \alpha_s < 0.215$*, one can show the following:*

$$\lim_{n\to\infty}\alpha(X,Y,n) \le \alpha_s, \text{ and}$$
$$\sup_{X,Y}\left\{\lim_{n\to\infty}\alpha(X,Y,n) : \mathcal{V}(X,Y) = 0\right\} = \alpha_s.$$

Note that the quantity $\mathcal{V}_n^2(X,Y)$ that is used above as in [80] differs from the one that will be used in our proposed method. As mentioned, we use the above as an illustration for distance-based tests of independence, as well as the theoretical (or asymptotic) properties that such a test can achieve.

3.2 Numerically Efficient Method for Random Vectors

This section is made of two components. We present a random-projection-based distance covariance estimator that will be proven to be unbiased with a computational complexity that is $O(Kn\log n)$ in Section 3.2.1. In Section 3.2.2, we describe how the test of independence can be done by utilizing the above estimator. For user's conveniences, stand-alone algorithms are furnished in the appendix.

3.2.1 Random Projection Based Methods for Approximating Distance Covariance

We consider how to use a fast algorithm for univariate random variables to compute or approximate the sample distance covariance of random vectors. The main idea works as

follows: first, projecting the multivariate observations on some random directions; then, using the fast algorithm to compute the distance covariance of the projections; finally, averaging distance covariances from different projecting directions.

More specifically, our estimator can be computed as follows. For potentially multivariate $X_1, \ldots, X_n \in \mathbb{R}^p, p \geq 1$ and $Y_1, \ldots, Y_n \in \mathbb{R}^q, q \geq 1$, let K be a predetermined number of iterations, we do:

(1) For each k ($1 \leq k \leq K$), randomly generate u_k and v_k from Uniform($\mathcal{S}^{p-1}$) and Uniform($\mathcal{S}^{q-1}$), respectively. Here $\mathcal{S}^{p-1}$ and $\mathcal{S}^{q-1}$ are the unit spheres in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. Uniform($\mathcal{S}^{p-1}$) is a uniform measure (or distribution) on $\mathcal{S}^{p-1}$.

(2) Let $u_k^t X$ and $v_k^t Y$ denote the projections of X and Y to the spaces that are spanned by vector u_k and v_k, respectively. That is we have

$$u_k^t X = (u_k^t X_1, \ldots, u_k^t X_n), \text{ and } v_k^t Y = (v_k^t Y_1, \ldots, v_k^t Y_n).$$

Note that samples $u_k^t X$ and $v_k^t Y$ are now univariate.

(3) Utilize the fast (i.e., order $O(n \log n)$) algorithm that was mentioned in Theorem 3.1.2 to compute for the unbiased estimator in (3.1.15) with respect to $u_k^t X$ and $v_k^t Y$. Formally, we denote

$$\Omega_n^{(k)} = C_p C_q \Omega_n(u_k^t X, v_k^t Y),$$

where C_p and C_q have been defined at the beginning of this chapter.

(4) The above three steps are repeated for K times. The final estimator is the average:

$$\overline{\Omega}_n = \frac{1}{K} \sum_{k=1}^{K} \Omega_n^{(k)}. \tag{3.2.16}$$

To emphasize the dependency of the above quantity with K, we sometimes use a notation $\overline{\Omega}_{n,K} \triangleq \overline{\Omega}_n$.

See Algorithm 1 in the appendix for a stand-alone presentation of the above method. In the light of Theorem 3.1.2, we can handily declare the following.

Theorem 3.2.1. *For potentially multivariate $X_1, \ldots, X_n \in \mathbb{R}^p$ and $Y_1, \ldots, Y_n \in \mathbb{R}^q$, the order of computational complexity of computing the aforementioned $\overline{\Omega}_n$ is $O(Kn \log n)$ with storage $O(\max\{n, K\})$, where K is the number of random projections.*

The proof of the above theorem is omitted, because it is straightforward from Theorem 3.1.2. The statistical properties of the proposed estimator $\overline{\Omega}_n$ will be studied in the subsequent section (specifically in Section 3.3.4).

3.2.2 Test of Independence

By a later result (cf. Theorem 3.3.18), we can apply $\overline{\Omega}_n$ in the independence testing. The corresponding asymptotic distribution of the test statistic $\overline{\Omega}_n$ can be approximated by a Gamma(α, β) distribution with α and β given in (3.3.24). We can compute the significant level of the test statistic by permutation and conduct the independence test accordingly. Recall that we have potentially multivariate $X_1, \ldots, X_n \in \mathbb{R}^p$ and $Y_1, \ldots, Y_n \in \mathbb{R}^q$. Recall that K denotes the number of Monte Carlo iterations in our previous algorithm. Let α_s denote the prescribed significance level of the independence test. Let L denote the number of random permutations that we will adopt. We would like to test the null hypothesis $\mathcal{H}_0$—X and Y are independent—against its alternative. Recall $\overline{\Omega}_n$ is our proposed estimator in (3.2.16). The following algorithm describes an independence test which applies permutations to generate a threshold.

(1) For each $\ell, 1 \leq \ell \leq L$, one generates a random permutation of Y: $Y^{\star,\ell} = (Y_1^\star, \ldots Y_n^\star)$;

(2) Using the algorithm in Section 3.2.1, one can compute the estimator $\overline{\Omega}_n$ as in (3.2.16) for X and $Y^{\star,\ell}$; denote the outcome to be $V_\ell = \overline{\Omega}_n(X, Y^{\star,\ell})$. Note under the random permutations, X and $Y^{\star,\ell}$ are independent.

(3) The above two steps are executed for all $\ell = 1, \ldots, L$. One rejects $\mathcal{H}_0$ if and only if we have

$$\frac{1 + \sum_{\ell=1}^{L} I(\overline{\Omega}_n > V_\ell)}{1 + L} > \alpha_s.$$

See Algorithm 2 in the appendix for a stand-alone description.

One can also use the information of an approximate asymptotic distribution to estimate a threshold in the aforementioned independence test. The following describes such an approach. Recall that we have random vectors $X_1, \ldots, X_n \in \mathbb{R}^p, p \geq 1$ and $Y_1, \ldots, Y_n \in \mathbb{R}^q, q \geq 1$, the number of random projections K, and a prescribed significance level α_s that has been mentioned earlier.

(1) For each k ($1 \leq k \leq K$), randomly generate u_k and v_k from uniform($\mathcal{S}^{p-1}$) and uniform($\mathcal{S}^{q-1}$), respectively.

(2) Use the fast algorithm in Theorem 3.1.2 to compute the following quantities:

$$\Omega_n^{(k)} = C_p C_q \Omega_n(u_k^t X, v_k^t Y),$$

$$S_{n,1}^{(k)} = C_p^2 C_q^2 \Omega_n(u_k^t X, u_k^t X)\Omega_n(v_k^t Y, v_k^t Y),$$

$$S_{n,2}^{(k)} = C_p \frac{a_{\cdot\cdot}^{u_k}}{n(n-1)}, \quad S_{n,3}^{(k)} = C_q \frac{b_{\cdot\cdot}^{v_k}}{n(n-1)},$$

where C_p and C_q have been defined at the beginning of this chapter and in the last

equation, the $a_{\cdot\cdot}^{u_k}$ and $b_{\cdot\cdot}^{v_k}$ are defined as follows:

$$a_{ij}^{u_k} = |u_k^t(X_i - X_j)|, \quad b_{ij}^{v_k} = |v_k^t(Y_i - Y_j)|,$$
$$a_{\cdot\cdot}^{u_k} = \sum_{k,l=1}^{n} a_{kl}^{u_k}, \quad b_{\cdot\cdot}^{v_k} = \sum_{k,l=1}^{n} b_{kl}^{v_k}.$$

(3) For the aforementioned k, one randomly generates u_k' and v_k' from uniform($\mathcal{S}^{p-1}$) and uniform($\mathcal{S}^{q-1}$), respectively. Use the fast algorithm that is mentioned in Theorem 3.1.2 to compute the following.

$$\Omega_{n,X}^{(k)} = C_p^2 \Omega_n(u_k^t X, u'^t_k X), \quad \Omega_{n,Y}^{(k)} = C_p^2 \Omega_n(v_k^t Y, v'^t_k Y).$$

where C_p and C_q have been defined at the beginning of this chapter.

(4) Repeat the previous steps for all $k = 1, \ldots, K$. Then we compute the following quantities:

$$\overline{\Omega}_n = \frac{1}{K}\sum_{k=1}^{K} \Omega_n^{(k)}, \quad \bar{S}_{n,1} = \frac{1}{K}\sum_{k=1}^{K} S_{n,1}^{(k)}, \quad \bar{S}_{n,2} = \frac{1}{K}\sum_{k=1}^{K} S_{n,2}^{(k)},$$
$$\bar{S}_{n,3} = \frac{1}{K}\sum_{k=1}^{K} S_{n,3}^{(k)}, \quad \overline{\Omega}_{n,X} = \frac{1}{K}\sum_{k=1}^{K} \Omega_{n,X}^{(k)}, \quad \overline{\Omega}_{n,Y} = \frac{1}{K}\sum_{k=1}^{K} \Omega_{n,Y}^{(k)},$$
$$\alpha = \frac{1}{2}\frac{\bar{S}_{n,2}^2 \bar{S}_{n,3}^2}{\frac{K-1}{K}\overline{\Omega}_{n,X}\overline{\Omega}_{n,Y} + \frac{1}{K}\bar{S}_{n,1}}, \tag{3.2.17}$$
$$\beta = \frac{1}{2}\frac{\bar{S}_{n,2}\bar{S}_{n,3}}{\frac{K-1}{K}\overline{\Omega}_{n,X}\overline{\Omega}_{n,Y} + \frac{1}{K}\bar{S}_{n,1}}. \tag{3.2.18}$$

(5) Reject $\mathcal{H}_0$ if $n\overline{\Omega}_n + \bar{S}_{n,2}\bar{S}_{n,3} >$ Gamma$(\alpha, \beta; 1 - \alpha_s)$; otherwise, accept it. Here Gamma$(\alpha, \beta; 1 - \alpha_s)$ is the $1 - \alpha_s$ quantile of the distribution Gamma(α, β).

The above procedure is motivated by the observation that the asymptotic distribution of the test statistic $n\overline{\Omega}_n$ can be approximated by a Gamma distribution, whose parameters can be estimated by (3.2.17) and (3.2.18). A stand-alone description of the above procedure can be found in Algorithm 3 in the appendix.

3.3 Theoretical Properties of Distance Covariance and Random Projections

In this section, we establish the theoretical foundation of the proposed method. In Section 3.3.1, we study some properties of the random projections and the subsequent average estimator. These properties will be needed in studying the properties of the proposed estimator. We study the properties of the proposed distance covariance estimator (Ω_n) in Section 3.3.2, taking advantage of the fact that Ω_n is a U-statistic. It turns out that the properties of eigenvalues of a particular operator plays an important role. We present the relevant results in Section 3.3.3. The main properties of the proposed estimator ($\overline{\Omega}_n$) is presented in Section 3.3.4.

3.3.1 Using Random Projections in Distance-Based Methods

In this section, we will study some properties of distance covariances of randomly projected random vectors. We begin with a necessary and sufficient condition of independence.

Lemma 3.3.1. *Suppose u and v are points on the hyper-spheres: $u \in \mathcal{S}^{p-1} = \{u \in \mathbb{R}^p : |u| = 1\}$ and $v \in \mathcal{S}^{q-1}$. We have*

$$\textit{random vectors } X \in \mathbb{R}^p \textit{ and } Y \in \mathbb{R}^q \textit{ are independent}$$

if and only if

$$\mathcal{V}^2(u^tX, v^tY) = 0, \textit{ for any } u \in \mathcal{S}^{p-1}, v \in \mathcal{S}^{q-1}.$$

The proof is relatively straightforward. We relegate a formal proof to the appendix. This lemmas indicates that the independence is somewhat preserved under projections. The main contribution of the above result is to motivate us to think of using random projection,

to reduce the multivariate random vectors into univariate random variables. As mentioned earlier, there exist fast algorithms of distance-based methods for univariate random variables.

The following result allows us to regard the distance covariance of random vectors of any dimension as an integral of distance covariance of univariate random variables, which are the projections of the aforementioned random vectors. The formulas in the following lemma provides foundation for our proposed method: the distance covariances in the multivariate cases can be written as integrations of distance covariances in the univariate cases. our proposed method essentially adopts the principle of Monte Carlo to approximate such integrals. We again relegate the proof to the appendix.

Lemma 3.3.2. *Suppose u and v are points on unit hyper-spheres: $u \in \mathcal{S}^{p-1} = \{u \in \mathbb{R}^p : |u| = 1\}$ and $v \in \mathcal{S}^{q-1}$. Let μ and ν denote the uniform probability measure on $\mathcal{S}^{p-1}$ and $\mathcal{S}^{q-1}$, respectively. Then, we have for random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$,*

$$\mathcal{V}^2(X,Y) = C_p C_q \int_{\mathcal{S}^{p-1}\times\mathcal{S}^{q-1}} \mathcal{V}^2(u^tX, v^tY) d\mu(u) d\nu(v),$$

where C_p and C_q are two constants that are defined at the beginning of this chapter. Moreover, a similar result holds for the sample distance covariance:

$$\mathcal{V}_n^2(X,Y) = C_p C_q \int_{\mathcal{S}^{p-1}\times\mathcal{S}^{q-1}} \mathcal{V}_n^2(u^tX, v^tY) d\mu(u) d\nu(v).$$

Besides the integral equations in the above lemma, we can also establish the following result for the unbiased estimator. Such a result provides direct foundation of our proposed method. Recall that Ω_n, which is in (3.1.15), is an unbiased estimator of the distance covariance $\mathcal{V}^2(X,Y)$. A proof is provided in the appendix.

Lemma 3.3.3. *Suppose u and v are points on the hyper-spheres: $u \in \mathcal{S}^{p-1} = \{u \in \mathbb{R}^p : |u| = 1\}$ and $v \in \mathcal{S}^{q-1}$. Let μ and ν denote the measure corresponding to the uniform densities on the surfaces $\mathcal{S}^{p-1}$ and $\mathcal{S}^{q-1}$, respectively. Then, we have*

$$\Omega_n(X,Y) = C_p C_q \int_{\mathcal{S}^{p-1}\times\mathcal{S}^{q-1}} \Omega_n(u^tX, v^tY) d\mu(u) d\nu(v),$$

where C_p and C_q are constants that were mentioned at the beginning of this chapter.

From the above lemma, recalling the design of our proposed estimator $\overline{\Omega}_n$ as in (3.2.16), it is straightforward to see that the proposed estimator $\overline{\Omega}_n$ is an unbiased estimator of $\Omega_n(X,Y)$. For completeness, we state the following without a proof.

Corollary 3.3.4. *The proposed estimator $\overline{\Omega}_n$ in* (3.2.16) *is an unbiased estimator of the estimator $\Omega_n(X,Y)$ that was defined in* (3.1.15)*.*

Note that the estimator $\overline{\Omega}_n$ in (3.2.16) evidently depends on the number of random projections K. Recall that to emphasize such a dependency, we sometimes use a notation $\overline{\Omega}_{n,K} \triangleq \overline{\Omega}_n$. The following concentration inequality shows the speed that $\overline{\Omega}_{n,K}$ can converge to Ω_n as $K \to \infty$.

Lemma 3.3.5. *Suppose $\mathbf{E}[|X|^2] < \infty$ and $\mathbf{E}[|Y|^2] < \infty$. For any $\epsilon > 0$, we have*

$$\mathbf{P}\left(|\overline{\Omega}_{n,K} - \Omega_n| > \epsilon\right) \leq 2\exp\left\{-\frac{CK\epsilon^2}{Tr[\Sigma_X]Tr[\Sigma_Y]}\right\},$$

where Σ_X and Σ_Y are the covariance matrices of X and Y, respectively, $Tr[\Sigma_X]$ and $Tr[\Sigma_Y]$ are their matrix traces, and $C = \frac{2}{25C_p^2C_q^2}$ is a constant.

The proof is a relatively standard application of the Hoeffding's inequality [31], which has been relegated to the appendix. The above lemma essentially indicates that the quantity $|\overline{\Omega}_{n,K} - \Omega_n|$ converges to zero at a rate no worse than $O(1/\sqrt{K})$.

3.3.2 Asymptotic Properties of the Sample Distance Covariance Ω_n

The asymptotic behavior of a range of sample distance covariance, such as Ω_n in (3.1.15) of this chapter, has been studied in many places, seeing [50, 34, 77, 69]. We found that it is still worthwhile to present them here, as we will use them to establish the statistical properties of our proposed estimator. The asymptotic distributions of Ω_n will be studied under two situations: (1) a general case and (2) when X and Y are assumed to be independent. We will see that the asymptotic distributions are different in these two situations.

It has been showed in [34, Theorem 3.2] that Ω_n is a U-statistic. In the following, we state the result without a formal proof. We will need the following function, denoted by h_4, which takes four pairs of input variables:

$$\begin{aligned} &h_4((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)) \\ &\quad= \frac{1}{4}\sum_{1\le i,j\le 4, i\neq j}|X_i-X_j||Y_i-Y_j| - \frac{1}{4}\sum_{i=1}^{4}\left(\sum_{j=1,j\neq i}^{4}|X_i-X_j|\sum_{j=1,j\neq i}^{4}|Y_i-Y_j|\right) \\ &\qquad\qquad + \frac{1}{24}\sum_{1\le i,j\le 4, i\neq j}|X_i-X_j|\sum_{1\le i,j\le 4, i\neq j}|Y_i-Y_j|. \end{aligned} \tag{3.3.19}$$

Note that the definition of h_4 coincides with Ω_n when the number of observations $n = 4$.

Lemma 3.3.6 (U-statistics)**.** *Let Ψ_4 denote all distinct 4-subset of $\{1,\dots,n\}$ and let us define $X_\psi = \{X_i | i \in \psi\}$ and $Y_\psi = \{Y_i | i \in \psi\}$, then Ω_n is a U-statistic and can be expressed as*

$$\Omega_n = \binom{n}{4}^{-1}\sum_{\psi\in\Psi_4} h_4\left(X_\psi, Y_\psi\right).$$

From the literature of the U-statistics, we know that the following quantities play critical

roles. We state them here:

$$h_1((X_1, Y_1)) = \mathbb{E}_{2,3,4}[h_4((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4))],$$

$$h_2((X_1, Y_1), (X_2, Y_2)) = \mathbb{E}_{3,4}[h_4((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4))],$$

$$h_3((X_1, Y_1), (X_2, Y_2), (X_3, Y_3)) = \mathbb{E}_4[h_4((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4))],$$

where $\mathbb{E}_{2,3,4}$ stands for taking expectation over $(X_2, Y_2), (X_3, Y_3)$ and (X_4, Y_4); $\mathbb{E}_{3,4}$ stands for taking expectation over (X_3, Y_3) and (X_4, Y_4); and $\mathbb{E}_4$ stands for taking expectation over (X_4, Y_4); respectively.

One immediate application of the above notations is the following result, which quantifies the variance of Ω_n. Since the formula is a known result, seeing [70, Chapter 5.2.1, Lemma A], we state it without a proof.

Lemma 3.3.7 (Variance of the U-statistic)**.** *The variance of Ω_n could be written as*

$$\begin{aligned} \mathit{Var}(\Omega_n) &= \binom{n}{4}^{-1} \sum_{l=1}^{4} \binom{4}{l} \binom{n-4}{4-l} \mathit{Var}(h_l) \\ &= \frac{16}{n} \mathit{Var}(h_1) + \frac{240}{n^2} \mathit{Var}(h_1) + \frac{72}{n^2} \mathit{Var}(h_2) + O\left(\frac{1}{n^3}\right), \end{aligned}$$

where $O(\cdot)$ is the standard big O notation in mathematics.

From the above lemma, we can see that $\mathrm{Var}(h_1)$ and $\mathrm{Var}(h_2)$ play indispensable roles in determining the variance of Ω_n. The following lemma shows that under some conditions, we can ensure that $\mathrm{Var}(h_1)$ and $\mathrm{Var}(h_2)$ are bounded. A proof has been relegated to the appendix.

Lemma 3.3.8. *If we have $\mathbb{E}[|X|^2] < \infty$, $\mathbb{E}[|Y|^2] < \infty$ and $\mathbb{E}[|X|^2|Y|^2] < \infty$, then we have $\mathit{Var}(h_4) < \infty$. Consequently, we also have $\mathit{Var}(h_1) < \infty$ and $\mathit{Var}(h_2) < \infty$.*

Even though as indicated in Lemma 3.3.7, the quantities $h_1(X_1, Y_1)$ and $h_2((X_1, Y_1), (X_2, Y_2))$ play important roles in determine the variance of Ω_n, in a generic case, they do not have a simple formula. The following lemma gives the generic formulas for $h_1(X_1, Y_1)$ and $h_2((X_1, Y_1), (X_2, Y_2))$. Its calculation can be found in the appendix.

Lemma 3.3.9 (Generic h_1 and h_2)**.** *In the general case, assuming* (X_1, Y_1)*,* (X, Y)*,* (X', Y')*, and* (X'', Y'') *are independent and identically distributed, we have*

$$\begin{aligned} h_1((X_1, Y_1)) = \quad & \frac{1}{2}\mathbb{E}[|X_1 - X'||Y_1 - Y'|] - \frac{1}{2}\mathbb{E}[|X_1 - X'||Y_1 - Y''|] \\ & + \frac{1}{2}\mathbb{E}[|X_1 - X'||Y - Y''|] - \frac{1}{2}\mathbb{E}[|X_1 - X'||Y' - Y''|] \\ & + \frac{1}{2}\mathbb{E}[|X - X''||Y_1 - Y'|] - \frac{1}{2}\mathbb{E}[|X' - X''||Y_1 - Y'|] \\ & + \frac{1}{2}\mathbb{E}[|X - X'||Y - Y'|] - \frac{1}{2}\mathbb{E}[|X - X'||Y - Y''|]. \end{aligned}$$

We have a similar formula for $h_2((X_1, Y_1), (X_2, Y_2))$ *in* (C.47)*. Due to its length, we do not display it here.*

If one assumes that X and Y are independent, we can have simpler formula for h_1, h_2, as well as their corresponding variances. We list the results below, with detailed calculation relegated to the appendix. One can see that under independence, the corresponding formulas are much simpler.

Lemma 3.3.10. *When* X *and* Y *are independent, we have the following. For* (X, Y) *and*

(X', Y') that are independent and identically distributed as (X_1, Y_1) and (X_2, Y_2), we have

$$h_1((X_1, Y_1)) = 0, \tag{3.3.20}$$

$$h_2((X_1, Y_1), (X_2, Y_2)) = \frac{1}{6}\left(|X_1 - X_2| - \mathbb{E}[|X_1 - X|] - \mathbb{E}[|X_2 - X|] + \mathbb{E}[|X - X'|]\right) \tag{3.3.21}$$

$$(|Y_1 - Y_2| - \mathbb{E}[|Y_1 - Y|] - \mathbb{E}[|Y_2 - Y|] + \mathbb{E}[|Y - Y'|]),$$

$$\mathit{Var}(h_2) = \frac{1}{36}\mathcal{V}^2(X, X)\mathcal{V}^2(Y, Y), \tag{3.3.22}$$

where $\mathbb{E}$ stands for the expectation operators with respect to X, X and X', Y, or Y and Y', whenever appropriate, respectively.

If we have $0 < \text{Var}(h_1) < \infty$, it is known that the asymptotic distribution of Ω_n is normal, as stated in the following. Note that based on Lemma 3.3.10, X and Y cannot be independent; otherwise one should have $h_1 = 0$ almost surely. The following theorem is based on a known result on the convergence of U-statistics, seeing [70, Chapter 5.5.1 Theorem A]. We state it without a proof.

Theorem 3.3.11. *Suppose $n \geq 7$, $0 < \mathit{Var}(h_1) < \infty$ and $\mathit{Var}(h_4) < \infty$, then we have*

$$\Omega_n \xrightarrow{P} \mathcal{V}^2(X, Y)$$

moreover, we have

$$\sqrt{n}(\Omega_n - \mathcal{V}^2(X, Y)) \xrightarrow{D} N(0, 16\mathit{Var}(h_1)), \text{ as } n \to \infty.$$

When X and Y are independent, the asymptotic distribution of $\sqrt{n}\Omega_n$ is no longer normal. In this case, from Lemma 3.3.10, we have

$$h_1((X_1, Y_1)) = 0 \text{ almost surely, and } \text{Var}[h_1((X_1, Y_1))] = 0.$$

The following theorem, which applies a result in [70, Chapter 5.5.2], indicates that $n\Omega_n$ converges to a weighted sum of (possibly infinitely many) independent χ_1^2 random variables.

Theorem 3.3.12. *If X and Y are independent, the asymptotic distribution of Ω_n is*

$$n\Omega_n \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1) = \sum_{i=1}^{\infty} \lambda_i Z_i^2 - \sum_{i=1}^{\infty} \lambda_i,$$

where $Z_i^2 \sim \chi_1^2$ i.i.d, λ_i's are the eigenvalues of operator G that is defined as

$$Gg(x_1, y_1) = \mathbb{E}_{x_2,y_2}[6h_2((x_1, y_1), (x_2, y_2))g(x_2, y_2)],$$

where function $h_2((\cdot,\cdot),(\cdot,\cdot))$ was defined in (3.3.21).

Proof. The asymptotic distribution of Ω_n is from the result in [70, Chapter 5.5.2]. □

See Subsection 3.3.3 for more details on methods for computing the value of λ_i's. In particular, we will show that we have $\sum_{i=1}^{\infty} \lambda_i = \mathbb{E}[|X - X'|]\mathbb{E}[|Y - Y'|]$ (Corollary 3.3.15) and $\sum_{i=1}^{\infty} \lambda_i^2 = \mathcal{V}^2(X, X)\mathcal{V}^2(Y, Y)$ (which is essentially from (3.3.22) and Lemma 3.3.7).

3.3.3 Properties of Eigenvalues λ_i's

From Theorem 3.3.12, we see that the eigenvalues λ_i's play important role in determining the asymptotic distribution of Ω_n. We study its properties here. Throughout this subsection, we assume that X and Y are independent. Let us recall that the asymptotic distribution of sample distance covariance Ω_n,

$$n\Omega_n \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1) = \sum_{i=1}^{\infty} \lambda_i Z_i^2 - \sum_{i=1}^{\infty} \lambda_i,$$

where λ_i's are the eigenvalues of the operator G that is defined as

$$Gg(x_1, y_1) = \mathbb{E}_{x_2,y_2}[6h_2((x_1, y_1), (x_2, y_2))g(x_2, y_2)],$$

where function $h_2((\cdot,\cdot),(\cdot,\cdot))$ was defined in (3.3.21). By definition, eigenvalues $\lambda_1, \lambda_2, \ldots$ corresponding to distinct solutions of the following equation

$$Gg(x_1, y_1) = \lambda g(x_1, y_1). \tag{3.3.23}$$

We now study the properties of λ_i's. Utilizing the Lemma 12 and equation (4.4) in [69], we can verify the following result. We give details of verifications in the appendix.

Lemma 3.3.13. *Both of the following two functions are positive definite kernels:*

$$h_X(X_1, X_2) = -|X_1 - X_2| + \mathbb{E}[|X_1 - X|] + \mathbb{E}[|X_2 - X|] - \mathbb{E}[|X - X'|]$$

and

$$h_Y(Y_1, Y_2) = -|Y_1 - Y_2| + \mathbb{E}[|Y_1 - Y|] + \mathbb{E}[|Y_2 - Y|] - \mathbb{E}[|Y - Y'|].$$

The above result gives us a foundation to apply the equivalence result that has been articulated thoroughly in [69]. Equipped with the above lemma, we have the following result, which characterizes a property of λ_i's. The detailed proof can be found in the appendix.

Lemma 3.3.14. *Suppose $\{\lambda_1, \lambda_2, \ldots\}$ are the set of eigenvalues of kernel $6h_2((x_1, y_1), (x_2, y_2))$, $\{\lambda_1^X, \lambda_2^X, \ldots\}$ and $\{\lambda_1^Y, \lambda_2^Y, \ldots\}$ are the sets of eigenvalues of the positive definite kernels h_X and h_Y, respectively. We have the following:*

$$\{\lambda_1, \lambda_2, \ldots\} = \{\lambda_1^X, \lambda_2^X, \ldots\} \otimes \{\lambda_1^Y, \lambda_2^Y, \ldots\};$$

that is, each λ_i satisfying (3.3.23) *can be written as, for some j, j',*

$$\lambda_i = \lambda_j^X \cdot \lambda_{j'}^Y$$

where λ_j^X and $\lambda_{j'}^Y$ are the eigenvalues corresponding to kernel functions $h_X(X_1, X_2)$ and $h_Y(Y_1, Y_2)$, respectively.

Above lemma implies that eigenvalues of h_2 could be obtained immediately after knowing the eigenvalues of h_X and h_Y. But, in practice, there usually does not exist analytic solution for even the eigenvalues of h_X or h_Y. Instead, given the observations $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$, we can compute the eigenvalues of matrices $\widetilde{K}_X = (h_X(X_i, X_j))_{n\times n}$ and $\widetilde{K}_Y = (h_Y(Y_i, Y_j))_{n\times n}$ and use those empirical eigenvalues to approximate $\lambda_1^X, \lambda_2^X, \ldots$ and $\lambda_1^Y, \lambda_2^Y, \ldots$, and then consequently $\lambda_1, \lambda_2, \ldots$

We end this subsection with the following corollary on the summations of eigenvalues, which is necessary for the proof of Theorem 3.3.12. The proof can be found in the appendix.

Corollary 3.3.15. *The aforementioned eigenvalues $\lambda_1^X, \lambda_2^X, \ldots$ and $\lambda_1^Y, \lambda_2^Y, \ldots$ satisfy*

$$\sum_{i=1}^{\infty} \lambda_i^X = \mathbb{E}[|X - X'|], \text{ and } \sum_{i=1}^{\infty} \lambda_i^Y = \mathbb{E}[|Y - Y'|].$$

As a result, we have

$$\sum_{i=1}^{\infty} \lambda_i = \mathbb{E}[|X - X'|]\mathbb{E}[|Y - Y'|],$$

and

$$\sum_{i=1}^{\infty} \lambda_i^2 = \mathcal{V}^2(X, X)\mathcal{V}^2(Y, Y).$$

3.3.4 Asymptotic Properties of Averaged Projected Sample Distance Covariance $\overline{\Omega}_n$

We have reviewed the properties of the statistics Ω_n in a previous section (Section 3.3.2). The disadvantage of directly applying Ω_n (which is defined in (3.1.15)) is that for multivariate X and Y, the implementation may require at least $O(n^2)$ operations. Recall that for univariate X and Y, an $O(n \log n)$ algorithm exists, cf. Theorem 3.1.2. The proposed estimator ($\overline{\Omega}_n$ in (3.2.16)) is the averaged distance covariances, after randomly projecting

X and Y to one-dimensional spaces, respectively. In this section, we will study the asymptotic behavior of $\overline{\Omega}_n$. It turns out that the analysis will be similar to the works in Section 3.3.2. The asymptotic distribution of $\overline{\Omega}_n$ will differ in two cases: (1) the dependent case and (2) the case when X and Y are independent.

As a preparation of presenting the main result, we recall and introduce some notations. Recall the definition of $\overline{\Omega}_n$:

$$\overline{\Omega}_n = \frac{1}{K}\sum_{k=1}^{K}\Omega_n^{(k)},$$

where

$$\Omega_n^{(k)} = C_p C_q \Omega_n(u_k^t X, v_k^t Y)$$

and constants C_p, C_q have been defined at the beginning of Chapter 3. By Corollary 3.3.4, we have $\mathbb{E}\left[\Omega_n^{(k)}\right] = \Omega_n$, where $\mathbb{E}$ stands for the expectation with respect to the random projection. Note that from the work in Section 3.3.2, estimator $\Omega_n^{(k)}$ is a U-statistic. The following equation reveals that estimator $\overline{\Omega}_n$ is also a U-statistic,

$$\overline{\Omega}_n = \binom{n}{4}^{-1}\sum_{\psi\in\Psi_4}\frac{C_p C_q}{K}\sum_{k=1}^{K}h_4(u_k^t X_\psi, v_k^t Y_\psi) \triangleq \binom{n}{4}^{-1}\sum_{\psi\in\Psi_4}\bar{h}_4(X_\psi, Y_\psi),$$

where

$$\bar{h}_4(X_\psi, Y_\psi) = \frac{1}{K}\sum_{k=1}^{K}C_p C_q h_4(u_k^t X_\psi, v_k^t Y_\psi).$$

We have seen that quantities h_1 and h_2 play significant roles in the asymptotic behavior

of statistic Ω_n. Let us define the counterpart notations as follows:

$$\begin{aligned}
\bar{h}_1((X_1,Y_1)) &= \mathbb{E}_{2,3,4}[\bar{h}_4((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4))] \\
&\triangleq \frac{1}{K}\sum_{k=1}^{K} h_1^{(k)} \\
\bar{h}_2((X_1,Y_1),(X_2,Y_2)) &= \mathbb{E}_{3,4}[\bar{h}_4((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4))] \\
&\triangleq \frac{1}{K}\sum_{k=1}^{K} h_2^{(k)},
\end{aligned}$$

where $\mathbb{E}_{2,3,4}$ stands for taking expectation over (X_2,Y_2), (X_3,Y_3) and (X_4,Y_4); $\mathbb{E}_{3,4}$ stands for taking expectation over (X_3,Y_3) and (X_4,Y_4); as well as the following:

$$\begin{aligned}
h_1^{(k)} &= \mathbb{E}_{2,3,4}[C_p C_q h_4(u_k^t X_\psi, v_k^t Y_\psi)], \\
h_2^{(k)} &= \mathbb{E}_{3,4}[C_p C_q h_4(u_k^t X_\psi, v_k^t Y_\psi)].
\end{aligned}$$

In the general case, we do not assume that X and Y are independent. Let $U = (u_1, \ldots, u_K)$ and $V = (v_1, \ldots, v_K)$ denote the collection of random projections. We can write the variance of $\overline{\Omega}_n$ as follows. The proof is an application of Lemma 3.3.7 and the law of total covariance. We relegate it to the appendix.

Lemma 3.3.16. *Suppose* $\mathbb{E}_{U,V}[\mathit{Var}_{X,Y}(\bar{h}_1|U,V)] > 0$ *and* $\mathit{Var}_{u,v}(\mathcal{V}^2(u^tX, v^tY)) > 0$*, then, the variance of* $\overline{\Omega}_n$ *is*

$$\begin{aligned}
\mathit{Var}(\overline{\Omega}_n) &= \frac{1}{K}\mathit{Var}_{u,v}(\mathcal{V}^2(u^tX, v^tY)) + \frac{16}{n}\mathbb{E}_{U,V}[\mathit{Var}_{X,Y}(\bar{h}_1|U,V)] \\
&\quad + \frac{72}{n^2}\mathbb{E}_{U,V}[\mathit{Var}_{X,Y}(\bar{h}_2|U,V)] + O\left(\frac{1}{n^3}\right).
\end{aligned}$$

With above preparation, we will derive the asymptotic distribution of proposed estimator $\overline{\Omega}_n$ in two different cases: (1) X and Y are dependent; (2) X and Y are independent. It

is worthing noting that the second case is of more interest for hypotheses testing while the first case is for theoretical completeness.

3.3.4.1 Asymptotic Properties under Depedence

Equipped with Lemma 3.3.16, we can summarize the asymptotic properties of proposed estimator in the following theorem. We state it without a proof as it is an immediate result from Lemma 3.3.16 as well as the contents in [70, Chapter 5.5.1 Theorem A].

Theorem 3.3.17. *Suppose* $0 < \mathbb{E}_{U,V}[\mathit{Var}_{X,Y}(\bar{h}_1|U,V)] < \infty$, $\mathbb{E}_{U,V}[\mathit{Var}_{X,Y}(\bar{h}_4|U,V)] < \infty$. *Also, let us assume that* $K \to \infty$, $n \to \infty$, *then we have*

$$\overline{\Omega}_n \xrightarrow{P} \mathcal{V}^2(X,Y).$$

And, the asymptotic distribution of $\overline{\Omega}_n$ *could differ under different conditions.*

(1) If $K \to \infty$ *and* $K/n \to 0$, *then*

$$\sqrt{K}\left(\overline{\Omega}_n - \mathcal{V}^2(X,Y)\right) \xrightarrow{D} N\left(0, \mathit{Var}_{u,v}(\mathcal{V}^2(u^tX, v^tY))\right).$$

(2) If $n \to \infty$ *and* $K/n \to \infty$, *then*

$$\sqrt{n}\left(\overline{\Omega}_n - \mathcal{V}^2(X,Y)\right) \xrightarrow{D} N\left(0, 16\mathbb{E}_{U,V}[\mathit{Var}_{X,Y}(\bar{h}_1|U,V)]\right).$$

(3) If $n \to \infty$ *and* $K/n \to C$, *where* C *is some constant, then*

$$\sqrt{n}\left(\overline{\Omega}_n - \mathcal{V}^2(X,Y)\right) \xrightarrow{D} N\left(0, \frac{1}{C}\mathit{Var}_{u,v}(\mathcal{V}^2(u^tX, v^tY)) + 16\mathbb{E}_{U,V}[\mathit{Var}_{X,Y}(\bar{h}_1|U,V)]\right).$$

Since our main idea is to utilize $\overline{\Omega}_n$ to approximate the quantity Ω_n, it is of interests to compare the asymptotic variance of Ω_n in Theorem 3.3.11 with the asymptotic variances in the above theorem. We present some discussions in the following remark.

Remark. *Let us recall the asymptotic properties of* Ω_n *,*

$$\sqrt{n}(\Omega_n - \mathcal{V}^2(X,Y)) \xrightarrow{D} N(0, 16\mathit{Var}(h_1)).$$

Then, we make the comparison in the following different scenarios.

(1) If $K \to \infty$ *and* $K/n \to 0$*, then the convergence rate of* $\overline{\Omega}_n$ *is much slower than* Ω_n *as* $K \ll n$*, which implies a high price to pay in terms of efficiency.*

(2) If $n \to \infty$ *and* $K/n \to \infty$*, then the convergence rate of* $\overline{\Omega}_n$ *is the same with* Ω_n *and but there is virtually no gain in terms of computational complexity.*

(3) If $n \to \infty$ *and* $K/n \to C$*, where* C *is some constant, then the convergence rate of* $\overline{\Omega}_n$ *is the same with* Ω_n *but the variance of* $\overline{\Omega}_n$ *is larger than that of* Ω_n*. In this case,* $\overline{\Omega}_n$ *loses statistical efficiency compared with* Ω_n*. The benefit for this loss of efficiency, however, is only a marginal improvement in terms of computational complexity.*

Theorem 3.3.17 is a general theoretical result with limited application in test of independence. First, we do not assume that X and Y are independent while, in test of independence, the asymptotic behavior of test statistics under the null hypotheses is of more interest. Second, we let number of random projections K be sufficiently large in Theorem 3.3.17. However, in practice, we must limit the value of K to achieve computational efficiency.

3.3.4.2 Asymptotic Properties under Independence

Generally, when X is not independent of Y, $\overline{\Omega}_n$ is not as good as Ω_n in terms of convergence rate. However, asymptotic distribution when X is independent of Y is of more interest for

hypotheses testing. In the following context of this section, we will show that $\overline{\Omega}_n$ has the same convergence rate with Ω_n when X is independent of Y.

By Lemma 3.3.10, we have

$$\bar{h}_1^{(k)} = 0, \bar{h}_1 = 0, \text{almost surely, and , Var}(\bar{h}_1) = 0.$$

And, by Lemma 3.3.1, we know that

$$\mathcal{V}^2(u^t X, v^t Y) = 0, \forall u, v,$$

which implies

$$\text{Var}_{u,v}\left(\mathcal{V}^2(u^t X, v^t Y)\right) = 0.$$

Therefore, we only need to consider $\text{Var}_{X,Y}(\bar{h}_2|U, V)$. Suppose (U, V) is given, a result in [70, Chapter 5.5.2], together with Lemma 3.3.16, indicates that $n\overline{\Omega}_n$ converges to a weighted sum of (possibly infinitely many) independent χ_1^2 random variables. The proof can be found in appendix.

Theorem 3.3.18. *If X and Y are independent, given the value of $U = (u_1, \ldots, u_K)$ and $V = (v_1, \ldots, v_K)$, the asymptotic distribution of $\overline{\Omega}_n$ is*

$$n\overline{\Omega}_n \xrightarrow{D} \sum_{i=1}^{\infty} \bar{\lambda}_i (Z_i^2 - 1) = \sum_{i=1}^{\infty} \bar{\lambda}_i Z_i^2 - \sum_{i=1}^{\infty} \bar{\lambda}_i,$$

where $Z_i^2 \sim \chi_1^2$ i.i.d, and

$$\sum_{i=1}^{\infty} \bar{\lambda}_i = \frac{C_p C_q}{K} \sum_{k=1}^{K} \mathbb{E}[|u_k^t (X - X')|]\mathbb{E}[|v_k^t (Y - Y')|],$$

$$\sum_{i=1}^{\infty} \bar{\lambda}_i^2 = \frac{C_p^2 C_q^2}{K^2} \sum_{k,k'=1}^{K} \mathcal{V}^2\left(u_k^t X, u_{k'}^t X\right) \mathcal{V}^2\left(v_k^t Y, v_{k'}^t Y\right).$$

Remark 3.3.19. *Let us recall that if X and Y are independent, the asymptotic distribution of Ω_n is*

$$n\Omega_n \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1).$$

Theorem 3.3.18 shows that under the null hypotheses, $\overline{\Omega}_n$ enjoys the same convergence rate with Ω_n.

It is also worth noting that in Theorem 3.3.18, the number of random projections to be fixed in order to achieve computational efficiency.

There usually does not exist a close-form expression for $\sum_{i=1}^{\infty} \bar{\lambda}_i Z_i^2$, but we can approximate it with the Gamma distribution whose first two moments matched. Thus, we have that $\sum_{i=1}^{\infty} \bar{\lambda}_i Z_i^2$ could be approximated by Gamma(α, β) with probability density function

$$\frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, x > 0,$$

where

$$\alpha = \frac{1}{2} \frac{(\sum_{i=1}^{\infty} \bar{\lambda}_i)^2}{\sum_{i=1}^{\infty} \bar{\lambda}_i^2}, \beta = \frac{1}{2} \frac{\sum_{i=1}^{\infty} \bar{\lambda}_i}{\sum_{i=1}^{\infty} \bar{\lambda}_i^2}. \tag{3.3.24}$$

See [12, Section 3] for an empirical justification on this Gamma approximation. See [11] for a survey on different approximation methods of weighted sum of chi-square distribution.

The following result shows that both $\sum_{i=1}^{\infty} \bar{\lambda}_i$ and $\sum_{i=1}^{\infty} \bar{\lambda}_i^2$ could be estimated from data, see appendix for the corresponding justification.

Proposition 3.3.20. *One can approximate* $\sum_{i=1}^{\infty} \bar{\lambda}_i$ *and* $\sum_{i=1}^{\infty} \bar{\lambda}_i$ *as follows:*

$$\begin{aligned}\sum_{i=1}^{\infty} \bar{\lambda}_i \approx & \frac{C_p C_q}{K n^2 (n-1)^2} \sum_{k=1}^{K} a_{\cdot\cdot}^{u_k} b_{\cdot\cdot}^{v_k}, \\ \sum_{i=1}^{\infty} \bar{\lambda}_i^2 \approx & \frac{K-1}{K} \Omega_n(X,X)\Omega_n(Y,Y) \\ & + \frac{C_p^2 C_q^2}{K} \sum_{k=1}^{K} \Omega_n(u_k^t X, u_k^t X)\Omega_n(v_k^t Y, v_k^t Y).\end{aligned}$$

3.4 Simulations for Randomly Projected Distance Covariance

Our numerical studies follow the works of [69, 29, 80]. In Section 3.4.1, we study how the performance of the proposed estimator is influenced by some parameters, including the sample size, the dimensions of the data, as well as the number of random projections in our algorithm. We also study and compare the computational efficiency of the direct method and the proposed method in Section 3.4.2. The comparison of the corresponding independence test with other existing methods will be included in Section 3.4.3.

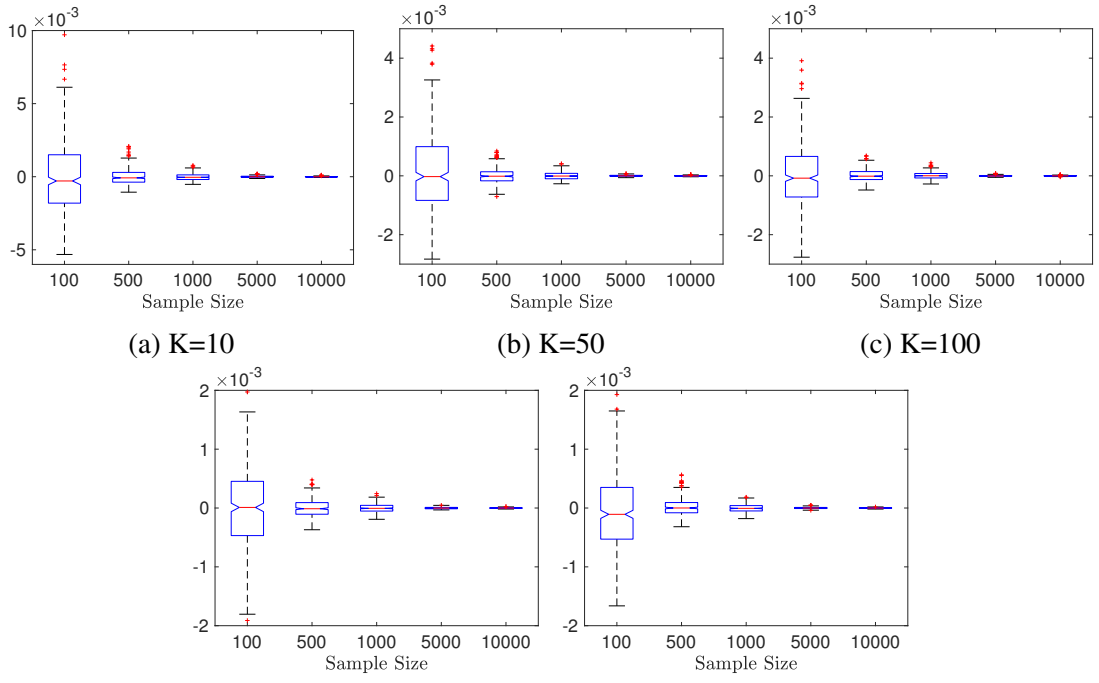
3.4.1 Impact of Sample Size, Data Dimensions and the Number of Monte Carlo Iterations

In this part, we will use some synthetic data to study impact of sample size n, data dimensions (p, q) and the number of the Monte Carlo iterations K on the convergence and test power of our proposed test statistic $\overline{\Omega}_n$. The significance level is set to be $\alpha_s = 0.05$. Each experiment is repeated for $N = 400$ times to get reliable mean and variance of estimators.

In first two examples, we fix data dimensions $p = q = 10$ and let the sample size n vary in 100, 500, 1000, 5000, 10000 and let the number of the Monte Carlo iterations K vary in 10, 50, 100, 500, and 1000. The data generation mechanism is described as follows, and it generates independent variables.

Example 3.4.1. *We generate random vectors* $X \in \mathbb{R}^{10}$ *and* $Y \in \mathbb{R}^{10}$. *Each entry* X_i *follows Unif*$(0, 1)$, *independently. Each entry* $Y_i = Z_i^2$, *where* Z_i *follows Unif*$(0, 1)$, *independently.*

See Figure 5 for the boxplots of the outcomes of Example 3.4.1. In each subfigure, we fix the Monte Carlo iteration number K and let the number of observations n grow. It is worth noting that the scale of each subfigure could be different in order to display the entire boxplots. This experiment shows that the estimator converges to 0 regardless of the number of the Monte Carlo iterations. It also suggests that $K = 50$ Monte Carlo iterations should suffice in the independent cases.



(a) K=10 (b) K=50 (c) K=100

(d) K=500 (e) K=1000

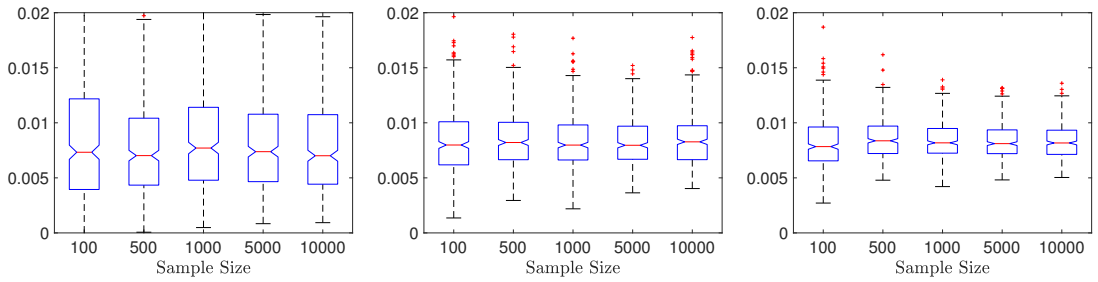
Figure 5: Boxplots of estimators in Example 3.4.1. Dimensions of X and Y are fixed to be $p = q = 10$; the result is based on 400 repeated experiments.

The following example is to study dependent random variables.

Example 3.4.2. *We generate random vectors* $X \in \mathbb{R}^{10}$ *and* $Y \in \mathbb{R}^{10}$. *Each entry* X_i

follows Unif$(0,1)$*, independently. Let* Y_i *denote the* i*-th entry of* Y*. We let* $Y_1 = X_1^2$ *and* $Y_2 = X_2^2$*. For the rest entry of* Y*, we have* $Y_i = Z_i^2$*,* $i = 3, \ldots, 10$*, where* Z_i *follows Unif*$(0,1)$*, independently.*

See Figure 6 for the boxplots of the outcomes of Example 3.4.2. In each subfigure, we fix the number of the Monte Carlo iterations K and let the number of observations n grow. This example shows that when K is fixed, the variation of the estimator remains regardless of the sample size n. In the dependent cases, the number of the Monte Carlo iterations K plays a more important role in estimator convergence than sample size n.



(a) K=10 (b) K=50 (c) K=100

(d) K=500 (e) K=1000

Figure 6: Boxplots of our estimators in Example 3.4.2. Dimension of X and Y are fixed to be $p = q = 10$; the result is based on 400 repeated experiments.

The outcomes of Example 3.4.1 and 3.4.2 confirm the theoretical results that the proposed estimator converges to 0 as sample size n grows in the independent case; and converges to some nonzero number as the number of the Monte Carlo iterations K grows in the dependent case.

In the following two examples, we fix the sample size $n = 2000$ as we noticed that our method is more efficient than direct method when n is large. We fix the number of the Monte Carlo iterations $K = 50$ and relax the restriction on the data dimensions to allow $p \neq q$ and let p and q vary in $(10, 50, 100, 500, 1000)$. We continue on with an independent case as follows.

Example 3.4.3. *We generate random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$. Each entry of X follows Unif$(0, 1)$, independently. Each entry $Y_i = Z_i^2$, where Z_i follows Unif$(0, 1)$, independently.*

See Figure 7 for the boxplots of the outcomes of Example 3.4.3. In each subfigure, we fix the dimension of X and let the dimension of Y grow. It is worth noting that the scale of each subfigure could be different in order to display the entire boxplots. It shows that the proposed estimator converges fairly fast in the independent case regardless of the dimension of the data.

The following presents a dependent case. In this case, only a small number of entries in X and Y are dependent, which means that the dependency structure between X and Y is low-dimensional though X or Y could be of high dimensions.

Example 3.4.4. *We generate random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$. Each entry of X follows Unif$(0, 1)$, independently. We let the first 5 entries of Y to be the square of first 5 entries of X and let the rest entries of Y to be the square of some independent Unif$(0, 1)$ random variables. Specifically, we let $Y_i = X_i^2, i = 1, \ldots, 5$, and, $Y_i = Z_i^2, i = 6, \ldots, q$, where Z_i's are drawn independently from Unif$(0, 1)$.*

See Figure 8 for the boxplots of the outcomes of Example 3.4.4. In each subfigure,

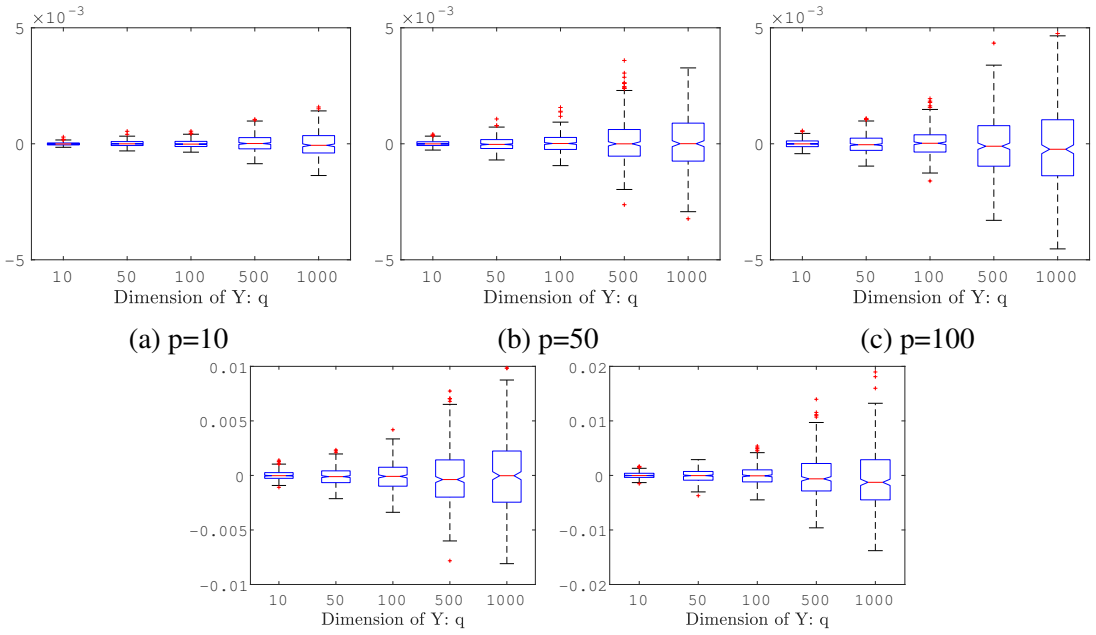
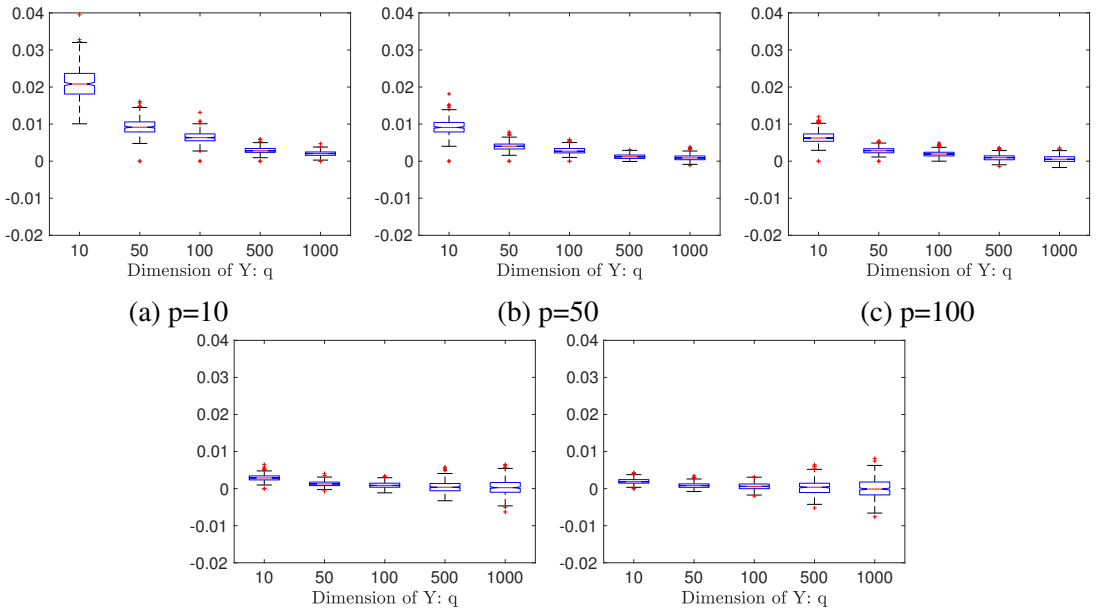


Figure 7: Boxplot of Estimators in Example 3.4.3: both sample size and the number of Monte Carlo iterations is fixed, $n = 2000$, $K = 50$; the result is based on 400 repeated experiments.

we fix the dimension of X and let the dimension of Y grow. The test power of proposed test against data dimensions can be seen in Table 6. It is worth noting that when sample size is fixed, the test power of our method decays as the dimension of X and Y increase. We use the Direct Distance Covariance (DDC) defined in (3.1.15) on the same data. As a contrast, the test power of DDC is 1.000 even $p = q = 1000$. This example raises a limitation of random projection: it may fail to detect the low dimensional dependency in high dimensional data. A possible remedy for this issue is performing dimension reduction before applying the proposed method. We do not research further along this direction since it is beyond the scope of this dissertation.

Note this chapter focuses on independence testing. Therefore the independent case is of more relevance.



(a) p=10 (b) p=50 (c) p=100

(d) p=500 (e) p=1000

Figure 8: Boxplots of the proposed estimators in Example 3.4.4: both sample size and the number of the Monte Carlo iterations are fixed: $n = 2000$ and $K = 50$; the result is based on 400 repeated experiments.

Table 6: Test Power in Example 3.4.4: this result is based 400 repeated experiments; the significant level is 0.05.

Dimension of X: p	Dimension of Y: q				
	10	50	100	500	1000
10	1.0000	1.0000	1.0000	1.0000	0.9975
50	1.0000	1.0000	1.0000	0.7775	0.4650
100	1.0000	1.0000	0.9925	0.4875	0.1800
500	0.9950	0.8150	0.4425	0.1225	0.0975
1000	0.9900	0.4000	0.2125	0.0900	0.0475

3.4.2 Comparison with Direct Method

In this section, we would like to illustrate the computational and space efficiency of the proposed method (RPDC). RPDC is much faster than the direct method (DDC, eq. (3.1.15)) when the sample size is large. It is worth noting that DDC is infeasible when the sample size is too large as its space complexity is $O(n^2)$. See Table 7 for a comparison of computing time (unit: second) against the sample size n. This experiment is run on a laptop (MacBook

Pro Retina, 13-inch, Early 2015, 2.7 GHz Intel Core i5, 8 GB 1867 MHz DDR3) with MATLAB R2016b (9.1.0.441655).

Table 7: Speed Comparison: the Direct Distance Covariance (Ω_n) versus the Randomly Projected Distance Covariance ($\overline{\Omega}_n$). This table is based on 100 repeated experiments, the dimensions of X and Y are fixed to be $p = q = 10$ and the number of Monte Carlo iterations in RPDC is $K = 50$. The numbers outside the parentheses are the average and the numbers inside the parentheses are the sample standard deviations.

Sample size	Ω_n	$\overline{\Omega}_n$
100	0.0043 (0.0047)	0.0207 (0.0037)
500	0.0210 (0.0066)	0.0770 (0.0086)
1000	0.0624 (0.0047)	0.1685 (0.0141)
2000	0.2349 (0.0133)	0.3568 (0.0169)
4000	0.9184 (0.0226)	0.7885 (0.0114)
8000	7.2067 (0.4669)	1.7797 (0.0311)
16000	—	3.7539 (0.0289)

3.4.3 Comparison with Other Independence Tests

In this part, we compare the statistical test power of the proposed test (RPDC) with Hilbert-Schmidt Independence Criterion (HSIC) ([29]) as HSIC is gaining attention in machine learning and statistics communities. We also compare with Randomized Dependence Coefficient (RDC) ([49]), which utilizes the technique of random projection as we do. Two classical tests for multivariate independence, which are described below, are included in the comparison, as well as the Direct Distance Covariance (DDC) defined in (3.1.15).

- Wilks Lambda (WL): the likelihood ratio test of hypotheses $\Sigma_{12} = 0$ with μ unknown is based on

$$\frac{\det(S)}{\det(S_{11})\det(S_{22})} = \frac{\det(S_{22} - S_{21}S_{11}^{-1}S_{12})}{\det(S_{22})},$$

  where $\det(\cdot)$ is the determinant, S, S_{11} and S_{22} denote the sample covariances of (X, Y), X and Y, respectively, and S_{12} is the sample covariance $\hat{\text{Cov}}(X, Y)$. Under

multivariate normality, the test statistic

$$W = -n \log \det(I - S_{22}^{-1} S_{21} S_{11}^{-1} S_{12})$$

has the Wilks Lambda distribution $\Lambda(q, n-1-p, p)$, see [88].

- Puri-Sen (PS) statistics: [59], Chapter 8, proposed similar tests based on more general sample dispersion matrices T. In that test S, S_{11}, S_{12} and S_{22} are replaced by T, T_{11}, T_{12} and T_{22}, where T could be a matrix of Spearman's rank correlation statistics. Then, the test statistic becomes

$$W = -n \log \det(I - T_{22}^{-1} T_{21} T_{11}^{-1} T_{12}).$$

The critical values of the Wilks Lambda (WL) and Puri-Sen (PS) statistics are given by Bartlett's approximation ([53], Section 5.3.2b): if n is large and $p, q > 2$, then

$$-(n - \frac{1}{2}(p+q+3)) \log \det(I - S_{22}^{-1} S_{21} S_{11}^{-1} S_{12})$$

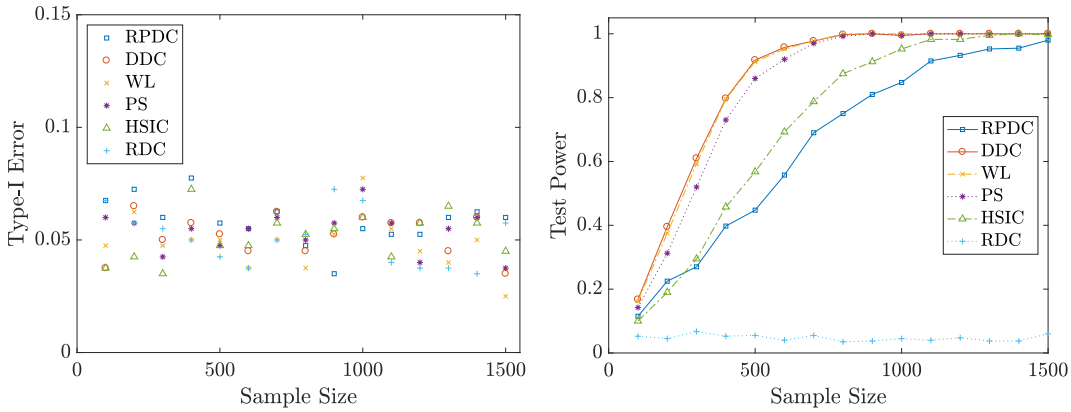
has an approximate $\chi^2(pq)$ distribution.

The reference distributions of RDC and HSIC are approximated by 200 permutations. And the reference distributions of DDC and RPDC are approximated by the Gamma Distribution. The significant level is set to be $\alpha_s = 0.05$ and each experiment is repeated for $N = 400$ times to get reliable type-I error / test power.

We start with an example that (X, Y) is multivariate normal. In this case, WL and PS are expected to be optimal as the distributional assumptions of these two classical tests are satisfied. Surprisingly, DDC has comparable performance with the aforementioned two methods. RPDC can achieve satisfactory performance when sample size is a reasonably large.

Example 3.4.5. *We set the dimension of the data to be* $p = q = 10$. *We generate random vectors* $X \in \mathbb{R}^{10}$ *and* $Y \in \mathbb{R}^{10}$ *from the standard multivariate normal distribution* $\mathcal{N}(0, \mathbf{I}_{10})$. *The joint distribution of* (X, Y) *is also normal and we have* $Cor(X_i, Y_i) = \rho, i = 1, \ldots, 10$, *and the rest correlation are all* 0. *We set the value of* ρ *to be* 0 *and* 0.1 *to represent independent and correlated scenarios, respectively. The sample size* n *is set to be from* 100 *to* 1500 *with an increment of* 100.

Figure 9 plots the type-I error in subfigure (a) and test power in subfigure (b) against sample size. In the independence case ($\rho = 0.0$), the type-I error of each test is always around the significance level $\alpha_s = 0.05$, which implies the Gamma approximation works well for the asymptotic distributions. In the dependent case ($\rho = 0.1$), the overall performance of RPDC is close to HSIC and RPDC outperforms when sample size is smaller and underperforms when sample size is larger. Unfortunately, RDC's test power is unsatisfactory.



(a) Independence: $\rho = 0.0$ (b) Dependence: $\rho = 0.1$

Figure 9: Type-I Error/Test Power vs Sample Size n in Example 3.4.5. The result is based on 400 repeated experiments.

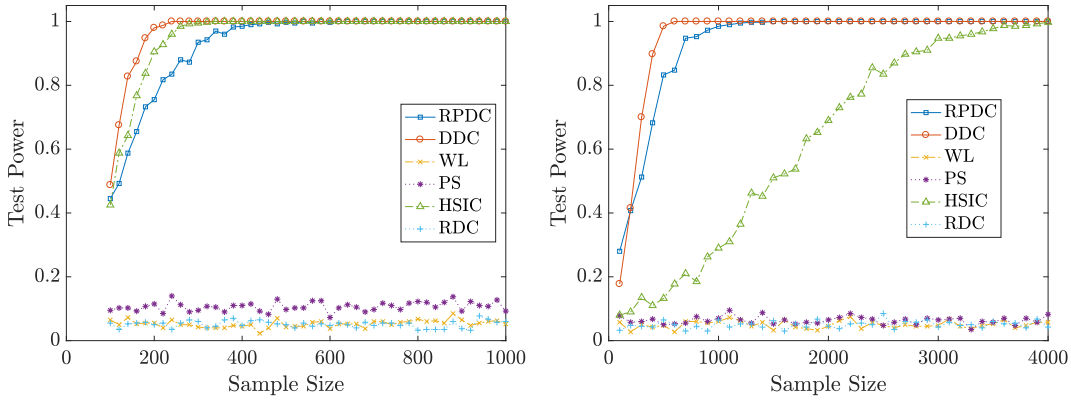
Next, we compare those methods when (X, Y) is no longer multivariate normal and

the dependency between X and Y is non-linear. We add a noise term to compare their performance in both the low and the high noise-to-signal ratio scenarios. In this case, DDC and RPDC are much better than WL, PS and RDC. The performance of HSIC is close to DDC and RPDC when the noise level is low but much worse than those two when the noise level is high.

Example 3.4.6. *We set the dimension of data to be* $p = q = 10$*. We generate random vector* $X \in \mathbb{R}^{10}$ *from the standard multivariate normal distribution* $\mathcal{N}(0, \mathbf{I}_{10})$*. Let the* i*-th entry of* Y *be* $Y_i = \log(X_i^2) + \epsilon_i, i = 1, \ldots, q$*, where* ϵ_i*'s are independent random errors,* $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$*. We set the value of* σ *to be* 1 *and* 3 *to represent low and high noise ratios, respectively. In the* $\sigma = 1$ *case, the sample size* n *is from* 100 *to* 1000 *with an increment* 20*; and in the* $\sigma = 3$ *case, the sample size* n *is from* 100 *to* 4000 *with an increment* 100*.*

Figure 10 plots the test power of each test against sample size. In both low and high noise cases, none of WL, PS and RDC has any test power. In the low noise case, all of RPDC, DDC and HSIC have satisfactory test power (> 0.9) when sample size is greater than 300. In the high noise case, RPDC and DDC could achieve more than 0.8 in test power once sample size is greater than 500 while the test power of HSIC reaches 0.8 when the sample size is more than 2000.

In the following example, we generate the data in the similar way with Example 3.4.6 but the difference is that the dependency is changing over time. Specifically, X and Y are independent at the beginning but they become dependent after some time point. Since all those tests are invariant with the order of the observations, this experiment simply means that only a proportion of observations are dependent while the rest are not.

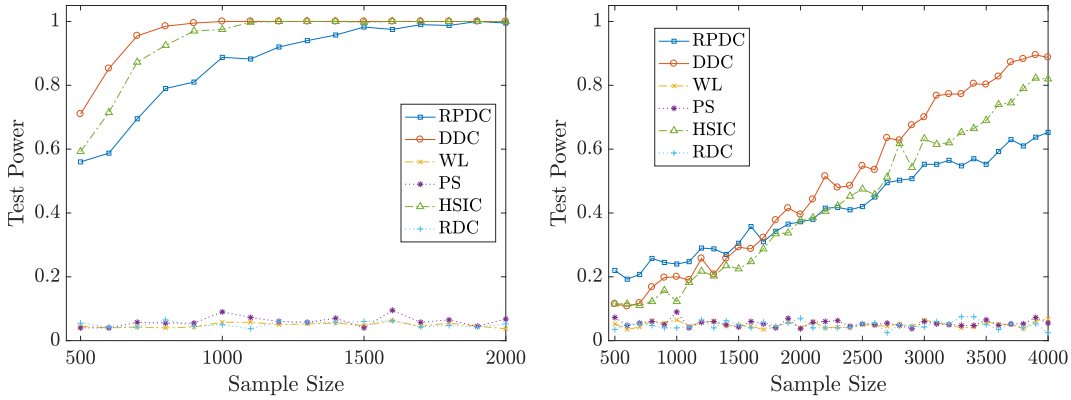


(a) Low Noise: $\sigma = 1$ (b) High Noise: $\sigma = 3$

Figure 10: Test Power vs Sample Size n in Example 3.4.6. The significance level is $\alpha_s = 0.05$. The result is based on $N = 400$ repeated experiments.

Example 3.4.7. *We set the dimension of data to be* $p = q = 10$. *We generate random vector* $X_t \in \mathbb{R}^{10}, t = 1, \ldots, n$, *from the standard multivariate normal distribution* $\mathcal{N}(0, \mathbf{I}_{10})$. *Let the* i*-th entry of* Y_t *be* $Y_{t,i} = \log(Z_{t,i}^2) + \epsilon_{t,i}, t = 1, \ldots, T$ *and* $Y_{t,i} = \log(X_{t,i}^2) + \epsilon_{t,i}, t = T+1, \ldots, n$, *where* Z_t $i.i.d.$ $\sim \mathcal{N}(0, \mathbf{I}_{10})$ *and* $\epsilon_{t,i}$*'s are independent random errors,* $\epsilon_{t,i} \sim \mathcal{N}(0, 1)$. *We set the value of* T *to be* $0.5n$ *and* $0.8n$ *to represent early and late dependency transition, respectively. In the early change case, the sample size* n *is from* 500 *to* 2000 *with an increment* 100*; and in the late change case, the sample size* n *is from* 500 *to* 4000 *with an increment* 100.

Figure 11 plots the test power of each test against sample size. In both early and late change cases, none of WL, PS and RDC has any test power. In the early change case, all of RPDC, DDC and HSIC have satisfactory test power (> 0.9) when sample size is greater than 1500. In the late change case, DDC and HSIC could achieve more than 0.8 in test power once sample size reaches 4000 while the test power of RPDC is only 0.6 when the sample size is 4000. As expected, the performance of DDC is better than RPDC in both cases and the performance of HSIC is between DDC and RPDC.



(a) Early Change: $T = 0.5n$ (b) Late Change: $T = 0.8n$

Figure 11: Test Power vs Sample Size n in Example 3.4.7. The significance level is $\alpha_s = 0.05$. The result is based on $N = 400$ repeated experiments.

Remark 3.4.8. *The experiments in this subsection show that though the RPDC underperforms the DDC when the sample size is relatively small, the RPDC could achieve the same test power with the DDC when the sample size is sufficiently large. Considering the computational advantage of the RPDC (it has a lower order of computational complexity), when the sample size is large enough, RPDC can be superior over the DDC.*

3.5 Discussions on Randomly Projected Distance Covariance

3.5.1 A Discussion on the Computational Efficiency

We compare the computational efficiency of proposed method (RPDC) and direct method (DDC) in Section 3.4.2. We will discuss this issue here.

As $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ are multivariate random variables, the effect of p and q on computing time could be significant when p and q are not negligible comparing to sample size n. Now, we analyze the computational efficiency of DDC and RPDC by taking p and q into consideration. The computational complexity of DDC becomes $O(n^2(p+q))$ and that of RPDC becomes $O(nK(\log n + p + q))$. Let us denote the total number of operations

in DDC by O_1 and that in RPDC by O_2. Then, by sacrificing the technical rigor, one may assume that there exist constants L_1 and L_2 such that

$$O_1 \approx L_1 n^2 (p+q), \text{ and } O_2 \approx L_2 n K (\log n + p + q).$$

There is no doubt that O_2 will eventually much less than O_1 as the sample size n grows. Due to the complexity of the fast algorithm, we may expect $L_2 > L_1$, which means that the computational time of the RPDC can be even larger than the one for the DDC when the sample size is relatively small. Then we need to study the problem: what is the break-even point in terms of sample size n when the RPDC and the DDC has the same computational time?

Let $n_0 = n_0(p+q, K)$ denote the break-even point, which is a function of $p+q$ and number of Monte Carlo iterations K. For simplicity, we fix $K = 50$ since 50 iterations could achieve satisfactory test power as we showed in Example 3.4.4. Consequently n_0 becomes a function solely depending on $p+q$. Since it is hard to derive the close form of n_0, we derive it numerically instead. For fixed $p+q$, we let the sample size vary and record the difference between the running time of two methods. We fit the difference of running time against sample size with smoothing spline. The root of this spline is the numerical value of n_0 at $p+q$.

We plot the n_0 against $p+q$ in Figure 12. As the figure predicts, the break-even sample size decreases as the data dimension increases, which implies that our proposed method is more advantageous than the direct method when random variables are of high dimension. However, as showed in Example 3.4.4, the random projection based method does not perform well when high dimensional data have low dimensional dependency structure. This

indicates that one need to be cautious to use the proposed method when the dimension is high.

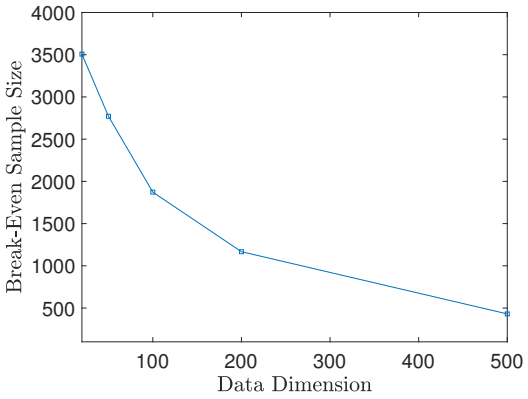


Figure 12: Break-Even Sample Size n_0 against Data Dimension $p + q$. This figure is based on 100 repeated experiments.

3.5.2 Connections with Existing Literature

It turns out that distance-based methods are not the only choices in independence testing. See [43] and the references therein to see alternatives. On the other hand, in our numerical experiments, it is evident that the distance-correlated-based approaches compare favorably against many other popular contemporary alternatives. Therefore it is meaningful to study the improvements of the distance-correlated-based approaches.

Our proposed method utilizes random projections, which bears similarity with the randomized feature mapping strategy [60] that was developed in the machine learning community. Such an approach has been proven to be effective in kernel-related methods [1, 10, 26, 24]. However, a closer examination will reveal the following difference: most of the aforementioned work are rooted on the Bochner's theorem [66] from harmonic analysis, which states that a continuous kernel in the Euclidean space is positive definite if and only if the kernel function is the Fourier transform of a non-negative measure. In this chapter,

we will deal with distance function which is not a positive definite kernel. We managed to derive a counterpart to the randomized feature mapping, which was the influential idea that has been used in [60].

Random projections have been used in [48] to develop a powerful two-sample test in high dimensions. They derived an asymptotic power function for their proposed test, and then provide sufficient conditions for their test to achieve greater power than other state-of-the-art tests. They then used the receiver operating characteristic (ROC) curves (that are generated from their simulated data) to evaluate its performance against competing tests. The derivation of the asymptotic relative efficiency (ARE) is of its own interests. Despite the usage of random projection, the details of their methodology is very different from the one that is studied in the present chapter.

Several distribution-free tests that are based on sample space partitions were suggested in [30] for univariate random variables. They proved that all suggested tests are consistent and showed the connection between their tests and the mutual information (MI). Most importantly, they derived fast (polynomial-time) algorithms, which are essential for large sample size, since the computational complexity of the naive algorithm is exponential in sample size. Efficient implementations of all statistics and tests described in the aforementioned paper are available in the R package HHG, which can be freely downloaded from the Comprehensive R Archive Network, http://cran.r-project.org/. Null tables can be downloaded from the first author's web site.

Distance-based independence/dependence measurements sometimes have been utilized in performing a greedy feature selection, often via dependence maximization [34], [95] and [45], and it has been effective on some real-world datasets. This paper simply mentions

such a potential research line, without pursuing it.

Paper [87] derives an efficient approach to compute for the conditional distance correlations. We noted that there are strong resemblances between the distance covariances and its conditional counterpart. The search for a potential extension of the work in this paper to conditional distance correlation can be a meaningful future topic of research.

3.6 Conclusions on Randomly Projected Distance Covariance

A significant contribution of this chapter is that we demonstrated that the multivariate variables in the independence tests need not imply the higher-order computational desideratum of the distance-based methods.

Distance-based methods are indispensable in statistics, particular in test of independence. When the random variables are univariate, efficient numerical algorithms exist. It is an open question when the random variables are multivariate. We study the random projection approach to tackle the above problem. It first turn the multivariate calculation problem into univariate calculation one via random projections. Then they study how the average of those statistics out of the projected (therefore univariate) samples can approximate the distance-based statistics that were intended to use. Theoretical analysis was carried out, which shows that the loss of asymptotic efficiency (in the form of the asymptotic variance of the test statistics) is likely insignificant. The new method can be numerically much more efficient, when the sample size is large; considering large sample sizes are well-expected under this information (or big-date) era. Simulation studies validate the theoretical statements. The theoretical analysis takes advantage of some newly available results, such as the equivalence of the distance-based methods with the reproducible kernel Hilbert spaces

[69]. The numerical methods utilizes a recently appeared fast algorithm in [34].

CHAPTER IV

ENERGY STATISTICS AND TWO-SAMPLE TESTING

This chapter is organized as follows. We will review the definition and property of energy distance and energy statistics in Section 4.1. In Section 4.2, we describe the details of fast algorithms and corresponding two-sample tests. Asymptotic properties of proposed test statistic will be studied in Section 4.3. In Section 4.4, we will some numerical examples with simulated data to illustrate the computational and statistical efficiency of the proposed test. Discussions could be found in Section 4.5 and we will conclude in Section 4.6.

Throughout this chapter, we adopt the following notations. We denote $c_p = \frac{\pi^{(p+1)/2}}{\Gamma((p+1)/2)}$ and $C_p = \frac{c_1 c_{p-1}}{c_p} = \frac{\sqrt{\pi}\Gamma((p+1)/2)}{\Gamma(p/2)}$ as two constants, where $\Gamma(\cdot)$ denotes the Gamma function. We also denote $|\cdot|$ as the Euclidian norm. For any vector v, v^T is its transpose.

4.1 Review of Energy Distance and Energy Statistics

Energy distance is initially proposed by [75] to measure the distance between two multivariate distributions. We follow the definition of energy distance in [78].

Definition 4.1.1. *[78, Definition 1] Suppose $X, Y \in \mathbb{R}^p$ are two real-valued independent random variables with finite means, i.e., $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|Y|] < \infty$, then the energy distance between X and Y is defined as*

$$\mathcal{E}(X, Y) = 2\mathbb{E}[|X - Y|] - \mathbb{E}[|X - X'|] - \mathbb{E}[|Y - Y'|],$$

where X' and Y' are independent and identical copies of X and Y, respectively.

[78] also show that energy distance is equivalent to the weighted L_2-distance of the characteristic functions.

Proposition 4.1.2. *[78, Proposition 1] Suppose $X, Y \in \mathbb{R}^p$ are two real-valued independent random variables with finite means and X' and Y' are independent identical copies of X and Y. Let $\tilde{f}_X(\cdot)$ and $\tilde{f}_Y(\cdot)$ denote the characteristic function of X and Y, respectively, we have*

$$\frac{1}{c_p}\int_{\mathbb{R}^p}\frac{|\tilde{f}_X(t)-\tilde{f}_Y(t)|^2}{|t|^2}dt = 2\mathbb{E}[|X-Y|]-\mathbb{E}[|X-X'|]-\mathbb{E}[|Y-Y'|]=\mathcal{E}(X,Y).$$

Thus, $\mathcal{E}(X,Y)\geq 0$ with equality to zero if and only if X and Y are identically distributed, i.e., $\tilde{f}_X \equiv \tilde{f}_Y$.

Suppose that we observe samples $X_1,\ldots,X_n \overset{i.i.d.}{\sim} F$ and $Y_1,\ldots,Y_m \overset{i.i.d.}{\sim} G$, the energy statistics is usually defined as follows (see [78], (6.1)).

$$\frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}|X_i-Y_j| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i-X_j| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|Y_i-Y_j|.$$

However, above estimator is NOT an unbiased estimator of $\mathcal{E}(X,Y)$. To mitigate this issue, let $h(X_1,X_2,Y_1,Y_2)=\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}|X_i-Y_j|-|X_1-X_2|-|Y_1-Y_2|$ be a two-sample kernel (see [84, Chapter 12.2]), which an unbiased estimator, i.e., $\mathbb{E}[h]=\mathcal{E}(X,Y)$, then it is easy to verify that

$$\begin{aligned}&\frac{1}{\binom{n}{2}\binom{m}{2}}\sum_{i_1<i_2,j_1<j_2}h(X_{i_1},X_{i_2},Y_{j_1},Y_{j_2})\\ =&\frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}|X_i-Y_j| - \frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n}|X_i-X_j| - \frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}|Y_i-Y_j|.\end{aligned}$$

is a U-statistic and an unbiased estimator of $\mathcal{E}(X,Y)$. Thus, we will use the following definition of energy statistics throughout this chapter.

Definition 4.1.3 (Unbiased Energy Statistics)**.** *Given samples* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$ *and* $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} G$*, the energy statistics between* X *and* Y *could be defined as*

$$\mathcal{E}_{n,m}(X,Y) = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}|X_i - Y_j| - \frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n}|X_i - X_j| - \frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}|Y_i - Y_j|.$$

4.2 Efficient Computational Methods for Energy Statistics

In this section, we will describe the efficient algorithms for energy statistics of both univariate and multivariate random variables in Section 4.2.1 and Section 4.2.2, respectively. We will also propose two different methods based on the efficient algorithm of multivariate random variables for two-sample test in Section 4.2.3.

4.2.1 A Fast Algorithm for Univariate Random Variables

We will start with the fast algorithm for univariate random variables. Let us recall the definition of energy statistics first. Given univariate random variables $X_1, \ldots, X_n \in \mathbb{R}$ and $Y_1, \ldots, Y_m \in \mathbb{R}$, the energy statistic of X and Y is defined below:

$$\mathcal{E}_{n,m}(X,Y) = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}|X_i - Y_j| - \frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n}|X_i - X_j| - \frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}|Y_i - Y_j|.$$

For simplicity of notation, we denote above term with $\mathcal{E}_{n,m}$. The following algorithm can compute $\mathcal{E}_{n,m}$ with an average order of complexity $O(N \log N)$, where $N = n + m$. The main idea of this algorithm is sorting the observations first and use a linear-time algorithm to compute the energy statistic with sorted observations.

(1) Sort X_i's and Y_j's, so that we have order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ and $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(m)}$.

(2) Compute the second term of $\mathcal{E}_{n,m}$ as follows:

$$E_2 = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i(n-i) \left| X_{(i+1)} - X_{(i)} \right|.$$

(3) Compute the third term of $\mathcal{E}_{n,m}$ as follows:

$$E_3 = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} i(m-i) \left| Y_{(i+1)} - Y_{(i)} \right|.$$

(4) In this step, we will compute the first term of $\mathcal{E}_{n,m}$.

(a) Merge two ordered series $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ and $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(m)}$ into a single ordered series $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n+m)}$, where each $Z_{(k)}$ is either from $X_{(i)}$'s or from $Y_{(j)}$'s. At the same time, one can generate a sequence $I_i, i = 1, 2, \ldots, n+m$, where I_i records the size of the subset of $Z_{(1)}$ through $Z_{(i)}$ that are from $X_{(i)}$'s.

(b) Compute the first term of $\mathcal{E}_{n,m}$,

$$E_1 = \frac{2}{nm} \sum_{i=1}^{n+m-1} \left[I_i(m - i + I_i) + (i - I_i)(n - I_i) \right] \left| Z_{(i+1)} - Z_{(i)} \right|.$$

(5) Compute the energy statistic,

$$\mathcal{E}_{n,m} = E_1 - E_2 - E_3.$$

A stand-alone description of above algorithm can be found in Algorithm 4 of Appendix A.2. Our result could be summarized in the following theorem.

Theorem 4.2.1. *Given univariate random variables $X_1, \ldots, X_n \in \mathbb{R}$ and $Y_1, \ldots, Y_m \in \mathbb{R}$, there exists an algorithm with complexity $O(N \log N)$, where $N = n + m$, for computing the energy statistic defined in Definition 4.1.3.*

See Appendix D.1 for the proof and detailed explanations.

4.2.2 A Fast Algorithm for Multivariate Random Variables

In this part, we will introduce a fast algorithm for the energy statistics of multivariate random variables. We will show later in Theorem 4.3.9 that the estimator produced by this algorithm converges fairly fast. The main idea works as follows: first, projecting the multivariate observations along some random directions; then, using the fast algorithm described in Section 4.2.1 to compute the energy statistics of projections; last, averaging those energy statistics from different projecting directions.

Formally, suppose we have observations $X_1, \ldots, X_n \in \mathbb{R}^p$ and $Y_1, \ldots, Y_m \in \mathbb{R}^p$ and let K denote the pre-determined number of random projections, the algorithm is as follows:

(1) For each k ($1 \leq k \leq K$), randomly generate projecting direction u_k from Uniform($\mathcal{S}_p$), where $\mathcal{S}_p$ is the unit sphere in $\mathbb{R}^p$.

(2) Let $u_k^T X$ and $u_k^T Y$ denote the projections of X and Y. That is,

$$u_k^T X = (u_k^T X_1, \ldots, u_k^T X_n), \text{ and } u_k^T Y = (u_k^T Y_1, \ldots, u_k^T Y_m).$$

Note that $u_k^T X$ and $u_k^T Y$ are now univariate.

(3) Utilize the fast algorithm described in Section 4.2.1 to compute the energy statistic of $u_k^T X$ and $u_k^T Y$. Formally, we denote

$$\mathcal{E}_{n,m}^{(k)} = C_p \mathcal{E}_{n,m}(u_k^T X, u_k^T Y),$$

where C_p is the constant defined at the beginning of Chapter 4.

(4) Repeat above steps for K times. The final estimator is

$$\overline{\mathcal{E}}_{n,m} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{E}_{n,m}^{(k)},$$

which is refered as Randomly Projected Energy Statistics (RPES). To emphasize the dependency of the above quantity with number of random projections K, we sometimes use another notation $\overline{\mathcal{E}}_{n,m;K} \triangleq \overline{\mathcal{E}}_{n,m}$.

A stand-alone description of above algorithm can be found in Algorithm 5 of Appendix A.2. The following theorem summarizes above result.

Theorem 4.2.2. *For multivariate random variables $X_1, \ldots, X_n \in \mathbb{R}$ and $Y_1, \ldots, Y_m \in \mathbb{R}$, there exists an algorithm with complexity $O(KN \log N)$, where $N = n+m$, for computing aforementioned $\overline{\mathcal{E}}_{n,m}$, where K is a pre-determined number of random projections.*

We omit the proof since above theorem is a straight-forward conclusion from Theorem 4.2.1.

4.2.3 Two-Sample Test based on Randomly Projected Energy Statistics (RPES)

The randomly projected energy statistic $\overline{\mathcal{E}}_{n,m}$ could be applied in the two-sample test. Let us recall that we would like to test the null hypotheses $\mathcal{H}_0$ — X and Y are identically distributed — against its alternative. The threshold of the test statistic could be determined by either permutation or the Gamma approximation of asymptotic distribution. Let us recall that we observe $X_1, \ldots, X_n \in \mathbb{R}^p$ and $Y_1, \ldots, Y_m \in \mathbb{R}^p$. Let $Z = (Z_1, \ldots, Z_{n+m}) = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ denote the collection of all observations. Let $\overline{\mathcal{E}}_{n,m}$ denote the proposed estimator defined in Section 4.2.2. Suppose α_s is the pre-specified significance level of the test and L is the pre-determined number of permutations. The following algorithm describes a two-sample test using permutation to generate the threshold.

(1) For each l, $1 \leq l \leq L$, generate a random permutation of observations: let

$$(X^{\star,l}, Y^{\star,l}) = (X_1^{\star,l}, \ldots, X_n^{\star,l}, Y_1^{\star,l}, \ldots, Y_m^{\star,l})$$

be a random permutation of $(Z_1, \ldots, Z_{n+m})$.

(2) Using the algorithm in Section 4.2.2, we compute the estimator for $X^{\star,l}$ and $Y^{\star,l}$: $D^{(l)} = \overline{\mathcal{E}}_{n,m}(X^{\star,l}, Y^{\star,l})$. Note that under null hypotheses, $X^{\star,l}$ and $Y^{\star,l}$ are identically distributed.

(3) Reject null hypotheses $\mathcal{H}_0$ if and only if

$$\frac{1 + \sum_{l=1}^{L} I(\overline{\mathcal{E}}_{n,m} > D^{(l)})}{1 + L} > \alpha_s.$$

See Algorithm 6 of Appendix A.2 for a stand-alone description of above algorithm.

We can also find the threshold for test statistic based on the Gamma approximation of its asymptotic distribution. Let K denote the pre-determined number of random projections. The algorithm is as follows:

(1) For each k, $1 \leq k \leq K$, randomly generate u_k independently from Unif$(\mathcal{S}^{p-1})$.

(2) Use the univariate fast algorithm in Section 4.2.1 to compute the following quantities:

$$\mathcal{E}_{n,m}^{(k)} = C_p \mathcal{E}_{n,m}(u_k^T X, u_k^T Y),$$
$$S_{1;n,m}^{(k)} = C_p \binom{n+m}{2}^{-1} \sum_{i<j}^{n} |u^T(Z_i - Z_j)|,$$

where constant C_p has been defined at the beginning of Chapter 4.

(3) Use the univariate fast algorithm for distance covariance in [34] to compute:

$$S_{2;n,m}^{(k)} = C_p^2 \text{SDC}(u_k^T Z, u_k^T Z),$$

where SDC stands for Sample Distance Covariance defined in [34, eq (3.3)]. Randomly generate v_k from Unif($\mathcal{S}^{p-1}$) and use aforementioned algorithm to compute

$$S^{(k)}_{3;n,m} = C_p^2 \text{SDC}(u_k^T Z, v_k^T Z).$$

(4) Repeat above steps for $k = 1, \ldots, K$ and aggregate the results as follows:

$$\overline{\mathcal{E}}_{n,m} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{E}^{(k)}_{n,m}, \qquad \overline{S}_{1;n,m} = \frac{1}{K}\sum_{k=1}^{K} S^{(k)}_{1;n,m},$$
$$\overline{S}_{2;n,m} = \frac{1}{K}\sum_{k=1}^{K} S^{(k)}_{2;n,m}, \qquad \overline{S}_{3;n,m} = \frac{1}{K}\sum_{k=1}^{K} S^{(k)}_{3;n,m},$$
$$\hat{\alpha} = \frac{1}{2}\frac{\overline{S}^2_{1;n,m}}{\frac{1}{K}\overline{S}_{2;n,m} + \frac{K-1}{K}\overline{S}_{3;n,m}}, \tag{4.2.25}$$
$$\hat{\beta} = \frac{1}{2}\frac{\overline{S}_{1;n,m}}{\frac{1}{K}\overline{S}_{2;n,m} + \frac{K-1}{K}\overline{S}_{3;n,m}}. \tag{4.2.26}$$

(5) Reject null hypothses $\mathcal{H}_0$ if and only if $(n+m)\overline{\mathcal{E}}_{n,m} + \overline{S}_{1;n,m} > \text{Gamma}(1-\alpha_s; \hat{\alpha}, \hat{\beta})$, where Gamma$(1-\alpha_s; \hat{\alpha}, \hat{\beta})$ is the $1-\alpha_s$ percentile of Gamma distribution with shape parameter $\hat{\alpha}$ and rate parameter $\hat{\beta}$; Otherwise, accept it.

See Algorithm 7 of Appendix A.2 for a stand-alone description of above algorithm.

4.3 Theoretical Properties of Energy Statistics and Random Projections

Firstly, we will show some nice properties of random projections in energy distance and energy statistics in Section 4.3.1. Then, we will study the asymptotic properties of energy statistics $\mathcal{E}_{n,m}$ and randomly projected energy statistics $\overline{\mathcal{E}}_{n,m}$ in Section 4.3.2 and 4.3.3, respectively.

4.3.1 Properties of Random Projections in Energy Distance

We will study some properties of randomly projected energy distance and energy statistics in this part. We begin a sufficient and necessary condition of equality of distributions.

Lemma 4.3.1. *Suppose u is some random point on unit sphere $\mathcal{S}^{p-1}$: $u \in \mathcal{S}^{p-1} := \{u \in \mathbb{R}^p : |u| = 1\}$. We have*

random vector $X \in \mathbb{R}^p$ has the same distribution with random vector $Y \in \mathbb{R}^p$

if and only if

$$\mathcal{E}(u^T X, u^T Y) = 0 \text{ for any } u \in \mathcal{S}^{p-1}.$$

The following result allows us to regard energy distance / energy statistics of multivariate random variables as the integration of energy distance / energy statistics of univariate random variables. This result provides the foundation of our proposed method in Section 4.2.2.

Lemma 4.3.2. *Suppose u is some random point on unit sphere $\mathcal{S}^{p-1}$. Let μ denote the uniform probability measure on $\mathcal{S}^{p-1}$. Then, for random vectors $X, Y \in \mathbb{R}^p$ with $\mathbb{E}[|X|] < \infty, \mathbb{E}[|Y|] < \infty$, we have*

$$\mathcal{E}(X, Y) = C_p \int_{\mathcal{S}^{p-1}} \mathcal{E}(u^T X, u^T Y) d\mu(u),$$

where C_p is the constant defined at the beginning of Chapter 4. Similarly, for energy statistics, we have

$$\mathcal{E}_{n,m}(X, Y) = C_p \int_{\mathcal{S}^{p-1}} \mathcal{E}_{n,m}(u^T X, u^T Y) d\mu(u).$$

4.3.2 Asymptotic Properties of Energy Statistics $\mathcal{E}_{n,m}$

As showed in Section 4.1, the energy statistics $\mathcal{E}_{n,m}$ is a two-sample u-statistics with respect to kernel

$$h(X_1, X_2, Y_1, Y_2) = \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}|X_i - Y_j| - |X_1 - X_2| - |Y_1 - Y_2|$$

which is a two-sample kernel. Before analyzing the asymptotic properties of $\mathcal{E}_{n,m}$, let us define the following quantities that will play important roles in subsequent studies:

$$h_{10} = h_{10}(X_1) = \mathbb{E}_{X_2,Y_1,Y_2}[h(X_1, X_2, Y_1, Y_2)],$$

$$h_{01} = h_{01}(Y_1) = \mathbb{E}_{X_1,X_2,Y_2}[h(X_1, X_2, Y_1, Y_2)],$$

$$h_{20} = h_{20}(X_1, X_2) = \mathbb{E}_{Y_1,Y_2}[h(X_1, X_2, Y_1, Y_2)],$$

$$h_{02} = h_{02}(Y_1, Y_2) = \mathbb{E}_{X_1,X_2}[h(X_1, X_2, Y_1, Y_2)],$$

$$h_{11} = h_{11}(X_1, Y_1) = \mathbb{E}_{X_2,Y_2}[h(X_1, X_2, Y_1, Y_2)],$$

where the two subindexes represent how many X's and Y's in the functions, respectively.

Lemma 4.3.3 (Generic Formula)**.** *If* $\mathbb{E}[|X|] + \mathbb{E}[|Y|] < \infty$*, for independent* X_1*,* X_2*,* X*,* X'*,* Y_1*,* Y_2*,* Y *and* Y'*, w we have*

$$h_{10}(X_1) = \mathbb{E}_Y[|X_1 - Y|] + \mathbb{E}_{X,Y}[|X - Y|] - \mathbb{E}_X[|X_1 - X|] - \mathbb{E}_{Y,Y'}[|Y - Y'|], \tag{4.3.27}$$

$$h_{01}(Y_1) = \mathbb{E}_X[|X - Y_1|] + \mathbb{E}_{X,Y}[|X - Y|] - \mathbb{E}_{X,X'}[|X - X'|] - \mathbb{E}_Y[|Y_1 - Y|], \tag{4.3.28}$$

$$h_{20}(X_1, X_2) = \mathbb{E}_Y[|X_1 - Y|] + \mathbb{E}_Y[|X_2 - Y|] - |X_1 - X_2| - \mathbb{E}_{Y,Y'}[|Y - Y'|], \quad (4.3.29)$$

$$h_{02}(Y_1, Y_2) = \mathbb{E}_X[|X - Y_1|] + \mathbb{E}_X[|X - Y_1|] - |Y_1 - Y_2| - \mathbb{E}_{X,X'}[|X - X'|], \quad (4.3.30)$$

$$h_{11}(X_1, Y_1) = \frac{1}{2}|X_1 - Y_1| + \frac{1}{2}\mathbb{E}_X[|X - Y_1|] + \frac{1}{2}\mathbb{E}_Y[|X_1 - Y|] + \frac{1}{2}\mathbb{E}_{X,Y}[|X - Y|] - \mathbb{E}_X[|X_1 - X|] - \mathbb{E}_Y[|Y_1 - Y|]. \quad (4.3.31)$$

We can also define h_{21}, h_{12} and h_{22} in a similar way but we do not list them here as they are not important in subsequent analysis. The corresponding variance of $h_{i,j}$ is denoted by

$$\sigma_{ij}^2 = \text{Var}[h_{ij}], 1 \leq i + j \leq 2, 1 \leq i, j \leq 2.$$

Then, by the result [42] Section 2.2 Theorem 2, the variance of $\mathcal{E}_{n,m}$ can be represented as follows.

Lemma 4.3.4 (Variance of two-sample U-statistics)**.** *Suppose Var*$[h(X_1, X_2, Y_1, Y_2)] < \infty$ *and* $n, m \geq 4$*, then the variance* $\mathcal{E}_{n,m}(X, Y)$ *is*

$$\begin{aligned} \textit{Var}[\mathcal{E}_{n,m}] &= \frac{1}{\binom{n}{2}\binom{m}{2}} \sum_{i,j=0,i+j\geq 1}^{2} \binom{2}{i}\binom{2}{j}\binom{n-2}{2-i}\binom{m-2}{2-j}\sigma_{ij}^2 \\ &= \frac{4}{m}\sigma_{01}^2 + \frac{4}{n}\sigma_{10}^2 + (\frac{16}{mn} + \frac{1}{m^2})\sigma_{01}^2 + (\frac{16}{mn} + \frac{1}{n^2})\sigma_{10}^2 \\ &\quad + \frac{2}{m^2}\sigma_{02}^2 + \frac{2}{n^2}\sigma_{20}^2 + \frac{16}{nm}\sigma_{11}^2 + O(\frac{1}{n^2 m}) + O(\frac{1}{nm^2}) \end{aligned}$$

[42] also shows that $\mathcal{E}_{n,m}$ is asymptotically normal under mild conditions.

Theorem 4.3.5. *([42, Section 3.7, Theorem 1]) Let* $N = n + m$ *denote the total number of observations. Suppose there exists constant* $0 < \eta < 1$ *such that* $n/N \to \eta$ *and* $m/N \to 1 - \eta$ *as* $n, m \to \infty$*. If Var*$[h(X_1, X_2, Y_1, Y_2)] < \infty$ *and* $\sigma_{10}^2 + \sigma_{01}^2 > 0$*, then* $\sqrt{N}(\mathcal{E}_{n,m} - \mathcal{E})$

converges in distribution to a normal distribution with mean zero and variance $4\sigma_{10}^2/\eta + 4\sigma_{01}^2/(1-\eta)$*, i.e.,*

$$\sqrt{N}(\mathcal{E}_{n,m} - \mathcal{E}) \xrightarrow{D} \mathcal{N}(0, 4\sigma_{10}^2/\eta + 4\sigma_{01}^2/(1-\eta)),$$

where $\mathcal{E}$ *is the energy distance* $\mathcal{E} = \mathbb{E}[\mathcal{E}_{n,m}]$.

Now, we assume that X has the same distribution with Y. Then, the formulas of h_{ij} could be simplified.

Lemma 4.3.6. *If* X *and* Y *are identically distributed, then we have*

$$h_{10}(X_1) = 0, \qquad h_{01}(Y_1) = 0, \tag{4.3.32}$$

$$h_{20}(X_1, X_2) = \mathbb{E}_X[|X_1 - X|] + \mathbb{E}_X[|X_2 - X|] - |X_1 - X_2| - \mathbb{E}_{X,X'}[|X - X'|], \tag{4.3.33}$$

$$h_{02}(Y_1, Y_2) = \mathbb{E}_Y[|Y_1 - Y|] + \mathbb{E}_Y[|Y_2 - Y|] - |Y_1 - Y_2| - \mathbb{E}_{Y,Y'}[|Y - Y'|], \tag{4.3.34}$$

$$h_{11}(X_1, Y_1) = \frac{1}{2}\left(|X_1 - Y_1| - \mathbb{E}_X[|X_1 - X|] - \mathbb{E}_X[|Y_1 - X|] + \mathbb{E}_{X,X'}[|X - X'|]\right). \tag{4.3.35}$$

The proof of this lemma is straightforward by noting the fact that the usage of X and Y is interchangeable as they are identically independently distributed.

When X has the same distribution with Y, $\mathcal{E}_{n,m}$ is no longer asymptotically normal. Instead, $(n+m)\mathcal{E}_{n,m}$ converges to a sum of (possibly infinite) independent chi-squared random variables.

Theorem 4.3.7. *Let* $N = n + m$ *denote the total number of observations. Suppose there exists constant* $0 < \eta < 1$ *such that* $n/N \to \eta$ *and* $m/N \to 1 - \eta$ *as* $n, m \to \infty$*. If* X *and*

Y are identically distributed, the asymptotic distribution of $\mathcal{E}_{n,m}$ is

$$N\mathcal{E}_{n,m} \xrightarrow{D} \sum_{l=1}^{\infty} \frac{\lambda_l}{\eta(1-\eta)}(Z_l^2 - 1),$$

where $Z_1, Z_2, \ldots$ are independent standard normal random variables and λ_l's are defined in Lemma D.4.1 and

$$\sum_{l=1}^{\infty} \lambda_l = \mathbb{E}_{X,X'}[|X - X'|], \qquad \sum_{l=1}^{\infty} \lambda_l^2 = DC(X, X),$$

where $DC(X, X)$ is the distance covariance of X, see [33].

See appendix for a proof.

4.3.3 Asymptotic Properties of Randomly Projected Energy Statistics $\overline{\mathcal{E}}_{n,m}$

Let us recall some notations. The randomly projected energy statistics $\overline{\mathcal{E}}_{n,m}$ is defined as

$$\overline{\mathcal{E}}_{n,m} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{E}_{n,m}^{(k)} = \frac{1}{K}\sum_{k=1}^{K} C_p \mathcal{E}_{n,m}(u_k^T X, u_k^T Y),$$

where constant C_p has been defined at the beginning of Chapter 4 and u_k's are independent samples from Unif($\mathcal{S}^{p-1}$). Note that $\mathcal{E}_{n,m}(u_k^T X, u_k^T Y)$ is a U-statistic for any k and $\overline{\mathcal{E}}_{n,m}$ is also a U-statistic as

$$\begin{aligned}
\overline{\mathcal{E}}_{n,m} &= \frac{1}{K}\sum_{k=1}^{K} \frac{C_p}{\binom{n}{2}\binom{m}{2}} \sum_{i_1<i_2, j_1<j_2} h(u_k^T X_{i_1}, u_k^T X_{i_2}, u_k^T Y_{j_1}, u_k^T Y_{j_2}) \\
&= \frac{1}{\binom{n}{2}\binom{m}{2}} \sum_{i_1<i_2, j_1<j_2} \frac{1}{K}\sum_{k=1}^{K} C_p h(u_k^T X_{i_1}, u_k^T X_{i_2}, u_k^T Y_{j_1}, u_k^T Y_{j_2}) \\
&\triangleq \frac{1}{\binom{n}{2}\binom{m}{2}} \sum_{i_1<i_2, j_1<j_2} \overline{h}(X_{i_1}, X_{i_2}, Y_{j_1}, Y_{j_2}),
\end{aligned}$$

where

$$\overline{h}(X_{i_1}, X_{i_2}, Y_{j_1}, Y_{j_2}) = \frac{1}{K}\sum_{k=1}^{K} C_p h(u_k^T X_{i_1}, u_k^T X_{i_2}, u_k^T Y_{j_1}, u_k^T Y_{j_2})$$

is the kernel of $\overline{\mathcal{E}}_{n,m}$. Let us define the following notations that will be essential in analyzing the asymptotic properties of $\overline{\mathcal{E}}_{n,m}$:

$$\overline{h}_{10} = \overline{h}_{10}(X_1) = \mathbb{E}_{X_2,Y_1,Y_2}[\overline{h}(X_1, X_2, Y_1, Y_2)],$$

$$\overline{h}_{01} = \overline{h}_{01}(Y_1) = \mathbb{E}_{X_1,X_2,Y_2}[\overline{h}(X_1, X_2, Y_1, Y_2)],$$

$$\overline{h}_{20} = \overline{h}_{20}(X_1, X_2) = \mathbb{E}_{Y_1,Y_2}[\overline{h}(X_1, X_2, Y_1, Y_2)],$$

$$\overline{h}_{02} = \overline{h}_{02}(Y_1, Y_2) = \mathbb{E}_{X_1,X_2}[\overline{h}(X_1, X_2, Y_1, Y_2)],$$

$$\overline{h}_{11} = \overline{h}_{11}(X_1, Y_1) = \mathbb{E}_{X_2,Y_2}[\overline{h}(X_1, X_2, Y_1, Y_2)],$$

where the expecations are taken with respect to (X, Y) given random projections U. We also let $\overline{\sigma}^2_{ij}$ denote the conditional variance of $\overline{h}_{ij}$ given all projection directions $U = (u_1, \ldots, u_K)$,

$$\overline{\sigma}^2_{ij} = \overline{\sigma}^2_{ij}(U) = \text{Var}_{X,Y}[\overline{h}_{ij}|U].$$

4.3.3.1 Asymptotic Properties in Inequality of Distribution

By Lemma 4.3.4 and the Law of Total Variance, we have the following result on the variance of $\overline{\mathcal{E}}_{n,m}$.

Lemma 4.3.8 (Variance of $\overline{\mathcal{E}}_{n,m}$)**.** *Suppose Var*$[h(X_1, X_2, Y_1, Y_2)] < \infty$ *and* $n, m \geq 4$*, then the variance* $\overline{\mathcal{E}}_{n,m}$ *is*

$$\begin{aligned} \textit{Var}[\overline{\mathcal{E}}_{n,m}] = \frac{1}{K}\textit{Var}_u\left[\mathcal{E}(u^T X, u^T Y)\right] &+ \mathbb{E}_U\left[\frac{4}{m}\overline{\sigma}^2_{01} + \frac{4}{n}\overline{\sigma}^2_{10}\right] \\ &+ \mathbb{E}_U\left[(\frac{16}{mn} + \frac{1}{m^2})\overline{\sigma}^2_{01} + (\frac{16}{mn} + \frac{1}{n^2})\overline{\sigma}^2_{10}\right] \\ &+ \mathbb{E}_U\left[\frac{2}{m^2}\overline{\sigma}^2_{02} + \frac{2}{n^2}\overline{\sigma}^2_{20} + \frac{16}{nm}\overline{\sigma}^2_{11}\right] + O(\frac{1}{n^2 m}) + O(\frac{1}{nm^2}). \end{aligned}$$

As an immediate result from Lemma 4.3.8, we have the following theorem on the asymptotic properties of $\overline{\mathcal{E}}_{n,m}$.

Theorem 4.3.9. *Suppose Var*$[h(X_1, X_2, Y_1, Y_2)] < \infty$. *Let* $N = n+m$ *and assume* $n/N \to \eta$ *as* $N \to \infty$, *where* $0 < \eta < 1$, *then we have*

$$\overline{\mathcal{E}}_{n,m} \xrightarrow{p} \mathcal{E}(X, Y) \text{ as } N \to \infty, K \to \infty.$$

The asymptotic distribution of $\overline{\mathcal{E}}_{n,m}$ *could differ under different conditions.*

(1) If $K \to \infty$ *and* $K/N \to 0$, *then*

$$\sqrt{K}(\overline{\mathcal{E}}_{n,m} - \mathcal{E}(X, Y)) \xrightarrow{D} \mathcal{N}(0, \textit{Var}_u[\mathcal{E}(u^T X, u^T Y)]).$$

(2) If $N \to \infty$ *and* $K/N \to \infty$, *then*

$$\sqrt{N}(\overline{\mathcal{E}}_{n,m} - \mathcal{E}(X, Y)) \xrightarrow{D} \mathcal{N}(0, \frac{4}{\eta}\mathbb{E}_U[\overline{\sigma}^2_{10}] + \frac{4}{1-\eta}\mathbb{E}_U[\overline{\sigma}^2_{01}]).$$

(3) If $N \to \infty$ *and* $K/N \to C$, *where* $0 < C < \infty$, *then*

$$\sqrt{N}(\overline{\mathcal{E}}_{n,m} - \mathcal{E}(X, Y)) \xrightarrow{D} \mathcal{N}(0, \frac{1}{C}\textit{Var}_u[\mathcal{E}(u^T X, u^T Y)] + \frac{4}{\eta}\mathbb{E}_U[\overline{\sigma}^2_{10}] + \frac{4}{1-\eta}\mathbb{E}_U[\overline{\sigma}^2_{01}]).$$

4.3.3.2 Asymptotic Properties in Equality of Distribution

It is of more interest to study the asymptotic properties of $\overline{\mathcal{E}}_{n,m}$ under the condition that X has the same distribution with Y. We have the following lemma under this condition.

Lemma 4.3.10. *If* X *has the same distribution with* Y, *we have*

$$\textit{Var}_u[\mathcal{E}(u^T X, u^T Y)] = 0,$$

and,

$$\overline{h}_{10} = 0, \overline{h}_{01} = 0 \textit{ with probability } 1,$$

which implies

$$\overline{\sigma}^2_{10} = \textit{Var}[\overline{h}_{10}|U] = 0, \overline{\sigma}^2_{01} = \textit{Var}[\overline{h}_{01}|U] = 0.$$

Therefore, the variance of $\overline{\mathcal{E}}_{n,m}$ could be expressed as

$$\textit{Var}[\overline{\mathcal{E}}_{n,m}] = \mathbb{E}_U \left[\frac{2}{m^2}\overline{\sigma}^2_{02} + \frac{2}{n^2}\overline{\sigma}^2_{20} + \frac{16}{nm}\overline{\sigma}^2_{11} \right] + O(\frac{1}{n^2 m}) + O(\frac{1}{nm^2}).$$

See appendix for the proof.

We should also be aware of a result, which is similar with Lemma D.4.1. This result will play an important role for our main theorem and its proof.

Lemma 4.3.11. *The kernel $\overline{\mathbf{k}}(\cdot,\cdot)$ defined as*

$$\overline{\mathbf{k}}(X_1, X_2) = \frac{C_p}{K} \sum_{k=1}^{K} \mathbb{E}_X[|u_k^T(X_1 - X)|] + \mathbb{E}_X[|u_k^T(X_2 - X)|] - |u_k^T(X_1 - X_2)| - \mathbb{E}_{X,X'}[|u_k^T(X - X')|]$$

is a positive kernel and thus there exists $\overline{\phi}_1(\cdot), \overline{\phi}_2(\cdot), \ldots$ such that

$$\overline{\mathbf{k}}(X_1, X_2) = \sum_{i=1}^{\infty} \overline{\lambda}_i \overline{\phi}_i(X_1) \overline{\phi}_i(X_2),$$

where $\overline{\lambda}_1 \geq \overline{\lambda}_2 \geq \ldots \geq 0$, $\mathbb{E}[\phi_i(X)] = 0$, $\mathbb{E}[\overline{\phi}_i(X)^2] = 1$ and $\mathbb{E}[\overline{\phi}_i(X)\overline{\phi}_j(X)] = 0$, $i = 1, 2, \ldots, \infty$, $i \neq j$.

Proof. It is worth noting that $\overline{\mathbf{k}}(\cdot,\cdot)$ a positive kernel as it is the sum of a collection of positive kernel. The rest follows by Mercer's Theorem. □

Equipped with above two lemmas, we can conclude that $\overline{\mathcal{E}}_{n,m}$ also converges to a weighted sum of chi-square random variables when the collection of random projections U is given.

Theorem 4.3.12. *Let $N = n + m$ denote the total number of observations. Suppose there exists constant $0 < \eta < 1$ such that $n/N \to \eta$ and $m/N \to 1 - \eta$ as $n, m \to \infty$. If X and Y are identically distributed and all projection directions $U = (u_1, \ldots, u_K)$ are given, the asymptotic distribution of $\mathcal{E}_{n,m}$ is*

$$N\overline{\mathcal{E}}_{n,m} \xrightarrow{D} \sum_{l=1}^{\infty} \frac{\overline{\lambda}_l}{\eta(1-\eta)}(Z_l^2 - 1) = \frac{1}{\eta(1-\eta)} \sum_{l=1}^{\infty} \overline{\lambda}_l Z_l^2 - \frac{1}{\eta(1-\eta)} \sum_{l=1}^{\infty} \overline{\lambda}_l,$$

where $Z_1, Z_2, \ldots$ are independent standard normal random variables and $\overline{\lambda}_l$'s are the eigenvalues associated with kernel $\overline{\mathbf{k}}(\cdot, \cdot)$ in Lemma 4.3.11. We also have

$$\sum_{l=1}^{\infty} \overline{\lambda}_l = \frac{C_p}{K} \sum_{k=1}^{K} \mathbb{E}_{X,X'}[|u_k^T(X - X')|], \qquad \sum_{l=1}^{\infty} \overline{\lambda}_l^2 = \frac{C_p^2}{K^2} \sum_{k,k'=1}^{K} \textit{DC}(u_k^T X, u_{k'}^T X),$$

where DC$(u_k^T X, u_{k'}^T X)$ is the distance covariance between $u_k^T X$ and $u_{k'}^T X$.

See appendix for the proof.

Usually, $\sum_{l=1}^{\infty} \overline{\lambda}_l Z_l^2$ is a weighted sum of infinite many chi-squared random variables. As a result, there is no close form for the asymptotic distribution of $\overline{\mathcal{E}}_{n,m}$. But, we can approximate it by a gamma distribution with first two moments matched, see [12]. As a result, $\sum_{l=1}^{\infty} \overline{\lambda}_l Z_l^2$ could be approximated by Gamma(α, β) with density function

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, x > 0,$$

where

$$\alpha = \frac{1}{2} \frac{(\sum_{l=1}^{\infty} \overline{\lambda}_l)^2}{\sum_{l=1}^{\infty} \overline{\lambda}_l^2}, \quad \beta = \frac{1}{2} \frac{\sum_{l=1}^{\infty} \overline{\lambda}_l}{\sum_{l=1}^{\infty} \overline{\lambda}_l^2}.$$

The following proposition gives a specific way to approximate $\sum_{l=1}^{\infty} \overline{\lambda}_l$ and $\sum_{l=1}^{\infty} \overline{\lambda}_l^2$ from data.

Proposition 4.3.13. *Let Z denote the collection of all observations,*

$$Z = (Z_1, \ldots, Z_{n+m}) = (X_1, \ldots, X_n, Y_1, \ldots, Y_m).$$

When X and Y have the same distribution, we can approximate $\sum_{l=1}^{\infty} \overline{\lambda}_l$ and $\sum_{l=1}^{\infty} \overline{\lambda}_l^2$ as follows:

$$\sum_{l=1}^{\infty} \overline{\lambda}_l \approx \frac{C_p}{K} \sum_{k=1}^{K} \frac{1}{(n+m)(n+m-1)} \sum_{i \neq j}^{n+m} |u_k^T (Z_i - Z_j)|$$
$$\sum_{l=1}^{\infty} \overline{\lambda}_l^2 \approx \frac{C_p^2}{K^2} \sum_{k=1}^{K} \mathit{SDC}(u_k^T Z, u_k^T Z) + \frac{(K-1)C_p^2}{K^2} \sum_{k=1}^{K} \mathit{SDC}(u_k^T Z, v_k^T Z),$$

where SDC$(\cdot, \cdot)$ denotes the sample distance covariance and $v_1, \ldots, v_K$ are all independent random variables from Unif$(\mathcal{S}^{p-1})$.

See appendix for the reasoning and justification.

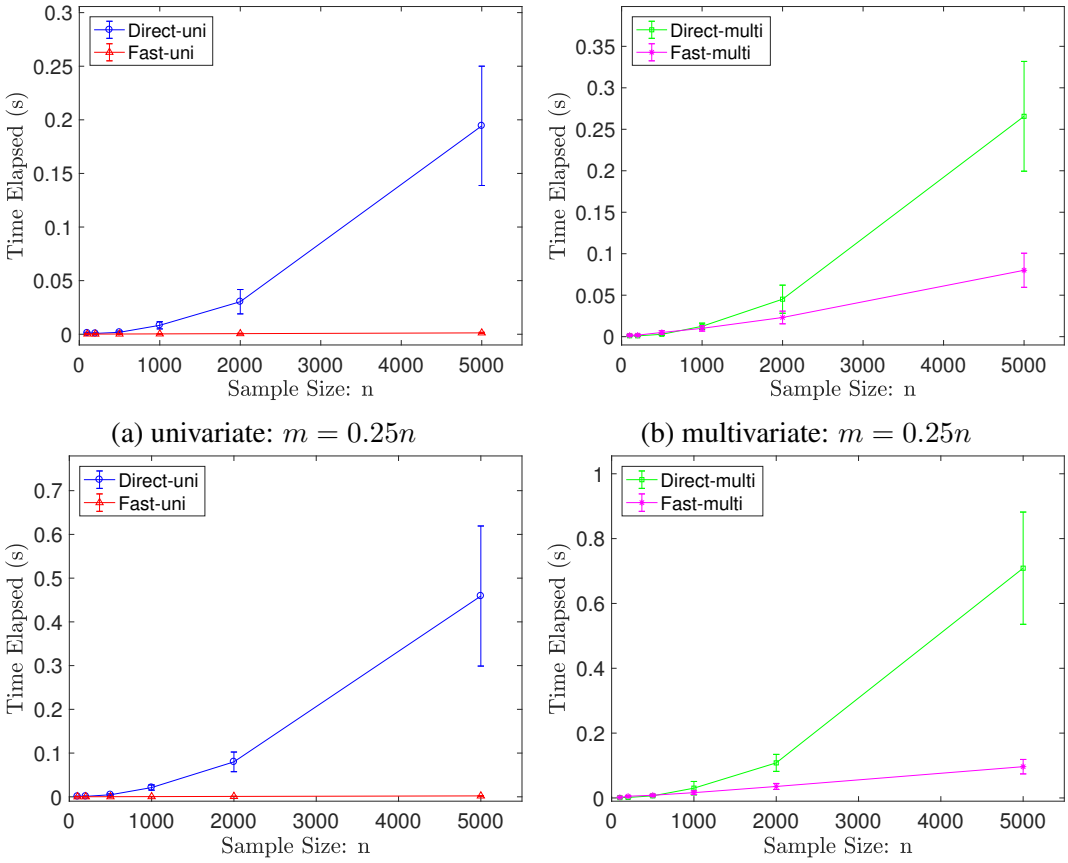
4.4 Simulations on Randomly Projected Energy Statistics

4.4.1 Speed Comparison with Direct Method

In this section, we compare the computing speed of the proposed algorithms for univariate random variables and multivariate random variables with direct method by Definition 4.1.3. This experiment is run on a laptop (MacBook Pro Retina, 13-inch, Early 2015, 2.7 GHz Intel Core i5, 8 GB 1867 MHz DDR3) with MATLAB R2016b (9.1.0.441655). Figure 13 summarizes the time cost of each method against sample size. Note that the scale of time elapsed is different in each subfigure. The result demonstrates the computational advantage of the fast algorithm when sample size is large.

4.4.2 Impact of Sample Size, Data Dimension and Number of Random Projections

In this section, we will use synthetic data to study the impact of sample size (n, m), data dimension p and Number of Random Projections K on the convergence and test power of



(a) univariate: $m = 0.25n$

(b) multivariate: $m = 0.25n$

(c) univariate: $m = n$

(d) multivariate: $m = n$

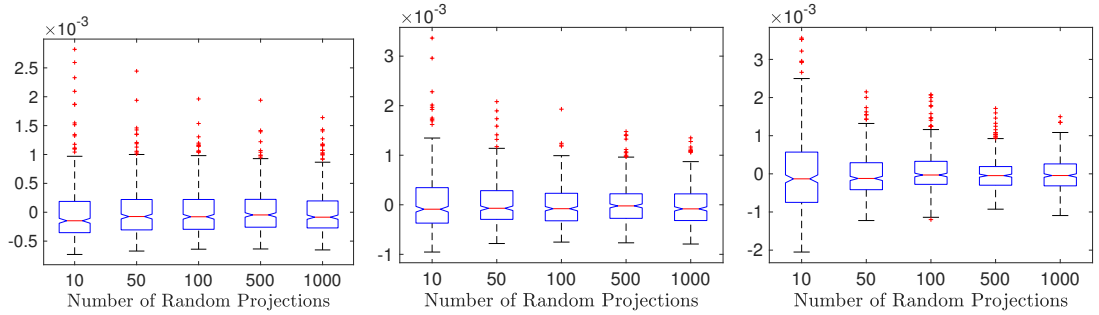
Figure 13: Speed Comparison:"Direct-uni" and "Direct-multi" represent the direct method for univariate and multivariate random variables, respectively; "Fast-uni" represents the fast algorithm for univariate random variables described in Section 4.2.1; "Fast-multi" represents the fast algorithm for multivariate random variables described in Section 4.2.2 and the number of Monte Carlo iterations is chosen to be $K = 50$. The dimension of the multivariate random varialbes is fixed to be $p = 10$. We let the ratio of sample size of Y over sample size of X be either 0.25 or 1. The experiment is repeated for 400 times.

multivariate energy statistics. The significance level is set to be $\alpha_s = 0.05$. Each experiment will be repeated for 400 times to achieve reliable means and variances.

In the following two examples, we will fix sample size $n = 5000, m = 5000$ and let data dimension p vary in $(5, 10, 50, 100, 500)$ and number of random projections K vary in $(10, 50, 100, 500, 1000)$. In Example 4.4.1, X and Y are identically distributed while they are not in Example 4.4.2. The result in these two examples suggests that $K = 50$ should

suffice when sample size is sufficiently large, regardless of the data dimension.

Example 4.4.1. *We generate random vector $X, Y \sim \mathcal{N}(0, I_p)$, which implies X and Y are identically distributed.*



(a) p=5 (b) p=10 (c) p=50

(d) p=100 (e) p=500

Figure 14: Boxplots of estimators in Example 4.4.1. Sample size of X and Y are fixed to be $n = 2000$, $m = 2000$, respectively; the result is based on 400 repeated experiments.

Example 4.4.2. *We generate random vector $X \sim \mathcal{N}(0, I_p)$, $Y \sim t(5)^{(p)}$, where each entry of Y follows t-distribution with degrees of freedom 5. In this case, the distribution of X is different from the distribution of Y.*

4.4.3 Compare with Other Two-Sample Tests

We compare our method — Randomly Projected Energy Statistics (RPES) with direct method of Energy Statistics (ES) as well as the most popular alternative in recent literature — the Maximum Mean Discrepancy (MMD) proposed by [28]. Specifically, we use

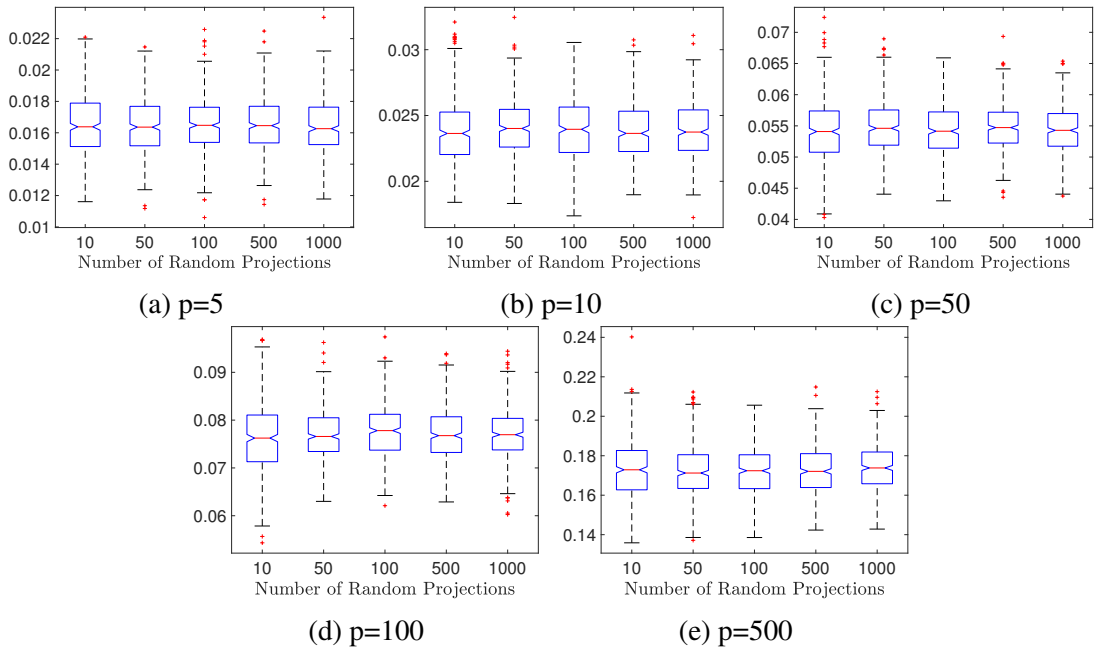


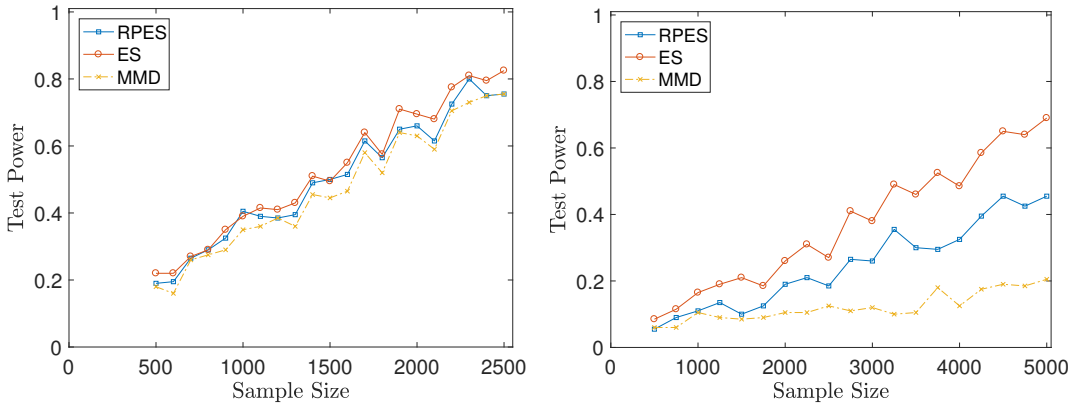
Figure 15: Boxplots of estimators in Example 4.4.2. Sample size of X and Y are fixed to be $n = 2000$, $m = 2000$, respectively; the result is based on 400 repeated experiments.

the MMD with Gaussian kernels in our implementation. To obtain reliable estimate of test power, the experiments will be repeated for 200 times.

In the following example, we will measure the power of those tests in distinguishing minor difference in mean of two multivariate normal distribution.

Example 4.4.3. *We generate random vector* $X \sim \mathcal{N}(0, I_p)$, $Y \sim \mathcal{N}(\mu, I_p)$. *We let* $\mu = (0.1, 0, \ldots, 0)^t$, *where the first entry of* μ *is 0.1 while the rest entries are all* 0.*We let* $p = 5$ *and* $p = 50$ *to represent low dimensional case and moderate dimensional case, respectively. In the* $p = 5$ *case, the sample sizes* $n = m$ *is from 500 to 2500 with an increment 100; and in the* $p = 50$ *case, the sample size n is from 500 to 5000 with an increment 250.*

Figure 16 plots the test power of each test against sample size in Example 4.4.3. In the low dimensional case, RPES, ES and MMD have similar performance. In higher dimensional case, RPES is less effective than ES since random projection may lose some



(a) $p = 5$ (b) $p = 50$

Figure 16: Test Power vs Sample Size in Example 4.4.3

efficiency when the mean of two distributions only differ in a single dimension. But, RPES still outperforms MMD by a significant margin.

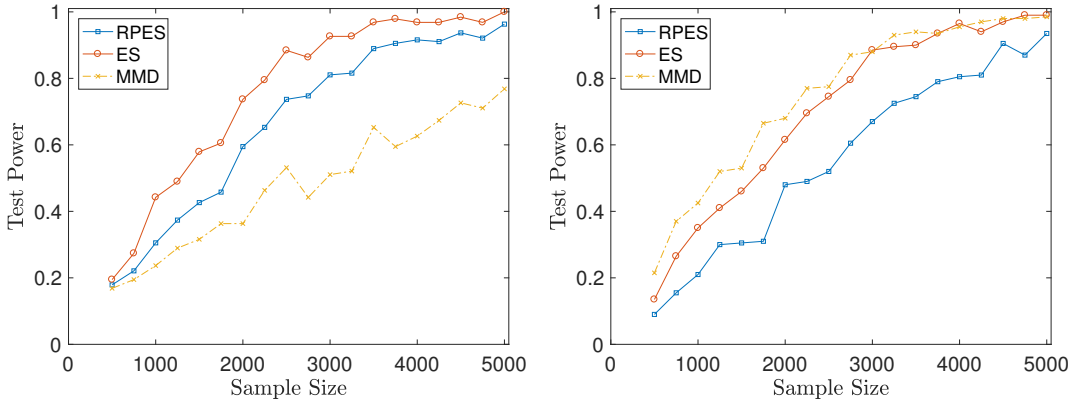
In the next example, we will check how those tests perform when there is only a minor difference in degrees of freedom of two multivariate student t-distribution.

Example 4.4.4. *We generate random vector* $X \sim t_{\nu_1}^{(50)}$, $Y \sim t_{\nu_2}^{(50)}$, *where each entry of* X *follows t-distribution with degree of freedom,* $X_i \sim t_{\nu_1}$, *and* $Y_i \sim t_{\nu_2}$. *We let* $(\nu_1, \nu_2) = (4, 5)$ *and* $(\nu_1, \nu_2) = (7, 10)$, *respectively. In both cases, the sample size n is from 500 to 5000 with an increment 250.*

Figure 17 plots the test power of each test against sample size in Example 4.4.4. In the first case, both RPES and ES outperforms MMD. In the second case, ES and MMD achieve similar performance while RPES underperforms slightly.

In the last example of this section, we will compare the performance of those tests in uniform distributions.

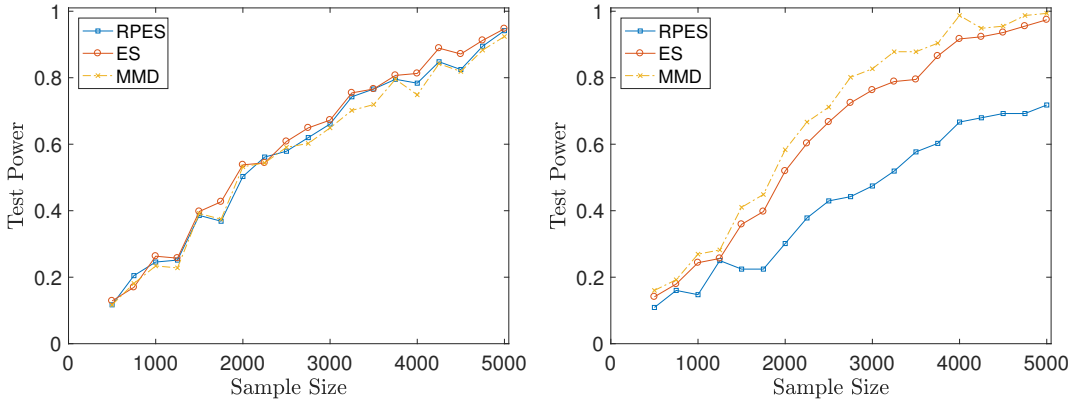
Example 4.4.5. *We generate random vector in the following two scenarios: (1)* $X \sim$ *Unif*$(0, 1)^{(5)}$, *which means each entry of* X *is drawn independently from Unif*$(0, 1)$, *and*



(a) $t_4^{(50)}$ v.s. $t_5^{(50)}$ (b) $t_7^{(50)}$ v.s. $t_{10}^{(50)}$

Figure 17: Test Power vs Sample Size in Example 4.4.4

$Y \sim$ *Unif*$(0, 0.98)^{(5)}$*; (2)* $X \sim$ *Unif*$(0, 1)^{(50)}$*, and* $Y \sim$ *Unif*$(0, 0.99)^{(50)}$*. In both cases, the sample size n is from 500 to 5000 with an increment 250.*



(a) Unif$(0, 1)^{(5)}$ v.s. Unif$(0, 0.98)^{(5)}$ (b) Unif$(0, 1)^{(50)}$ v.s. Unif$(0, 0.99)^{(50)}$

Figure 18: Test Power vs Sample Size in Example 4.4.5

Figure 18 plots the test power of each test against sample size in Example 4.4.5. Similar with the result of Example 4.4.3, the performance of RPES, ES and MMD are quite close in the lower dimensional case. In higher dimensional case, RPES and MMD are also very close in performance while RPES underperforms the aforementioned two methods.

The experiments results in this part show that ES performs best in nearly all the cases. Although RPES tends to be slightly less effective than ES when the data dimension is high

and sample size is relatively small, their performances are quite close when the dimension is moderate or the size is sufficiently large.

4.5 Discussions on Randomly Projected Energy Statistics

There are plenty existing work on graph-based two-sample tests. For instance, [16], [17] propose a graph-based two-sample test based on minimum spanning tree for multivariate data and categorical data, respectively. However, like aforementioned graph-based methods, they still suffer from the high computational complexity — $O(N^2 \log N)$ with Kruskal's algorithm. It is worth noting that [7] introduce a general notion of graph-based two-sample tests, and provide a unified framework for analyzing their asymptotic properties.

The kernel two-sample test statistic proposed by [28] has a very similar form with energy statistics. Though the Euclidean distance $f(x, y) = |x - y|$ is not a positive definite kernel, [69] show that distance-based methods and kernel-based methods might be unified under the same framework.

A possible application of the proposed two-sample tests is change-point detection. [67] develop a change-point detection method based on the minimum non-bipartite matching, which could be regared as an extension of [63]. So, it might be of interest to extend energy distance based method for change-detection problems.

The technique of random projection could be beneficial in reducing the computational complexity without significant compromise in statistical efficiency. [33] propose an computationally and statistically efficient test of independence with the random projection and distance covariance, which reveals the potential of random projection in all distance-based

methods.

Another interesting application of energy distance is distribution representation. [52] introduce a new way to compact a continuous probability distribution into a set of representative points called support points, which are obtained by minimizing the energy distance.

4.6 Conclusions on Randomly Projected Energy Statistics

This work makes three major contributions. First, we develop an efficient algorithm based on sorting and rearrangement to compute energy statistics of univariate random variables. Second, we propose an efficient scheme for computing the energy statistics of multivariate random variables with random projections and univariate fast algorithm. Third, we carry out a two-sample test based on the efficient algorithms and derive its asymptotic properties.

The theoretical analysis shows that the proposed test has nearly the same asymptotic efficiency (in terms of asymptotic variance) with the energy statistics. Numerical examples validate the theoretical results in computational and statistical efficiency.

APPENDIX A

ALGORITHMS

A.1 Algorithms in Distance Covariance

For readers' convenience, we present all the numerical algorithms here.

- The Algorithm 1 summarizes how to compute the proposed distance covariance for multivariate inputs.
- The Algorithm 2 describe an independence testing which applies permutation to generate a threshold.
- The Algorithm 3 describes an independence test that is based on the approximate asymptotic distribution.

In the following algorithms, recall that C_p and C_q have been defined at the beginning of Chapter 3.

Algorithm 1: An Approximation of Sample Distance Covariance $\overline{\Omega}_n$

Data: Observations $X_1, \ldots, X_n \in \mathbb{R}^p$, $Y_1, \ldots, Y_n \in \mathbb{R}^q$; Number of Monte Carlo Iterations K

Result: Approximation of Sample Distance Covariance $\overline{\Omega}_n$

for *k = 1,. . . , K* **do**

Randomly generate u_k from uniform($\mathcal{S}^{p-1}$); randomly generate v_k from uniform($\mathcal{S}^{q-1}$);

Compute the projection of X_i's on u_k: $u_k^t X = (u_k^t X_1, \ldots, u_k^t X_n)$;

Compute the projection of Y_i's on v_k: $v_k^t Y = (v_k^t Y_1, \ldots, v_k^t Y_n)$;

Compute $\Omega_n^{(k)} = C_p C_q \Omega_n(u_k^t X, v_k^t Y)$ with the Fast Algorithm in [34];

end

Return $\overline{\Omega}_n = \frac{1}{K} \sum_{k=1}^{K} \Omega_n^{(k)}$.

Algorithm 2: Independence Test Based on Permutations

Data: Observations $X_1, \ldots, X_n \in \mathbb{R}^p$, $Y_1, \ldots, Y_n \in \mathbb{R}^q$; Number of Monte Carlo Iterations K; Significance Level α_s; Number of Permutation: L

Result: Accept or Reject the Null Hypothesis $\mathcal{H}_0$: X and Y are independent

for *l = 1,. . . , L* **do**

Generate a random permutation of Y: $Y^{\star,l} = (Y_1^\star, \ldots Y_n^\star)$;

Compute $V_l = \overline{\Omega}_n(X, Y^{\star,l})$, using the approach in Algorithm 1;

end

Reject $\mathcal{H}_0$ if $\frac{1+\sum_{l=1}^{L} I(\overline{\Omega}_n > V_l)}{1+L} > \alpha_s$; otherwise, accept.

Algorithm 3: Independence Test Based on Asymptotic Distribution

Data: Observations $X_1, \ldots, X_n \in \mathbb{R}^p$, $Y_1, \ldots, Y_n \in \mathbb{R}^q$; Number of Monte Carlo Iterations K; Significance Level α_s

Result: Accept or Reject the Null Hypothesis $\mathcal{H}_0$: X and Y are independent

for *k = 1,..., K* **do**

Randomly generate u_k from uniform($\mathcal{S}^{p-1}$); randomly generate v_k from uniform($\mathcal{S}^{q-1}$);

Use the Fast Algorithm in [34] to compute:

$$\Omega_n^{(k)} = C_p C_q \Omega_n(u_k^t X, v_k^t Y),$$

$$S_{n,1}^{(k)} = C_p^2 C_q^2 \Omega_n(u_k^t X, u_k^t X)\Omega_n(v_k^t Y, v_k^t Y),$$

$$S_{n,2}^{(k)} = \frac{C_p a_{\cdot\cdot}^{u_k}}{n(n-1)},$$

$$S_{n,3}^{(k)} = \frac{C_q b_{\cdot\cdot}^{v_k}}{n(n-1)};$$

Randomly generate u_k' from uniform($\mathcal{S}^{p-1}$); randomly generate v_k' from uniform($\mathcal{S}^{q-1}$);

Use the Fast Algorithm in [34] to compute:

$$\Omega_{n,X}^{(k)} = C_p^2 \Omega_n(u_k^t X, u'^t_k X),$$

$$\Omega_{n,Y}^{(k)} = C_q^2 \Omega_n(v_k^t Y, v'^t_k Y);$$

end

$\overline{\Omega}_n = \frac{1}{K}\sum_{k=1}^K \Omega_n^{(k)}$; $\bar{S}_{n,1} = \frac{1}{K}\sum_{k=1}^K S_{n,1}^{(k)}$; $\bar{S}_{n,2} = \frac{1}{K}\sum_{k=1}^K S_{n,2}^{(k)}$;

$\bar{S}_{n,3} = \frac{1}{K}\sum_{k=1}^K S_{n,2}^{(k)}$;

$\overline{\Omega}_{n,X} = \frac{1}{K}\sum_{k=1}^K \Omega_{n,X}^{(k)}$; $\overline{\Omega}_{n,Y} = \frac{1}{K}\sum_{k=1}^K \Omega_{n,Y}^{(k)}$;

$\alpha = \frac{1}{2}\frac{\bar{S}_{n,2}^2 \bar{S}_{n,3}^2}{\frac{K-1}{K}\overline{\Omega}_{n,X}\overline{\Omega}_{n,Y} + \frac{1}{K}\bar{S}_{n,1}}$; $\beta = \frac{1}{2}\frac{\bar{S}_{n,2}\bar{S}_{n,3}}{\frac{K-1}{K}\overline{\Omega}_{n,X}\overline{\Omega}_{n,Y} + \frac{1}{K}\bar{S}_{n,1}}$;

Reject $\mathcal{H}_0$ if $n\overline{\Omega}_n + \bar{S}_{n,2}\bar{S}_{n,3} >$ Gamma$(\alpha, \beta; 1-\alpha_s)$; otherwise, accept it. Here Gamma$(\alpha, \beta; 1-\alpha_s)$ is the $1-\alpha_s$ quantile of the distribution Gamma(α, β).

A.2 Algorithms in Energy Statistics

We present all numerical algorithms of Chapter 4 here.

- Algorithm 4 summarizes how to compute the energy statistics of univariate random variables in $O(N \log N)$ time.

- Algorithm 5 describes how to approximate the energy statistics of random variables of any dimension in $O(KN \log N)$ time.

- Algorithm 6 describes a two-sample test that applies permutations to determine the threshold.

- Algorithm 7 describes a two-sample test using approximation of asymptotic distribution to determine the threshold.

Algorithm 4: A Fast Algorithm for Energy Statistics of Univariate Random Variables: $\mathcal{E}_{n,m}(X, Y)$

Data: Observations $X_1, \ldots, X_n \in \mathbb{R}$, $Y_1, \ldots, Y_m \in \mathbb{R}$;

Result: Energy Statistics $\mathcal{E}_{n,m}(X, Y)$

Sort $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ and

$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(m)}$ denote the order statistics.

Compute $E_2 = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} i(n-i) \left| X_{(i+1)} - X_{(i)} \right|$ and

$E_3 = \frac{1}{m(m-1)} \sum_{i=1}^{m-1} i(m-i) \left| Y_{(i+1)} - Y_{(i)} \right|$.

Merge two ordered series $X_{(i)}$'s and $Y_{(j)}$'s into a single ordered series

$Z_{(1)} \leq \cdots \leq Z_{(n+m)}$. Let I_i record the size of the subset of $Z_{(1)}$ through $Z_{(i)}$ that are

from $X_{(i)}$'s.

Compute $E_1 = \frac{2}{nm} \sum_{i=1}^{n+m-1} [I_i(m - i + I_i) + (i - I_i)(n - I_i)] \left| Z_{(i+1)} - Z_{(i)} \right|$.

Return $\mathcal{E}_{n,m}(X, Y) = E_1 - E_2 - E_3$.

Algorithm 5: A Fast Algorithm for Energy Statistics of Multivariate Random Variables: $\overline{\mathcal{E}}_{m,n}$

Data: Observations $X_1, \ldots, X_n \in \mathbb{R}^p$, $Y_1, \ldots, Y_m \in \mathbb{R}^p$; Number of Random Projections K

Result: Average Randomly Projected Energy Statistics $\overline{\mathcal{E}}_{m,n}$

for *k = 1,..., K* **do**

Randomly generate u_k from Uniform($\mathcal{S}^{p-1}$);

Compute the projection of X_i's on u_k: $u_k^T X = (u_k^T X_1, \ldots, u_k^T X_n)$;

Compute the projection of Y_j's on u_k: $u_k^T Y = (u_k^T Y_1, \ldots, u_k^T Y_m)$;

Compute the energy statistics of $u_k^T X$ and $u_k^T Y$ with Algorithm 4:

$\mathcal{E}_{n,m}^{(k)} = C_p \mathcal{E}_{n,m}(u_k^T X, u_k^T Y)$;

end

Return $\overline{\mathcal{E}}_{m,n} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{E}_{n,m}^{(k)}$.

Algorithm 6: Two-Sample Test Based on Permutations

Data: Observations $X_1, \ldots, X_n \in \mathbb{R}^p$, $Y_1, \ldots, Y_m \in \mathbb{R}^p$; Number of Random Projections K; Significance Level α_s; Number of Permutations L

Result: Accept or Reject the Null Hypotheses $\mathcal{H}_0$: X and Y have the same distribution

Compute $\overline{\mathcal{E}}_{m,n}$ with Algorithm 5;

for *l = 1,..., L* **do**

> Generate a random permutation of the observations: $(X^{\star,l}, Y^{\star,l})$;
>
> Use Algorithm 5 to compute $D^{(l)} = \overline{\mathcal{E}}_{m,n}(X^{\star,l}, Y^{\star,l})$ with permutated observations;

end

Reject $\mathcal{H}_0$ if and only if $\frac{1+\sum_{l=1}^{L} I(\overline{\mathcal{E}}_{n,m} > D^{(l)})}{1+L} > \alpha_s$; otherwise, accept it.

Algorithm 7: Two-Sample Test Based on Approximated Asymptotic Distribution

Data: Observations $X_1, \ldots, X_n \in \mathbb{R}^p$, $Y_1, \ldots, Y_m \in \mathbb{R}^p$,

$Z = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$; Number of Random Projections K;

Significance Level α_s

Result: Accept or Reject the Null Hypotheses $\mathcal{H}_0$: X and Y have the same distribution

for *k = 1,..., K* **do**

Randomly generate u_k from Uniform($\mathcal{S}^{p-1}$);

Use Algorithm 4 to Compute:

$$\mathcal{E}_{n,m}^{(k)} = C_p \mathcal{E}_{n,m}(u_k^T X, u_k^T Y)$$

$$S_{1;n,m}^{(k)} = C_p \binom{n+m}{2}^{-1} \sum_{i<j}^{n} |u^T (Z_i - Z_j)|;$$

Use the fast algorithm for distance covariance in [34] to compute:

$$S_{2;n,m}^{(k)} = C_p^2 \text{SDC}(u_k^T Z, u_k^T Z);$$

Randomly generate v_k from Uniform($\mathcal{S}^{p-1}$);

Use the fast algorithm for distance covariance in [34] to compute:

$$S_{3;n,m}^{(k)} = C_p^2 \text{SDC}(u_k^T Z, v_k^T Z);$$

end

$\overline{\mathcal{E}}_{n,m} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{E}_{n,m}^{(k)}$; $\quad \overline{S}_{1;n,m} = \frac{1}{K} \sum_{k=1}^{K} S_{1;n,m}^{(k)}$;

$\overline{S}_{2;n,m} = \frac{1}{K} \sum_{k=1}^{K} S_{2;n,m}^{(k)}$; $\quad \overline{S}_{3;n,m} = \frac{1}{K} \sum_{k=1}^{K} S_{3;n,m}^{(k)}$;

$\hat{\alpha} = \frac{1}{2} \frac{\overline{S}_{1;n,m}^2}{\frac{1}{K}\overline{S}_{2;n,m} + \frac{K-1}{K}\overline{S}_{3;n,m}}$; $\quad \hat{\beta} = \frac{1}{2} \frac{\overline{S}_{1;n,m}}{\frac{1}{K}\overline{S}_{2;n,m} + \frac{K-1}{K}\overline{S}_{3;n,m}}$;

Reject null hypothses $\mathcal{H}_0$ if and only if

$(n+m)\overline{\mathcal{E}}_{n,m} + \overline{S}_{1;n,m} > \text{Gamma}(1 - \alpha_s; \hat{\alpha}, \hat{\beta})$; otherwise, accept it.

APPENDIX B

PROOFS OF DISTRIBUTED STATISTICAL INFERENCE

This appendix is organized as follows. In Section B.1, we analyze the upper bounds of sum of i.i.d. random vectors and random matrices, which will be useful in later proofs. In Section B.2, we derive the upper bounds of the local M-estimators and the simple averaging estimator. We present the proofs to Theorem 2.2.4 and Theorem 2.2.5 in Section B.3 and Section B.4, respectively. A proof of Corollary 2.2.7 will be in Section B.5.

B.1 Bounds on Gradient and Hessian

In order to establish the convergence of gradients and Hessians of the empirical criterion function to those of population criterion function, which is essential for the later proofs, we will present some results on the upper bound of sums of i.i.d. random vectors and random matrices. We start with stating a useful inequality on the sum of independent random variables from [65].

Lemma B.1.1 (Rosenthal's Inequality, [65], Theorem 3)**.** *For $q > 2$, there exists constant $C(q)$ depending only on q such that if $X_1, \ldots, X_n$ are independent random variables with $\mathbb{E}[X_j] = 0$ and $\mathbb{E}[|X_j|^q] < \infty$ for all j, then*

$$(\mathbb{E}[|\sum_{j=1}^{n} X_j|^q])^{1/q} \leq C(q) \max\left\{ (\sum_{j=1}^{n} \mathbb{E}[|X_j|^q])^{1/q}, (\sum_{j=1}^{n} \mathbb{E}[|X_j|^2])^{1/2} \right\}$$

Equipped with the above lemma, we can bound the moments of mean of random vectors.

Lemma B.1.2. *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be i.i.d. random vectors with $\mathbb{E}[X_i] = \mathbf{0}$. And there exists some constants $G > 0$ and $q_0 \geq 2$ such that $\mathbb{E}[\|X_i\|^{q_0}] < G^{q_0}$. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, then for $1 \leq q \leq q_0$, we have*

$$\mathbb{E}[\|\overline{X}\|^q] \leq \frac{C_v(q,d)}{n^{q/2}} G^q,$$

where $C(q,d)$ is a constant depending solely on q and d.

Proof. The main idea of this proof is to transform the sum of random vectors into the sum of random variables and then apply Lemma B.1.1. Let $X_{i,j}$ denote the j-th component of X_i and $\overline{X}_j$ denote the j-th component of $\overline{X}$.

(1) Let us start with a simpler case in which $q = 2$.

$$\begin{aligned}\mathbb{E}[\|\overline{X}\|^2] &= \sum_{j=1}^d \mathbb{E}[|\overline{X}_j|^2] = \sum_{j=1}^d \sum_{i=1}^n \mathbb{E}[|X_{i,j}/n|^2] \\ &= \sum_{j=1}^d \mathbb{E}[|X_{1,j}|^2]/n = n^{-1}\mathbb{E}[\|X_1\|^2] \leq n^{-1}G^2,\end{aligned}$$

The last inequality holds because $\mathbb{E}[\|X_1\|^q] \leq (\mathbb{E}[\|X_1\|^{q_0}])^{q/q_0} \leq G^q$ for $1 \leq q \leq q_0$ by Hölder's inequality.

(2) When $1 \leq q < 2$, we have

$$\mathbb{E}[\|\overline{X}\|^q] \leq (\mathbb{E}[\|\overline{X}\|^2])^{q/2} \leq n^{-q/2}G^q.$$

(3) For $2 < q \leq q_0$, with some simple algebra, we have

$$\begin{aligned}\mathbb{E}[\|\overline{X}\|^q] &= \mathbb{E}\left[(\sum_{j=1}^d |\overline{X}_j|^2)^{q/2}\right] \\ &\leq \mathbb{E}\left[(d \max_{1\leq j\leq d} |\overline{X}_j|^2)^{q/2}\right] = d^{q/2}\mathbb{E}\left[\max_{1\leq j\leq d} |\overline{X}_j|^q\right] \\ &\leq d^{q/2}\mathbb{E}\left[\sum_{j=1}^d |\overline{X}_j|^q\right] = d^{q/2}\sum_{j=1}^d \mathbb{E}\left[|\overline{X}_j|^q\right].\end{aligned}$$

As a continuation, we have

$$\begin{aligned}
&\mathbb{E}[\|\overline{X}\|^q] \leq d^{q/2} \sum_{j=1}^{d} \mathbb{E}[|\overline{X}_j|^q] \\
&\leq d^{q/2} \sum_{j=1}^{d} [C(q)]^q \max\left\{ \sum_{i=1}^{n} \mathbb{E}[|X_{i,j}/n|^q], (\sum_{i=1}^{n} \mathbb{E}[|X_{i,j}/n|^2])^{q/2} \right\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Lemma B.1.1)} \\
&= d^{q/2} [C(q)]^q \sum_{j=1}^{d} \max\left\{ \frac{\mathbb{E}[|X_{1,j}|^q]}{n^{q-1}}, \frac{(\mathbb{E}[|X_{1,j}|^2])^{q/2}}{n^{q/2}} \right\} \\
&\leq d^{q/2+1} [C(q)]^q \max\left\{ \frac{\mathbb{E}[\|X_1\|^q]}{n^{q-1}}, \frac{(\mathbb{E}[\|X_1\|^2])^{q/2}}{n^{q/2}} \right\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(since } \mathbb{E}[|X_{1,j}|^q] \leq \mathbb{E}[\|X_1\|^q]) \\
&\leq d^{q/2+1} [C(q)]^q \max\left\{ \frac{G^q}{n^{q-1}}, \frac{G^q}{n^{q/2}} \right\} \qquad \text{(Hölder's inequality)} \\
&= \frac{d^{q/2+1} [C(q)]^q}{n^{q/2}} G^q. \qquad (q - 1 > q/2 \text{ when } q > 2)
\end{aligned}$$

To complete this proof, we just need to set $C_v(q, d) = d^{q/2+1}[C(q)]^q$.

□

To bound the moment of the mean of i.i.d. random matrices, let us consider another matrix norm – Frobenius norm $|||\cdot|||_F$, i.e.,

$$|||A|||_F = \sqrt{\sum_{i,j} |a_{ij}|^2}, \forall A \in \mathbb{R}^{d \times d}.$$

Note that

$$|||A|||_F = \sqrt{\sum_{i,j} |a_{ij}|^2} = \sqrt{\text{trace}(A^t A)} \geq \sqrt{\sup_{u \in \mathbb{R}^d : \|u\| \leq 1} \|A^t A u\|} = |||A|||,$$

and

$$|||A|||_F \leq \sqrt{d \sup_{u \in \mathbb{R}^d : \|u\| \leq 1} \|A^t A u\|} = \sqrt{d}\, |||A|||.$$

With Frobenius norm, we can regard a random matrix $X \in \mathbb{R}^{d\times d}$ as a random vector in $\mathbb{R}^{d^2}$ and apply Lemma B.1.2 to obtain the following lemma.

Lemma B.1.3. *Let $X_1, \ldots, X_n \in \mathbb{R}^{d\times d}$ be i.i.d. random matrices with $\mathbb{E}[X_i] = \mathbf{0}_{d\times d}$. Let $|||X_i|||$ denote the norm of X_i, which is defined as its maximal singular value. Suppose $\mathbb{E}[|||X_i|||^{q_0}] \leq H^{q_0}$, where $q_0 \geq 2$ and $H > 0$. Then for $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $1 \leq q \leq q_0$, we have*

$$\mathbb{E}\left[\left|\left|\left|\overline{X}\right|\right|\right|^q\right] \leq \frac{C_m(q,d)}{n^{q/2}} H^q,$$

where $C_m(q,d)$ is a constant depending on q and d only.

Proof. By the fact $|||A|||_F \leq \sqrt{d}|||A|||$, we have

$$\mathbb{E}\left[|||X_i|||_F^{q_0}\right] \leq \mathbb{E}\left[\left|\left|\left|\sqrt{d}X_i\right|\right|\right|^{q_0}\right] \leq (\sqrt{d}H)^{q_0}.$$

Then by the fact $|||A||| \leq |||A|||_F$ and Lemma B.1.2, we have

$$\mathbb{E}\left[\left|\left|\left|\overline{X}\right|\right|\right|^q\right] \leq \mathbb{E}\left[\left|\left|\left|\overline{X}\right|\right|\right|_F^q\right] \leq \frac{C_v(q,d^2)}{n^{q/2}}(\sqrt{d}H)^q = \frac{C_v(q,d^2)d^{\frac{q}{2}}}{n^{q/2}}H^q.$$

In the second inequality, we treat $\overline{X}$ as a d^2-dimensional random vector and then apply Lemma B.1.2. Then the proof can be completed by setting $C_m(q,d) = C_v(q,d^2)d^{\frac{q}{2}}$. □

B.2 Error Bound of Local M-estimator and Simple Averaging Estimator

Since the simple averaging estimator is the average of all local estimators and the one-step estimator is just a single Newton-Raphson update from the simple averaging estimator. Thus, it is natural to study the upper bound of the mean squared error (MSE) of a local M-estimator and the upper bound of the MSE of the simple averaging estimator. The main idea in the following proof is similar to the thread in the proof of **Theorem 1** in [93], but

the conclusions are different. Besides, in the following proof, we use a correct analogy of mean value theorem for vector-valued functions.

B.2.1 Bound the Error of Local M-estimators $\theta_i, i = 1, \ldots, k$

In this subsection, we would like to analyze the mean squared error of a local estimator $\theta_i = \arg\max_{\theta \in \Theta} M_i(\theta)$, $i = 1, \ldots, k$ and prove the following lemma in the rest of this subsection.

Lemma B.2.1. *Let $\Sigma = \ddot{M}_0(\theta_0)^{-1}\mathbb{E}[\dot{m}(X;\theta_0)\dot{m}(X;\theta_0)^t]\ddot{M}_0(\theta_0)^{-1}$, where the expecation is taken with respect to X. Under Assumption 2.2.1, 2.2.2 and 2.2.3, for each $i = 1, \ldots, k$, we have*

$$\mathbb{E}[\|\theta_i - \theta_0\|^2] \leq \frac{2}{n} Tr(\Sigma) + O(n^{-2}).$$

Since $\dot{M}_i(\theta_i) = 0$, by Theorem 4.2 in Chapter XIII of [40], we have

$$\begin{aligned}
0 = \dot{M}_i(\theta_i) &= \dot{M}_i(\theta_0) + \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho\,[\theta_i - \theta_0] \\
&= \dot{M}_i(\theta_0) + \ddot{M}_0(\theta_0)[\theta_i - \theta_0] + \left[\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho - \ddot{M}_0(\theta_0)\right][\theta_i - \theta_0] \\
&= \dot{M}_i(\theta_0) + \ddot{M}_0(\theta_0)[\theta_i - \theta_0] + \left[\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho - \ddot{M}_i(\theta_0)\right][\theta_i - \theta_0] \\
&\quad + [\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i - \theta_0] \qquad \text{(subtract and add } \ddot{M}_i(\theta_0)),
\end{aligned}$$

Remark. *Here, it is worth noting that there is no analogy of mean value theorem for vector-valued functions, which implies that there does not necessarily exist θ' lying on the line between θ_i and θ_0 satisfying $\dot{M}_i(\theta_i) - \dot{M}_i(\theta_0) = \ddot{M}_i(\theta')(\theta_i - \theta_0)$. Numerous papers make errors by claiming such θ' lies between θ_i and θ_0.*

If last two terms in above equation are reasonably small, this lemma follows immediately. So, our strategy is as follows. First, we show that the mean squared error of both

$[\int_0^1 \ddot{M}_i((1-\rho)\theta_0+\rho\theta_i)d\rho - \ddot{M}_i(\theta_0)][\theta_i-\theta_0]$ and $[\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i-\theta_0]$ is small under some "good" events. Then we will show the probability of "bad" events is small enough. And Lemma B.2.1 will follow by the fact that Θ is compact.

Suppose $S_i = \{x_1, \ldots, x_n\}$ is the data set on local machine i. Let us define some good events:

$$\begin{aligned} E_1 &= \left\{ \frac{1}{n}\sum_{j=1}^{n} L(x_j) \le 2L \right\}, \\ E_2 &= \left\{ \left|\left|\left| \ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0) \right|\right|\right| \le \lambda/4 \right\}, \\ E_3 &= \left\{ \|\dot{M}_i(\theta_0)\| \le \frac{\lambda}{2}\delta' \right\}, \end{aligned}$$

where $\delta' = \min(\delta, \frac{\lambda}{8L})$, λ is the constant in Assumption 2.2.2 and L and δ are the constants in Assumption 2.2.3. We will show that event E_1 and E_2 ensure that $M_i(\theta)$ is strictly concave at a neighborhood of θ_0. And we will also show that in event E_3, θ_i is fairly close to θ_0. Let $E = E_1 \cap E_2 \cap E_3$, then we have the following lemma:

Lemma B.2.2. *Under event E, we have*

$$\|\theta_i - \theta_0\| \le \frac{4}{\lambda}\|\dot{M}_i(\theta_0)\|$$

Proof. First, we will show $\ddot{M}_i(\theta)$ is a negative definite matrix over a ball centered at θ_0: $B_{\delta'} = \{\theta \in \Theta : \|\theta - \theta_0\| \le \delta'\} \subset B_\delta$. For any fixed $\theta \in B_{\delta'}$, we have

$$\begin{aligned} \left|\left|\left| \ddot{M}_i(\theta) - \ddot{M}_0(\theta_0) \right|\right|\right| &\le \left|\left|\left| \ddot{M}_i(\theta) - \ddot{M}_i(\theta_0) \right|\right|\right| + \left|\left|\left| \ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0) \right|\right|\right| \\ &\le 2L\|\theta - \theta_0\| + \frac{\lambda}{4} \le \lambda/4 + \lambda/4 = \lambda/2, \end{aligned}$$

where we apply event E_1, Assumption 2.2.3 and the fact that $\delta' = \min(\delta, \frac{\lambda}{8L})$ on the first term and event E_2 on the second term. Since $\ddot{M}_0(\theta_0)$ is negative definite by Assumption

2.2.2, above inequality implies that $\ddot{M}_i(\theta)$ is negative definite for all $\theta \in B_{\delta'}$ and

$$\sup_{u\in\mathbb{R}^d:\|u\|\leq 1} u^t \ddot{M}_i(\theta) u \leq -\lambda/2. \tag{B.36}$$

With negative definiteness of $\ddot{M}_i(\theta), \theta \in B_{\delta'}$, event E_3 and concavity of $M_i(\theta), \theta \in \Theta$, we have

$$\frac{\lambda}{2}\delta' \overset{E_3}{\geq} \|\dot{M}_i(\theta_0)\| = \|\dot{M}_i(\theta_0) - \dot{M}_i(\theta_i)\| \overset{(B.36)}{\geq} \frac{\lambda}{2}\|\theta_i - \theta_0\|.$$

Thus, we know $\|\theta_i - \theta_0\| \leq \delta'$, or equivalently, $\theta_i \in B_{\delta'}$. Then by applying Taylor's Theorem on $M_i(\theta)$ at θ_0, we have

$$M_i(\theta_i) \overset{(B.36)}{\leq} M_i(\theta_0) + \dot{M}_i(\theta_0)^t(\theta_i - \theta_0) - \frac{\lambda}{4}\|\theta_i - \theta_0\|^2.$$

Thus, as $M_i(\theta_0) \leq M_i(\theta_i)$ by definiton,

$$\begin{aligned}\frac{\lambda}{4}\|\theta_i - \theta_0\|^2 &\leq M_i(\theta_0) - M_i(\theta_i) + \dot{M}_i(\theta_0)^t(\theta_i - \theta_0) \\ &\leq \|\dot{M}_i(\theta_0)\|\|\theta_i - \theta_0\|,\end{aligned}$$

which implies

$$\|\theta_i - \theta_0\| \leq \frac{4}{\lambda}\|\dot{M}_i(\theta_0)\|.$$

□

For $1 \leq q \leq 8$, we can bound $\mathbb{E}[\|\dot{M}_i(\theta_0)\|^q]$ by Lemma B.1.2 and Assumption 2.2.3,

$$\mathbb{E}[\|\dot{M}_i(\theta_0)\|^q] \leq \frac{C_v(q,d)}{n^{q/2}} G^q,$$

where $C_v(q,d)$ is a constant depending on q and d only. Then by conditioning on event E,

we have

$$\begin{aligned}\mathbb{E}[\|\theta_i-\theta_0\|^q] &= \mathbb{E}[\|\theta_i-\theta_0\|^q 1_{(E)}]+\mathbb{E}[\|\theta_i-\theta_0\|^q 1_{(E^c)}]\\ &\le \frac{4^q}{\lambda^q}\mathbb{E}[\|\dot{M}_i(\theta_0)\|^q]+D^q\mathbf{Pr}(E^c)\\ &\le \frac{4^q}{\lambda^q}\frac{C_v(q,d)}{n^{q/2}}G^q+D^q\mathbf{Pr}(E^c).\end{aligned}$$

If we can show $\mathbf{Pr}(E^c)=O(n^{-\frac{q}{2}})$, then $\mathbb{E}[\|\theta_i-\theta_0\|^q]=O(n^{-\frac{q}{2}})$ follows immediately.

Lemma B.2.3. *Under Assumption 2.2.3, we have*

$$\boldsymbol{Pr}(E^c)=O(n^{-4}).$$

Proof. Under Assumption 2.2.3, by applying Lemma B.1.2 and B.1.3, we can bound the moments of $\dot{M}_i(\theta_0)$ and $\ddot{M}_i(\theta_0)-\ddot{M}_0(\theta_0)$. Rigorously, for $1\le q\le 8$, we have

$$\begin{aligned}\mathbb{E}\left[\|\dot{M}_i(\theta_0)\|^q\right] &\le \frac{C_v(q,d)}{n^{q/2}}G^q,\\ \mathbb{E}\left[\left|\left|\left|\ddot{M}_i(\theta_0)-\ddot{M}_0(\theta_0)\right|\right|\right|^q\right] &\le \frac{C_m(q,d)}{n^{q/2}}H^q.\end{aligned}$$

Therefore, by Markov's inequality, we have

$$\begin{aligned}&\mathbf{Pr}(E^c)=\mathbf{Pr}(E_1^c\cup E_2^c\cup E_3^c)\le \mathbf{Pr}(E_1^c)+\mathbf{Pr}(E_2^c)+\mathbf{Pr}(E_3^c)\\ \le\ & \frac{\mathbb{E}\left[|\frac{1}{n}\sum_{j=1}^n L(x_j)-\mathbb{E}[L(x)]|^8\right]}{L^8}+\frac{\mathbb{E}\left[\left|\left|\left|\ddot{M}_i(\theta_0)-\ddot{M}_0(\theta_0)\right|\right|\right|^8\right]}{(\lambda/4)^8}+\frac{\mathbb{E}\left[\|\dot{M}_i(\theta_0)\|^8\right]}{(\lambda\delta'/2)^8}\\ \le\ & O(\frac{1}{n^4})+O(\frac{1}{n^4})+O(\frac{1}{n^4})=O(n^{-4}).\end{aligned}$$

□

Now, we have showed that for $1\le q\le 8$,

$$\mathbb{E}[\|\theta_i-\theta_0\|^q]\le\frac{4^q}{\lambda^q}\frac{C_v(q,d)}{n^{q/2}}G^q+O(n^{-4})=O(n^{-\frac{q}{2}}). \tag{B.37}$$

Until now, $[\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i - \theta_0]$ has been well bounded. Next, we will consider the moment bound of $\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho - \ddot{M}_i(\theta_0)$.

Lemma B.2.4. *Under assumption 2.2.3, for* $1 \leq q \leq 4$*,*

$$\begin{aligned}&\mathbb{E}\left[\left|\left|\left|\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho - \ddot{M}_i(\theta_0)\right|\right|\right|^q\right]\\ \leq& L^q \frac{4^q}{\lambda^q}\frac{\sqrt{C_v(2q,d)}}{n^{q/2}}G^q + O(n^{-2}) = O(n^{-q/2}).\end{aligned}$$

Proof. By Minkowski's integral inequality, we have

$$\begin{aligned}&\mathbb{E}\left[\left|\left|\left|\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho - \ddot{M}_i(\theta_0)\right|\right|\right|^q\right]\\ \leq& \mathbb{E}\left[\int_0^1 \left|\left|\left|\ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) - \ddot{M}_i(\theta_0)\right|\right|\right|^q d\rho\right]\\ =& \int_0^1 \mathbb{E}\left[\left|\left|\left|\ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) - \ddot{M}_i(\theta_0)\right|\right|\right|^q\right] d\rho.\end{aligned}$$

For simplicity of notation, we use $\theta' = (1-\rho)\theta_0 + \rho\theta_i$ in this proof. When event E holds, we have

$$\|\theta' - \theta_0\| = \|\rho(\theta_i - \theta_0)\| \leq \rho\delta' \leq \delta,$$

which means that $\theta' \in B_\delta, \forall \rho \in [0,1]$. Thus, because of the convexity of the matrix norm $|||\cdot|||$, we can apply Jensen's inequality and Assumption 2.2.3 and get

$$\left|\left|\left|\ddot{M}_i(\theta') - \ddot{M}_i(\theta_0)\right|\right|\right|^q \leq \frac{1}{n}\sum_{j=1}^n |||\ddot{m}(x_j;\theta') - \ddot{m}(x_j;\theta_0)|||^q \leq \frac{1}{n}\sum_{j=1}^n L(x_i)^q\|\theta' - \theta_0\|^q.$$

Then apply Hölder's inequality,

$$\begin{aligned}\mathbb{E}\left[\left|\left|\left|\ddot{M}_i(\theta') - \ddot{M}_i(\theta_0)\right|\right|\right|^q 1_{(E)}\right] \quad &\leq \quad \left\{\mathbb{E}[(\frac{1}{n}\sum_{j=1}^n L(x_i)^q)^2]\right\}^{1/2}\left\{\mathbb{E}[\|\theta' - \theta_0\|^{2q}]\right\}^{1/2}\\ &\overset{\text{Jensen's}}{\leq} \quad C(q)L^q\rho^q\left\{\mathbb{E}[\|\theta_i - \theta_0\|^{2q}]\right\}^{1/2}\\ &\overset{\text{(B.37)}}{\leq} \quad C(q)L^q\frac{4^q\sqrt{C(2q,d)}G^q}{\lambda^q n^{q/2}} + O(n^{-2}).\end{aligned}$$

When event E does not hold, we know that $\left|\left|\left|\ddot{M}_i(\theta') - \ddot{M}_i(\theta_0)\right|\right|\right|^q$ must be finite by the assumption that Θ is compact and $\ddot{M}_i(\theta)$ is continuous. By Lemma B.2.3, the probability that event E does not hold is bounded by $O(n^{-4})$, which implies,

$$\mathbb{E}\left[\left|\left|\left|\ddot{M}_i(\theta') - \ddot{M}_i(\theta_0)\right|\right|\right|^q\right] \leq C(q)L^q \frac{4^q\sqrt{C(2q,d)}G^q}{\lambda^q n^{q/2}} + O(n^{-2}) + O(n^{-4}) = O(n^{-q/2}).$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\left|\left|\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho - \ddot{M}_i(\theta_0)\right|\right|\right|^q\right] \\
\leq &\int_0^1 \mathbb{E}\left[\left|\left|\left|\ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) - \ddot{M}_i(\theta_0)\right|\right|\right|^q\right] d\rho \\
\leq &C(q)L^q \frac{4^q\sqrt{C(2q,d)}G^q}{\lambda^q n^{q/2}} + O(n^{-2}) + O(n^{-4}) = O(n^{-q/2}).
\end{aligned}$$

□

Now, recall that we have

$$\begin{aligned}
0 = \dot{M}_i(\theta_0) + \ddot{M}_0(\theta_0)[\theta_i - \theta_0] + \left[\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho - \ddot{M}_i(\theta_0)\right][\theta_i - \theta_0] \\
+ [\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i - \theta_0]. \quad \text{(B.38)}
\end{aligned}$$

For the sum of last two terms, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|[\int_0^1 \ddot{M}_i((1-\rho)\theta_0+\rho\theta_i)d\rho-\ddot{M}_i(\theta_0)][\theta_i-\theta_0]+[\ddot{M}_i(\theta_0)-\ddot{M}_0(\theta_0)][\theta_i-\theta_0]\right\|^2\right] \\
\leq\ & 2\mathbb{E}\left[\left\|[\int_0^1 \ddot{M}_i((1-\rho)\theta_0+\rho\theta_i)d\rho-\ddot{M}_i(\theta_0)][\theta_i-\theta_0]\right\|^2\right] \\
&+2\mathbb{E}\left[\|[\ddot{M}_i(\theta_0)-\ddot{M}_0(\theta_0)][\theta_i-\theta_0]\|^2\right] \qquad \text{(since } (a+b)^2\leq 2a^2+2b^2\text{)} \\
\leq\ & 2(\mathbb{E}\left[\left\|\int_0^1 \ddot{M}_i((1-\rho)\theta_0+\rho\theta_i)d\rho-\ddot{M}_i(\theta_0)\right\|^4\right])^{1/2}(\mathbb{E}[\|\theta_i-\theta_0\|^4])^{1/2} \\
&+2(\mathbb{E}[\|\ddot{M}_i(\theta_0)-\ddot{M}_0(\theta_0)\|^4])^{1/2}(\mathbb{E}[\|\theta_i-\theta_0\|^4])^{1/2} \qquad \text{(Hölder's inequality)} \\
=\ & O(n^{-2})+O(n^{-2}) \qquad \text{(Lemma B.1.3 \& B.2.4 and (B.37))} \\
=\ & O(n^{-2}).
\end{aligned}
$$

Unitl now, we have established the upper bound for the mean squared error of local M-estimators,

$$\mathbb{E}[\|\theta_i-\theta_0\|^2]\leq\frac{2}{n}\text{Tr}(\Sigma)+O(n^{-2}),$$

for $i=1,\ldots,k$.

B.2.2 Bound the Error of Simple Averaging Estimator $\theta^{(0)}$

Next, we will study the mean squared error of simple averaging estimator,

$$\theta^{(0)}=\frac{1}{k}\sum_{i=1}^k\theta_i.$$

We start with a lemma, which bounds the bias of local M-estimator $\theta_i, i=1,\ldots,k$.

Lemma B.2.5. *There exists some constant $\tilde{C}>0$ such that for $i=1,\ldots,k$, we have*

$$\|\mathbb{E}[\theta_i-\theta_0]\|\leq\frac{\tilde{C}}{n}+O(n^{-2}),$$

where $\tilde{C}=16[C_v(4,d)]^{\frac{1}{4}}\sqrt{C_v(2,d)}\lambda^{-3}G^2L+4\sqrt{C_m(2,d)}\sqrt{C_v(2,d)}\lambda^{-2}GH$.

Proof. The main idea of this proof is to use equation (B.38) and apply the established error bounds of Hessian and the aforementioned local m-estimators. By equation (B.38) and fact $\mathbb{E}[M_i(\theta_0)] = 0$, we have

$$
\begin{aligned}
& \|\mathbb{E}[\theta_i - \theta_0]\| \\
= \quad & \|\mathbb{E}\{\ddot{M}_0(\theta_0)^{-1}\left[\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho - \ddot{M}_i(\theta_0)\right][\theta_i - \theta_0] \\
& \quad + \ddot{M}_0(\theta_0)^{-1}[\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i - \theta_0]\}\| \\
\leq \quad & \|\mathbb{E}\{\ddot{M}_0(\theta_0)^{-1}\left[\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho - \ddot{M}_i(\theta_0)\right][\theta_i - \theta_0]\}\| \\
& \quad + \|\mathbb{E}\{\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i - \theta_0]\}\| \\
\overset{\text{Jensen's}}{\leq} \quad & \mathbb{E}\left[\|\ddot{M}_0(\theta_0)^{-1}\left[\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho - \ddot{M}_i(\theta_0)\right][\theta_i - \theta_0]\|\right] \\
& \quad + \mathbb{E}\left[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i - \theta_0]\|\right] \\
\overset{\text{Assumption 2.2.2}}{\leq} \quad & \lambda^{-1}\mathbb{E}\left[\left\|\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho - \ddot{M}_i(\theta_0)\right\| \|\theta_i - \theta_0\|\right] \\
& \quad + \lambda^{-1}\mathbb{E}\left[\|\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)\|\|\theta_i - \theta_0\|\right] \\
\overset{\text{Hölder's}}{\leq} \quad & \lambda^{-1}\mathbb{E}\left[\left\|\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i)d\rho - \ddot{M}_i(\theta_0)\right\|^2\right]^{1/2} \mathbb{E}[\|\theta_i - \theta_0\|^2]^{1/2} \\
& \quad + \lambda^{-1}\mathbb{E}[\|\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)\|^2]^{1/2}\mathbb{E}[\|\theta_i - \theta_0\|^2]^{1/2}.
\end{aligned}
$$

Then we can apply Lemma B.1.3 & B.2.4, and (B.37) to bound each term, thus, we have

$$\begin{aligned}
\|\mathbb{E}[\theta_i - \theta_0]\|] &\leq \lambda^{-1}\sqrt{L^2\frac{4^2}{\lambda^2}\frac{\sqrt{C_v(4,d)}}{n}G^2 + O(n^{-2})}\sqrt{\frac{4^2}{\lambda^2}\frac{C_v(2,d)}{n}G^2 + O(n^{-4})} \\
&\quad + \lambda^{-1}\sqrt{\frac{C_m(2,d)}{n}H^2}\sqrt{\frac{4^2}{\lambda^2}\frac{C_v(2,d)}{n}G^2 + O(n^{-4})} \\
&\leq \lambda^{-1}\left[L\frac{4}{\lambda}\frac{C_v(4,d)^{1/4}}{\sqrt{n}}G + O(n^{-\frac{3}{2}})\right]\left[\frac{4}{\lambda}\frac{\sqrt{C_v(2,d)}}{\sqrt{n}}G + O(n^{-\frac{7}{2}})\right] \\
&\quad + \lambda^{-1}\frac{\sqrt{C_m(2,d)}}{\sqrt{n}}H\left[\frac{4}{\lambda}\frac{\sqrt{C_v(2,d)}}{\sqrt{n}}G + O(n^{-\frac{7}{2}})\right] \\
&= L\frac{4^2}{\lambda^3}\frac{C_v(4,d)^{1/4}\sqrt{C_v(2,d)}}{n}G^2 + O(n^{-2}) \\
&\quad + \frac{4}{\lambda^2}\frac{\sqrt{C_m(2,d)}\sqrt{C_v(2,d)}}{n}GH + O(n^{-4}).
\end{aligned}$$

Let $\tilde{C} = 16[C_v(4,d)]^{\frac{1}{4}}\sqrt{C_v(2,d)}\lambda^{-3}G^2L + 4\sqrt{C_m(2,d)}\sqrt{C_v(2,d)}\lambda^{-2}GH$, then we have

$$\|\mathbb{E}[\theta_i - \theta_0]\| \leq \frac{\tilde{C}}{n} + O(n^{-2}).$$

□

Then we can show that the MSE of $\theta^{(0)}$ could be bounded as follows.

Lemma B.2.6. *There exists some constant $\tilde{C} > 0$ such that*

$$\mathbb{E}[\|\theta^{(0)} - \theta_0\|^2] \leq \frac{2}{N}\textit{Tr}(\Sigma) + \frac{\tilde{C}^2k^2}{N^2} + O(kN^{-2}) + O(k^3N^{-3}),$$

where $\tilde{C} = 16[C_v(4,d)]^{\frac{1}{4}}\sqrt{C_v(2,d)}\lambda^{-3}G^2L + 4\sqrt{C_m(2,d)}\sqrt{C_v(2,d)}\lambda^{-2}GH$.

Proof. The mean squared error of $\theta^{(0)}$ could be decomposed into two parts: covariance and bias. Thus,

$$\begin{aligned}
\mathbb{E}[\|\theta^{(0)} - \theta_0\|^2] &= \text{Tr}(\text{Cov}[\theta^{(0)}]) + \|\mathbb{E}[\theta^{(0)} - \theta_0]\|^2 \\
&= \frac{1}{k}\text{Tr}(\text{Cov}[\theta_1]) + \|\mathbb{E}[\theta_1 - \theta_0]\|^2 \\
&\leq \frac{1}{k}\mathbb{E}[\|\theta_1 - \theta_0\|^2] + \|\mathbb{E}[\theta_1 - \theta_0]\|^2,
\end{aligned}$$

where the first term is well bounded by Lemma B.2.1 and the second term could be bounded by Lemma B.2.5. Thus, we know

$$\mathbb{E}[\|\theta^{(0)} - \theta_0\|^2] \leq \frac{2}{N}\text{Tr}(\Sigma) + \frac{\tilde{C}^2 k^2}{N^2} + O(kN^{-2}) + O(k^3 N^{-3}).$$

More generally, for $1 \leq q \leq 8$, we have

$$\begin{aligned}
& \mathbb{E}[\|\theta^{(0)} - \theta_0\|^q] = \mathbb{E}[\|(\theta^{(0)} - \mathbb{E}[\theta^{(0)}]) + (\mathbb{E}[\theta^{(0)}] - \theta_0)\|^q] \\
\leq \quad & 2^q \mathbb{E}[\|\theta^{(0)} - \mathbb{E}[\theta^{(0)}]\|^q] + 2^q \|\mathbb{E}[\theta^{(0)}] - \theta_0\|^q \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(since } (a+b)^q \leq 2^q a^q + 2^q b^q\text{)} \\
= \quad & 2^q \mathbb{E}[\|\theta^{(0)} - \mathbb{E}[\theta^{(0)}]\|^q] + 2^q \|\mathbb{E}[\theta_1] - \theta_0\|^q \qquad \text{(since } \mathbb{E}[\theta^{(0)}] = \mathbb{E}[\theta_1]\text{)} \\
\leq \quad & 2^q \mathbb{E}[\|\theta^{(0)} - \theta_0\|^q] + 2^q \frac{\tilde{C}^q}{n^q} + O(n^{-q-1}) \\
\overset{\text{Lemma B.1.2}}{\leq} \quad & 2^q \frac{C_v(q,d)}{k^{q/2}} \mathbb{E}[\|\theta_1 - \theta_0\|^q]) + 2^q \frac{\tilde{C}^q}{n^q} + O(n^{-q-1}) \\
\overset{\text{(B.37)}}{\leq} \quad & 2^q \frac{C_v(q,d)}{k^{q/2}} \left[\frac{4^q}{\lambda^q} \frac{C(q,d)}{n^{q/2}} G^q + O(n^{-4})\right] + 2^q \frac{\tilde{C}^q}{n^q} + O(n^{-q-1}) \\
= \quad & 8^q [C_v(q,d)]^2 \lambda^{-q} G^q N^{-\frac{q}{2}} + O(N^{-\frac{q}{2}} n^{\frac{q}{2}-4}) + 2^q \frac{\tilde{C}^q}{n^q} + O(n^{-q-1}).
\end{aligned}$$

In summary, we have

$$\mathbb{E}[\|\theta^{(0)} - \theta_0\|^q] \leq O(N^{-\frac{q}{2}}) + \frac{2^q \tilde{C}^q k^q}{N^q} + O(k^{q+1} N^{-q-1}) = O(N^{-\frac{q}{2}}) + O(k^q N^{-q}). \quad \text{(B.39)}$$

□

B.3 Proof of Theorem 2.2.4

The whole proof could be completed in two steps: first, show simple averaging estimator $\theta^{(0)}$ is $\sqrt{N}$-consistent when $k = O(\sqrt{N})$; then show the consistency and asymptotic normality of the one-step estimator $\theta^{(1)}$. In the first step, we need to show the following.

Lemma B.3.1. *Under Assumption 2.2.1, 2.2.2 and 2.2.3, when $k = O(\sqrt{N})$, the simple averaging estimator $\theta^{(0)}$ is $\sqrt{N}$-consistent estimator of θ_0, i.e.,*

$$\sqrt{N}\|\theta^{(0)} - \theta_0\| = O_P(1) \text{ as } N \to \infty.$$

Proof. If k is finite and does not grow with N, the proof is trivial. So, we just need to consider the case that $k \to \infty$. We know that $\|\mathbb{E}[\sqrt{n}(\theta_i - \theta_0)]\| \leq O(\frac{1}{\sqrt{n}})$ by Lemma B.2.5 and $\mathbb{E}[\|\sqrt{n}(\theta_i - \theta_0)\|^2] \leq 2\text{Tr}(\Sigma) + O(n^{-1})$ by Lemma B.2.1. By applying Lindeberg-Lévy Central Limit Theorem, we have

$$\begin{aligned}
&\sqrt{N}(\theta^{(0)} - \theta_0) = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \sqrt{n}(\theta_i - \theta_0) \\
=&\frac{1}{\sqrt{k}} \sum_{i=1}^{k} \{\sqrt{n}(\theta_i - \theta_0) - \mathbb{E}[\sqrt{n}(\theta_i - \theta_0)]\} + \sqrt{nk}\mathbb{E}[\theta_1 - \theta_0] \\
\xrightarrow{d}&\mathbf{N}(0, \Sigma) + \lim_{N \to \infty} \sqrt{nk}\mathbb{E}[\theta_1 - \theta_0],
\end{aligned}$$

It suffices to show $\lim_{N\to\infty} \sqrt{nk}\mathbb{E}[\theta_1 - \theta_0]$ is finite. By Lemma B.2.5, we have

$$\|\mathbb{E}[\theta_i - \theta_0]\| = O(\frac{1}{n}), \forall i \in \{1, 2, \ldots, k\},$$

which means that $\|\sqrt{nk}\mathbb{E}[\theta_i - \theta_0]\| = O(1)$ if $k = O(\sqrt{N}) = O(n)$. Thus, when $k = O(\sqrt{N})$, $\sqrt{N}(\theta^{(0)} - \theta_0)$ is bounded in probability. □

Now, we can prove Theorem 2.2.4.

Proof. By the definition of the one-step estimator

$$\theta^{(1)} = \theta^{(0)} - \ddot{M}(\theta^{(0)})^{-1}\dot{M}(\theta^{(0)}),$$

and by Theorem 4.2 in Chapter XIII of [40], we have

$$\sqrt{N}\ddot{M}(\theta^{(0)})(\theta^{(1)}-\theta_0) = \ddot{M}(\theta^{(0)})\sqrt{N}(\theta^{(0)}-\theta_0) - \sqrt{N}(\dot{M}(\theta^{(0)})-\dot{M}(\theta_0)) - \sqrt{N}\dot{M}(\theta_0)$$
$$= \ddot{M}(\theta^{(0)})\sqrt{N}(\theta^{(0)}-\theta_0) - \sqrt{N}\int_0^1 \ddot{M}((1-\rho)\theta_0+\rho\theta^{(0)})d\rho\,(\theta^{(0)}-\theta_0) - \sqrt{N}\dot{M}(\theta_0)$$
$$= \left[\ddot{M}(\theta^{(0)}) - \int_0^1 \ddot{M}((1-\rho)\theta_0+\rho\theta^{(0)})d\rho\right]\sqrt{N}(\theta^{(0)}-\theta_0) - \sqrt{N}\dot{M}(\theta_0),$$

As it is shown in (B.39), for any $\rho \in [0,1]$, when $k = O(\sqrt{N})$, we have

$$\|(1-\rho)\theta_0+\rho\theta^{(0)}-\theta_0\| \le \rho\|\theta^{(0)}-\theta_0\| \xrightarrow{P} 0.$$

Since $\ddot{M}(\cdot)$ is a continuous function, $\left\|\left|\ddot{M}(\theta^{(0)}) - \int_0^1 \ddot{M}((1-\rho)\theta_0+\rho\theta^{(0)})d\rho\right|\right\| \xrightarrow{P} 0$.

Thus,

$$\sqrt{N}\ddot{M}(\theta^{(0)})(\theta^{(1)}-\theta_0) = -\sqrt{N}\dot{M}(\theta_0) + o_P(1).$$

And, $\ddot{M}(\theta^{(0)}) \xrightarrow{P} \ddot{M}_0(\theta_0)$ because of $\theta^{(0)} \xrightarrow{P} \theta_0$ and Law of Large Number. Therefore, we can obatain

$$\sqrt{N}(\theta^{(1)}-\theta_0) \xrightarrow{d} \mathbf{N}(0,\Sigma) \text{ as } N \to \infty$$

by applying Slutsky's Lemma. □

B.4 Proof of Theorem 2.2.5

Let us recall the formula for one-step estimator,

$$\theta^{(1)} = \theta^{(0)} - \ddot{M}(\theta^{(0)})^{-1}\dot{M}(\theta^{(0)}).$$

Then by Theorem 4.2 in Chapter XIII of [40], we have

$$\begin{aligned}
&\ddot{M}_0(\theta_0)(\theta^{(1)}-\theta_0) = [\ddot{M}_0(\theta_0)-\ddot{M}(\theta^{(0)})](\theta^{(1)}-\theta_0)+\ddot{M}(\theta^{(0)})(\theta^{(1)}-\theta_0)\\
=\;& [\ddot{M}_0(\theta_0)-\ddot{M}(\theta^{(0)})](\theta^{(1)}-\theta_0)+\ddot{M}(\theta^{(0)})(\theta^{(0)}-\theta_0)-[\dot{M}(\theta^{(0)})-\dot{M}(\theta_0)]-\dot{M}(\theta_0)\\
=\;& [\ddot{M}_0(\theta_0)-\ddot{M}(\theta^{(0)})](\theta^{(1)}-\theta_0)\\
&\qquad +\left[\ddot{M}(\theta_0)-\int_0^1 \ddot{M}((1-\rho)\theta_0+\rho\theta^{(0)})d\rho\right](\theta^{(0)}-\theta_0)-\dot{M}(\theta_0).
\end{aligned}$$

Then we have

$$\begin{aligned}
\theta^{(1)}-\theta_0 = &-\ddot{M}_0(\theta_0)^{-1}\dot{M}(\theta_0)+\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_0(\theta_0)-\ddot{M}(\theta^{(0)})](\theta^{(1)}-\theta_0)\\
&+\ddot{M}_0(\theta_0)^{-1}\left[\ddot{M}(\theta_0)-\int_0^1 \ddot{M}((1-\rho)\theta_0+\rho\theta^{(0)})d\rho\right](\theta^{(0)}-\theta_0) \quad \text{(B.40)}
\end{aligned}$$

We will show the last two terms are small enough. Similar to the proof of Lemma B.2.1, we define a "good" event:

$$E_4 \;=\; \{\|\theta^{(0)}-\theta_0\|\le\delta\}.$$

The probability of above event is close to 1 when N is large.

$$\Pr(E_4^c)\le\frac{\mathbb{E}[\|\theta^{(0)}-\theta_0\|^8]}{\delta^8}\le O(N^{-4})+O(k^8N^{-8}).$$

Lemma B.4.1. *If event E_4 holds, for $1\le q\le 4$, we have*

$$\begin{aligned}
\mathbb{E}\left[\left|\left|\left|\ddot{M}_0(\theta_0)-\ddot{M}(\theta^{(0)})\right|\right|\right|^q\right] &\le O(N^{-\frac{q}{2}})+O(k^qN^{-q}),\\
\mathbb{E}\left[\left|\left|\left|\ddot{M}(\theta_0)-\int_0^1\ddot{M}((1-\rho)\theta_0+\rho\theta^{(0)})d\rho\right|\right|\right|^q\right] &\le O(N^{-\frac{q}{2}})+O(k^qN^{-q}).
\end{aligned}$$

Proof. By Lemma B.1.3, we know

$$\mathbb{E}\left[\left|\left|\left|\ddot{M}_0(\theta_0)-\ddot{M}(\theta_0)\right|\right|\right|^q\right]\le\frac{C(q,d)}{N^{q/2}}H^q.$$

Under event E_4 and Assumption 2.2.3, by applying Jensen's inequality, we have

$$\begin{aligned}\left|\left|\left|\ddot{M}(\theta_0) - \ddot{M}(\theta^{(0)})\right|\right|\right|^q &\leq \frac{1}{N}\sum_{i=1}^{k}\sum_{x\in S_i}\left|\left|\left|\ddot{m}(x;\theta^{(0)}) - \ddot{m}(x;\theta_0)\right|\right|\right|^q \\ &\leq \frac{1}{N}\sum_{i=1}^{k}\sum_{x\in S_i} L(x)^q\|\theta^{(0)} - \theta_0\|^q.\end{aligned}$$

Thus, for $1 \leq q \leq 4$, we have

$$\begin{aligned}\mathbb{E}\left[\left|\left|\left|\ddot{M}(\theta_0) - \ddot{M}(\theta^{(0)})\right|\right|\right|^q\right] &\overset{\text{Hölder's}}{\leq} \mathbb{E}\left[(\frac{1}{N}\sum_{i=1}^{k}\sum_{x\in S_i} L(x)^q)^2\right]^{\frac{1}{2}}\mathbb{E}\left[\|\theta^{(0)} - \theta_0\|^{2q}\right]^{\frac{1}{2}} \\ &\overset{\text{(B.39)}}{\leq} O(N^{-\frac{q}{2}}) + \frac{2^q\tilde{C}^qL^qk^q}{N^q} + O(\frac{k^{q+1}}{N^{q+1}}).\end{aligned}$$

As a result, we have, for $1 \leq q \leq 4$,

$$\begin{aligned}&\mathbb{E}\left[\left|\left|\left|\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})\right|\right|\right|^q\right] \\ \leq\ & 2^q\mathbb{E}\left[\left|\left|\left|\ddot{M}_0(\theta_0) - \ddot{M}(\theta_0))\right|\right|\right|^q\right] + 2^q\mathbb{E}\left[\left|\left|\left|\ddot{M}(\theta_0) - \ddot{M}(\theta^{(0)})\right|\right|\right|^q\right] \\ \leq\ & O(N^{-\frac{q}{2}}) + \frac{4^q\tilde{C}^qL^qk^q}{N^q} + O(\frac{k^{q+1}}{N^{q+1}}).\end{aligned}$$

In this proof, we let $\theta' = (1-\rho)\theta_0 + \rho\theta^{(0)}$ for the simplicity of notation. Note that $\theta' - \theta_0 = \rho(\theta^{(0)} - \theta_0)$, then by event E_4, Assumption 2.2.3 and inequality (B.39), we have

$$\begin{aligned}\mathbb{E}\left[\left|\left|\left|\ddot{M}(\theta_0) - \ddot{M}(\theta')\right|\right|\right|^q\right] &\overset{\text{Jensen's}}{\leq} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{k}\sum_{x\in S_i} L(x)^q\|\theta' - \theta_0\|^q\right] \\ &\overset{\text{Hölder's}}{\leq} \mathbb{E}\left[(\frac{1}{N}\sum_{i=1}^{k}\sum_{x\in S_i} L(x)^q)^2\right]^{\frac{1}{2}}\rho^q\mathbb{E}\left[\|\theta^{(0)} - \theta_0\|^{2q}\right]^{\frac{1}{2}} \\ &\leq O(N^{-\frac{q}{2}}) + \frac{2^q\tilde{C}^qL^qk^q}{N^q} + O(\frac{k^{q+1}}{N^{q+1}}).\end{aligned}$$

So, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\left|\left|\ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho\right|\right|\right|^q\right] \\
\leq &\mathbb{E}\left[\int_0^1 \left|\left|\left|\ddot{M}(\theta_0) - \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})\right|\right|\right|^q d\rho\right] \\
= &\int_0^1 \mathbb{E}\left[\left|\left|\left|\ddot{M}(\theta_0) - \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})\right|\right|\right|^q\right] d\rho \\
\leq &O(N^{-\frac{q}{2}}) + \frac{2^q\tilde{C}^q L^q k^q}{N^q} + O(\frac{k^{q+1}}{N^{q+1}}).
\end{aligned}$$

□

Therefore, under event E_4, for $1 \leq q \leq 4$, we can bound $\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0)$ and $\ddot{M}_0(\theta_0)^{-1}[\ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho](\theta^{(0)} - \theta_0)$ as follows:

$$\begin{aligned}
&\mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0)\|^q] \\
\leq \;& \lambda^{-q}\mathbb{E}\left[\left|\left|\left|\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})\right|\right|\right|^q\right] D^q \\
\leq \;& O(N^{-\frac{q}{2}}) + O(k^q N^{-q}).
\end{aligned}$$

and,

$$\begin{aligned}
&\mathbb{E}\left[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho](\theta^{(0)} - \theta_0)\|^q\right] \\
\leq \;& \lambda^{-q}\mathbb{E}\left[\left|\left|\left|\ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho\right|\right|\right|^q\right] D^q \\
\leq \;& O(N^{-\frac{q}{2}}) + O(k^q N^{-q}).
\end{aligned}$$

And by Lemma B.1.2, for $1 \leq q \leq 8$, we have

$$\mathbb{E}[\|\dot{M}(\theta_0)\|^q] = O(N^{-\frac{q}{2}}).$$

Therefore, combining above three bounds and equation (B.40), we have, for $1 \leq q \leq 4$,

$$\begin{aligned}
& \mathbb{E}[\|\theta^{(1)} - \theta_0\|^q] \\
\leq\ & 3^q \mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0)\|^q] \\
& + 3^q \mathbb{E}[|\ddot{M}_0(\theta_0)^{-1} \left[\ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho\right] (\theta^{(0)} - \theta_0)\|^q] \\
& + 3^q \mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}\dot{M}(\theta_0)\|^q] + Pr(E_4^c)D^q \\
=\ & O(N^{-\frac{q}{2}}) + O(k^q N^{-q}).
\end{aligned}$$

Now, we can give tighter bounds for the first two terms in equation (B.40) by Hölder's inequality.

$$\begin{aligned}
& \mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0)\|^2] \\
\leq\ & \lambda^{-2} \sqrt{\mathbb{E}\left[\left\|\left\|\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})\right\|\right\|^4\right]} \sqrt{\mathbb{E}[\|\theta^{(1)} - \theta_0\|^4]} \\
=\ & O(N^{-2}) + O(k^4 N^{-4}),
\end{aligned}$$

and,

$$\begin{aligned}
& \mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho](\theta^{(0)} - \theta_0)\|^2] \\
\leq\ & \lambda^{-2} \sqrt{\mathbb{E}\left[\left\|\left\|\ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho\right\|\right\|^4\right]} \sqrt{\mathbb{E}[\|\theta^{(0)} - \theta_0\|^4]} \\
=\ & O(N^{-2}) + O(k^4 N^{-4}).
\end{aligned}$$

Now, we can finalize our proof by using equation (B.38) again,

$$\begin{aligned}
& \mathbb{E}[\|\theta^{(1)} - \theta_0\|^2] \\
\leq\ & 2\mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}\dot{M}(\theta_0)\|^2] + 4\mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0)\|^2] \\
& + 4\mathbb{E}\left[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho](\theta^{(0)} - \theta_0)\|^2\right] \\
\leq\ & 2\frac{\mathrm{Tr}(\Sigma)}{N} + O(N^{-2}) + O(k^4 N^{-4}).
\end{aligned}$$

B.5 Proof of Corollary 2.2.7

At first, we will present a lemma on the negative moments of a Binomial random variable, i.e., $\mathbb{E}\left[\frac{1}{Z}1_{(Z>0)}\right]$ and $\mathbb{E}\left[\frac{1}{Z^2}1_{(Z>0)}\right]$, where $Z \sim \mathbf{B}(k,p)$ and $\mathbf{B}(k,p)$ denotes Binomial distribution with k independent trials and a success probability p for each trial. We believe that $\mathbb{E}\left[\frac{1}{Z}1_{(Z>0)}\right]$ and $\mathbb{E}\left[\frac{1}{Z^2}1_{(Z>0)}\right]$ should have been well studied. However, we did not find any appropriate reference on their upper bounds that we need. So, we derive the upper bounds as follows, which will be useful in the proof of Corollary 2.2.7.

Lemma B.5.1. *Suppose $Z \sim \mathbf{B}(k,p)$, when $z > 0$, we have*

$$\mathbb{E}\left[\frac{1}{Z}1_{(Z>0)}\right] < \frac{1}{kp} + \frac{3}{k^2p^2} \quad \textit{and} \quad \mathbb{E}\left[\frac{1}{Z^2}1_{(Z>0)}\right] < \frac{6}{k^2p^2}.$$

Proof. By definition, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{Z}1_{(Z>0)}\right] = \sum_{z=1}^{k}\frac{1}{z}\binom{k}{z}p^z(1-p)^{k-z} = \sum_{z=1}^{k}\frac{1}{z}\frac{k!}{z!(k-z)!}p^z(1-p)^{k-z}\\
=\ &\sum_{z=1}^{k}\frac{z+1}{z}\frac{1}{(k+1)p}\frac{(k+1)!}{(z+1)!(k-z)!}p^{z+1}(1-p)^{k-z}\\
=\ &\sum_{z=1}^{k}\frac{1}{(k+1)p}\binom{k+1}{z+1}p^{z+1}(1-p)^{k-z} + \sum_{z=1}^{k}\frac{1}{z}\frac{1}{(k+1)p}\binom{k+1}{z+1}p^{z+1}(1-p)^{k-z}\\
<\ &\frac{1}{(k+1)p} + \sum_{z=1}^{k}\frac{z+2}{z}\frac{1}{(k+1)(k+2)p^2}\binom{k+2}{z+2}p^{z+2}(1-p)^{k-z}\\
<\ &\frac{1}{(k+1)p} + \frac{3}{(k+1)(k+2)p^2} < \frac{1}{kp} + \frac{3}{k^2p^2}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{Z^2}1_{(Z>0)}\right] = \sum_{z=1}^{k}\frac{1}{z^2}\binom{k}{z}p^z(1-p)^{k-z} = \sum_{z=1}^{k}\frac{1}{z^2}\frac{k!}{z!(k-z)!}p^z(1-p)^{k-z} \\
= &\sum_{z=1}^{k}\frac{(z+1)(z+2)}{z^2}\frac{1}{(k+1)(k+2)p^2}\frac{(k+2)!}{(z+2)!(k-z)!}p^{z+2}(1-p)^{k-z} \\
\leq &\sum_{z=1}^{k}\frac{6}{(k+1)(k+2)p^2}\binom{k+2}{z+2}p^{z+2}(1-p)^{k-z} \\
< &\frac{6}{(k+1)(k+2)p^2} < \frac{6}{k^2p^2}.
\end{aligned}$$

□

Now, we can prove Corollary 2.2.7 could be as follows.

Proof. Let the random variable Z denote the number of machines that successfully communicate with the central machine, which means that Z follows Binomial distribution, $B(k, 1-r)$. By Law of Large Number, $\frac{Z}{(1-r)k} \xrightarrow{P} 1$ as $k \to \infty$. If Z is known, the size of available data becomes Zn. By Theorem 2.2.4, the one-step estimator $\theta^{(1)}$ is still asymptotic normal when $k = O(\sqrt{N})$,

$$\sqrt{Zn}(\theta^{(1)} - \theta_0) \xrightarrow{d} N(0, \Sigma) \text{ as } n \to \infty.$$

Therefore, when $k \to \infty$, we have

$$\sqrt{(1-r)N}(\theta^{(1)} - \theta_0) = \sqrt{\frac{(1-r)N}{Zn}}\sqrt{Zn}(\theta^{(1)} - \theta_0) \xrightarrow{d} \sqrt{\frac{(1-r)N}{Zn}}N(0, \Sigma).$$

Since $\frac{(1-r)N}{Zn} = \frac{(1-r)k}{Z} \xrightarrow{P} 1$, by Slutsky's Lemma, we have

$$\sqrt{(1-r)N}(\theta^{(0)} - \theta_0) \xrightarrow{d} N(0, \Sigma).$$

This result indicates that when the local machines could lose communication independently with central machine with probability q, the one-step estimator $\theta^{(1)}$ shares the same asymptotic properties with the oracle M-estimator using $(1-r) \times 100\%$ of the total samples.

Next, we will analyze the mean squared error of one-step estimator with the presence of local machine failures. Note that, when Z is fixed and known, by Theorem 2.2.5, we have

$$\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2 \big| Z] \leq \frac{2\text{Tr}[\Sigma]}{nZ} + O(n^{-2}Z^{-2}) + O(n^{-4}).$$

By Rule of Double Expectation and Lemma B.5.1,

$$\begin{aligned}
& \mathbb{E}[\|\theta^{(1)} - \theta_0\|^2 1_{(Z>0)}] = \mathbb{E}[\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2 \big| Z] 1_{(Z>0)}] \\
\leq \; & \mathbb{E}\left[\frac{2\text{Tr}[\Sigma]}{nZ} 1_{(Z>0)}\right] + \mathbb{E}[(O(n^{-2}Z^{-2}) + O(n^{-4}))1_{(Z>0)}] \\
\leq \; & 2\text{Tr}[\Sigma]\left\{\frac{1}{nk(1-r)} + \frac{3}{nk^2(1-r)^2}\right\} + O(n^{-2}k^{-2}(1-r)^{-2}) + O(n^{-4}) \\
= \; & \frac{2\text{Tr}[\Sigma]}{N(1-r)} + \frac{6\text{Tr}[\Sigma]}{Nk(1-r)^2} + O(N^{-2}(1-r)^{-2}) + O(k^2N^{-2}).
\end{aligned}$$

□

APPENDIX C

PROOFS OF DISTANCE COVARIANCE

We present all proofs for Chapter 3 here. For reader's convenience, we restate some constants that we have defined at the beginning of Chapter 3. We denote $c_p = \frac{\pi^{(p+1)/2}}{\Gamma((p+1)/2)}$ and $c_q = \frac{\pi^{(q+1)/2}}{\Gamma((q+1)/2)}$ as two constants, where $\Gamma(\cdot)$ denotes the Gamma function. We will also need the following constants: $C_p = \frac{c_1 c_{p-1}}{c_p} = \frac{\sqrt{\pi}\Gamma((p+1)/2)}{\Gamma(p/2)}$ and $C_q = \frac{c_1 c_{q-1}}{c_q} = \frac{\sqrt{\pi}\Gamma((q+1)/2)}{\Gamma(q/2)}$.

C.1 Proof of Lemma 3.3.1

Proof. The proof is straightforward as follows. It is known that X and Y are independent if and only if $\phi_{X,Y}(t,s) = \phi_X(t)\phi_Y(s), \forall t \in \mathbb{R}^p, s \in \mathbb{R}^q$, which by definition of the characteristic functions is equivalent to

$$\mathbb{E}[e^{iX^t t + iY^t s}] = \mathbb{E}[e^{iX^t t}]\mathbb{E}[e^{iY^t s}], \forall t \in \mathbb{R}^p, s \in \mathbb{R}^q.$$

Changing of variables $t = ut'$ and $s = vs'$ in the above expression results in the following:

$$\mathbb{E}[e^{iX^t ut' + iY^t vs'}] = \mathbb{E}[e^{iX^t ut'}]\mathbb{E}[e^{iY^t vs'}], \forall u \in \mathcal{S}^{p-1}, v \in \mathcal{S}^{q-1}, t', s' \in \mathbb{R},$$

or equivalently, the following

$$\mathbb{E}[e^{iu^t Xt' + iv^t Ys'}] = \mathbb{E}[e^{iu^t Xt'}]\mathbb{E}[e^{iv^t Ys'}], \forall u \in \mathcal{S}^{p-1}, v \in \mathcal{S}^{q-1}, t', s' \in \mathbb{R}.$$

Note the above, again by the definitions of the characteristic functions, is equivalent to

$$\phi_{u^t X, v^t Y}(t', s') = \phi_{u^t X}(t')\phi_{v^t Y}(s'), \forall u \in \mathcal{S}^{p-1}, v \in \mathcal{S}^{q-1}, t', s' \in \mathbb{R}.$$

From the definition and the properties of the distance covariance $\mathcal{V}^2$ (Theorem 3.1.1), we know that the previous is equivalent to

$$\mathcal{V}^2(u^tX, v^tY) = 0, \forall u \in \mathcal{S}^{p-1}, v \in \mathcal{S}^{q-1}.$$

From all the above, we have proved Lemma 3.3.1. □

C.2 Proof of Lemma 3.3.2

We prove Lemma 3.3.2.

Proof. We will use the following change of variables: $t = r_1 \cdot u, s = r_2 \cdot v$, where $r_1, r_2 \in (-\infty, +\infty)$ and $u \in \mathcal{S}^{p-1}, v \in \mathcal{S}^{q-1}$. As the surface area of $\mathcal{S}^{p-1}$ is equal to $\frac{2\pi^{p/2}}{\Gamma(p/2)} = 2c_{p-1}$, we have

$$\begin{aligned}
& \mathcal{V}^2(X, Y) \\
= & \int_{R^{p+q}} \frac{|\mathbb{E}[e^{iX^tt+iY^ts}] - \mathbb{E}[e^{iX^tt}]\mathbb{E}[e^{iY^ts}]|^2}{c_pc_q|t|^{p+1}|s|^{q+1}} dtds \\
= & c_{p-1}c_{q-1} \int_{\mathcal{S}_+^{p-1}} \int_{-\infty}^{+\infty} \int_{\mathcal{S}_+^{q-1}} \int_{-\infty}^{+\infty} \frac{|\mathbb{E}[e^{ir_1u^tX+ir_2v^tY}] - \mathbb{E}[e^{ir_1u^tX}]\mathbb{E}[e^{ir_2v^tY}]|^2}{c_pc_q|r_1|^{p+1}|r_2|^{q+1}} \\
& \qquad |r_1|^{p-1}|r_2|^{q-1} d\mu(u) dr_1 d\nu(v) dr_2 \\
= & c_{p-1}c_{q-1} \int_{\mathcal{S}_+^{p-1}} \int_{\mathcal{S}_+^{q-1}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{|\mathbb{E}[e^{ir_1u^tX+ir_2v^tY}] - \mathbb{E}[e^{ir_1u^tX}]\mathbb{E}[e^{ir_2v^tY}]|^2}{c_pc_q|r_1|^2|r_2|^2} \\
& \qquad d\mu(u) d\nu(v) dr_1 dr_2 \\
= & \frac{c_1^2 c_{p-1}c_{q-1}}{c_pc_q} \int_{\mathcal{S}_+^{p-1}} \int_{\mathcal{S}_+^{q-1}} \mathcal{V}^2(u^tX, v^tY) d\mu(u) d\nu(v) \\
= & C_pC_q \int_{\mathcal{S}^{p-1}} \int_{\mathcal{S}^{q-1}} \mathcal{V}^2(u^tX, v^tY) d\mu(u) d\nu(v).
\end{aligned}$$

In the above, the first and fourth equations are due to the definition of $\mathcal{V}^2(\cdot, \cdot)$; the second equation reflects the aforementioned change of variables; the third equation is a reorganization; the last equation is from the definition of constants C_p and C_q. From all the above,

we establish the first part of Lemma 3.3.2.

For the sample distance covariance part, we just need to replace the population characteristic function $\phi_X(t) = \mathbb{E}[e^{iX^t t}]$ with the sample characteristic function $\hat{\phi}_X(t) = \frac{1}{n}\sum_{j=1}^{n} e^{iX_j^t t}$, the rest reasoning part is nearly identical. We omit the details here. □

C.2.1 Proof of Lemma 3.3.3

We will need the following lemma.

Lemma C.2.1. *Suppose v is a fixed unit vector in $\mathbb{R}^{p-1}$ and $u \in \mathcal{S}^{p-1}$. Let μ be the uniform probability measure on $\mathcal{S}^{p-1}$. We have*

$$C_p \int_{\mathcal{S}^{p-1}} |u^t v| d\mu(u) = 1,$$

where constant C_p has been mentioned at the beginning of this chapter.

Proof. Since both u and v are unit vector, we have

$$|u^t v| = \left| \frac{\langle u, v \rangle}{\sqrt{|u||v|}} \right| = |\cos\theta|,$$

where θ is the angle between vectors u and v. As we know, the angle between two random vectors on $\mathcal{S}^{p-1}$ follows distribution with density, (see [15]) for $\theta \in [0, \pi]$,

$$h(\theta) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(p/2)}{\Gamma((p-1)/2)} (\sin\theta)^{p-2}. \tag{C.41}$$

Therefore, we have

$$\begin{aligned}
\int_{\mathcal{S}^{p-1}} |u^t v| d\mu(u) &= \int_0^{\pi} h(\theta) |\cos\theta| d\theta \\
&= 2\int_0^{\pi/2} h(\theta)\cos\theta d\theta \\
&\overset{(C.41)}{=} 2\int_0^{\pi/2} \frac{1}{\sqrt{\pi}} \frac{\Gamma(p/2)}{\Gamma((p-1)/2)} (\sin\theta)^{p-2}\cos\theta d\theta \\
&= 2\int_0^{1} \frac{1}{\sqrt{\pi}} \frac{\Gamma(p/2)}{\Gamma((p-1)/2)} x^{p-2} dx \\
&= \frac{2}{\sqrt{\pi}} \frac{\Gamma(p/2)}{\Gamma((p-1)/2)} \int_0^1 x^{p-2} dx \\
&= \frac{\Gamma(p/2)}{\sqrt{\pi}\Gamma((p+1)/2)} = \frac{1}{C_p}.
\end{aligned}$$

The second equation is due to the symmetry of the function on $[0, \pi]$; the third equation is a change of random variable; the sixth equation is from the fact that $\Gamma((p+1)/2) = \frac{p-1}{2}\Gamma((p-1)/2)$. □

We now prove Lemma 3.3.3

Proof. We will need the following notations:

$$\begin{aligned}
&a_{ij}^u = |u^t(X_i - X_j)|, \quad b_{ij}^v = |v^t(Y_i - Y_j)|, \\
&a_{i\cdot}^u = \sum_{l=1}^n a_{il}^u, \quad b_{i\cdot}^v = \sum_{l=1}^n b_{il}^v, \\
&a_{\cdot\cdot}^u = \sum_{k,l=1}^n a_{kl}^u, \text{ and } b_{\cdot\cdot}^v = \sum_{k,l=1}^n b_{kl}^v.
\end{aligned} \tag{C.42}$$

Recall the definition of $\Omega_n(\cdot,\cdot)$ in (3.1.15), we have

$$\begin{aligned}
\Omega_n(u^t X, v^t Y) = &\frac{1}{n(n-3)} \sum_{i\neq j} a_{ij}^u b_{ij}^v \\
&- \frac{2}{n(n-2)(n-3)} \sum_{i=1}^n a_{i\cdot}^u b_{i\cdot}^v + \frac{a_{\cdot\cdot}^u b_{\cdot\cdot}^v}{n(n-1)(n-2)(n-3)}.
\end{aligned} \tag{C.43}$$

By Lemma C.2.1, we have the following: $\forall 1 \leq i, j \leq n$,

$$C_p \int_{\mathcal{S}^{p-1}} |u^t(X_i - X_j)| d\mu(u) = |X_i - X_j| \quad \text{and} \tag{C.44}$$

$$C_q \int_{\mathcal{S}^{q-1}} |v^t(Y_i - Y_j)| d\nu(v) = |Y_i - Y_j|. \tag{C.45}$$

By integrating $\Omega_n(u^t X, v^t Y)$ on u and v, we have

$$\begin{aligned}
& C_p C_q \int_{\mathcal{S}^{p-1} \times \mathcal{S}^{q-1}} \Omega_n(u^t X, v^t Y) d\mu(u) d\nu(v) \\
\overset{\text{(C.43)}}{=} & \frac{1}{n(n-3)} \sum_{i \neq j} C_p \int_{\mathcal{S}^{p-1}} a_{ij}^u d\mu(u) C_q \int_{\mathcal{S}^{q-1}} b_{ij}^v d\nu(v) \\
& - \frac{2}{n(n-2)(n-3)} \sum_{i=1}^{n} C_p \int_{\mathcal{S}^{p-1}} a_{i\cdot}^u d\mu(u) C_q \int_{\mathcal{S}^{q-1}} b_{i\cdot}^v d\nu(v) \\
& + \frac{C_p \int_{\mathcal{S}^{p-1}} a_{\cdot\cdot}^u d\mu(u) C_q \int_{\mathcal{S}^{q-1}} b_{\cdot\cdot}^v d\nu(v)}{n(n-1)(n-2)(n-3)} \\
\overset{\text{(C.44)(C.45)}}{=} & \frac{1}{n(n-3)} \sum_{i \neq j} a_{ij} b_{ij} - \frac{2}{n(n-2)(n-3)} \sum_{i=1}^{n} a_{i\cdot} b_{i\cdot} \\
& + \frac{a_{\cdot\cdot} b_{\cdot\cdot}}{n(n-1)(n-2)(n-3)} = \Omega_n(X, Y).
\end{aligned}$$

From all the above, the equation in the lemma is established. □

C.2.2 Proof of Lemma 3.3.5

Proof. We can regard $\Omega_n(u^t X, v^t Y)$ as a real-valued function on $\mathbb{R}^p \times \mathbb{R}^q$. It is easy to find that $\Omega_n(u^t X, v^t Y)$ is a continuous differentiable function by its definition. Since $\mathcal{B}^p \times \mathcal{B}^q$ is a convex compact set, $\Omega_n(u^t X, v^t Y)$ must be bounded on this set. Let $L_{X,Y} = \sup_{u \in \mathcal{B}^p, v \in \mathcal{B}^q} \Omega_n(u^t X, v^t Y)$ denote this upper bound, which is constant depending on the distribution of X and Y only. Since $a_{ij}^u = |u^t(X_i - X_j)| \leq |u||X_i - X_j| = |X_i - X_j| = a_{ij}$,

then we have

$$
\begin{aligned}
L_{X,Y} &\leq \frac{1}{n(n-3)} \sum_{i \neq j} a_{ij} b_{ij} + \frac{a_{..} b_{..}}{n(n-1)(n-2)(n-3)} \\
&\leq \mathbb{E}[|X - X'||Y - Y'|] + \mathbb{E}[|X - X'|]\mathbb{E}[|Y - Y'|] + o_P(1) \\
&\leq 2\sqrt{\mathbb{E}[|X - X'|^2]\mathbb{E}[|Y - Y'|^2]} + o_P(1) \\
&\leq 2\sqrt{2\text{Tr}[\Sigma_X]2\text{Tr}[\Sigma_Y]} + o_P(1) \\
&\leq 5\sqrt{\text{Tr}[\Sigma_X]\text{Tr}[\Sigma_Y]} \text{ for sufficiently large } n.
\end{aligned}
$$

We can get the first inequality from the definition in (2.5) by removing the negative term. It is worth noting that $\frac{1}{n(n-3)} \sum_{i \neq j} a_{ij} b_{ij}$ and $\frac{a_{..} b_{..}}{n(n-1)(n-2)(n-3)}$ are the U-statistics for $\mathbb{E}[|X - X'||Y - Y'|]$ and $\mathbb{E}[|X - X'|]\mathbb{E}[|Y - Y'|]$, respectively. So, the second inequality is due to almost sure convergence of U-statistics, see [70, Chapter 5.4 Theorem A], where $o_P(1)$ represents a small error that converges to 0 as $n \to \infty$. The third inequality is an immediate result from Hölder's inequality. The fourth inequality holds as

$$
\begin{aligned}
\mathbb{E}[|X - X'|^2] &= \sum_{i=1}^{p} \mathbb{E}[(X_{(i)} - X'_{(i)})^2] = \sum_{i=1}^{p} (\mathbb{E}[X_{(i)}^2] + \mathbb{E}[X_{(i)}^2] - 2\mathbb{E}[X_{(i)} X'_{(i)}]) \\
&= 2 \sum_{i=1}^{p} (\mathbb{E}[X_{(i)}^2] - \mathbb{E}^2[X_{(i)}]) = 2 \sum_{i=1}^{p} \text{Var}(X_{(i)}) = 2\text{Tr}[\Sigma_X],
\end{aligned}
$$

where $X_{(i)}$ and $X'_{(i)}$ are the i-th component of X and X', respectively.

Since $(u_1, v_1), \ldots, (u_K, v_K)$ are draw i.i.d. from uniform distribution on $\mathcal{S}^{p-1} \times \mathcal{S}^{q-1}$. $h_1, \ldots, \Omega_K$ are i.i.d. random variables with $\mathbb{E}[\Omega^{(k)}] = \Omega_n, \forall k$. And, we know that $\Omega^{(k)} \leq$

$C_pC_qL_{X,Y}$. By Chernoff-Hoeffding's inequality [31], we have

$$\begin{aligned}
\mathbf{P}\left(\left|\overline{\Omega}_n - \Omega_n\right| > \epsilon\right) &= \mathbf{P}\left(\left|\sum_{k=1}^{K}\Omega^{(k)} - K\Omega_n\right| > K\epsilon\right) \\
&\leq 2\exp\left\{\frac{-2K^2\epsilon^2}{KC_p^2C_q^2L_{X,Y}^2}\right\} \\
&\leq 2\exp\left\{-\frac{2K\epsilon^2}{25C_p^2C_q^2\mathrm{Tr}[\Sigma_X]\mathrm{Tr}[\Sigma_Y]}\right\}.
\end{aligned}$$

□

C.2.3 Proof of Lemma 3.3.8

Proof. Recall that Ω_n is an unbiased estimator of $\mathcal{V}^2(X,Y)$ and $\Omega_4 = h_4$, we have $\mathbb{E}[h_4] = \mathcal{V}^2(X,Y) \geq 0$, consequently, we have the following:

$$
\begin{aligned}
&\text{Var}(h_4) \leq \mathbb{E}[h_4^2] \\
=&\mathbb{E}\left[\frac{1}{4}\sum_{1\leq i,j\leq 4, i\neq j}|X_i - X_j||Y_i - Y_j| \right. \\
&\quad -\frac{1}{4}\sum_{i=1}^{4}\left(\sum_{1\leq j\leq 4, j\neq i}|X_i - X_j|\sum_{1\leq j\leq 4, j\neq i}|Y_i - Y_j|\right) \\
&\quad \left. +\frac{1}{24}\sum_{1\leq i,j\leq 4, i\neq j}|X_i - X_j|\sum_{1\leq i,j\leq 4, i\neq j}|Y_i - Y_j|\right]^2 \\
\leq& C_1\mathbb{E}[|X_1 - X_2|^2|Y_1 - Y_2|^2] + C_2\mathbb{E}[|X_1 - X_2|^2|Y_1 - Y_2||Y_1 - Y_3|] \\
&\quad + C_3\mathbb{E}[|X_1 - X_2|^2|Y_1 - Y_2||Y_3 - Y_4|] \\
&\quad + C_4\mathbb{E}[|X_1 - X_2||X_1 - X_3||Y_1 - Y_2|^2] \\
&\quad + C_5\mathbb{E}[|X_1 - X_2||X_1 - X_3||Y_1 - Y_2||Y_1 - Y_3|] \\
&\quad + C_6\mathbb{E}[|X_1 - X_2||X_1 - X_3||Y_1 - Y_2||Y_3 - Y_4|] \\
&\quad + C_7\mathbb{E}[|X_1 - X_2||X_3 - X_4||Y_1 - Y_2|^2] \\
&\quad + C_8\mathbb{E}[|X_1 - X_2||X_3 - X_4||Y_1 - Y_2||Y_1 - Y_3|] \\
&\quad + C_9\mathbb{E}[|X_1 - X_2||X_3 - X_4||Y_1 - Y_2||Y_3 - Y_4|] \\
\leq& C_1'\mathbb{E}[|X_1 - X_2|^2|Y_1 - Y_2|^2] + C_2'\mathbb{E}[|X_1 - X_2|^2|Y_1 - Y_3|^2] \\
&\quad + C_3'\mathbb{E}[|X_1 - X_2|^2|Y_3 - Y_4|^2] \\
\leq& C_4'\mathbb{E}[|X|^2|Y|^2] \leq \infty,
\end{aligned}
$$

where $C_1, \ldots, C_9, C_1', \ldots, C_4' \geq 0$ are some constants. The second inequality is due to computing the squared term and set all coefficients to their absolution value, the third inequality is by Cauchy's inequality $ab \leq \frac{1}{2}a^2 + b^2$, and the fourth inequality is because of $|X_1 - X_2|^2 \leq 2|X_1|^2 + 2|X_2|^2$.

By the law of total variance, both h_1 and h_2 must have variances no more than the variance of h_4. We can have $\text{Var}(h_1) \leq \text{Var}(h_4) < \infty$ and $\text{Var}(h_2) \leq \text{Var}(h_4) < \infty$. □

C.2.4 Proof of Lemma 3.3.9

Proof. Under the general case, we derive the formulas of $h_1((X_1, Y_1))$ and $h_2((X_1, Y_1), (X_2, Y_2))$. Recall that

$$h_1((X_1, Y_1)) = \mathbb{E}_{2,3,4}[h_4((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4))],$$

$$h_2((X_1, Y_1), (X_2, Y_2)) = \mathbb{E}_{3,4}[h_4((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4))],$$

where

$$\begin{aligned}
&h_4((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4)) \\
=&\frac{1}{4} \sum_{1 \leq i,j \leq 4, i \neq j} |X_i - X_j||Y_i - Y_j| - \frac{1}{4} \sum_{i=1}^{4} \left(\sum_{j=1, j \neq i}^{4} |X_i - X_j| \sum_{j=1, j \neq i}^{4} |Y_i - Y_j| \right) \\
&+ \frac{1}{24} \sum_{1 \leq i,j \leq 4, i \neq j} |X_i - X_j| \sum_{1 \leq i,j \leq 4, i \neq j} |Y_i - Y_j|.
\end{aligned}$$

To facilitate the calculation, we introduce the notations $a_{ij} = |X_i - X_j|$ and $b_{ij} = |Y_i - Y_j|$, and then utilize them to expand quantity $h_4((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4))$ as

follows:

$$
\begin{aligned}
& h_4((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)) \\
= \quad & \frac{1}{6}a_{12}b_{12} - \frac{1}{12}a_{12}b_{13} - \frac{1}{12}a_{12}b_{14} - \frac{1}{12}a_{12}b_{23} - \frac{1}{12}a_{12}b_{24} + \frac{1}{6}a_{12}b_{34} \\
& - \frac{1}{12}a_{13}b_{12} + \frac{1}{6}a_{13}b_{13} - \frac{1}{12}a_{13}b_{14} - \frac{1}{12}a_{13}b_{23} + \frac{1}{6}a_{13}b_{24} - \frac{1}{12}a_{13}b_{34} \\
& - \frac{1}{12}a_{14}b_{12} - \frac{1}{12}a_{14}b_{13} + \frac{1}{6}a_{14}b_{14} + \frac{1}{6}a_{14}b_{23} - \frac{1}{12}a_{14}b_{24} - \frac{1}{12}a_{14}b_{34} \\
& - \frac{1}{12}a_{23}b_{12} - \frac{1}{12}a_{23}b_{13} + \frac{1}{6}a_{23}b_{14} + \frac{1}{6}a_{23}b_{23} - \frac{1}{12}a_{23}b_{24} - \frac{1}{12}a_{23}b_{34} \\
& - \frac{1}{12}a_{24}b_{12} + \frac{1}{6}a_{24}b_{13} - \frac{1}{12}a_{24}b_{14} - \frac{1}{12}a_{24}b_{23} + \frac{1}{6}a_{24}b_{24} - \frac{1}{12}a_{24}b_{34} \\
& + \frac{1}{6}a_{34}b_{12} - \frac{1}{12}a_{34}b_{13} - \frac{1}{12}a_{34}b_{14} - \frac{1}{12}a_{34}b_{23} - \frac{1}{12}a_{34}b_{24} + \frac{1}{6}a_{34}b_{34}.
\end{aligned}
$$

One may verify the correctness of the above by brute force. The following is a matrix that consists of the terms of $h_4((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4))$. In the same matrix, we highlighted the terms, which will become equal after taking the expectation with respect to random variables (X_2,Y_2), (X_3,Y_3) and (X_4,Y_4).

$$
\begin{pmatrix}
+\frac{1}{6}a_{12}b_{12} & -\frac{1}{12}a_{12}b_{13} & -\frac{1}{12}a_{12}b_{14} & -\frac{1}{12}a_{12}b_{23} & -\frac{1}{12}a_{12}b_{24} & +\frac{1}{6}a_{12}b_{34} \\
-\frac{1}{12}a_{13}b_{12} & +\frac{1}{6}a_{13}b_{13} & -\frac{1}{12}a_{13}b_{14} & -\frac{1}{12}a_{13}b_{23} & +\frac{1}{6}a_{13}b_{24} & -\frac{1}{12}a_{13}b_{34} \\
-\frac{1}{12}a_{14}b_{12} & -\frac{1}{12}a_{14}b_{13} & +\frac{1}{6}a_{14}b_{14} & +\frac{1}{6}a_{14}b_{23} & -\frac{1}{12}a_{14}b_{24} & -\frac{1}{12}a_{14}b_{34} \\
-\frac{1}{12}a_{23}b_{12} & -\frac{1}{12}a_{23}b_{13} & +\frac{1}{6}a_{23}b_{14} & +\frac{1}{6}a_{23}b_{23} & -\frac{1}{12}a_{23}b_{24} & -\frac{1}{12}a_{23}b_{34} \\
-\frac{1}{12}a_{24}b_{12} & +\frac{1}{6}a_{24}b_{13} & -\frac{1}{12}a_{24}b_{14} & -\frac{1}{12}a_{24}b_{23} & +\frac{1}{6}a_{24}b_{24} & -\frac{1}{12}a_{24}b_{34} \\
+\frac{1}{6}a_{34}b_{12} & -\frac{1}{12}a_{34}b_{13} & -\frac{1}{12}a_{34}b_{14} & -\frac{1}{12}a_{34}b_{23} & -\frac{1}{12}a_{34}b_{24} & +\frac{1}{6}a_{34}b_{34}
\end{pmatrix}
$$

Thus, $h_1((X_1, Y_1))$ could be expressed as follows.

$$
\begin{aligned}
&h_1((X_1,Y_1)) = \mathbb{E}_{2,3,4}[h_4((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4))] \\
=&\frac{1}{2}\mathbb{E}[|X_1 - X'||Y_1 - Y'|] - \frac{1}{2}\mathbb{E}[|X_1 - X'||Y_1 - Y''|] \\
&+ \frac{1}{2}\mathbb{E}[|X_1 - X'||Y - Y''|] - \frac{1}{2}\mathbb{E}[|X_1 - X'||Y' - Y''|] \\
&+ \frac{1}{2}\mathbb{E}[|X - X''||Y_1 - Y'|] - \frac{1}{2}\mathbb{E}[|X' - X''||Y_1 - Y'|] \\
&+ \frac{1}{2}\mathbb{E}[|X - X'||Y - Y'|] - \frac{1}{2}\mathbb{E}[|X - X'||Y - Y''|].
\end{aligned}
\tag{C.46}
$$

We may notice that the four above lines are equal to the expectations of sums of terms in the upper left, upper right, bottom left, and bottom right quadrants of the aforementioned matrix, respectively.

Similarly, we can highlight the entries, which will be the same after taking expectation with respect to (X_3, Y_3) and (X_4, Y_4). We do it in the following:

$$
\begin{pmatrix}
+\frac{1}{6}a_{12}b_{12} & -\frac{1}{12}a_{12}b_{13} & -\frac{1}{12}a_{12}b_{14} & -\frac{1}{12}a_{12}b_{23} & -\frac{1}{12}a_{12}b_{24} & +\frac{1}{6}a_{12}b_{34} \\
-\frac{1}{12}a_{13}b_{12} & +\frac{1}{6}a_{13}b_{13} & -\frac{1}{12}a_{13}b_{14} & -\frac{1}{12}a_{13}b_{23} & +\frac{1}{6}a_{13}b_{24} & -\frac{1}{12}a_{13}b_{34} \\
-\frac{1}{12}a_{14}b_{12} & -\frac{1}{12}a_{14}b_{13} & +\frac{1}{6}a_{14}b_{14} & +\frac{1}{6}a_{14}b_{23} & -\frac{1}{12}a_{14}b_{24} & -\frac{1}{12}a_{14}b_{34} \\
-\frac{1}{12}a_{23}b_{12} & -\frac{1}{12}a_{23}b_{13} & +\frac{1}{6}a_{23}b_{14} & +\frac{1}{6}a_{23}b_{23} & -\frac{1}{12}a_{23}b_{24} & -\frac{1}{12}a_{23}b_{34} \\
-\frac{1}{12}a_{24}b_{12} & +\frac{1}{6}a_{24}b_{13} & -\frac{1}{12}a_{24}b_{14} & -\frac{1}{12}a_{24}b_{23} & +\frac{1}{6}a_{24}b_{24} & -\frac{1}{12}a_{24}b_{34} \\
+\frac{1}{6}a_{34}b_{12} & -\frac{1}{12}a_{34}b_{13} & -\frac{1}{12}a_{34}b_{14} & -\frac{1}{12}a_{34}b_{23} & -\frac{1}{12}a_{34}b_{24} & +\frac{1}{6}a_{34}b_{34}
\end{pmatrix}
$$

Therefore, the expression of $h_2((X_1, Y_1), (X_2, Y_2))$ can be written as follows.

$$h_2((X_1, Y_1), (X_2, Y_2)) = \mathbb{E}_{3,4}[h_4((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4))] \tag{C.47}$$

$$\begin{aligned}
=&\frac{1}{6}|X_1 - X_2||Y_1 - Y_2| + \frac{1}{3}\mathbb{E}[|X_1 - X'||Y_1 - Y'|] + \frac{1}{3}\mathbb{E}[|X_2 - X'||Y_2 - Y'|] \\
&+ \frac{1}{6}\mathbb{E}[|X - X'||Y - Y'|] + \frac{1}{6}|X_1 - X_2|\mathbb{E}[|Y - Y'|] + \frac{1}{3}\mathbb{E}[|X_1 - X||Y_2 - Y'|] \\
&+ \frac{1}{3}\mathbb{E}[|X_2 - X||Y_1 - Y'|] + \frac{1}{6}|Y_1 - Y_2|\mathbb{E}[|X - X'|] - \frac{1}{6}|X_1 - X_2|\mathbb{E}[|Y_1 - Y'|] \\
&- \frac{1}{6}|X_1 - X_2|\mathbb{E}[|Y_2 - Y'|] - \frac{1}{6}|Y_1 - Y_2|\mathbb{E}[|X_1 - X|] - \frac{1}{6}\mathbb{E}[|X_1 - X||Y_1 - Y'|] \\
&- \frac{1}{6}\mathbb{E}[|X_1 - X||Y_2 - Y|] - \frac{1}{6}\mathbb{E}[|X_1 - X||Y - Y'|] - \frac{1}{6}|Y_1 - Y_2|\mathbb{E}[|X_2 - X|] \\
&- \frac{1}{6}\mathbb{E}[|X_2 - X||Y_1 - Y'|] - \frac{1}{6}\mathbb{E}[|X_2 - X||Y_2 - Y'|] - \frac{1}{6}\mathbb{E}[|X_2 - X||Y - Y'|] \\
&- \frac{1}{6}\mathbb{E}[|X - X'||Y_1 - Y|] - \frac{1}{6}\mathbb{E}[|X - X'||Y_2 - Y|].
\end{aligned}$$

□

C.2.5 Proof of Lemma 3.3.10

Proof. In the rest of this section, let us assume that X's are independent of Y's. The following notations will be utilized to simplify our calculations.

$$\begin{aligned}
a_{12} &= |X_1 - X_2|, & b_{12} &= |Y_1 - Y_2|, \\
a_1 &= \mathbb{E}[|X_1 - X|], & b_1 &= \mathbb{E}[|Y_1 - Y|], \\
a_2 &= \mathbb{E}[|X_2 - X|], & b_2 &= \mathbb{E}[|Y_2 - Y|], \\
a &= \mathbb{E}[|X - X'|], \text{ and} & b &= \mathbb{E}[|Y - Y'|],
\end{aligned}$$

where the expectation operator $\mathbb{E}$ is taken with respect to X, X', Y, Y', or any combination of them, whenever it is appropriate. Then, when X's are independent of Y's, one can easily

verify the following:

$$h_1((X_1, Y_1)) = \frac{1}{2}a_1b_1 + \frac{1}{2}ab + \frac{1}{2}a_1b + \frac{1}{2}ab_1 - \frac{1}{2}a_1b_1 - \frac{1}{2}a_1b - \frac{1}{2}ab_1 - \frac{1}{2}ab = 0,$$

as well as the following:

$$\begin{aligned}
&h_2((X_1, Y_1), (X_2, Y_2)) \\
=&\frac{1}{6}a_{12}b_{12} + \frac{1}{3}a_1b_1 + \frac{1}{3}a_2b_2 + \frac{1}{6}ab + \frac{1}{6}a_{12}b + \frac{1}{3}a_1b_2 + \frac{1}{3}a_2b_1 + \frac{1}{6}ab_{12} \\
&- \frac{1}{6}a_{12}b_1 - \frac{1}{6}a_{12}b_2 - \frac{1}{6}a_1b_{12} - \frac{1}{6}a_1b_1 - \frac{1}{6}a_1b_2 - \frac{1}{6}a_1b \\
&- \frac{1}{6}a_2b_{12} - \frac{1}{6}a_2b_1 - \frac{1}{6}a_2b_2 - \frac{1}{6}a_2b - \frac{1}{6}ab_1 - \frac{1}{6}ab_2 \\
=&\frac{1}{6}(a_{12}b_{12} + a_1b_1 + a_2b_2 + ab + a_{12}b + a_1b_2 + a_2b_1 + ab_{12} \\
&- a_{12}b_1 - a_{12}b_2 - a_1b_{12} - a_1b - a_2b_{12} - a_2b - ab_1 - ab_2) \\
=&\frac{1}{6}(a_{12} - a_1 - a_2 + a)(b_{12} - b_1 - b_2 + b).
\end{aligned}$$

Note that the above two are essentially (3.3.20) and (3.3.21) in Lemma 3.3.10. As we have had $\mathbb{E}[h_2] = \mathbb{E}[h_4] = 0$ when X and Y are independent, we have $\mathrm{Var}(h_2) = \mathbb{E}[h_2^2]$. Let us compute $\mathbb{E}[(a_{12} - a_1 - a_2 + a)^2]$ first. It is worth noting that

$$\begin{aligned}
&\mathbb{E}[a_{12}^2] = \mathbb{E}[|X - X'|^2], \\
&\mathbb{E}[a^2] = \mathbb{E}[a_1 a] = \mathbb{E}[a_2 a] = \mathbb{E}[a_{12} a] = \mathbb{E}^2[|X - X'|], \text{ and} \\
&\mathbb{E}[a_1^2] = \mathbb{E}[a_2^2] = \mathbb{E}[a_{12} a_1] = \mathbb{E}[a_{12} a_1] = \mathbb{E}[|X - X'||X - X''|].
\end{aligned}$$

As a result, we have

$$
\begin{aligned}
&\mathbb{E}[(a_{12}-a_1-a_2+a)^2]\\
=&\mathbb{E}[a_{12}^2+a_1^2+a_2^2+a^2-2a_{12}a_1-2a_{12}a_2+2a_{12}a+2a_1a_2-2a_1a-2a_2a]\\
=&\mathbb{E}[|X-X'|^2]+2\mathbb{E}[|X-X'||X-X''|]+\mathbb{E}^2[|X-X'|]\\
&-2\mathbb{E}[|X-X'||X-X''|]-2\mathbb{E}[|X-X'||X-X''|]\\
&+2\mathbb{E}^2[|X-X'|]+2\mathbb{E}^2[|X-X'|]-2\mathbb{E}^2[|X-X'|]-2\mathbb{E}^2[|X-X'|]\\
=&\mathbb{E}[|X-X'|^2]-2\mathbb{E}[|X-X'||X-X''|]+\mathbb{E}^2[|X-X'|]=\mathcal{V}^2(X,X).
\end{aligned}
$$

Similarly, we have $\mathbb{E}[(b_{12}-b_1-b_2+b)^2]=\mathcal{V}^2(Y,Y)$. In summary, we have

$$
\text{Var}(h_2)=\mathbb{E}[h_2^2]=\frac{1}{36}\mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y),
$$

which is (3.3.22) in Lemma 3.3.10. □

C.2.6 Proof of Lemma 3.3.13

Proof. By [69, Lemma 12], it is known that

$$
\tilde{k}(x,x')=|x-x_0|+|x'-x_0|-|x-x'|
$$

is a positive definite kernel. Due to [69, equation (4.4)], we have the following:

$$
\begin{aligned}
\tilde{k}_P(x,x')&=\tilde{k}(x,x')+\mathbb{E}_{W,W'}\tilde{k}(W,W')-\mathbb{E}_{W'}\tilde{k}(x,W')-\mathbb{E}_W\tilde{k}(W,x')\\
&=|x-x_0|+|x'-x_0|-|x-x'|+\mathbb{E}_x|x-x_0|+\mathbb{E}_{x'}|x'-x_0|\\
&\quad-\mathbb{E}_{x,x'}|x-x'|-|x-x_0|-\mathbb{E}_{x'}|x'-x_0|\\
&\quad+\mathbb{E}_{x'}|x-x'|-\mathbb{E}_x|x-x_0|-|x'-x_0|+\mathbb{E}_x|x-x'|\\
&=-|x-x'|-\mathbb{E}_{x,x'}|x-x'|+\mathbb{E}_{x'}|x-x'|+\mathbb{E}_x|x-x'|\\
&=h_X(x,x')
\end{aligned}
$$

is also a positive definite kernel. Similarly, $h_Y(Y_1, Y_2)$ is also a positive definite kernel. $\square$

C.2.7 Proof of Lemma 3.3.14

Proof. Since h_X is a positive definite kernel, by Mercer's Theorem, there exists a function sequence $\psi_1^X, \psi_1^X, \ldots$ and eigenvalues $\lambda_1^X \geq \lambda_2^X \geq \ldots \geq 0$ such that

$$h_X(x, x') = \sum_{l=1}^{\infty} \lambda_l^X \psi_l^X(x) \psi_l^X(x'),$$

where $\mathbb{E}[\psi_l^X(x)] = 0$, $\mathbb{E}[\psi_l^X(x)^2] = 1$ and $\mathbb{E}[\psi_l^X(x)\psi_{l'}^X(x)] = 0$ for $l \neq l'$. Similarly, we have

$$h_Y(y, y') = \sum_{l=1}^{\infty} \lambda_l^Y \psi_l^Y(y) \psi_l^Y(y').$$

By [69] equation (3.5), we that know

$$h_2((X_1, Y_1), (X_2, Y_2)) = \frac{1}{6} h_X(X_1, X_2) h_Y(Y_1, Y_2)$$

is a kernel with Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ isometrically isomorphic to the tensor product $\mathcal{H}_X \otimes \mathcal{H}_Y$. Thus,

$$6h_2((X_1, Y_1), (X_2, Y_2)) = \sum_{l,l'=1}^{\infty} \lambda_l^X \lambda_{l'}^Y [\psi_l^X(X_1)\psi_{l'}^Y(Y_1)][\psi_l^X(X_2)\psi_{l'}^Y(Y_2)],$$

which implies

$$\{\lambda_1, \lambda_2, \ldots\} = \{\lambda_1^X, \lambda_2^X, \ldots\} \otimes \{\lambda_1^Y, \lambda_2^Y, \ldots\}.$$

$\square$

C.2.8 Proof of Corollary 3.3.15

Proof. In this proof, we follow the notations in the proof of Lemma 3.3.14. It is worth noting that

$$\sum_{l=1}^{\infty} \lambda_l^X = \mathbb{E}[h_X(x, x)] = \mathbb{E}_x[-\mathbb{E}_{x,x'}|x - x'| + \mathbb{E}_{x'}|x - x'| + \mathbb{E}_{x'}|x - x'|] = \mathbb{E}[|X - X'|].$$

As an immediate result of Lemma 3.3.14, we have

$$\sum_{i=1}^{\infty}\lambda_i = \sum_{i=1}^{\infty}\lambda_i^X \sum_{i=1}^{\infty}\lambda_i^Y = \mathbb{E}[|X - X'|]\mathbb{E}[|Y - Y'|].$$

Similarly, we verify that

$$\sum_{l=1}^{\infty}(\lambda_l^X)^2 = \mathbb{E}[h_X(x, x')^2] = \mathcal{V}^2(X, X).$$

Then, we have

$$\sum_{i=1}^{\infty}\lambda_i^2 = \sum_{i=1}^{\infty}(\lambda_i^X)^2 \sum_{i=1}^{\infty}(\lambda_i^Y)^2 = \mathcal{V}^2(X, X)\mathcal{V}^2(Y, Y).$$

□

C.2.9 Proof of Lemma 3.3.16

Proof. By the law of total variance, we have

$$\text{Var}(\overline{\Omega}_n) = \mathbb{E}_{U,V}[\text{Var}_{X,Y}(\overline{\Omega}_n|U, V)] + \text{Var}_{U,V}[\mathbb{E}_{X,Y}(\overline{\Omega}_n|U, V)].$$

For the first term, when the random projections $U = (u_1, \ldots, u_K)$ and $V = (v_1, \ldots, v_K)$ are given, then by Lemma 3.3.7, we have

$$\text{Var}_{X,Y}(\overline{\Omega}_n|U, V) = \frac{16}{n}\text{Var}_{X,Y}(\bar{h}_1|U, V) + \frac{72}{n^2}\text{Var}_{X,Y}(\bar{h}_2|U, V) + O\left(\frac{1}{n^3}\right),$$

thus,

$$\mathbb{E}_{U,V}[\text{Var}_{X,Y}(\overline{\Omega}_n|U, V)] = \frac{16}{n}\mathbb{E}_{U,V}[\text{Var}_{X,Y}(\bar{h}_1|U, V)] + \frac{72}{n^2}\mathbb{E}_{U,V}[\text{Var}_{X,Y}(\bar{h}_2|U, V)] + O\left(\frac{1}{n^3}\right).$$

For the second term, we have

$$\mathbb{E}_{X,Y}(\overline{\Omega}_n|U, V) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{V}^2(u_k^t X, v_k^t Y)$$

thus, since $(u_k, v_k), k = 1, \ldots, K$ are independent,

$$\begin{aligned}\text{Var}_{U,V}[\mathbb{E}_{X,Y}(\overline{\Omega}_n|U,V)] &= \text{Var}_{U,V}\left(\frac{1}{K}\sum_{k=1}^{K}\mathcal{V}^2(u_k^t X, v_k^t Y)\right) \\ &= \frac{1}{K}\text{Var}_{u,v}(\mathcal{V}^2(u^t X, v^t Y)),\end{aligned}$$

where (u, v) stands for random projection vectors from Unif($\mathcal{S}^{p-1}$) and Unif($\mathcal{S}^{q-1}$), respectively. In summary, the variance of $\overline{\Omega}_n$ is

$$\begin{aligned}\text{Var}(\overline{\Omega}_n) &= \frac{1}{K}\text{Var}_{u,v}(\mathcal{V}^2(u^t X, v^t Y)) + \frac{16}{n}\mathbb{E}_{U,V}[\text{Var}_{X,Y}(\bar{h}_1|U,V)] \\ &\quad + \frac{72}{n^2}\mathbb{E}_{U,V}[\text{Var}_{X,Y}(\bar{h}_2|U,V)] + O\left(\frac{1}{n^3}\right).\end{aligned}$$

□

C.2.10 Proof of Theorem 3.3.18

Proof. For simplicity of notation, in this proof, without explicit statement, Var$(\cdot)$ and Cov$(\cdot)$ are with respect to (X, Y). By the definition of $\bar{h}_2$, we have

$$\text{Var}(\bar{h}_2|U,V) = \frac{1}{K^2}\sum_{k,k'=1}^{K}\text{Cov}(h_2^{(k)}, h_2^{(k')}|U,V).$$

To simplify the notation, we define the following:

$$\begin{aligned} a_{12}^u &= |u^t(X_1 - X_2)|, & b_{12}^v &= |v^t(Y_1 - Y_2)|, \\ a_1^u &= \mathbb{E}[|u^t(X_1 - X)|], & b_1^v &= \mathbb{E}[|v^t(Y_1 - Y)|], \\ a_2^u &= \mathbb{E}[|u^t(X_2 - X)|], & b_2^v &= \mathbb{E}[|v^t(Y_2 - Y)|], \\ a^u &= \mathbb{E}[|u^t(X - X')|], \text{ and} & b^v &= \mathbb{E}[|v^t(Y - Y')|]. \end{aligned}$$

Thus, by (3.3.21), we have

$$
\begin{aligned}
&\mathrm{Cov}(h_2^{(k)}, h_2^{(k')}|U,V)\\
=&\frac{C_p^2C_q^2}{36}\mathbb{E}_{X,Y}[(a_{12}^{u_k}-a_1^{u_k}-a_2^{u_k}+a^{u_k})(b_{12}^{v_k}-b_1^{v_k}-b_2^{v_k}+b^{v_k})\\
&\quad(a_{12}^{u_{k'}}-a_1^{u_{k'}}-a_2^{u_{k'}}+a^{u_{k'}})(b_{12}^{v_{k'}}-b_1^{v_{k'}}-b_2^{v_{k'}}+b^{v_{k'}})]\\
=&\frac{C_p^2C_q^2}{36}\mathbb{E}_{X,Y}[(a_{12}^{u_k}-a_1^{u_k}-a_2^{u_k}+a^{u_k})(a_{12}^{u_{k'}}-a_1^{u_{k'}}-a_2^{u_{k'}}+a^{u_{k'}})]\\
&\quad\mathbb{E}_{X,Y}[(b_{12}^{v_k}-b_1^{v_k}-b_2^{v_k}+b^{v_k})(b_{12}^{v_{k'}}-b_1^{v_{k'}}-b_2^{v_{k'}}+b^{v_{k'}})]\\
=&\frac{C_p^2C_q^2}{36}\mathcal{V}^2(u_k^tX,u_{k'}^tX)\mathcal{V}^2(v_k^tY,v_{k'}^tY),
\end{aligned}
$$

where the second equation holds by the assumption that X and Y are independent and the last equation holds by the definition of distance covariance in (3.1.13).

To summarize, the variance of $\overline{\Omega}_n$ with respect to (X,Y) is

$$
\mathrm{Var}(\overline{\Omega}_n|U,V)=\frac{2C_p^2C_q^2}{n^2}\frac{1}{K^2}\sum_{k,k'=1}^{K}\mathcal{V}^2(u_k^tX,u_{k'}^tX)\mathcal{V}^2(v_k^tY,v_{k'}^tY)+O(\frac{1}{n^3}),
$$

which implies

$$
\sum_{i=1}^{\infty}\bar{\lambda}_i^2=36\mathrm{Var}(\bar{h}_2|U,V)=\frac{C_p^2C_q^2}{K^2}\sum_{k,k'=1}^{K}\mathcal{V}^2(u_k^tX,u_{k'}^tX)\mathcal{V}^2(v_k^tY,v_{k'}^tY).
$$

By Corollary 3.3.15, we know that

$$
\sum_{i=1}^{\infty}\bar{\lambda}_i=\mathbb{E}[6\bar{h}_4(x,x)]=\frac{C_pC_q}{K}\sum_{k=1}^{K}\mathbb{E}[|u_k^t(X-X')|]\mathbb{E}[|v_k^t(Y-Y')|].
$$

□

C.2.11 Proof of Proposition 3.3.20

Proof. Let us recall the definition,

$$
\sum_{i=1}^{\infty}\bar{\lambda}_i=\mathbb{E}[6\bar{h}_4(x,x)]=\frac{C_pC_q}{K}\sum_{k=1}^{K}\mathbb{E}[|u_k^t(X-X')|]\mathbb{E}[|v_k^t(Y-Y')|],
$$

$$\sum_{i=1}^{\infty} \bar{\lambda}_i^2 = \frac{C_p^2 C_q^2}{K^2} \sum_{k,k'=1}^{K} \mathcal{V}^2(u_k^t X, u_{k'}^t X)\mathcal{V}^2(v_k^t Y, v_{k'}^t Y).$$

To estimate $\sum_{i=1}^{\infty} \bar{\lambda}_i^2$, we can use

$$\frac{C_p^2 C_q^2}{K^2} \sum_{k,k'=1}^{K} \Omega_n(u_k^t X, u_{k'}^t X)\Omega_n(v_k^t Y, v_{k'}^t Y),$$

which takes $O(K^2 n \log n)$ time and is costly when K is large. It is worth noting that if $k \neq k'$ and (u_k, v_k) is independent of $(u_{k'}, v_{k'})$, by Lemma 3.3.2, we know that

$$C_p^2 C_q^2 \mathbb{E}_{U,V}[\mathcal{V}^2(u_k^t X, u_{k'}^t X)\mathcal{V}^2(v_k^t Y, v_{k'}^t Y)] = \mathcal{V}^2(X, X)\mathcal{V}^2(Y, Y).$$

Thus, $\sum_{i=1}^{\infty} \bar{\lambda}_i^2$ could be estimated by

$$\frac{K-1}{K}\Omega_n(X, X)\Omega_n(Y, Y) + \frac{C_p^2 C_q^2}{K} \sum_{k=1}^{K} \Omega_n(u_k^t X, u_k^t X)\Omega_n(v_k^t Y, v_k^t Y),$$

which takes only $O(Kn \log n)$ time.

And, $\sum_{i=1}^{\infty} \bar{\lambda}_i$ could be estimated by:

$$\frac{C_p C_q}{Kn^2(n-1)^2} \sum_{k=1}^{K} a_{\cdot\cdot}^{u_k} b_{\cdot\cdot}^{v_k},$$

where

$$a_{\cdot\cdot}^{u_k} = \sum_{i,j=1}^{n} |u_k^t(X_i - X_j)| \text{ and } b_{\cdot\cdot}^{v_k} = \sum_{i,j=1}^{n} |v_k^t(Y_i - Y_j)|.$$

So, in summary, we have

$$\sum_{i=1}^{\infty} \bar{\lambda}_i \approx \frac{C_p C_q}{Kn^2(n-1)^2} \sum_{k=1}^{K} a_{\cdot\cdot}^{u_k} b_{\cdot\cdot}^{v_k},$$
$$\sum_{i=1}^{\infty} \bar{\lambda}_i^2 \approx \frac{K-1}{K}\Omega_n(X, X)\Omega_n(Y, Y) + \frac{C_p^2 C_q^2}{K} \sum_{k=1}^{K} \Omega_n(u_k^t X, u_k^t X)\Omega_n(v_k^t Y, v_k^t Y).$$

□

APPENDIX D

PROOFS OF ENERGY STATISTICS

We present all the proofs of Chapter 4 here. For reader's convenience, we restate the following notations. We denote $c_p = \frac{\pi^{(p+1)/2}}{\Gamma((p+1)/2)}$ and $C_p = \frac{c_1 c_{p-1}}{c_p} = \frac{\sqrt{\pi}\Gamma((p+1)/2)}{\Gamma(p/2)}$ as two constants, where $\Gamma(\cdot)$ denotes the Gamma function.

D.1 Proof of Theorem 4.2.1

Proof. The detailed explanations and corresponding complexity analysis of the fast algorithm in Section 4.2.1 is as follows.

(1) Sort X_i's and Y_j's, so that we have order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ and $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(m)}$. By adopting the merge sort [38, 36], the average computational complexity in this step is $O(\max(n,m)\log\max(n,m))$. In addition, it is easy to verify the following:

$$\begin{aligned}\mathcal{E} :=& \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}|X_{(i)}-Y_{(j)}| - \frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n}|X_{(i)}-X_{(j)}| \\ &- \frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}|Y_{(i)}-Y_{(j)}| \\ =& \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}|X_{(i)}-Y_{(j)}| - \frac{2}{n(n-1)}\sum_{i<j}^{n}|X_{(i)}-X_{(j)}| \\ &- \frac{2}{m(m-1)}\sum_{i<j}^{m}|Y_{(i)}-Y_{(j)}|\end{aligned}$$

That is, we can compute $\mathcal{E}$ through merely the order statistics. The rest of algorithmic description will be based on the above formula.

(2) We can verify the following:

$$\frac{2}{n(n-1)}\sum_{i<j}^{n}|X_{(i)}-X_{(j)}|=\frac{2}{n(n-1)}\sum_{i=1}^{n-1}i(n-i)\left|X_{(i+1)}-X_{(i)}\right|.$$

Given order statistics $X_{(i)}$'s, the computational complexity of implementing the above is $O(n)$.

(3) Essentially identical to the previous item, one can verify the following:

$$\frac{2}{m(m-1)}\sum_{i<j}^{m}|Y_{(i)}-Y_{(j)}|=\frac{2}{m(m-1)}\sum_{i=1}^{m-1}i(m-i)\left|Y_{(i+1)}-Y_{(i)}\right|.$$

Given order statistics $Y_{(i)}$'s, the computational complexity of implementing the above is $O(m)$.

(4) For the first term in $\mathcal{E}$, one can computer it in two sub-steps as below.

(a) One can merge two ordered series $X_{(1)}\leq X_{(2)}\leq\cdots\leq X_{(n)}$ and $Y_{(1)}\leq Y_{(2)}\leq\cdots\leq Y_{(m)}$ into a single ordered series $Z_{(1)}\leq Z_{(2)}\leq\cdots\leq Z_{(n+m)}$, where each $Z_{(k)}$ is either from $X_{(i)}$'s or from $Y_{(j)}$'s. At the same time, one can generate a sequence $I_i, i=1,2,\ldots,n+m$, where I_i records the size of the subset of $Z_{(1)}$ through $Z_{(i)}$ that are from $X_{(i)}$'s. It is evident to show that quantity $i-I_i$ is the size of the subset of $Z_{(1)}$ through $Z_{(i)}$ that are from $Y_{(j)}$'s.

Note the computational complexity in this step is $O(n+m)$.

(b) Given the above preparation, we can verify the following:

$$\frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left|X_{(i)}-Y_{(j)}\right|$$
$$=\frac{2}{nm}\sum_{i=1}^{n+m-1}\left[I_i(m-i+I_i)+(i-I_i)(n-I_i)\right]\left|Z_{(i+1)}-Z_{(i)}\right|.$$

Note that the term $I_i(m-i+I_i)+(i-I_i)(n-I_i)$ on the right hand side is equal to the number of times the length $|Z_{(i+1)} - Z_{(i)}|$ has been counted in the double summation on the left hand side. Through this, we can establish the equality. The computational complexity of implementing the above is $O(n+m)$.

From all the above, we show that the complexity of computing $\mathcal{E}$ is dominated by the sorting step, thus the average total complexity is $O(\max(n,m)\log\max(n,m))$. □

D.2 Proof of Lemma 4.3.1

Proof. The proof is straightforward. First, by Proposition 4.1.2, we know that

random vector $X \in \mathbb{R}^p$ has the same distribution with random vector $Y \in \mathbb{R}^p$

if and only if

$$\Phi_X = \Phi_Y, \text{ almost everywhere,}$$

where Φ_X and Φ_Y are the characteristic functions of X and Y, respectively. That becomes

$$\mathbb{E}\left[e^{iX^Tt}\right] = \mathbb{E}\left[e^{iY^Tt}\right], \forall t \in \mathbb{R}^p.$$

By variable change $t = ut'$, where $u \in \mathcal{S}^{p-1}$ and $t' \in [0,\infty)$, we have

$$\mathbb{E}\left[e^{iu^TXt'}\right] = \mathbb{E}\left[e^{iu^TYt'}\right], \forall u \in \mathcal{S}^{p-1} \text{ and } t' \in [0,\infty),$$

or equivalently,

$$\Phi_{u^TX} = \Phi_{u^TY}, \forall u \in \mathcal{S}^{p-1}.$$

By Proposition 4.1.2, we know that

$$\Phi_{u^TX} = \Phi_{u^TY}, \forall u \in \mathcal{S}^{p-1},$$

is equivalent with

$$\mathcal{E}(u^T X, u^T Y) = 0, \forall u \in \mathcal{S}^{p-1}.$$

□

D.3 Proof of Lemma 4.3.2

First, let us state a result from [33], which shows relationship between the norm of random projections and the norm of original vector.

Lemma D.3.1. *[33, Lemma B.1] Suppose v is a fixed unit vector in $\mathbb{R}^p$ and $u \in \mathcal{S}^{p-1}$. Let μ be the uniform probability measure on $\mathcal{S}^{p-1}$. We have*

$$C_p \int_{\mathcal{S}^{p-1}} |u^T v| d\mu(u) = C_p \mathbb{E}_u[|u^T v|] = 1,$$

where constant C_p has been mentioned at the beginning of this chapter.

Equipped with above lemma, we can prove Lemma 4.3.2 as follows.

Proof. By Lemma D.3.1, we have

$$C_p \mathbb{E}_u \left[\left| u^T \frac{(X-Y)}{|X-Y|} \right| \right] = 1, \text{ thus, } |X-Y| = C_p \mathbb{E}_u \left[\left| u^T (X-Y) \right| \right].$$

Therefore, the energy distance could be written as

$$\begin{aligned}
\mathcal{E}(X,Y) &= 2\mathbb{E}[|X-Y|] - \mathbb{E}[|X-X'|] - \mathbb{E}[|Y-Y'|] \\
&= 2\mathbb{E}_{X,Y}[C_p \mathbb{E}_u[|u^T(X-Y)|]] - \mathbb{E}_{X,X'}[C_p \mathbb{E}_u[|u^T(X-X')|]] \\
&\qquad\qquad - \mathbb{E}_{Y,Y'}[C_p \mathbb{E}_u[|u^T(Y-Y')|]] \\
&= C_p \mathbb{E}_u \left[2\mathbb{E}_{X,Y}[|u^T(X-Y)|] - \mathbb{E}_{X,X'}[|u^T(X-X')|] - \mathbb{E}_{Y,Y'}[|u^T(Y-Y')|] \right] \\
&= C_p \mathbb{E}_u[\mathcal{E}(u^T X, u^T Y)] = C_p \int_{\mathcal{S}^{p-1}} \mathcal{E}(u^T X, u^T Y) d\mu(u),
\end{aligned}$$

where u is a uniformly distributed random variable on $\mathcal{S}^{p-1}$, the second equality is by Lemma D.3.1, the third equality is by exchanging the order of expectation, and the fourth equality is by the definition of energy distance.

We can reach a similar result for energy statistics simply by replacing $\mathbb{E}_{X,Y}[\cdot]$, $\mathbb{E}_{X,X'}[\cdot]$ and $\mathbb{E}_{Y,Y'}[\cdot]$ with summation. The rest reasoning is almost the same with above reasoning for energy distance. □

D.4 Proof of Theorem 4.3.7

First, let us introduce a lemma that will be used in later proof.

Lemma D.4.1. *[33, Lemma 4.13] If $\mathbb{E}[|X|^2] < \infty$, we have that kernel*

$$\mathbf{k}(X_1, X_2) = \mathbb{E}_X[|X_1 - X|] + \mathbb{E}_X[|X_2 - X|] - |X_1 - X_2| - \mathbb{E}_{X,X'}[|X - X'|]$$

is a positive definite kernel. As a result, if X and Y have the same distirbution, $h_{20}(\cdot,\cdot)$, $h_{02}(\cdot,\cdot)$ and $-h_{11}(\cdot,\cdot)$ in Lemma 4.3.6 are all positive definite kernels. Also, there exist functions $\phi_1(\cdot), \phi_2(\cdot), \ldots$ such that

$$\mathbf{k}(X_1, X_2) = \sum_{i=1}^{\infty} \lambda_i \phi_i(X_1)\phi_i(X_2),$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$, $\mathbb{E}[\phi_i(X)] = 0$, $\mathbb{E}[\phi_i(X)^2] = 1$ and $\mathbb{E}[\phi_i(X)\phi_j(X)] = 0$, $i = 1, 2, \ldots, \infty$, $i \neq j$.

Now, let us prove Theorem 4.3.7.

Proof. By Lemma 4.3.6 and [42, Section 2.2, Theorem 3], we have

$$\mathcal{E}_{n,m} = \binom{n}{2}^{-1} \sum_{i_1<i_2} h_{20}(X_{i_1}, X_{i_2}) + 4(nm)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} h_{11}(X_i, Y_j) + \binom{m}{2}^{-1} \sum_{j_1<j_2} h_{02}(Y_{j_1}, Y_{j_2}) + \mathcal{R}_{n,m},$$

where $\mathcal{R}_{n,m}$ is the residual with $N\mathcal{R}_{n,m} \xrightarrow{P} 0$. By Lemma D.4.1, we know that

$$h_{20}(X_{i_1}, X_{i_2}) = \sum_{l=1}^{\infty} \lambda_l \phi_l(X_{i_1})\phi_l(X_{i_2}), \qquad h_{02}(Y_{j_1}, Y_{j_2}) = \sum_{l=1}^{\infty} \lambda_l \phi_l(Y_{j_1})\phi_l(Y_{j_2}),$$

and

$$h_{11}(X_i, Y_j) = -\frac{1}{2}\sum_{l=1}^{\infty} \lambda_l \phi_l(X_i)\phi_l(Y_j).$$

Therefore, we have

$$\begin{aligned}
\mathcal{E}_{n,m} = &\sum_{l=1}^{\infty} \lambda_l \left[\left(\frac{1}{n}\sum_{i=1}^{n} \phi_l(X_i) - \frac{1}{m}\sum_{j=1}^{m} \phi_l(Y_j) \right)^2 - \frac{1}{n^2}\sum_{i=1}^{n} \phi_l(X_i)^2 - \frac{1}{m^2}\sum_{j=1}^{m} \phi_l(Y_j)^2 \right] \\
&+ \mathcal{R}_{n,m} + \left(\frac{2}{n(n-1)} - \frac{2}{n^2} \right) \sum_{i_1<i_2} h_{20}(X_{i_1}, X_{i_2}) \\
&+ \left(\frac{2}{m(m-1)} - \frac{2}{m^2} \right) \sum_{j_1<j_2} h_{02}(Y_{j_1}, Y_{j_2}) \\
= &\frac{1}{N}\sum_{l=1}^{\infty} \lambda_l \left[\left(\sqrt{N/n}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \phi_l(X_i) - \sqrt{N/m}\frac{1}{\sqrt{m}}\sum_{j=1}^{m} \phi_l(Y_j) \right)^2 \right. \\
&\left. - \frac{N}{n^2}\sum_{i=1}^{n} \phi_l(X_i)^2 - \frac{N}{m^2}\sum_{j=1}^{m} \phi_l(Y_j)^2 \right] + \tilde{\mathcal{R}}_{n,m},
\end{aligned}$$

where

$$\tilde{\mathcal{R}}_{n,m} = \mathcal{R}_{n,m} + \frac{2}{n^2(n-1)} \sum_{i_1<i_2} h_{20}(X_{i_1}, X_{i_2}) + \frac{2}{m^2(m-1)} \sum_{j_1<j_2} h_{02}(Y_{j_1}, Y_{j_2}).$$

It is worth noting that $N\tilde{\mathcal{R}}_{n,m} \xrightarrow{P} 0$. Therefore, as $N \to \infty$, we have

$$N\mathcal{E}_{n,m} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l [(\sqrt{1/\eta} Z_{l,1} - \sqrt{1/(1-\eta)} Z_{l,2})^2 - \frac{1}{\eta} - \frac{1}{1-\eta}] = \sum_{l=1}^{\infty} \frac{\lambda_l}{\eta(1-\eta)} (Z_l^2 - 1),$$

where $Z_{l,1}, Z_{l,2}, l = 1, 2, \ldots$ are all independent standard normal random variables and $Z_l = \sqrt{1-\eta} Z_{l,1} + \sqrt{\eta} Z_{l,2}$. It is worth noting that

$$\sum_{l=1}^{\infty} \lambda_l = \mathbb{E}[h_{20}(X, X)] = \mathbb{E}[|X - X'|].$$

Similarly, we know that

$$\begin{aligned}\sum_{l=1}^{\infty}\lambda_l^2 &= \mathbb{E}_{X_1,X_2}\left[\sum_{l=1}^{\infty}\lambda_l\phi_l(X_1)\phi_l(X_2)\right]^2\\ &= \mathbb{E}_{X_1,X_2}\left[h_{20}(X_1,X_2)^2\right]\\ &= \mathbb{E}_{X_1,X_2}\left[\left(\mathbb{E}_X[|X_1-X|]+\mathbb{E}_X[|X_2-X|]-|X_1-X_2|-\mathbb{E}_{X,X'}[|X-X'|]\right)^2\right]\\ &= DC(X,X),\end{aligned}$$

where the last equality is by the definition of distance covariance. □

D.5 Proof of Lemma 4.3.10

Proof. It is worth noting that when X and Y have the same distribution, u^TX and u^TY also should have the same distribution for any u, thus

$$\mathcal{E}(u^TX,u^TY)=0,\forall u\in\mathbb{R}^p,$$

which indicates that

$$\text{Var}_u[\mathcal{E}(u^TX,u^TY)]=0.$$

Moreover, we have

$$\overline{h}_{10}(X_1)=\frac{1}{K}\sum_{k=1}^{K}C_p h_{10}(u_k^TX_1).$$

By the definition of $h_{10}(\cdot)$, we know $h_{10}(u_k^TX_1)=0$ when X and Y are identically distributed, which sugguests

$$\overline{h}_{10}=0,\ \text{and Var}[\overline{h}_{10}|U]=0.$$

Similarly, we have

$$\overline{h}_{01}=0,\ \text{and Var}[\overline{h}_{01}|U]=0.$$

Combining above results and Lemma 4.3.8, we have the formula of the variance of $\overline{\mathcal{E}}_{n,m}$ in this lemma. □

D.6 Proof of Theorem 4.3.12

Proof. This proof is almost identical with the proof of Theorem 4.3.7. We can simply replace the notations like $h_{20}, h_{02}, h_{11}, \lambda_i, \phi_i(\cdot)$ with corresponding notations like $\overline{h}_{20}, \overline{h}_{02}, \overline{h}_{11}, \overline{\lambda}_i, \overline{\phi}_i(\cdot)$. The rest reasoning is the same.

For $\sum_{l=1}^{\infty} \overline{\lambda}_l$, it is easy to see that

$$\sum_{l=1}^{\infty} \overline{\lambda}_l = \mathbb{E}[\overline{\mathbf{k}}(X, X)] = \frac{C_p}{K} \sum_{k=1}^{K} \mathbb{E}_{X,X'}[|u_k^T(X - X')|].$$

For $\sum_{l=1}^{\infty} \overline{\lambda}_l^2$, we have

$$\begin{aligned}
\sum_{l=1}^{\infty} \overline{\lambda}_l^2 &= \mathbb{E}_{X_1,X_2}\left[\sum_{l=1}^{\infty} \overline{\lambda}_l \overline{\phi}_l(X_1)\overline{\phi}_l(X_2)\right]^2 = \mathbb{E}_{X_1,X_2}\left[C_p^2 \overline{\mathbf{k}}(u_k^T X_1, X_2)^2\right] \\
&= \mathbb{E}_{X_1,X_2}\left[\left(\frac{C_p^2}{K}\sum_{k=1}^{K} \mathbf{k}(u_k^T X_1, u_k^T X_2)\right)^2\right] \\
&= \frac{C_p^2}{K^2}\sum_{k,k'=1}^{K} \mathbb{E}_{X_1,X_2}\left[\mathbf{k}(u_k^T X_1, u_k^T X_2)\mathbf{k}(u_{k'}^T X_1, u_{k'}^T X_2)\right] \\
&= \frac{C_p^2}{K^2}\sum_{k,k'=1}^{K} DC(u_k^T X, u_{k'}^T X),
\end{aligned}$$

where the last equation is by the definition of distance covariance. □

D.7 Proof of Proposition 4.3.13

Proof. When X and Y are identically distributed, we know

$$\mathbb{E}[|u_k^T(Z_i - Z_j)|] = \mathbb{E}_{X,X'}[|u_k^T(X - X')|],$$

which implies

$$\frac{C_p}{K}\sum_{k=1}^{K} \frac{1}{(n+m)(n+m-1)} \sum_{i \neq j}^{n+m} |u_k^T(Z_i - Z_j)|$$

is an unbiased estimator for $\frac{C_p}{K}\sum_{k=1}^{K}\mathbb{E}_{X,X'}[|u_k^T(X-X')|] = \sum_{l=1}^{\infty}\overline{\lambda}_l$.

We have

$$\begin{aligned}\sum_{l=1}^{\infty}\overline{\lambda}_l^2 &= \frac{C_p^2}{K^2}\sum_{k,k'=1}^{K}\text{DC}(u_k^TX, u_{k'}^TX)\\ &= \frac{C_p^2}{K^2}\sum_{k=1}^{K}\text{DC}(u_k^TX, u_k^TX) + \frac{C_p^2}{K^2}\sum_{k\neq k'}^{K}\text{DC}(u_k^TX, u_{k'}^TX).\end{aligned}$$

For $\text{DC}(u_k^TX, u_k^TX)$ in the first term , it is natural to estimate it with $\text{SDC}(u_k^TZ, u_k^TZ)$. It is worth noting that u_k is independent of $u_{k'}$ for all $k' \neq k$. When the number of random projections K is sufficient large, by the Law of Large Number, we have

$$\frac{C_p^2}{K^2}\sum_{k\neq k'}^{K}\text{DC}(u_k^TX, u_{k'}^TX) \xrightarrow{P} \frac{(K-1)C_p^2}{K^2}\sum_{k=1}^{K}\text{DC}(u_k^TX, v_k^TX).$$

We can estimate the quantity on the right-hand-side by simply estimating distance covariance with the sample version. Thus, we have

$$\frac{C_p^2}{K^2}\sum_{k=1}^{K}\text{SDC}(u_k^TZ, u_k^TZ) + \frac{(K-1)C_p^2}{K^2}\sum_{k=1}^{K}\text{SDC}(u_k^TZ, v_k^TZ) \to \sum_{l=1}^{\infty}\overline{\lambda}_l^2 \text{ as } N, K \to \infty.$$

□

REFERENCES

[1] ACHLIOPTAS, D., MCSHERRY, F., and SCHÖLKOPF, B., "Sampling techniques for kernel methods," in *In Annual Advances In Neural Information Processing Systems 14: Proceedings Of The 2001 Conference*, 2001.

[2] ARJEVANI, Y. and SHAMIR, O., "Communication complexity of distributed convex learning and optimization," tech. rep., ArXiv, June 2015. http://arxiv.org/abs/1506.01900.

[3] BALCAN, M.-F., BLUM, A., FINE, S., and MANSOUR, Y., "Distributed learning, communication complexity and privacy," arxiv, Georgia Institute of Technology, April 2012.

[4] BALCAN, M.-F., KANCHANAPALLY, V., LIANG, Y., and WOODRUFF, D., "Improved distributed principal component analysis," tech. rep., ArXiv, August 2014. http://arxiv.org/abs/1408.5823.

[5] BARINGHAUS, L. and FRANZ, C., "On a new multivariate two-sample test," *Journal of multivariate analysis*, vol. 88, no. 1, pp. 190–206, 2004.

[6] BATTEY, H., FAN, J., LIU, H., LU, J., and ZHU, Z., "Distributed estimation and inference with statistical guarantees," *arXiv preprint arXiv:1509.05457*, 2015.

[7] BHATTACHARYA, B. B., "Power of graph-based two-sample tests," *arXiv preprint arXiv:1508.07530*, 2015.

[8] BICKEL, P. J., "One-step Huber estimates in the linear model," *Journal of the American Statistical Association*, vol. 70, no. 350, pp. 428–434, 1975.

[9] BISWAS, M., MUKHOPADHYAY, M., GHOSH, A. K., and OTHERS, "A distribution-free two-sample run test applicable to high-dimensional data," *Biometrika*, vol. 101, no. 4, pp. 913–926, 2014.

[10] BLUM, A., "Random projection, margins, kernels, and feature-selection," in *Subspace, Latent Structure and Feature Selection*, pp. 52–68, Springer, 2006.

[11] BODENHAM, D. A. and ADAMS, N. M., "A comparison of efficient approximations for a weighted sum of chi-squared random variables," *Statistics and Computing*, pp. 1–12, 2014.

[12] BOX, G. E. and OTHERS, "Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in the one-way classification," *The Annals of Mathematical Statistics*, vol. 25, no. 2, pp. 290–302, 1954.

[13] BOYD, S., PARIKH, N., CHU, E., PELEATO, B., and ECKSTEIN, J., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[14] BRADLEY, J. K., KYROLA, A., BICKSON, D., and GUESTRIN, C., "Parallel coordinate descent for l1-regularized loss minimization," *arXiv preprint arXiv:1105.5379*, 2011.

[15] CAI, T. T., FAN, J., and JIANG, T., "Distributions of angles in random packing on spheres.," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1837–1864, 2013.

[16] CHEN, H. and FRIEDMAN, J. H., "A new graph-based two-sample test for multivariate and object data," *Journal of the American statistical association*, no. just-accepted, pp. 1–41, 2016.

[17] CHEN, H. and ZHANG, N. R., "Graph-based tests for two-sample comparisons of categorical data," *Statistica Sinica*, pp. 1479–1503, 2013.

[18] CHEN, S., DONOHO, D. L., and SAUNDERS, M. A., "Atomic decomposition by basis pursuit," *SIAM J. Sci. Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[19] CHEN, X. and XIE, M.-G., "A split-and-conquer approach for analysis of extraordinarily large data," *Statistica Sinica*, vol. 24, pp. 1655–1684, 2014.

[20] CORBETT, J. C., DEAN, J., EPSTEIN, M., FIKES, A., FROST, C., FURMAN, J., GHEMAWAT, S., GUBAREV, A., HEISER, C., HOCHSCHILD, P., and ET AL., "Spanner: Googles globally distributed database," in *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation*, 2012.

[21] DAVIS, D. and YIN, W., "Convergence rates of relaxed peaceman-rachford and admm under regularity assumptions," *arXiv preprint arXiv:1407.5210*, 2014.

[22] DEKEL, O., GILAD-BACHRACH, R., SHAMIR, O., and XIAO, L., "Optimal distributed online prediction using mini-batches," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 165–202, 2012.

[23] DERIGS, U., "Solving non-bipartite matching problems via shortest path techniques," *Annals of Operations Research*, vol. 13, no. 1, pp. 225–261, 1988.

[24] DRINEAS, P. and MAHONEY, M. W., "On the Nyström method for approximating a Gram matrix for improved kernel-based learning," *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 2153–2175, 2005.

[25] FAN, J. and CHEN, J., "One-step local quasi-likelihood estimation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 4, pp. 927–943, 1999.

[26] FRIEZE, A., KANNAN, R., and VEMPALA, S., "Fast Monte-Carlo algorithms for finding low-rank approximations," *Journal of the ACM (JACM)*, vol. 51, no. 6, pp. 1025–1041, 2004.

[27] GAMAL, M. E. and LAI, L., "Are Slepian-Wolf rates necessary for distributed parameter estimation?," tech. rep., arXiv, August 2015. http://arxiv.org/abs/1508.02765.

[28] GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B., and SMOLA, A., "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.

[29] GRETTON, A., BOUSQUET, O., SMOLA, A., and SCHÖLKOPF, B., "Measuring statistical dependence with hilbert-schmidt norms," in *International Conference on Algorithmic Learning Theory*, pp. 63–77, Springer, 2005.

[30] HELLER, R., HELLER, Y., KAUFMAN, S., BRILL, B., and GORFINE, M., "Consistent distribution-free k-sample and independence tests for univariate random variables," *Journal of Machine Learning Research*, vol. 17, no. 29, pp. 1–54, 2016.

[31] HOEFFDING, W., "Probability inequalities for sums of bounded random variables," *Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.

[32] HUANG, C. and HUO, X., "A distributed one-step estimator," *arXiv preprint arXiv:1511.01443*, 2015.

[33] HUANG, C. and HUO, X., "A statistically and numerically efficient independence test based on random projections and distance covariance," *arXiv preprint arXiv:1701.06054*, 2017.

[34] HUO, X. and SZÉKELY, G. J., "Fast computing for distance covariance," *Technometrics*, vol. 58, no. 4, pp. 435–447, 2016.

[35] JAGGI, M., SMITH, V., TAKÁC, M., TERHORST, J., KRISHNAN, S., HOFMANN, T., and JORDAN, M. I., "Communication-efficient distributed dual coordinate ascent," in *Advances in Neural Information Processing Systems*, pp. 3068–3076, 2014.

[36] KATAJAINEN, J. and TRÄFF, J. L., "A meticulous analysis of mergesort programs," in *Italian Conference on Algorithms and Complexity*, pp. 217–228, Springer, 1997.

[37] KLEINER, A., TALWALKAR, A., SARKAR, P., and JORDAN, M. I., "A scalable bootstrap for massive data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 4, pp. 795–816, 2014.

[38] KNUTH, D., "Section 5.2. 4: Sorting by merging," *The Art of Computer Programming*, vol. 3, pp. 158–168, 1998.

[39] KRUSKAL, J. B., "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical society*, vol. 7, no. 1, pp. 48–50, 1956.

[40] LANG, S., *Real and functional analysis*, vol. 142. Springer Science & Business Media, 1993.

[41] LEE, J. D., SUN, Y., LIU, Q., and TAYLOR, J. E., "Communication-efficient sparse regression: a one-shot approach," *arXiv preprint arXiv:1503.04337*, 2015.

[42] LEE, J., *U-statistics: Theory and Practice*. Citeseer, 1990.

[43] LEE, K.-Y., LI, B., and ZHAO, H., "Variable selection via additive conditional independence," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. Part 5, pp. 1037–1055, 2016.

[44] LEHMANN, E. L. and CASELLA, G., "Theory of point estimation," vol. 31, 1998.

[45] LI, R., ZHONG, W., and ZHU, L., "Feature screening via distance correlation learning," *Journal of the American Statistical Association*, vol. 107, no. 499, pp. 1129–1139, 2012.

[46] LIN, X., PHAM, M., and RUSZCZYŃSKI, A., "Alternating linearization for structured regularization problems," *Journal of Machine Learning Research*, vol. 15, pp. 3447–3481, 2014.

[47] LIU, Q. and IHLER, A. T., "Distributed estimation, information loss and exponential families," in *Advances in Neural Information Processing Systems*, pp. 1098–1106, 2014.

[48] LOPES, M., JACOB, L., and WAINWRIGHT, M. J., "A more powerful two-sample test in high dimensions using random projection," in *Advances in Neural Information Processing Systems*, pp. 1206–1214, 2011.

[49] LOPEZ-PAZ, D., HENNIG, P., and SCHÖLKOPF, B., "The randomized dependence coefficient," in *Advances in Neural Information Processing Systems*, pp. 1–9, 2013.

[50] LYONS, R., "Distance covariance in metric spaces," *The Annals of Probability*, vol. 41, no. 5, pp. 3284–3305, 2013.

[51] MACKEY, L. W., TALWALKAR, A., and JORDAN, M. I., "Divide-and-conquer matrix factorization," *CoRR*, vol. abs/1107.0789, 2011.

[52] MAK, S. and JOSEPH, V. R., "Support points," *arXiv preprint arXiv:1609.01811*, 2016.

[53] MARDIA, K., BIBBY, J., and KENT, J., *Multivariate analysis*. Probability and Mathematical Statistics, Acad. Press, 1982.

[54] MCDONALD, R., HALL, K., and MANN, G., "Distributed training strategies for the structured perceptron," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2010.

[55] MITRA, S., AGRAWAL, M., YADAV, A., CARLSSON, N., EAGER, D., and MAHANTI, A., "Characterizing web-based video sharing workloads," *ACM Transactions on the Web*, vol. 5, May 2011.

[56] NEISWANGER, W., WANG, C., and XING, E., "Asymptotically exact, embarrassingly parallel MCMC," *arXiv preprint arXiv:1311.4780*, 2013.

[57] NOWAK, R. D., "Distributed EM algorithms for density estimation and clustering in sensor networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2245–2253, 2003.

[58] PAN, X., JEGELKA, S., GONZALEZ, J. E., BRADLEY, J. K., and JORDAN, M. I., "Parallel double greedy submodular maximization," in *Advances in Neural Information Processing Systems*, pp. 118–126, 2014.

[59] PURI, M. and SEN, P., *Nonparametric methods in multivariate analysis*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics, Wiley, 1971.

[60] RAHIMI, A. and RECHT, B., "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems*, pp. 1177–1184, 2007.

[61] REIMHERR, M. and NICOLAE, D. L., "On quantifying dependence: a framework for developing interpretable measures," *Statistical Science*, vol. 28, no. 1, pp. 116–130, 2013.

[62] RESHEF, D. N., RESHEF, Y. A., FINUCANE, H. K., GROSSMAN, S. R., MCVEAN, G., TURNBAUGH, P. J., LANDER, E. S., MITZENMACHER, M., and SABETI, P. C., "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[63] ROSENBAUM, P. R., "An exact distribution-free test comparing two multivariate distributions based on adjacency," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 4, pp. 515–530, 2005.

[64] ROSENBLATT, J. and NADLER, B., "On the optimality of averaging in distributed statistical learning," *arXiv preprint arXiv:1407.2724*, 2014.

[65] ROSENTHAL, H. P., "On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables," *Israel Journal of Mathematics*, vol. 8, no. 3, pp. 273–303, 1970.

[66] RUDIN, W., *Fourier Analysis on Groups*. John Wiley & Sons, 1990.

[67] RUTH, D. M. and KOYAK, R. A., "Nonparametric tests for homogeneity based on non-bipartite matching," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1615–1625, 2011.

[68] SCHWEIZER, B. and WOLFF, E. F., "On nonparametric measures of dependence for random variables," *The Annals of Statistics*, pp. 879–885, 1981.

[69] SEJDINOVIC, D., SRIPERUMBUDUR, B., GRETTON, A., and FUKUMIZU, K., "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," *The Annals of Statistics*, vol. 41, no. 5, pp. 2263–2291, 2013.

[70] SERFLING, R. J., *Approximation Theorems of Mathematical Statistics (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 1980.

[71] SHAMIR, O., SREBRO, N., and ZHANG, T., "Communication-efficient distributed optimization using an approximate Newton-type method," in *Proceedings of The 31st International Conference on Machine Learning*, pp. 1000–1008, 2014.

[72] SIBURG, K. F. and STOIMENOV, P. A., "A measure of mutual complete dependence," *Metrika*, vol. 71, no. 2, pp. 239–251, 2010.

[73] SONG, Q. and LIANG, F., "A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression," *J. R. Statist. Soc. B*, vol. 77, Part 5, pp. 947–972, 2015.

[74] SRIVASTAVA, R., LI, P., and RUPPERT, D., "Raptt: An exact two-sample test in high dimensions using random projections," *Journal of Computational and Graphical Statistics*, vol. 25, no. 3, pp. 954–970, 2016.

[75] SZÉKELY, G. J. and RIZZO, M. L., "Testing for equal distributions in high dimension," *InterStat*, vol. 5, pp. 1–6, 2004.

[76] SZÉKELY, G. J. and RIZZO, M. L., "A new test for multivariate normality," *Journal of Multivariate Analysis*, vol. 93, no. 1, pp. 58–80, 2005.

[77] SZÉKELY, G. J. and RIZZO, M. L., "Brownian distance covariance," *The Annals of Applied Statistics*, vol. 3, no. 4, pp. 1236–1265, 2009.

[78] SZÉKELY, G. J. and RIZZO, M. L., "Energy statistics: A class of statistics based on distances," *Journal of statistical planning and inference*, vol. 143, no. 8, pp. 1249–1272, 2013.

[79] SZEKELY, G. J. and RIZZO, M. L., "Partial distance correlation with methods for dissimilarities," *The Annals of Statistics*, vol. 42, no. 6, pp. 2382–2412, 2014.

[80] SZÉKELY, G. J., RIZZO, M. L., and BAKIROV, N. K., "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[81] TASKINEN, S., OJA, H., and RANDLES, R. H., "Multivariate nonparametric tests of independence," *Journal of the American Statistical Association*, vol. 100, no. 471, pp. 916–925, 2005.

[82] TIBSHIRANI, R., "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[83] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y., and DEZEURE, R., "On asymptotically optimal confidence regions and tests for high-dimensional models," *The Annals of Statistics*, vol. 42, no. 3, pp. 1166–1202, 2014.

[84] VAN DER VAART, A. W., *Asymptotic statistics (Cambridge series in statistical and probabilistic mathematics)*. Cambridge University Press, 2000.

[85] WAINWRIGHT, M., "Constrained forms of statistical minimax: Computation, communication, and privacy," in *Proceedings of International Congress of Mathematicians*, 2014.

[86] WANG, X., PENG, P., and DUNSON, D. B., "Median selection subset aggregation for parallel inference," in *Advances in Neural Information Processing Systems*, pp. 2195–2203, 2014.

[87] WANG, X., PAN, W., HU, W., TIAN, Y., and ZHANG, H., "Conditional distance correlation," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1726–1734, 2015.

[88] WILKS, S., "On the independence of k sets of normally distributed statistical variables," *Econometrica, Journal of the Econometric Society*, pp. 309–326, 1935.

[89] XU, M., LAKSHMINARAYANAN, B., TEH, Y. W., ZHU, J., and ZHANG, B., "Distributed bayesian posterior sampling via moment sharing,"

[90] YANG, Y. and BARRON, A., "Information-theoretic determination of minimax rates of convergence," *Annals of Statistics*, vol. 27, no. 5, pp. 1564–1599, 1999.

[91] ZHANG, Y., DUCHI, J., and WAINWRIGHT, M., "Divide and conquer kernel ridge regression," in *Conference on Learning Theory*, pp. 592–617, 2013.

[92] ZHANG, Y., DUCHI, J. C., JORDAN, M. I., and WAINWRIGHT, M. J., "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," technical report, UC Berkeley, 2013. Presented at the NIPS Conference 2013.

[93] ZHANG, Y., DUCHI, J. C., and WAINWRIGHT, M. J., "Communication-efficient algorithms for statistical optimization," *Journal of Machine Learning Research*, vol. 14, pp. 3321–3363, 2013.

[94] ZHAO, T., CHENG, G., and LIU, H., "A partially linear framework for massive heterogeneous data," *arXiv preprint arXiv:1410.8570*, 2014.

[95] ZHU, L.-P., LI, L., LI, R., and ZHU, L.-X., "Model-free feature screening for ultrahigh-dimensional data," *Journal of the American Statistical Association*, 2012.

[96] ZINKEVICH, M., WEIMER, M., LI, L., and SMOLA, A. J., "Parallelized stochastic gradient descent," in *Advances in Neural Information Processing Systems*, pp. 2595–2603, 2010.

[97] ZOU, H. and LI, R., "One-step sparse estimates in nonconcave penalized likelihood models," *The Annals of Statistics*, vol. 36, no. 4, pp. 1509–1533, 2008.