2017

Topics in statistical inference for massive data and high-dimensional data

Liuhua Peng
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd
Part of the Statistics and Probability Commons

Recommended Citation
Peng, Liuhua, "Topics in statistical inference for massive data and high-dimensional data" (2017). Graduate Theses and Dissertations. 15601.
https://lib.dr.iastate.edu/etd/15601

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Topics in statistical inference for massive data and high-dimensional data

by

Liuhua Peng

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Song Xi Chen, Co-major Professor
Dan Nettleton, Co-major Professor
Yehua Li
Raymond Wong
Zhengyuan Zhu

Iowa State University
Ames, Iowa
2017

Copyright © Liuhua Peng, 2017. All rights reserved.
DEDICATION

To Mingqi, Ethan and Zoe
# TABLE OF CONTENTS

**LIST OF TABLES** ................................................................. vi

**LIST OF FIGURES** ............................................................... viii

**ACKNOWLEDGEMENTS** ......................................................... xi

**ABSTRACT** ........................................................................ xii

**CHAPTER 1. GENERAL INTRODUCTION** ................................. 1

**CHAPTER 2. DISTRIBUTED STATISTICAL INFECTION FOR MASSIVE DATA** ................................................................. 4

  2.1 Introduction ................................................................. 4

  2.2 Symmetric statistics ...................................................... 7

  2.3 Distributed statistics ..................................................... 8

  2.4 Computing issues and selection of $K$ ............................ 11

  2.5 Asymptotic distribution of $T_{N,K}$ ............................. 13

  2.6 Bootstrap procedures ................................................... 17

    2.6.1 Entire sample bootstrap ....................................... 17

    2.6.2 U-statistics as an example ................................... 19

    2.6.3 BLB and SDB ....................................................... 21

  2.7 Distributed bootstrap ................................................ 25

  2.8 Pseudo-distributed bootstrap ....................................... 28

  2.9 Comparison of the distributed bootstrap, the pseudo-distributed bootstrap, the BLB and the SDB ........................................ 32

  2.10 Applications .............................................................. 33

  2.11 Numerical studies .................................................... 38
2.11.1 Gini’s mean difference .................................................. 38
2.11.2 Distance covariance ..................................................... 46
2.11.3 Real data analysis ......................................................... 47
2.12 Conclusion ................................................................. 51
2.13 Appendix ................................................................. 54
  2.13.1 Proof of Theorem 2.4 .................................................. 55
  2.13.2 Proof of Theorem 2.6 .................................................. 56
  2.13.3 Proof of Theorem 2.7 .................................................. 57
  2.13.4 Proof of Theorem 2.8 .................................................. 58
  2.13.5 Proof of Corollary 2.1 .................................................. 58
  2.13.6 Proof of Corollary 2.2 .................................................. 59
  2.13.7 Proof of Theorem 2.9 .................................................. 60

CHAPTER 3. MORE POWERFUL TESTS FOR SPARSE HIGH DIMENSIONAL COVARIANCES MATRICES ........................................... 65

  3.1 Introduction ............................................................... 65
  3.2 Motivations and preliminaries ........................................... 68
  3.3 Testing procedures ......................................................... 71
    3.3.1 Identity test .......................................................... 71
    3.3.2 Power of the identity test .......................................... 74
    3.3.3 Sphericity test ........................................................ 75
  3.4 Selection of $k$ ............................................................. 77
  3.5 Numerical results ........................................................ 80
    3.5.1 Simulation studies .................................................... 80
    3.5.2 Empirical study ....................................................... 87
  3.6 Discussion ............................................................... 89
  3.7 Appendix ................................................................. 90
    3.7.1 Critical lemmas ....................................................... 91
    3.7.2 Proof of Proposition 3.1 .......................................... 96
    3.7.3 Proof of Theorem 3.1 .............................................. 97
CHAPTER 4. VARIABLE IMPORTANCE ASSESSMENT AND BACKWARD VARIABLE SELECTION FOR HIGH-DIMENSIONAL DATA . . 103

4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104
4.2 Variable importance . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106
  4.2.1 The multi-response permutation procedure . . . . . . . . . . . . . . . . . . 106
  4.2.2 An MRPP-based importance measure . . . . . . . . . . . . . . . . . . . . . 107
  4.2.3 An alternative importance measure . . . . . . . . . . . . . . . . . . . . . . 110
4.3 Backward selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 112
4.4 A modified MRPP . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 114
4.5 Real data analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 115
4.6 Simulations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 118
  4.6.1 Backward selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 118
  4.6.2 Modified MRPP . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 124
4.7 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 126

CHAPTER 5. GENERAL CONCLUSIONS . . . . . . . . . . . . . . . . . . . . . . . . . . 130

BIBLIOGRAPHY . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 132
LIST OF TABLES

Table 2.1 Algorithm for the distributed bootstrap. ................................. 25
Table 2.2 Algorithm for the pseudo-distributed bootstrap. ....................... 29
Table 2.3 Comparison of the distributed bootstrap (DB), the pseudo-distributed bootstrap (PDB), the BLB and the SDB. ................................. 33
Table 2.4 Number of iterations completed in 10 seconds for each method. DB: the distributed bootstrap, PDB: the pseudo-distributed bootstrap, BLB: the bag of little bootstrap, SDB: the subsampled double bootstrap. . 43
Table 2.5 Coverage rates and confidence interval widths (in parentheses) of the 95% confidence intervals with 10 seconds time budget for Gaussian data. 43
Table 2.6 Sizes of independence tests based on $d_{cov}^2_{N,K}(\mathbf{Y}, \mathbf{Z})$. $T_{Var}$: test using variance estimation in (2.17); $T_{PDB}$: test using the pseudo-distributed bootstrap. .............................................. 48
Table 2.7 Powers of independence tests based on $d_{cov}^2_{N,K}(\mathbf{Y}, \mathbf{Z})$ for $\varrho = 0.05$. ................................. 49
Table 2.8 Powers of independence tests based on $d_{cov}^2_{N,K}(\mathbf{Y}, \mathbf{Z})$ for $\varrho = 0.1$. ................................. 50
Table 2.9 $U_{N,K}$ and $S^2_{U_{N,K}}$ for airline on-time dataset along with computing time (in seconds) given in the parentheses. ................................. 52
Table 2.10 Coverage rates and confidence interval widths (in parentheses) of the 95% confidence intervals with 10 seconds time budget for Gamma data. 62
Table 2.11 Coverage rates and confidence interval widths (in parentheses) of the 95% confidence intervals with 10 seconds time budget for Poisson data. 62
Table 3.1 Empirical power with respect to $k$ for the identity test. ............ 78
Table 3.2  Empirical sizes of the proposed test (IT) for the identity hypothesis, along with those of the tests by Chen, Zhang and Zhong (2010) (CZZ) and Ledoit and Wolf (2002) (LW). ......................................................... 82

Table 3.3  Number of identified gene-sets against null hypotheses. ST and IT stand for the proposed sphericity and identity tests, respectively. The last column reports the means and standard deviations for $k_{qc}$. ........................................... 89

Table 3.4  Computational time (in seconds) for a single sphericity test under the alternative $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p}$ with $\sigma_{ij} = I(i = j) + \theta|i-j|^{-\rho}I(i \neq j)$ with different $n$ and $p$. ................................................................. 90

Table 3.5  Empirical sizes of the proposed test (ST) for the sphericity hypothesis, along with those of the tests by Chen, Zhang and Zhong (2010) (CZZ) and Ledoit and Wolf (2002) (LW). ......................................................... 100

Table 4.1  Number of gene sets for which each method ranks 1, 2, or 3 with respect to false positive rate. (Rank 1 is best and 3 is worst). ........................................... 124

Table 4.2  Combinations of $n_1$ and $n_2$ considered in the simulations. ................. 124

Table 4.3  Empirical sizes of original MRPP (MRPP$_{Org}$) and modified MRPP with different choices of $R_0$. For Mod$_{S(L)}$, $R_0$ is chosen as the cardinality of $S(L)$. For Mod$_2$, Mod$_4$, Mod$_8$, Mod$_{16}$ and Mod$_{\sqrt{R}}$, $R_0$ is pre-specified as 2, 4, 8, 16 and $\sqrt{R}$, respectively. ........................................... 125
# LIST OF FIGURES

| Figure 2.1 | A diagram summary of the BLB procedure. | 23 |
| Figure 2.2 | $\text{MSE}(U_{N,K})/\text{MSE}(U_N)$ and $\text{MSE}(N^{-1}S^2_{U,N,K})/\text{MSE}(N^{-1}S^2_{U,N})$ against the number of blocks $K$ for each scenario. | 41 |
| Figure 2.3 | $|\text{Bias}(N^{-1}S^2_{U,N,K})|/|\text{Bias}(N^{-1}S^2_{U,N})|$ and $\text{Var}(N^{-1}S^2_{U,N,K})/\text{Var}(N^{-1}S^2_{U,N})$ against the number of blocks $K$ for each scenario. | 42 |
| Figure 2.4 | Time evolution of relative errors under Gaussian scenario. | 45 |
| Figure 2.5 | Time evolution of estimated confidence interval width. | 53 |
| Figure 2.6 | Time evolution of relative errors under Gamma scenario. | 63 |
| Figure 2.7 | Time evolution of relative errors under Poisson scenario. | 64 |
| Figure 3.1 | Empirical size of the 5\% sphericity test with respect to the banding width $k$ for various sample sizes and dimensions. | 78 |
| Figure 3.2 | Proportions of the banding width selected by the method of Qiu and Chen (2015). | 79 |
| Figure 3.3 | Empirical powers for the identity tests against the alternative in the diagonal form with the Gaussian distributed innovation. | 83 |
| Figure 3.4 | Empirical powers for the identity tests against the alternative in the diagonal form with the Gamma distributed innovation. | 83 |
| Figure 3.5 | Empirical powers for the identity tests against the alternatives in either the bandable or banded forms of the proposed test with either the Gaussian or Gamma distributed innovation. | 84 |
| Figure 3.6 | Empirical powers for the sphericity tests against the alternative in the diagonal form with the Gaussian distributed innovation. | 85 |
Figure 3.7  Empirical powers for the sphericity tests against the alternative in the diagonal form with the Gamma distributed innovation.

Figure 3.8  Empirical powers for the sphericity tests against the alternatives in either the bandable or banded forms of the proposed test with either the Gaussian or Gamma distributed innovation.

Figure 3.9  Comparison of heat maps of correlation matrices before and after reordering of a gene set, GO:0000086 from BP category in NEG group.

Figure 3.10  Histograms of p-values for identity hypotheses (right panels) and sphericity hypotheses (left panels) for the three gene ontologies by the proposed tests.

Figure 4.1  The MRPP p-values on the sets of selected and deleted genes in each iteration.

Figure 4.2  Summary of 859 gene sets selected for simulations: (a) Histogram of the number of genes in each gene set; (b) Histogram of MRPP test p-values for each gene set.

Figure 4.3  Number of genes selected by each method for each gene set.

Figure 4.4  Comparison of average false positive rate for three methods, $\rho_t$ for two-sample t-test, $\rho_l$ for limma and $\rho_{bs}$ for backward selection. (a) $\rho_t - \rho_{bs}$; (b) $\rho_l - \rho_{bs}$; (c) $\rho_t - \rho_l$; (d) the proportion of 1000 simulation replications in which each method has the lowest false positive rate across gene sets with of varying size.

Figure 4.5  (a): Histogram of number of genes in the 61 gene sets for which all three methods initially selected the same number of genes. (b) to (d): Comparison of average false positive rate for three methods when applied to the 61 gene sets.

Figure 4.6  Empirical powers of the original MRPP and modified MRPP with different choices of $R_0$ for $\nu = 0.5$. 
Figure 4.7 Empirical powers of the original MRPP and modified MRPP with different choices of $R_0$ for $\nu = 1$ \hfill 128

Figure 4.8 Empirical powers of the original MRPP and modified MRPP with different choices of $R_0$ for $\nu = 1.5$ \hfill 129
ACKNOWLEDGEMENTS

I would like to thank my major professor, Dr. Song Xi Chen and Dr. Dan Nettleton, for their extremely helpful guidance and support during my Ph.D. study. They showed me their great patience and encouragement, which have often inspired me for completing my graduate education and pursuing my future research. I have learned so much from them and will be indebted to them for everything they have taught me. I would also like to thank my remaining committee members, Dr. Yehua Li, Dr. Raymond Wong and Dr. Zhenyuan Zhu, for their interests and efforts for this work. Finally, I would like to thank my family and friends for their endless love and support.
ABSTRACT

This dissertation consists of three research papers that deal with three different problems in statistics concerning high-volume datasets. The first paper studies the distributed statistical inference for massive data. With the increasing size of the data, computational complexity and feasibility should be taken into consideration for statistical analyses. We investigate the statistical efficiency of the distributed version of a general class of statistics. Distributed bootstrap algorithms are proposed to approximate the distribution of the distributed statistics. These approaches relieve the computational burdens of conventional methods while preserving adequate statistical efficiency. The second paper deals with testing the identity and sphericity hypotheses problem regarding high-dimensional covariance matrices, with a focus on improving the power of existing methods. By taking advantage of the sparsity in the underlying covariance matrices, the power improvement is accomplished by utilizing the banding estimator for the covariance matrices, which leads to a significant reduction in the variance of the test statistics. The last paper considers variable selection for high-dimensional data. Distance-based variable importance measures are proposed to rank and select variables with dependence structures being taken into consideration. The importance measures are inspired by the multi-response permutation procedure (MRPP) and the energy distance. A backward selection algorithm is developed to discover important variables and to improve the power of the original MRPP for high-dimensional data.
CHAPTER 1. GENERAL INTRODUCTION

Datasets with high volume are becoming increasingly prevalent as advances in technology have allowed for the generation of such datasets. Two different types of high-volume datasets are considered in this thesis. The first type is the massive data with large sample size and the second category is the high-dimensional data with a large number of variables. Analyzing those type of datasets demands new technologies that meet the inherent properties of the datasets. For massive data, with the growth in the size of the datasets, classical statistical methods can be computationally expensive, or infeasible, even though they can achieve favorable statistical efficiency. Thus, computational efficiency is also an important criterion when comparing different statistical approaches for massive data.

Specifically, in the first paper, we consider distributed statistical inference for massive data. Among the techniques developed for analyzing massive datasets, there are two categories that are most important. One is the “split-and-conquer” method (Zhang, Duchi and Wainwright, 2013; Chen and Xie, 2014; Battey, et al., 2015), and the other one is the resampling-based methods (Kleiner, et al., 2014; Sengupta, Volgushev and Shao, 2016). In this paper, we consider a general class of symmetric statistics (Lai and Wang, 1993; Jing and Wang, 2010) that encompasses many commonly used statistics, for example, the U and L-statistics. By following the procedure of split-and-conquer, we formulate a distributed version of the statistics. The distributed statistic has advantages in computation over the statistic based on the whole dataset, including computing time, storage and memory requirement. Moreover, we study the statistical efficiency of the distributed statistic and compare it with the full sample one. It is shown that for non-degenerate statistics, with a properly selected number of data blocks, the distributed statistic can maintain the same leading order efficiency as the full sample statistic. By recognizing the appealing properties of the distributed statistic, approximating its distribu-
tion function is of great interest. Two bootstrap algorithms (the distributed bootstrap and the pseudo-distributed bootstrap) that can be implemented distributively are proposed to achieve this goal. Being distributed in computing makes the algorithms adaptive to massive datasets. It is also shown that the distributed approaches can preserve a sufficient level of statistical efficiency.

For high-dimensional data, due to the large amount of variables, classical statistical approaches are no longer powerful or even not applicable. In the second paper, we consider testing the identity and sphericity hypotheses for high-dimensional covariance matrices, where traditional likelihood based tests fail under the high-dimensional setting as the sample covariance matrix is no longer a consistent estimator of the population covariance matrix (Bai, Silverstein and Yin, 1988; Bai and Yin, 1993). Inspired by the existing U-statistic based tests for high-dimensional covariance matrices proposed in Chen, Zhang and Zhong (2010) and the natural feature of sparsity for high-dimensional data, we formulate two new test statistics by employing the banded covariance estimator (Bickel and Levina, 2008a) incorporate with two discrepancy measures for the identity and sphericity hypotheses. This enables us to take advantage of the sparsity of the underlying covariance matrices and thus leads to the power improvement. Theoretical results are established to ensure the consistency of the proposed tests and to reveal the reason for power improvement, which is due to the reduction in the variance of the test statistics.

Variable selection is in great need for high-dimensional data. For example, identifying differentially expressed genes for gene expression data is of great interest in genomic analysis. For this type of applications, an important feature of the data is that only a limit number of experimental subjects can be measured, which leads to the so-called “small $n$, large $p$” problem. In the third paper, we consider variable selection problem under the high-dimensional scenario. Existing methods for solving this problem have some limitations. The multi-response permutation procedure (MRPP) (Mielke and Berry, 2007) is a powerful tool for detecting differences between multivariate distributions. By imposing a hypothetical perturbation that tilts each dimension in the Euclidean distance between two data vectors, we introduce distance-based variable importance measures that takes the covariance structure into consideration. In
addition, a backward selection algorithm based on the importance measure is proposed to discover importance variables. By eliminating irrelevant variables in each iteration, we can focus on a subset of important variables. Besides identifying important variables, we modify the original MRPP based on the backward selection algorithm, with focusing on the important subset of the variables. This leads to a power improvement of the original MRPP when detecting differences between multivariate distributions.

The rest of this dissertation is organized as followings. We present the results concerning the distributed inference for massive data in Chapter 2, where theoretical supports are given; In Chapter 3, we propose two more powerful tests for testing the identity and sphericity hypothesis for high-dimensional covariance matrices; In Chapter 4, we introduce the concept of variable importance inspired by the MRPP and propose a backward selection algorithm, which can be used to discover importance variables and improve the original MRPP. A general conclusion is included in Chapter 5, with discussions on future studies.
CHAPTER 2. DISTRIBUTED STATISTICAL INFERENCE FOR
MASSIVE DATA

Liuhua Peng and Song Xi Chen

Abstract

This paper studies distributed statistical inference for a general type of statistics that en-
compasses U-statistics and M-estimator in the context of massive data. When the data are
stored on multiple platforms, it is usually expensive and slow to do data communication. To
deal with this issue, we formulate the distributed statistics which can be computed distribu-
tively and hence reduces computational time significantly. We investigate properties of the
distributed statistics from the perspective of mean square error of estimation and their asymp-
totic distributions. In addition, we propose two distributed bootstrap algorithms which are
computationally effective and consistent theoretically. Applications of our approaches and nu-
merical studies are provided to support them.

2.1 Introduction

In the past decade, massive data with rapidly increasing size are generated in many fields of
scientific studies that create needs for statistical analyses. Not only the sheer size of the data is
an issue, but also the fact that these datasets are often stored in multiple locations or clouds,
where each contains a subset of the data which can be massive in its own right. This implies
that a statistical procedure that involves the entire data has to involve data communication
between different storage facilities or clouds, which is usually slow and expensive. For statistical
analyzes involving a massive data set, the computational and data storage implications of a
statistical procedure should be taken into consideration in addition to the usual criteria of statistical inference.

Two kinds of methods have been developed to deal with these new challenges with massive data. One is the so-called “split-and-conquer” (SaC) method (Zhang, Duchi and Wainwright, 2013; Chen and Xie, 2014; Battey, et al., 2015); and the other one is the resampling-based methods advocated by Kleiner, et al. (2014) and Sengupta, Volgushev and Shao (2016). From the estimation point of view, the SaC first partitions the entire dataset into blocks with relatively smaller sizes, performs the estimation on each data block and then aggregate the estimators from each block to get the final estimator. SaC has been used in different settings under the massive data scenario, for instance the M-estimation by Zhang, Duchi and Wainwright (2013), the generalized linear models by Chen and Xie (2014), and other studies by Huang and Huo (2015), Battey, et al. (2015) and Lee, et al. (2015).

The other method involves making the bootstrap procedure adaptive to the massive data setting in order to obtain standard errors or confidence intervals for statistical inference. As the conventional bootstrap resampling of the entire dataset is not feasible for massive data due to its being computationally too expensive, Kleiner, et al. (2014) introduced the bag of little bootstrap (BLB) that incorporates subsampling and the $m$ out of $n$ bootstrap to assess the quality of estimators for various inference purposes. Sengupta, Volgushev and Shao (2016) proposed the subsampled double bootstrap (SDB) method based on the idea of the BLB and a fast double bootstrap (Davidson and MacKinnon, 2002; Chang and Hall, 2015). However, both the BLB and the SDB have some limitations. Firstly, the core idea of the BLB and the SDB is to construct subsets of small size from the entire dataset, and then resample these small subsets to obtain weighted resamples of the subset such that the resample size is equal to the size of the entire data. Hence, the computational efficiency of the BLB and the SDB heavily rely on an assumption that the estimator of interest admitted a weighted subsample representation (Kleiner, et al., 2014), which is usually satisfied for linear statistics. Secondly, the SDB requires resampling from the entire dataset, which is computationally expensive and hard to accomplish when the data are stored in different locations.
In this paper, we consider distributed statistical inference for a class of symmetric statistics that does the SaC step for a distributed formulation of the statistics followed by two versions of the bootstrap specifically designed for the distributed statistic formulation. We first divide the entire dataset into $K$ data blocks (which is similar to SaC) and then formulate a distributed statistic by averaging the statistics obtained on each data block. The distributed statistic has computational advantage over the statistics based on the entire (massive) data. Besides the computational issue, we consider achieving the best estimation efficiency for the distributed estimation, that touches on the issue of selecting the number of data blocks. After studying the relative efficiency of the distributed statistic relative to the full sample statistics, we device computationally efficient bootstrap algorithms to approximate the distribution of the distributed statistics, which can be used in various inferential tasks of the statistical inference. The latter is carried out by recognizing the limitations of the BLB and the SDB. Specifically, we develop a distributed bootstrap (DB) and a pseudo-distributed bootstrap (PDB) to approximate the distribution of the distributed statistics. Both strains of the bootstrap method can be implemented distributively and are suitable to parallel and distributed computing platforms. It is shown that being distributed in both the statistical formulation and the subsequent bootstrap makes the computation much suited to the massive data setting while maintaining sufficient level of statistical efficiency and accuracy in the inference.

The rest of the paper is organized as follows. The class of symmetric statistics considered in this paper is introduced in Section 2.2. The distributed statistic is formulated in Section 2.3 together with its statistical efficiency. Section 2.4 discusses the issues concerning computational complexity and the selection of the number of data blocks. In Section 2.5, asymptotic distributions of the distributed statistics are established for both non-degenerate and degenerate cases. Bootstrap procedures which are designed to approximate the distribution of the distributed statistics are discussed through Section 2.6 to Section 2.9. Application of the distributed inference procedures on the distance covariance is presented in Section 2.10. Section 2.11 provides numerical studies to support our theoretical results. All technical proofs along with additional simulation results are provided in the appendix.
2.2 Symmetric statistics

Let $\mathbf{X}_N = \{X_1, \ldots, X_N\}$ be a sequence of independent random vectors taking values in a measurable space $(\mathcal{X}, \mathcal{B})$ with a common distribution $F$. The statistic of interest $T_N = T(\mathbf{X}_N)$ is a statistic that admits a general non-linear form

$$T_N = \theta + N^{-1} \sum_{i=1}^{N} \alpha(X_i; F) + N^{-2} \sum_{1 \leq i < j \leq N} \beta(X_i, X_j; F) + R_N, \quad (2.1)$$

where $\theta = \theta(F)$ is the parameter of interest, $\alpha(x; F)$ and $\beta(x, y; F)$ are known measurable functions satisfying the conditions given in Condition 2.1. In (2.1), $R_N = R(\mathbf{X}_N; F)$ is a remainder term which may be viewed as the approximation error of the expansion that constitutes the first three terms on the right side. In summary, $T_N$ may be considered as an estimator of $\theta$ and $\mathbb{E}(R_N)$ is the estimation bias, while both the linear and quadratic terms contribute to the variation of $T_N$.

The statistic $T_N$ encompasses many commonly used statistics, for example, the U and L-statistics, and the smoothed functions of the sample means. The form (2.1) was introduced in Lai and Wang (1993) while Edgeworth expansions for $T_N$ were studied in Jing and Wang (2010). In both Lai and Wang (1993) and Jing and Wang (2010), it is assumed that $\alpha(X_1; F)$ is non-lattice distributed and $R_N$ is of smaller order of the quadratic term involving $\beta$. For instance, $T_N$ can be a U-statistic with $\alpha(X_i; F)$ and $\beta(X_i, X_j; F)$ terms to be the first and second order terms in its Hoeffding decomposition (Hoeffding, 1948; Serfling, 1980). In this case, $\mathbb{E}(R_N) = 0$ and $\text{Var}(R_N) = O(N^{-3})$. In this paper, we relax the conditions in Lai and Wang (1993) and Jing and Wang (2010) such that $T_N$ includes more types of statistics. The linear term involving $\alpha(X_i; F)$ can vanish, for instance in the case of the degenerate U-statistics. Furthermore, the M-estimators are included in (2.1) with $R_N$ having certain order of magnitude depending on the forms of the score functions. When the score function of the M-estimator is twice differentiable and its second derivative is Lipschitz continuous, the M-estimator can be expressed in the form of (2.1) with an explicit $\beta$ and the reminder term $R_N = \mathcal{O}(N^{-1})$ (Lahiri, 1994). However, when the score function is not smooth enough, the M-estimator may not be expanded to the second-order $\beta(X_i, X_j; F)$ term. In this situation, we can absorb the term involving $\beta$ into $R_N$, which is often of order $\mathcal{O}(N^{-3/4})$ (He and Shao, 1996).
We assume the following condition regarding \( \alpha \) and \( \beta \).

**Condition 2.1.** The functions \( \alpha(x; F) \) and \( \beta(x, y; F) \) are known measurable functions of \( x \) and \( y \), satisfying \( E\{\alpha(X_1; F)\} = 0 \) and \( \text{Var}\{\alpha(X_1; F)\} = \sigma^2_\alpha \in [0, \infty) \), and \( \beta(x, y; F) \) is symmetric in \( x \) and \( y \) such that \( E\{\beta(X_1, X_2; F)|X_1\} = 0 \) and \( E\{\beta(X_1, X_2; F)\}^2 = \sigma^2_\beta \in [0, \infty) \).

### 2.3 Distributed statistics

To improve the computation of \( T_N \) in the context of massive data, we divide the full dataset \( X_N \) into \( K \) data blocks. Let \( X_{N,K}^{(k)} = \{X_{k,1}, \ldots, X_{k,n_k}\} \) be the \( k \)-th data block of size \( n_k \), for \( k = 1, \ldots, K \). Sometimes, such division is not needed when \( X_N \) is naturally stored over \( K \) storage facilities. Otherwise, the blocks can be obtained by random sampling.

We have the following conditions concerning the sizes of data blocks \( \{n_k\}_{k=1}^K \) and the number of blocks \( K \).

**Condition 2.2.** There exist finite positive constants \( c_1 \) and \( c_2 \) such that \( c_1 \leq \inf_{k_1, k_2} n_{k_1}/n_{k_2} \leq \sup_{k_1, k_2} n_{k_1}/n_{k_2} \leq c_2 \). In addition, \( K \) can be either finite or diverging to infinity as long as it satisfies that \( K/N \to 0 \) as \( N \to \infty \).

Condition 2.2 implies that \( \{n_k\}_{k=1}^K \) are of the same order and the number of data blocks \( K \) should be of a smaller order of \( N \).

For \( k = 1, \ldots, K \), let \( T_{N,K}^{(k)} = T(X_{N,K}^{(k)}) \) be a version of \( T_N \), but is based on the \( k \)-th data block \( X_{N,K}^{(k)} \) such that

\[
T_{N,K}^{(k)} = \theta + n_k^{-1} \sum_{i=1}^{n_k} \alpha(X_{k,i}; F) + n_k^{-2} \sum_{1 \leq i < j \leq n_k} \beta(X_{k,i}, X_{k,j}; F) + R_{N,K}^{(k)},
\]

where \( R_{N,K}^{(k)} = R(X_{N,K}^{(k)}; F) \) is the remainder term specific to the \( k \)-th block.

By doing a weighted average of the \( K \) block-wise statistics, we arrive at the distributed statistic

\[
T_{N,K} = N^{-1} \sum_{k=1}^{K} n_k T_{N,K}^{(k)}, \tag{2.2}
\]

which can be expressed as

\[
T_{N,K} = \theta + N^{-1} \sum_{i=1}^{N} \alpha(X_i; F) + N^{-1} \sum_{k=1}^{K} n_k^{-1} \sum_{1 \leq i < j \leq n_k} \beta(X_{k,i}, X_{k,j}; F) + R_{N,K}, \tag{2.3}
\]
where \( R_{N,K} = N^{-1} \sum_{k=1}^{K} n_k R_{N,K}^{(k)} \). It is clear that the first two terms (containing \( \theta \) and \( \alpha \)) of \( T_{N,K} \) and \( T_N \) are the same. The difference between them occurs at the terms involving \( \beta \) and the remainders.

To facilitate analyzes on \( T_{N,K} \), we assume the following assumption.

**Condition 2.3.** If \( E(R_N) = b_1 N^{-\tau_1} + o(N^{-\tau_1}) \) and \( \text{Var}(R_N) = o(N^{-\tau_2}) \) for some \( b_1 \neq 0, \tau_1 \geq 1 \) and \( \tau_2 \geq 1 \), then \( E(R_{N,K}^{(k)}) = b_{1,k} n_k^{-\tau_1} + o(n_k^{-\tau_1}) \) for some \( b_{1,k} \neq 0 \) and \( \text{Var}(R_{N,K}^{(k)}) = o(n_k^{-\tau_2}) \), for \( k = 1, \ldots, K \).

Condition 2.3 describes a divisible property of \( R_N \) in its first two moments with respect to the sample size \( N \). That is, \( \{R_{N,K}^{(k)}\}_{k=1}^{K} \) inherit the first two moments properties of \( R_N \) with \( N \) substituted by \( n_k \).

First, we present results concerning \( T_N \) to provide baseline results.

**Proposition 2.1.** Under Condition 2.1, if \( E(R_N) = b_1 N^{-\tau_1} + o(N^{-\tau_1}) \) for some \( b_1 \neq 0, \tau_1 \geq 1 \), and \( \text{Var}(R_N) = o(N^{-2}) \), then

\[
\text{Bias}(T_N) = b_1 N^{-\tau_1} + o(N^{-\tau_1}) \quad \text{and} \quad \text{Var}(T_N) = \sigma^2 \alpha N^{-1} + 2^{-1} \sigma^2 \beta N^{-2} + N^{-1} \sum_{i=1}^{N} \text{Cov}\{\alpha(X_i; F), R_N\} + o(N^{-2}).
\]

The bias of \( T_N \) is totally determined by the expectation of \( R_N \). In the variance expression, we keep the term involving the covariance between \( \alpha(X_i, F) \) and \( R_N \) as it may not be negligible in a general case. For the special case when \( R_N \) is uncorrelated with \( \alpha(X_i; F) \), which is the case for the U-statistics, the covariance term is negligible. Then, the mean square error (MSE) of \( T_N \) is

\[
\text{MSE}(T_N) = \sigma^2 \alpha N^{-1} + 2^{-1} \sigma^2 \beta N^{-2} + b_1^2 N^{-2\tau_1} + N^{-1} \sum_{i=1}^{N} \text{Cov}\{\alpha(X_i; F), R_N\} + o(N^{-2} + N^{-2\tau_1}).
\]

The results of \( T_{N,K} \) is given in the following theorems.
Theorem 2.1. Under Conditions 2.1, 2.2, 2.3, if \( E(R_N) = b_1 N^{-\tau_1} + o(N^{-\tau_1}) \) for some \( b_1 \neq 0 \), \( \tau_1 \geq 1 \), and \( \text{Var}(R_N) = o(N^{-2}) \), then

\[
\text{Bias}(T_{N,K}) = N^{-1} \sum_{k=1}^{K} b_{1,k} n_k^{1-\tau_1} + o(K^{\tau_1} N^{-\tau_1}) \quad \text{and} \quad \\
\text{Var}(T_{N,K}) = \sigma^2_\alpha N^{-1} + 2^{-1} \sigma^2_\beta KN^{-2} + N^{-2} \sum_{k=1}^{K} n_k \sum_{i=1}^{n_k} \text{Cov}\{\alpha(X_{k,i}; F), R_{N,K}^{(k)}\} + o(KN^{-2}).
\]

If \( T_N \) is an unbiased estimator of \( \theta \), namely \( \tau_1 = \infty \), \( T_{N,K} \) is also unbiased so that \( E(T_{N,K}) = \theta \). Thus the distributed approach can preserve the unbiasedness of the statistics. For the biased case of \( \tau_1 < \infty \), according to Theorem 2.1, as \( N^{-1} \sum_{k=1}^{K} b_{1,k} n_k^{1-\tau_1} \) is of order \( O(K^{\tau_1} N^{-\tau_1}) \) under Condition 2.2, the bias is enlarged by a factor of \( K^{\tau_1} \) for \( T_{N,K} \) relative to that of \( T_N \). This indicates that the data blocking accumulates the biases from each data block, leading to an increase in the bias of the distributed statistic. For the more common case that the bias of \( T_N \) is at the order of \( O(N^{-1}) \), the bias of \( T_{N,K} \) is \( O(KN^{-1}) \).

For the variance, we note there is an increase in the factor of \( K \) in the term that involves \( \sigma^2_\beta N^{-2} \) due to the data blocking. For the covariance term in \( \text{Var}(T_{N,K}) \), under the conditions in Theorem 2.1, as \( \text{Var}\{N^{-1} \sum_{i=1}^{N} \alpha(X_i; F)\} = O(N^{-1}) \) and \( \text{Var}(R_N) = o(N^{-2}) \), by utilizing Cauchy-Schwarz inequality we have

\[
\text{Cov}\left\{N^{-1} \sum_{i=1}^{N} \alpha(X_i; F), R_N\right\} = o(N^{-3/2}).
\]

Similarly, \( N^{-2} \sum_{k=1}^{K} n_k \sum_{i=1}^{n_k} \text{Cov}\{\alpha(X_{k,i}; F), R_{N,K}^{(k)}\} = o(K^{1/2} N^{-3/2}) \) by the independence between each data block. Thus, an increase of order \( K^{1/2} \) occurs in the covariance term. Moreover, the covariance term in \( \text{Var}(T_{N,K}) \) is always at a smaller order of \( N^{-1} \) as \( K = o(N) \), while it may exceed the \( KN^{-2} \) order term. For the case of \( \sigma^2_\alpha > 0 \), the variance inflation happens in the second order term without altering the leading order term. However, for the degenerate case of \( \sigma^2_\alpha = 0 \), but \( \sigma^2_\beta > 0 \), the variance inflation appears in the leading order via a factor of \( K \).
Combining the bias and variance from Theorem 2.1, we have

\[
\text{MSE}(T_{N,K}) = \sigma_\alpha^2 N^{-1} + 2\sigma_\beta^2 KN^{-2} + N^{-2} \left( \sum_{k=1}^{K} b_{1,k} \tau_1^1 \right)^2 \\
+ N^{-2} \sum_{k=1}^{K} n_k \sum_{i=1}^{n_k} \text{Cov}\left\{ \alpha(X_{k,i}; F), R_{N,K}^{(k)} \right\} + o(KN^{-2} + K^2 \tau_1 N^{-2}).
\]

(2.5)

To compare the MSEs of \(T_N\) and \(T_{N,K}\), we first consider the non-degenerate case of \(\sigma_\alpha^2 > 0\).

By comparing (2.4) with (2.5). If

\[
K = o \left( N^{1-1/(2\tau_1)} \right)
\]

then \(K^{2\tau_1} N^{-2\tau_1} = o(N^{-1})\), and \(\text{MSE}(T_N)\) and \(\text{MSE}(T_{N,K})\) share the same leading order term \(\sigma_\alpha^2 N^{-1}\). We note that \(1 - 1/(2\tau_1)\) is an increasing function of \(\tau_1\), ranging from \(1/2\) (when \(\tau_1 = 1\)) to \(1\) (when \(\tau_1 \to \infty\)). For the case of \(\tau_1 = 1\) when \(T_N\) has a bias of order \(N^{-1}\), the number of blocks \(K\) is required to be at a smaller order of \(N^{1/2}\) such that \(T_{N,K}\) maintains the same leading order efficiency as \(T_N\) in terms of MSEs. The increase in the variance due to the data blocking is reflected in the MSE, although it is of the second order unless in the degenerate case.

In summary, we see that the data blocking increases the bias and variance of the distributed statistic \(T_{N,K}\). This may be viewed as a price paid to gain computational scalability for massive data. Despite this, there is room to select \(K\) properly to minimize the increases. For instance, by choosing \(K\) to satisfy (2.6), the increase in the bias is confined in the second order. Also, for non-degenerate situation, the impact of the data blocking can be largely limited to the second order as reflected in the MSE expression (2.5).

2.4 Computing issues and selection of \(K\)

The rationale for carrying out the distributed formulation is in reducing the computing time and data storage significantly in the context of massive data while maintaining the best level of the estimation accuracy given a computational and storage budget. In this section, we will discuss this issue with the U-statistics as an example.

First, the distributed approach can reduce the computing time significantly when the inferences involved is computationally intensive. For example, if the computational complexity
of calculating $T_N$ is at the order of $O(N^a)$ with $a > 1$, then computing $T_{N,K}$ only needs $K \times O((N/K)^a)$, that is $O(K^{1-a}N^a)$ steps. So calculating $T_{N,K}$ is a factor of $O(K^{a-1})$ faster than computing $T_N$ directly. Similar statement can be found from Theorem 5 in Chen and Xie (2014). In our distributed approach, the computing effort of the average step is only at the order of $K$, which is negligible.

Besides the saving in computing cost, the distributed approach requires less memory space. Suppose the memory needed for computing $T_N$ is $O(N^b)$ for some $b \geq 1$. Then for $T_{N,K}$, the requirement on the size of memory is at the order of $O((N/K)^b + K)$. This means a memory saving by a factor of $1 - K^{-b}$ when $K = O(N^{b/(1+b)})$, or $1 - KN^{-b}$ when $KN^{-b/(1+b)} \to \infty$, for the distributed approach.

Now we use the U-statistics as an example to demonstrate the savings in computing. Consider a U-statistic $U_N$ of degree $m \geq 2$ with a symmetric kernel function $h$ such that

$$ U_N = \binom{N}{m}^{-1} \sum_{1 \leq i_1 < \ldots < i_m \leq N} h(X_{i_1}, \ldots, X_{i_m}) $$

$$ = \theta_U + N^{-1} \sum_{i=1}^{N} \alpha_U(X_i; F) + N^{-2} \sum_{1 \leq i < j \leq N} \beta_U(X_i, X_j; F) + R_{UN}, $$

where $\theta_U = \mathbb{E}(U_N)$, $\alpha_U(x; F) = m \mathbb{E}\{h(x, X_2, \ldots, X_m)\} - \theta_U$,

$$ \beta_U(x, y; F) = (m - 1) \left[ m \mathbb{E}\{h(x, y, X_3, \ldots, X_m)\} - \alpha_U(x; F) - \alpha_U(y; F) - m\theta_U \right], $$

$R_{UN}$ is the remainder term satisfies $\mathbb{E}(R_{UN}) = 0$ and $\text{Var}(R_{UN}) = O(N^{-3})$ under the condition that $\mathbb{E}(h^2) < \infty$. It is clear that $U_N$ can be represented in the form of (2.1). This representation is called the Hoeffding’s decomposition of U-statistics and it is initially from Hoeffding (1948, 1961).

If we need $s$ steps to calculate the $h$ function, then the computing cost of $U_N$ is $(s + 1)\binom{N}{m}$, which is at the order of $O(N^m)$. Assume the entire dataset of size $N$ is divided evenly into $K$ subsets with each of size $n = N/K$, denote $U_{N,K}^{(k)}$ as the U-statistic using data in the $k$-th block. Then the distributed U-statistic is $U_{U,K} = K^{-1} \sum_{k=1}^{K} U_{N,K}^{(k)}$. Notice that the number of computing steps for $U_{N,K}$ is $K \times (s + 1)\binom{N}{m}$, which is of order $O(K^{1-m}N^m)$. Thus, the computing steps of the distributed statistic is a factor of $K^{m-1}$ less than that of the full sample U-statistic $U_N$. 
Now we consider the problem of selecting $K$. As we discussed earlier, using a larger $K$ would reduce the computing burden and make computation more feasible. However, the mean square error of $T_{N,K}$ is an increasing function of $K$, which indicates that there would be a loss in the estimation efficiency for using a large $K$. Thus, in practice, the selection of $K$ should be a result of a compromise between the statistical accuracy and the computational cost and feasibility.

First of all, the storage and memory bottleneck require $K$ to be big enough such that the data points in each data block are manageable. Suppose that $K \geq K_0$ is a hard requirement to meet the capacity of memory and storage. As the mean square error of $T_{N,K}$ is increasing with respect to $K$, so the minimum mean square error we can get is $\text{MSE}(T_{N,K_0})$, which is achieved when $K = K_0$.

This statement is based on the assumption that we have plenty enough time to fulfill all computing procedures. In practice, we might have fixed time budget. In this situation, denote the computing cost as $C(K)$ and the fixed time budget as $C_0$. It is reasonable to assume that the computing cost $C(K)$ is a decreasing function of $K$. Denote $K'_0 = \min\{K | C(K) \leq C_0\}$, then the optimal $K$ is $\min(K_0, K'_0)$. This result suggests us to select the smallest $K$ that meet the limits of memory, storage and computing time budget, to attain the best statistical efficiency.

### 2.5 Asymptotic distribution of $T_{N,K}$

Now we investigate the asymptotic distribution of $T_{N,K}$ and compare it with that of $T_N$. First, we have the following theorem presenting the asymptotic distribution of $T_N$ when $\sigma_\alpha^2 > 0$.

**Theorem 2.2.** Under Condition 2.1 and $\sigma_\alpha^2 > 0$, if $R_N = o_p(N^{-1/2})$, then as $N \to \infty$,

$$N^{1/2}\sigma_\alpha^{-1}(T_N - \theta) \overset{d}{\to} \mathcal{N}(0,1).$$

Here $\mathcal{N}(0,1)$ denotes the standard normal distribution. This theorem is a consequence of the central limit theorem and the assumption that $R_N = o_p(N^{-1/2})$. The requirement that $R_N = o_p(N^{-1/2})$ is easy to achieve and one sufficient assumption for it is $\mathbb{E}(R_N^2) = o(N^{-1})$. Next, we have the following theorem concerning the asymptotic distribution of $T_{N,K}$.
Theorem 2.3. Under Conditions 2.1, 2.2, 2.3 and $\sigma_\alpha^2 > 0$, if $E(R_N) = b_1 N^{-\tau_1} + o(N^{-\tau_1})$, $\text{Var}(R_N) = o(N^{-\tau_2})$ for some $\tau_1 \geq 1$, $\tau_2 \geq 1$, and $K = o\left(N^{1-1/(2\tau_1)}\right)$, then as $N \to \infty$,

$$N^{1/2} \sigma_\alpha^{-1}(T_{N,K} - \theta)^d \to \mathcal{N}(0,1).$$

The condition $K = o\left(N^{1-1/(2\tau_1)}\right)$ indicates that, under Conditions 2.1–2.3 and the assumption that $\sigma_\alpha^2 > 0$, the smaller order $R_N$ is, the higher order $K$ can be so that $T_{N,K}$ can attain the same asymptotic normal distribution as $T_N$. Thus the condition on $K$ is determined by the order of the remainder term $R_N$. The conditions on $K$ coincides with the requirement on $K$ (2.6) in order that $\text{MSE}(T_N)$ and $\text{MSE}(T_{N,K})$ have the same leading order term, and $T_{N,K}$ has the same efficiency as $T_N$.

Furthermore, we can attain the uniform convergence of the distribution of $T_{N,K}$ to that of $T_N$. Before that, we need the following assumption on $\alpha$.

Condition 2.4. There exists a positive constant $\delta \leq 1$ such that $E|\alpha(X_1; F)|^{2+\delta} < \infty$.

This condition is commonly used in obtaining the uniform convergence, for instance in Petrov (1998).

Theorem 2.4. Under Conditions 2.1, 2.2, 2.3, 2.4 and $\sigma_\alpha^2 > 0$, assume $E(R_N) = b_1 N^{-\tau_1} + o(N^{-\tau_1})$ and $\text{Var}(R_N) = o(N^{-\tau_2})$ for some $\tau_1 \geq 1$ and $\tau_2 > 1$, if $K = O\left(N^{r'}\right)$ for a positive constant $r'$ such that $r' < 1 - 1/(2\tau_1)$, then as $N \to \infty$,

$$\sup_{x \in \mathbb{R}} \left| P\left\{N^{1/2} \sigma_\alpha^{-1}(T_{N,K} - \theta) \leq x\right\} - P\left\{N^{1/2} \sigma_\alpha^{-1}(T_N - \theta) \leq x\right\} \right| = o(1).$$

It is noticed that the uniform convergence requires $K = O\left(N^{r'}\right)$ for $0 < r' < 1 - 1/(2\tau_1)$, which is slightly stronger than $K = o\left(N^{1-1/(2\tau_1)}\right)$ assumed in Theorem 2.3.

With all these results, for the case of $\sigma_\alpha^2 > 0$, as long as the number of data blocks $K$ does not diverges too fast, $T_{N,K}$ can maintain the same estimation efficiency as $T_N$ and they have the same asymptotic distribution. Thus $T_{N,K}$ is a good substitute of $T_N$ as an estimator of $\theta$.

For the degenerate case where $\sigma_\alpha^2 = 0$, an increase of order $K$ in the variance has been shown in Theorem 2.1. This means that $T_{N,K}$ can not achieve the same efficiency as $T_N$. We want to explore the asymptotic distribution of $T_{N,K}$ in the degenerate case and compare it with
the one of $T_N$. It is noticed that when $\sigma_\alpha^2 = 0$ and $\sigma_\beta^2 > 0$,

$$T_N - \theta = N^{-2} \sum_{1 \leq i < j \leq N} \beta(X_i, X_j; F) + R_N$$

and

$$T_{N,K} - \theta = N^{-1} \sum_{k=1}^{K} \sum_{1 \leq i < j \leq n_k} \beta(X_{k,i}, X_{k,j}; F) + R_{N,K}.$$  \hspace{1cm} (2.7)

Thus, the leading order term of $T_{N,K} - \theta$ is no longer the same as that of $T_N - \theta$, which leads to an order of $K$ increase in the variance as shown in Theorem 2.1. Before investigating the asymptotic distributions of $T_N$ and $T_{N,K}$, we introduce the notations in the spectral decomposition of $\beta$.

As $\beta$ is a symmetric function of two variables, when it has finite second moment, there exist sequences of eigenvalues $\{\lambda_\ell\}_{\ell=1}^\infty$ and eigenfunctions $\{\beta_\ell\}_{\ell=1}^\infty$ in connection with $\beta$ such that $\beta$ admits the following expansion (Dunford and Schwartz, 1963; Serfling, 1980)

$$\beta(x, y; F) = \sum_{\ell=1}^\infty \lambda_\ell \beta_\ell(x; F) \beta_\ell(y; F),$$  \hspace{1cm} (2.8)

such that

$$\lim_{L \to \infty} E \left\{ \left| \beta(X_1, X_2; F) - \sum_{\ell=1}^L \lambda_\ell \beta_\ell(X_1; F) \beta_\ell(X_2; F) \right|^2 \right\} = 0.$$  

Here $\beta_\ell$ satisfies $E \{\beta_\ell(X_1; F)\} = 0$ for all $\ell \in \mathbb{Z}$ and

$$E \{\beta_{\ell_1}(X_1; F)\beta_{\ell_2}(X_1; F)\} = \begin{cases} 1 & \ell_1 = \ell_2, \\ 0 & \ell_1 \neq \ell_2. \end{cases}$$  

Moreover, $\{\lambda_\ell\}_{\ell=1}^\infty$ satisfy $\sum_{\ell=1}^\infty \lambda_\ell^2 = \sigma_\beta^2 < \infty$. The following theorem establishes the asymptotic distribution of $T_N$ under the degenerate case.

**Theorem 2.5.** Under Condition 2.1, $\sigma_\alpha^2 = 0$ and $\sigma_\beta^2 > 0$, if $R_N = o_p(N^{-1})$, then as $N \to \infty$,

$$2N(T_N - \theta) \xrightarrow{d} \sum_{\ell=1}^\infty \lambda_\ell (\chi_{1\ell}^2 - 1),$$

where $\{\chi_{1\ell}^2\}_{\ell=1}^\infty$ are independent chi-square random variables with one degree of freedom.
This theorem is a direct result of the asymptotic distribution of degenerate U-statistics (Serfling, 1980) and the condition that $R_N = o_p(N^{-1})$.

Note that from (2.7), $T_{N,K}$ can be viewed as a weighted average of $K$ independent statistics, the asymptotic behavior of degenerate $T_{N,K}$ is given in the following theorem.

**Theorem 2.6.** Under Conditions 2.1, 2.2, $\sigma_\alpha^2 = 0$ and $\sigma_\beta^2 > 0$, then

(i) If $K$ is finite and $R_{N,K} = o_p(N^{-1})$, then as $N \to \infty$,

$$2N(T_{N,K} - \theta) \xrightarrow{d} \sum_{\ell=1}^{\infty} \lambda_\ell (\chi_{K\ell}^2 - K),$$

where $\{\chi_{K\ell}^2\}_{\ell=1}^{\infty}$ are independent chi-square random variables with $K$ degrees of freedom.

(ii) If $K \to \infty$, there exists a constant $\delta' > 0$ such that $E|\beta(X_1,X_2;F)|^{2+\delta'} < \infty$, and $R_{N,K} = o_p(K^{1/2}N^{-1})$, then as $N \to \infty$,

$$2^{1/2} K^{-1/2} N \sigma_\beta^{-1} (T_{N,K} - \theta) \xrightarrow{d} N(0,1).$$

From this theorem, we get two different types of limit distribution for $T_{N,K}$ depending on whether $K$ is finite or diverging. This is easy to interpret because $T_{N,K}$ may be viewed as an average of $K$ independent random variables. So with the data divided into increasingly many data blocks, the distribution of $T_{N,K}$ is asymptotically normal. When $K$ is finite, the condition $R_{N,K} = o_p(N^{-1})$ is a directly result of $R^{(k)}_{N,K} = o_p(n_k^{-1})$. When $K$ diverges to infinity, under Condition 2.3 which prescribes the divisibility of $R_N$ with respect to the sample size $N$, so that $E(R_{N,K}) = N^{-1} \sum_{k=1}^{K} b_{1,k} n_k^{1-\tau_1} + o(K^{\tau_1}N^{-\tau_1})$ and $\text{Var}(R_{N,K}) = o(KN^{-2})$. Thus in order that $R_{N,K} = o_p(K^{1/2}N^{-1})$, it requires that $K^{-1} N^{1-\tau_1} = o(K^{1/2}N^{-1})$, that is,

$$K = o\left(N^{1-1/(2\tau_1-1)}\right).$$

If $\tau_1 = 1$ with $b_1 \neq 0$, $E(R_{N,K}) = O(KN^{-1})$ which means that $R_{N,K} = o_p(K^{1/2}N^{-1})$ is not achievable in this case. It is of interest to mention that when $K \to \infty$, the normalizing factor is of order $K^{-1/2}N$, which is of smaller order of $N$, the normalizing factor for $T_N$. This indicates that the distributed statistic $T_{N,K}$ is less efficient than $T_N$, and this property coincides with the increase in the variance of $T_{N,K}$ as discussed in Theorem 2.1.
2.6 Bootstrap procedures

We have studied the asymptotic properties of $T_{N,K}$ by focusing on its mean square error and asymptotic distributions in the previous sections. An important remaining issue is how to approximate the distribution of $T_{N,K}$ under different conditions on $\sigma^2$ and $K$? This motivates the bootstrap. The bootstrap is a powerful tool for statistical analysis, especially in approximating distributions of statistics. First, we investigate the properties of the conventional bootstrap when applied to the distributed statistics for massive data. Then we review two novel bootstrap algorithms, the BLB and the SDB, designed for massive data. After recognizing the limitations of the BLB and the SDB, we propose two distributed bootstrap procedures that adapt the distributed statistics $T_{N,K}$ under the massive data scenario.

2.6.1 Entire sample bootstrap

In order to approximate the distribution of $T_{N,K}$, we consider resampling the entire sample to construct bootstrap resamples by following the procedures of the conventional bootstrap. By calculating the corresponding distributed statistic for each resample, we can use the empirical distribution of the bootstrap distributed statistics to estimate the distribution of $T_{N,K}$.

As $T_{N,K}$ is a distributed type statistic and the sample size $N$ is large, in order to construct resamples of size $N$ and get the bootstrap version of $T_{N,K}$, one may carry out the resampling procedure in the following two steps:

**Step 1.** For $k = 1, \ldots, K$, randomly sample a set $X_{N,K}^{(k)} = \{X_{k,1}^{*}, \ldots, X_{k,n_k}^{*}\}$ from $X_N$ with replacement and compute the corresponding statistic as $T_{N,K}^{*(k)} = T(X_{N,K}^{(k)})$.

**Step 2.** Obtain the bootstrap distributed statistic

$$T_{N,K}^* = N^{-1} \sum_{k=1}^{K} n_k^* T_{N,K}^{*(k)}.$$

By carrying out this two steps resampling procedure, it ensures that the resample $X_N^* = \{X_{N,K}^{*(1)}, \ldots, X_{N,K}^{*(K)}\}$ are conditionally independently and identically distributed from the empirical distribution of $X_N$, denoted as $F_N$, and they are split into $K$ subsets automatically. So the storage space requirement of the resampling procedure is scale in $K^{-1}N$, rather than $N$ and it is easy to be paralleled.
Next, we explore the asymptotic properties of the entire sample bootstrap procedure. To facilitate the working of the bootstrap procedure, we need $T_{N,K}^*$ to admit an expansion similar to $T_{N,K}$ in (2.3). Suppose that for $k = 1, \ldots, K$, $T_{N,K}^{(k)} = T(X_{N,K}^*)$ has the following form

$$T_{N,K}^{(k)} = \theta_{N}^{*} + n_{k}^{-1} \sum_{i=1}^{n_{k}} \alpha(X_{k,i}^{*}; F_{N}) + n_{k}^{-2} \sum_{1 \leq i < j \leq n_{k}} \beta(X_{k,i}^{*}, X_{k,j}^{*}; F_{N}) + R_{N,K}^{(k)},$$

where $\theta_{N}^{*} = \theta(F_{N})$, $R_{N,K}^{(k)} = R(X_{N,K}^{*}; F_{N})$ are the analog of $\theta$ and $R^{(k)}$ under $F_{N}$. Then

$$T_{N,K}^{*} = \theta_{N}^{*} + N^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_{k}} \alpha(X_{k,i}^{*}; F_{N}) + N^{-1} \sum_{k=1}^{K} n_{k}^{-1} \sum_{1 \leq i < j \leq n_{k}} \beta(X_{k,i}^{*}, X_{k,j}^{*}; F_{N}) + R_{N,K}^{*},$$

(2.9)

where $R_{N,K}^{*} = N^{-1} \sum_{k=1}^{K} n_{k} R_{N,K}^{(k)}$. Here, $\alpha(x; F_{N})$ and $\beta(x, y; F_{N})$ can be viewed as the empirical version of $\alpha(x; F)$ and $\beta(x, y; F)$. Denote $S(F)$ as the support of $F$, we need to regulate them as in the following condition.

**Condition 2.5.** $\alpha(x; F_{N})$ and $\beta(x, y; F_{N})$ are functions depend on $F_{N}$. $\alpha(x; F_{N})$ satisfies $\sum_{i=1}^{N} \alpha(X_{i}; F_{N}) = 0$ and $\sup_{x \in S(F)} |\alpha(x; F_{N}) - \alpha(x; F)| = o_{p}(1)$, and $\beta(x, y; F_{N})$ is symmetric in $x$ and $y$ such that $\sum_{i=1}^{N} \{\beta(X_{i}, y; F_{N})\} = 0$ for any $y \in S(F)$ and $\sup_{x,y \in S(F)} |\beta(x, y; F_{N}) - \beta(x, y; F)| = o_{p}(1)$.

Condition 2.5 indicates that for $X_{k,i}^{*}$ with distribution $F_{N}$, $E\{\alpha(X_{k,i}^{*}; F_{N})|F_{N}\} = 0$ and $E\{\beta(X_{k,i}^{*}, y; F_{N})|F_{N}\} = 0$ for any $y \in S(F)$. Moreover, it requires that $\alpha(x; F_{N})$ and $\beta(x, y; F_{N})$ are close to $\alpha(x; F)$ and $\beta(x, y; F)$. Explicit discussions on these conditions will be given shortly in the context of U-statistics.

The following theorem establishes the theoretical properties of the bootstrap distributed statistic which is inspired by Lai and Wang (1993). For specific statistics, we need to check the bootstrap distributed statistic admits the form (2.9) and satisfies other assumptions assumed in the theorem to get the corresponding results.

**Theorem 2.7.** Under the conditions of Theorem 2.4 and $E|\beta(X_{1}, X_{1}; F)|^2 < \infty$, suppose that there exist functions $\alpha(x; F_{N})$ and $\beta(x, y; F_{N})$ satisfy Condition 2.5, let

$$T_{N,K} = \theta_{N} + N^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_{k}} \alpha(X_{k,i}^{*}; F_{N}) + N^{-1} \sum_{k=1}^{K} n_{k}^{-1} \sum_{1 \leq i < j \leq n_{k}} \beta(X_{k,i}^{*}, X_{k,j}^{*}; F_{N}) + R_{N,K}^{*},$$

(2.10)
where \( R_{N,K}^* \) satisfies \( P\{|R_{N,K}^*| \geq N^{-1/2}(lnN)^{-1}|F_N\} = o_p(1) \), then as \( N \to \infty \),

\[
\sup_{x \in \mathbb{R}} \left| P\left\{ N^{1/2}\hat{\sigma}_{\alpha,N}^{-1}(T_{N,K}^* - \theta_N^*) \leq x \right| F_N \right\} - P\left\{ N^{1/2}\sigma_{\alpha}^{-1}(T_{N,K} - \theta) \leq x \right\} = o_p(1),
\]

and \( \hat{\sigma}_{\alpha,N}^2 - \sigma_{\alpha}^2 = o_p(1) \) where \( \hat{\sigma}_{\alpha,N}^2 = E\{\alpha^2(X_{1,i}^*; F_N)|F_N\} \).

This theorem offers theoretical support for the entire sample bootstrap method. Repeat step 1 to 2 a number \( B \) of times, we get \( B \) bootstrap distributed statistics, denote them as \( T_{N,K}^{*1}, \ldots, T_{N,K}^{*B} \). Then, the empirical distribution of \( \{ N^{1/2}\hat{\sigma}_{\alpha,N}^{-1}(T_{N,K}^{*b} - \theta_N^*) \}, b = 1, \ldots, B \) can be used to approximate the distribution of \( N^{1/2}\sigma_{\alpha}^{-1}(T_{N,K} - \theta) \).

### 2.6.2 U-statistics as an example

We would like to use U-statistics as an example to demonstrate the conditions in Theorem 2.7 are appropriate. Suppose \( U_N = U(\mathcal{X}_N) \) is a U-statistic of degree 2 with a symmetric kernel function \( h \), that is,

\[
U_N = 2\{N(N - 1)\}^{-1} \sum_{1 \leq i < j \leq N} h(X_i, X_j)
\]

\[
= \theta_U + N^{-1} \sum_{i=1}^{N} \alpha_U(X_i; F) + N^{-2} \sum_{1 \leq i < j \leq N} \beta_U(X_i, X_j; F) + R_{UN}, \tag{2.10}
\]

where \( \theta_U = E(U_N) \), \( \alpha_U(x; F) = 2[E\{h(x, X_2)\} - \theta_U] \), \( \beta_U(x, y; F) = 2h(x, y) - \alpha_U(x; F) - \alpha_U(y; F) - 2\theta_U \), and \( R_{UN} \) is the remainder term. Obviously \( U_N \) is in the form of (2.1). Under the condition that \( E\{h(X_1, X_2)\}^2 < \infty \), it can be shown that \( \alpha_U \) and \( \beta_U \) meet Condition 2.1 and \( R_{UN} \) satisfies \( E(R_{UN}) = 0 \) and \( E(R_{UN}^2) = O(N^{-4}) \) (Hoeffding, 1948; Serfling, 1980). Let \( U_{N,K}^{(k)} = U(\mathcal{X}_{N,K}^{(k)}) \) be the corresponding U-statistic obtained from the \( k \)-th data block, for \( k = 1, \ldots, K \), then the distributed U-statistic is

\[
U_{N,K} = N^{-1} \sum_{k=1}^{K} n_k^{-1} U_{N,K}^{(k)}
\]

\[
= \theta_U + N^{-1} \sum_{i=1}^{N} \alpha_U(X_i; F) + N^{-1} \sum_{k=1}^{K} n_k^{-1} \sum_{1 \leq i < j \leq n_k} \beta_U(X_{k,i}, X_{k,j}; F) + R_{UN,K},
\]

where \( R_{UN,K} \) satisfies \( E(R_{UN,K}) = 0 \) and \( E\left( R_{UN,K}^2 \right) = O(K^3N^{-4}) \) under Condition 2.2.

According to Theorem 2.4, if \( E|\alpha_U(X_1; F)|^{2+\delta} < \infty \) and \( K = O(N^{\tau'}) \) for some positive \( \delta \) and
$$\tau' < 1,$$ then as \( N \to \infty,$$ 

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left\{ N^{1/2} \sigma_{U}^{-1}(U_N - \theta_U) \leq x \right\} - \mathbb{P} \left\{ N^{1/2} \sigma_{U}^{-1}(U_N - \theta_U) \leq x \right\} \right| = o(1).$$

Now we consider the bootstrap version of the distributed U-statistic. Let \( \mathbf{x}^{(1)}_{N,K}, \ldots, \mathbf{x}^{(K)}_{N,K} \) be an i.i.d. sample from the empirical distribution \( F_N \) and \( U^{(k)}_{N,K} = U(\mathbf{x}^{(k)}_{N,K}) \) be the corresponding U-statistic based on \( \mathbf{x}^{(k)}_{N,K} \), for \( k = 1, \ldots, K \), then

$$U^{(k)}_{N,K} = 2\{n_k(n_k - 1)\}^{-1} \sum_{1 \leq i < j \leq n_k} h(X_{k,i}^*, X_{k,j}^*)$$

$$= \theta_{U,N}^* + n_k^{-1} \sum_{i=1}^{n_k} \alpha_U(X_{k,i}^*; F_N) + n_k^{-2} \sum_{1 \leq i < j \leq n_k} \beta_U(X_{k,i}^*, X_{k,j}^*; F_N) + R_{U,N,K}^*,$$

where \( \theta_{U,N}^* = \mathbb{E}(U^{(k)}_{N,K} | F_N) = N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} h(X_i, X_j), \)

$$\alpha_U(x; F_N) = 2 \left[ \mathbb{E} \left\{ h(x, X_{k,2}^*) | F_N \right\} - \theta_{U,N}^* \right],$$

and

$$\beta_U(x, y; F_N) = 2h(x, y) - \alpha_U(x; F_N) - \alpha_U(y; F_N) - 2\theta_{U,N}^*.$$ 

Thus, the bootstrap distributed statistic \( U_{N,K}^* = N^{-1} \sum_{k=1}^{K} n_k U_{N,K}^{(k)} \) is in the form of (2.9) with reminder term \( R_{U,N,K}^* = N^{-1} \sum_{k=1}^{K} n_k R_{U,N,K}^{(k)} \). Represent \( \alpha_U(x; F_N) \) and \( \beta_U(x, y; F_N) \) in term of \( \alpha_U(x; F) \) and \( \beta_U(x, y; F) \), we have

$$\alpha_U(x; F_N) = \alpha_U(x; F) - N^{-1} \sum_{i=1}^{N} \alpha_U(X_i; F) + N^{-1} \sum_{i=1}^{N} \beta_U(x, X_i; F)$$

$$- N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_U(X_i, X_j; F),$$

and

$$\beta_U(x, y; F_N) = \beta_U(x, y; F) - N^{-1} \sum_{i=1}^{N} \beta_U(x, X_i; F) - N^{-1} \sum_{i=1}^{N} \beta_U(X_i, y; F)$$

$$+ N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_U(X_i, X_j; F).$$

By simple algebra, it is checked that \( \sum_{i=1}^{N} \alpha_U(X_i; F_N) = 0, \beta_U(x, y; F_N) \) is symmetric in \( x \) and \( y \), and \( \sum_{i=1}^{N} \{ \beta_U(X_i, y; F_N) \} = 0 \) for any \( y \in S(F) \). Furthermore, according to the condition \( \mathbb{E}(h(X_1, X_2))^2 < \infty \), \( \sup_{x \in S(F)} |\alpha_U(x; F_N) - \alpha_U(x; F)| = o_p(1) \) and \( \sup_{x,y \in S(F)} |\beta_U(x, y; F_N) - \beta_U(x, y; F)| = o_p(1) \) are easily verifiable by law of large numbers.
Note that $R^{∗(k)}_{U,N,K} = 2n_k^{-2}(n_k - 1)^{-1}\sum_{1 \leq i < j \leq n_k} \beta_U(X_{k,i}^*, X_{k,j}^*; F_N)$, $E(R^{∗(k)}_{U,N,K}|F_N) = 0$ and $E(R^{∗(k)2}_{U,N,K}|F_N) = O_p(n_k^{-4})$, from which
\[P\left\{ |R^{∗}_{U,N,K}| \geq N^{-1/2}(lnN)^{-1}|F_N \right\} = o_p(1)\]
can be verified by Cauchy-Schwarz inequality and the condition that $K = O(N^{\tau'})$ for a positive $\tau' < 1$. Thus the conditions in Theorem 2.7 are all satisfied by the U-statistic. Finally, let $\hat{\sigma}^2_{\alpha,UN} = E\left\{ \alpha^2 U(X_1^*, X_1; F_N) | F_N \right\}$, if $E\{h(X_1, X_1)\} < \infty$, Theorem 2.7 implies that
\[
\sup_{x \in \mathbb{R}} P\left\{ N^{1/2} \hat{\sigma}^{-1}_{\alpha,UN}(U_{N,K}^* - \theta_{UN}) \leq x \mid F_N \right\} - P\left\{ N^{1/2} \sigma^{-1}_{\alpha,U}(U_{N,K} - \theta_U) \leq x \right\} = o_p(1),
\]
and $\hat{\sigma}^2_{\alpha,UN} - \sigma^2_{\alpha,U} = o_p(1)$.

Even the entire sample bootstrap enjoys nice theoretical properties, it has several limitations when applied to massive data. Data communication expenditure is the most serious issue. Resampling from the whole dataset $X_N$ is expensive for huge datasets. Moreover, the whole data may be stored in different locations, which makes the entire sample bootstrap require data communication among different data clusters at different locations. In addition, $\theta_{N}^* = \theta(F_N)$ has the same computational complexity as $T_N$. For instance, for the U-statistics discussed above, $\theta_{N}^* = N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} h(X_i, X_j)$ which is computationally expensive for massive data.

### 2.6.3 BLB and SDB

The bag of little bootstrap (Kleiner, et al., 2014) and the subsampled double bootstrap (Sengupta, Volgushev and Shao, 2016) are two novel resampling-based methods for massive data with emphasis on accessing the quality of estimators. Suppose $\hat{\theta}_N = \hat{\theta}_N(F_N)$ is an estimator of $\theta = \theta(F)$ based on sample $X_N = \{X_1, \ldots, X_N\}$. Let $u(\hat{\theta}_N, \theta)$ be some quantity concerning $\hat{\theta}_N$, and $Q_N(F)$ be its sampling distribution, which is unknown because it depends on the underlying distribution $F$. We are interested in an assessment $\xi\{Q_N(F)\}$ of $u(\theta_N, \theta)$, regarding an aspect of $Q_N(F)$. For example, if $u(\hat{\theta}_N, \theta) = \hat{\theta}_N - \theta$ is the estimation error of $\hat{\theta}_N$, $\xi\{Q_N(F)\}$ is the expectation of $u(\hat{\theta}_N, \theta)$ representing the bias of $\hat{\theta}_N$. Similarly, $\xi\{Q_N(F)\}$ can define the mean square error of $\hat{\theta}_N$ or a confidence interval of $\theta$.

Resampling methods like the bootstrap can be used to estimate $\xi\{Q_N(F)\}$. The basic idea of the bootstrap is, first estimate the unknown sampling distribution $Q_N(F)$ by an empirical
distribution, which can be derived by resampling the observed data, then estimate \( \xi \{ Q_N(F) \} \) by plug in the empirical distribution as an estimate of \( Q_N(F) \). For the conventional bootstrap, by resampling on the entire data, we can obtain a number (\( B \)) of resamples, denoted as \( \mathcal{X}_N^{(b)} = \{ X_1^{(b)}, \ldots, X_N^{(b)} \}, \ b = 1, \ldots, B \). For each resample, we calculate \( u(\hat{\theta}_N^{(b)}, \hat{\theta}_N) \) where \( \hat{\theta}_N^{(b)} = \hat{\theta}_N(F_N^{(b)}) \) with \( F_N^{(b)} \) be the empirical distribution of \( \mathcal{X}_N^{(b)} \). Then the empirical distribution of \( \{ u(\hat{\theta}_N^{(b)}, \hat{\theta}_N), \ b = 1, \ldots, B \} \), denoted as \( Q_N^* \), can be used to estimate \( Q_N(F) \). Finally, \( \xi \{ Q_N^* \} \) is used to estimate \( \xi \{ Q_N(F) \} \).

For massive data, calculating \( \hat{\theta}_N \) and \( \hat{\theta}_N^{(b)} \) for each resample in the conventional bootstrap can be computationally expansive. Kleiner, et al. (2014) proposed the bags of little bootstrap (BLB) to obtain an estimate of \( \xi \{ Q_N(F) \} \) for massive datasets. To avoid calculating the estimators \( \hat{\theta}_N \) and \( \hat{\theta}_N^{(b)} \) directly, the first step of the BLB is randomly sampling subsets of small size from the entire data. In the BLB, \( S \) subsets of size \( n \) (chosen to be \( n = N^\varsigma \) for some \( 0 < \varsigma < 1 \)) are sampled from the original dataset \( \mathcal{X}_N \) without replacement (these subsets can be disjoint predefined subsets of \( \mathcal{X}_N \)). Denote the \( s \)-th subset as \( \mathcal{X}_{s,n} = \{ X_{s,1}, \ldots, X_{s,n} \} \) and its empirical distribution as \( F_{s,n} \), for \( s = 1, \ldots, S \). Then calculating the corresponding estimator \( \hat{\theta}_{s,n} = \hat{\theta}_n(F_{s,n}) \) for each subset is computationally economical.

The next step is constructing estimates of \( Q_N(F) \) for each subset. Instead of using \( Q_n(F_{s,n}) \), the key idea of the BLB is the usage of \( Q_N(F_{s,n}) \), even though \( F_{s,n} \) has its support on only \( n \) points (Kleiner, et al., 2014). Then the BLB’s estimate of \( \xi \{ Q_N(F) \} \) is \( S^{-1} \sum_{s=1}^S \xi \{ Q_N(F_{s,n}) \} \).

For each subset \( \mathcal{X}_{s,n} \), in order to approximate \( \xi \{ Q_N(F_{s,n}) \} \), we generate \( B \) resamples of size \( N \) from the subset and denote them as \( \mathcal{X}_{s,n}^{(b)} = \{ X_{s,1}^{(b)}, \ldots, X_{s,n}^{(b)} \}, \ b = 1, \ldots, B \). Actually, in order to generate the resample \( \mathcal{X}_{s,n}^{(b)} \), it is equivalent to draw a multinomial random vector with index \( N \) and parameter \( n^{-1} \mathbf{1}_n \), denoted as \( \mathbf{W}_{s,n}^{(b)} = (w_{s,1}^{(b)}, \ldots, w_{s,n}^{(b)}) \), which represents the frequencies for each datum in \( \mathcal{X}_{s,n} \) in the resample. Thus \( \mathcal{X}_{s,n}^{(b)} \) can be represented as \( (\mathcal{X}_{s,n}, \mathbf{W}_{s,n}^{(b)}) \), for \( s = 1, \ldots, S \) and \( b = 1, \ldots, B \).

Let \( F_{s,n}^{(b)} \) be the empirical distribution of the resample \( (\mathcal{X}_{s,n}, \mathbf{W}_{s,n}^{(b)}) \). Calculate the estimator \( \hat{\theta}_{s,n}^{(b)} = \hat{\theta}_N(F_{s,n}^{(b)}) \) and compute \( u(\hat{\theta}_{s,n}^{(b)}, \hat{\theta}_{s,n}) \) for each resample. Then for each subset \( \mathcal{X}_{s,n} \), the empirical distribution of \( \{ u(\hat{\theta}_{s,n}^{(b)}, \hat{\theta}_{s,n}), \ b = 1, \ldots, B \} \), denoted as \( Q_N^*(F_{s,n}) \), is used to estimate \( Q_N(F_{s,n}) \). Finally, the estimate of \( \xi \{ Q_N(F) \} \) is obtained by averaging \( \xi \{ Q_N^*(F_{s,n}) \} \).
over $S$ different subsets, that is $S^{-1} \sum_{s=1}^{S} \xi \{Q_{N}^{s}(F_{s,n})\}$. A diagram summary of the BLB procedure is shown in Figure 2.1. Kleiner, et al. (2014) proved that under some conditions, $S^{-1} \sum_{s=1}^{S} \xi \{Q_{N}^{s}(F_{s,n})\}$ convergence to $\xi \{Q_{N}(F)\}$ in probability as $n$ goes into infinity.

![Figure 2.1 A diagram summary of the BLB procedure.](image)

Sengupta, Volgoshev and Shao (2016) proposed a method based on the idea of the BLB and a fast double bootstrap (Davidson and MacKinnon, 2002; Chang and Hall, 2015), called the subsampled double bootstrap (SDB). For the SDB, a large number ($S$) of random subsets of size $n$ are sampled from the original dataset, which is similar as in the BLB. Then for each subset $X_{s,n}^{*} = \{X_{s,1}^{*}, \ldots, X_{s,n}^{*}\}$, instead of using $B$ resamples, only one resample $X_{s,n}^{(1)} = \{X_{s,1}^{(1)}, \ldots, X_{s,n}^{(1)}\}$ is generated in the SDB. So for each subset, only one $u(\hat{\theta}_{s,n}^{(1)}, \hat{\theta}_{s,n}^{*})$ is obtained, here $\hat{\theta}_{s,n} = \hat{\theta}_{N}(F_{s,n})$ and $\hat{\theta}_{s,n}^{*} = \hat{\theta}_{N}(F_{s,n}^{(1)})$, where $F_{s,n}^{*}$ and $F_{s,n}^{(1)}$ are the empirical distributions of $X_{s,n}^{*}$ and $X_{s,n}^{(1)}$, respectively. Then the empirical distribution of $\{u(\hat{\theta}_{s,n}^{(1)}, \hat{\theta}_{s,n}^{*}), s = 1, \ldots, S\}$, denoted as $Q_{N,n,S}^{*}$, is used to estimate $Q_{N}(F)$, and $\xi \{Q_{N,n,S}^{*}\}$ is the SDB’s estimate of $\xi \{Q_{N}(F)\}$.
The key idea of the BLB and the SDB is generating resamples with sizes equal to the size of the original data by resampling a small subset, and this can be done by sampling a multinomial distributed random vector \((W_{s,n}^{*})\) of length \(n\). The BLB and the SDB have computational advantages only when the estimator can be calculated directly with the weighted data representation \((X_{s,n}, W_{s,n}^{*})\), for example, general M-estimators. In those cases, the computational requirement of calculating \(\hat{\theta}_{s,n}^{*}\) is scale in \(n\), rather than \(N\).

There are several differences between the BLB and the SDB. First, the number of subsets \(S\) in the BLB can be either finite or infinite, while for the SDB, \(S\) needs to diverge to infinity. In the BLB, we get an estimate of \(\xi\{Q_N(F)\}\) for each subset by generating \(B\) resamples. In contrast, only one \(u(\hat{\theta}_{s,n}^{*(1)}, \hat{\theta}_{s,n}^{*})\) is derived from the each subset \(X_{s,n}^{*}\) in the SDB and the estimate of \(\xi\{Q_N(F)\}\) is based on these \(X_{s,n}^{*}\)'s across the subsets. Thus the SDB requires resampling from the entire dataset \(X_N\) to get the random subsets \(\{X_{s,n}^{*}, s = 1, \ldots, S\}\). For massive data that stored in different locations, resampling from the entire dataset in the SDB would reduce its computational appeal. The BLB can be applied with the subsets \(\{X_{s,n}^{*}, s = 1, \ldots, S\}\) chosen as each data block and the following procedures can all be done within each data block. Data communication are only needed for the final average step which is cheap and convenient. So the BLB can be implemented distributively and thus is suitable for massive data that stored in different locations.

There are several limitations of the BLB, as mentioned in Sengupta, Volgushev and Shao (2016). First, Kleiner, et al. (2014) suggests using a small number of subsets (small \(S\)) but a large number of resamples for each subset (big \(B\)), which leads to that only a small portion of the whole data is used to get the BLB estimator. Second, using too many resamples for each subset can increase the computational burden. Last, it is unclear how to choose \(S\) and \(R\) for the BLB to achieve optimal statistical efficiency. To overcome these limitations, we propose a distributed bootstrap method which does not require communication among different data blocks.
2.7 Distributed bootstrap

We are interested in approximating the distribution of \( N^{1/2}(T_{N,K} - \theta) \) for massive datasets, especially when the data are stored in different locations. In this case, resampling from the entire dataset could be computationally costly. By recognizing the natural formation of the distributed statistics, we consider doing resampling within each data block.

Suppose the entire data have already been divided into \( K \) subsets: \( \mathcal{X}^{(1)}_{N,K}, \ldots, \mathcal{X}^{(K)}_{N,K} \). For each subset \( \mathcal{X}^{(k)}_{N,K} \), we generate \( B \) resamples of the same size \( n_k \), denote them \( \mathcal{X}^{*1(k)}_{N,K}, \ldots, \mathcal{X}^{*B(k)}_{N,K} \).

Compute the corresponding statistic \( T^{*b(k)}_{N,K} = T(\mathcal{X}^{*b(k)}_{N,K}) \) for each resampled data block, for \( k = 1, \ldots, K \). Average them over \( K \) subsets leads to

\[
T^{*b}_{N,K} = N^{-1} \sum_{k=1}^{K} n_k T^{*b(k)}_{N,K}
\]

for \( b = 1, \ldots, B \). Let \( F^{(k)}_{N,K} \) be the empirical distribution of \( \mathcal{X}^{(k)}_{N,K} \) and \( \theta^{*(k)}_{N,K} = \theta(F^{(k)}_{N,K}) \) be the analogy of \( \theta \) under \( \mathcal{X}^{(k)}_{N,K} \). Define \( \theta^{*}_{N,K} = N^{-1} \sum_{k=1}^{K} n_k \theta^{*(k)}_{N,K} \), then the empirical distribution of \( \{N^{1/2}(T^{*b}_{N,K} - \theta^{*}_{N,K}), b = 1, \ldots, B\} \) is used to approximate the distribution of \( N^{1/2}(T_{N,K} - \theta) \).

The algorithm is outlined in Table 2.1.

**Table 2.1 Algorithm for the distributed bootstrap.**

<table>
<thead>
<tr>
<th>Input: data ( \mathcal{X}^{(1)}<em>{N,K}, \ldots, \mathcal{X}^{(K)}</em>{N,K} ); ( n_1, \ldots, n_K ), subset sizes; ( B ), number of Monte Carlo iterations; ( T ), function deriving statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output: an estimate of the distribution of ( N^{-1/2}(T_{N,K} - \theta) )</td>
</tr>
</tbody>
</table>

For \( k \leftarrow 1 \) to \( K \) do

- compute \( \theta^{*(k)}_{N,K} = \theta(F^{(k)}_{N,K}) \)
- for \( b \leftarrow 1 \) to \( B \) do
  - generate resample \( \mathcal{X}^{*b(k)}_{N,K} \) from \( \mathcal{X}^{(k)}_{N,K} \)
  - compute \( T^{*b(k)}_{N,K} = T(\mathcal{X}^{*b(k)}_{N,K}) \)
- end

End

Compute \( \theta^{*}_{N,K} = N^{-1} \sum_{k=1}^{K} n_k \theta^{*(k)}_{N,K} \)

For \( b \leftarrow 1 \) to \( B \) do

- compute \( T^{*b}_{N,K} = N^{-1} \sum_{k=1}^{K} n_k T^{*b(m)}_{N,K} \) and \( N^{1/2}(T^{*b}_{N,K} - \theta^{*}_{N,K}) \)

End
We call this algorithm the distributed bootstrap because all the sampling procedures can be fulfilled within each data block. For each subset in each iteration, we can calculate $T_{N,K}^{*h(k)}$ and $\theta_{N,K}^{*h(k)}$ totally locally and avoid data communication between different data blocks. These make the distributed bootstrap suitable for the massive data setting.

Besides the advantages in computing, we want to investigate the asymptotic properties of the distributed bootstrap. Similar as the technique used in establishing theoretical properties of the entire sample bootstrap, we need to make assumptions on the statistics derived from each resampled data block. For $k = 1, \ldots, K$, denote $F_{N,K}^{(k)}$ as the empirical distribution of $X_{N,K}^{(k)}$; let $X_{N,K}^{(k)} = \{X_{k,1}^{*}, \ldots, X_{k,n_k}^{*}\}$ be i.i.d. resamples drawn from $F_{N,K}^{(k)}$. Suppose that for $k = 1, \ldots, K$, $T_{N,K}^{*(k)} = T(X_{N,K}^{(k)})$ has the following form

$$T_{N,K}^{*(k)} = \theta_{N,K}^{*(k)} + n_k \sum_{i=1}^{n_k} \alpha(X_{k,i}; F_{N,K}^{(k)}) + n_k^{-2} \sum_{1 \leq i < j \leq n_k} \beta(X_{k,i}, X_{k,j}; F_{N,K}^{(k)}) + R_{N,K}^{*(k)},$$

where $\theta_{N,K}^{*(k)} = \theta(F_{N,K}^{(k)})$, $R_{N,K}^{*(k)} = R(X_{N,K}^{(k)}; F_{N,K}^{(k)})$ are the analog of $\theta$ and $R_{N,K}^{(k)}$ under $F_{N,K}^{(k)}$. Then $T_{N,K}^{*} = N^{-1} \sum_{k=1}^{K} n_k T_{N,K}^{*(k)}$ can be written as

$$T_{N,K}^{*} = \theta_{N,K}^{*} + N^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \alpha(X_{k,i}; F_{N,K}^{(k)}) + N^{-1} \sum_{k=1}^{K} n_k^{-1} \sum_{1 \leq i < j \leq n_k} \beta(X_{k,i}, X_{k,j}; F_{N,K}^{(k)}) + R_{N,K}^{*},$$

where $\theta_{N,K}^{*} = N^{-1} \sum_{k=1}^{K} n_k \theta_{N,K}^{*(k)}$ and $R_{N,K}^{*} = N^{-1} \sum_{k=1}^{K} n_k R_{N,K}^{*(k)}$.

We have the following assumptions on $\alpha(x; F_{N,K}^{(k)})$ and $\beta(x, y; F_{N,K}^{(k)})$.

**Condition 2.6.** For $k = 1, \ldots, K$, the functions $\alpha(x; F_{N,K}^{(k)})$ and $\beta(x, y; F_{N,K}^{(k)})$, depending on $F_{N,K}^{(k)}$, satisfy $\sum_{i=1}^{n_k} \alpha(X_{k,i}; F_{N,K}^{(k)}) = 0$ and $\sup_{x \in S(F)} |\alpha(x; F_{N,K}^{(k)}) - \alpha(x; F)| = o_p(1)$; $\beta(x, y; F_{N,K}^{(k)})$ is symmetric in $x$ and $y$, $\sum_{i=1}^{n_k} \{\beta(X_{k,i}; F_{N,K}^{(k)})\} = 0$ for any $y \in S(F)$ and

$$\sup_{x, y \in S(F)} |\beta(x, y; F_{N,K}^{(k)}) - \beta(x, y; F)| = o_p(1).$$

The following theorem establishes the consistency of the distributed bootstrap.

**Theorem 2.8.** Under the conditions of Theorem 2.4 and $\sum_{x} |\beta(X_1, X_1; F)|^2 < \infty$, suppose there exist functions $\alpha(x; F_{N,K}^{(k)})$ and $\beta(x, y; F_{N,K}^{(k)})$, $k = 1, \ldots, K$ satisfy Condition 2.6, let

$$T_{N,K}^{*} = \theta_{N,K}^{*} + N^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \alpha(X_{k,i}; F_{N,K}^{(k)}) + N^{-1} \sum_{k=1}^{K} n_k^{-1} \sum_{1 \leq i < j \leq n_k} \beta(X_{k,i}, X_{k,j}; F_{N,K}^{(k)}) + R_{N,K}^{*},$$
where $R_{N,K}^*$ satisfies $P\left\{ |R_{N,K}^*| \geq N^{-1/2}(\ln N)^{-1}\left| F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)} \right| \right\} = o_p(1)$, then as $N \to \infty$,

$$
\sup_{x \in \mathbb{R}} \left| P\left\{ N^{1/2}\hat{\sigma}_{\alpha,N,K}^{-1}(T_{N,K} - \theta_{N,K}^*) \leq x \mid F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)} \right\} \right| \leq P\left\{ N^{1/2}\sigma_{\alpha}^{-1}(T_{N,K} - \theta) \leq x \right\} = o_p(1),
$$

and $\hat{\sigma}_{\alpha,N,K}^2 - \sigma_{\alpha}^2 = o_p(1)$ where $\hat{\sigma}_{\alpha,N,K}^2 = N^{-1} \sum_{k=1}^{K} n_k E \left\{ \alpha^2(X_{k,1}^*; F_{N,K}^{(k)}) \right\}$. 

Theorem 2.8 gives theoretical support for the distributed bootstrap, which works for both finite and infinite $K$, as long as $K$ is at a proper order of $N$. The condition on $K$ is hidden in the assumption $P\left\{ |R_{N,K}^*| \geq N^{-1/2}(\ln N)^{-1}\left| F_{N,K}^{(1)}, \ldots, F_{N,K}^{(K)} \right| \right\} = o_p(1)$. For example, for a U-statistic defined in (2.10), $R_{UN,K}^{(k)}$ satisfies $E \left( R_{UN,K}^{(k)} \mid F_{N,K}^{(k)} \right) = 0$ and $E \left( R_{UN,K}^{(k)2} \mid F_{N,K}^{(k)} \right) = O_p(n_k^{-4})$, for $k = 1, \ldots, K$. Thus the consistency of the distributed bootstrap for the U-statistic is ensured as long as $K = O(N^{\tau'})$ for a positive $\tau' < 1$.

Again, for specific statistics, we need to check the bootstrap distributed statistic satisfies the conditions assumed in Theorem 2.8. It is worth mentioning that the underlying empirical distribution for each subset are different, as we are resampling from each subset, the bootstrap resamples $\{ X_{N,K}^{(1)}, \ldots, X_{N,K}^{(K)} \}$ are not identically distributed. This theorem ensures that the distributed bootstrap procedure can be utilized in a broad range by combining it with the continuous mapping theorem and delta method, for instance in the variance estimation and confidence interval establishing.

Let

$$
\hat{\sigma}_{DB}^2 = B^{-1} \sum_{b=1}^{B} \left( T_{N,K}^{*b} - B^{-1} \sum_{i=1}^{B} T_{N,K}^{*i} \right)^2
$$

be the sample variance of $\{ T_{N,K}^{*b}, b = 1, \ldots, B \}$. Then $\hat{\sigma}_{DB}^2$ is a consistent estimator of $N^{-1}\sigma_{\alpha}^2$, hence it is a consistent estimator of $\text{Var}(T_{N,K})$. In addition, denote $u_{\alpha}^*$ as the lower $\tau$ quantile of the empirical distribution of $\{ N^{1/2}(T_{N,K}^{*b} - \theta_{N,K}^*), b = 1, \ldots, B \}$, then an equal-tail two-sided confidence interval for $\theta$ with level $1 - \tau$ can be constructed as

$$
\left( T_{N,K} - N^{-1/2}u_{1-\tau/2}, T_{N,K} - N^{-1/2}u_{\tau/2} \right).
$$
2.8 Pseudo-distributed bootstrap

Even the distributed bootstrap can save computing time substantially, it can still be computationally expensive when the size of the data is huge. The reason is that the distributed statistic need to be recalculated for each resample in the distributed bootstrap procedure. To avoid this issue, we consider another way for approximating the distribution of $T_{N,K}$ under the assumption that $K$ diverges ($K \to \infty$).

The idea comes from the expression

$$T_{N,K} = N^{-1} \sum_{k=1}^{K} n_k T_{N,K}^{(k)} = K^{-1} \sum_{k=1}^{K} (Kn_k/N) T_{N,K}^{(k)},$$

which indicates that $T_{N,K}$ is the average of $K$ independent random variables. So when $K$ is large enough, approximating the distribution of $T_{N,K}$ is similar to the problem that approximating the distribution of the sample mean of a sequence of independent but not necessary identically distributed samples. In view of this point, if $K \to \infty$, we can propose a computational more efficient bootstrap strategy, which directly resamples $\{T_{N,K}^{(k)}\}_{k=1}^{K}$ rather than the original data.

We state our approach in a more general way, not restricted to $T_{N,K}$ with the form (2.3). Denote $T_{N,K}^{(k)} = T_{X(N,K)}^{(k)}$, $k = 1, \ldots, K$ as a sequence of statistics obtained from each subset. Suppose $E(T_{N,K}^{(k)}) = \theta_k$ and $\text{Var}(T_{N,K}^{(k)}) = \sigma^2_k$ where $\sigma^2_k \in (0, \infty)$ for $k = 1, \ldots, K$, also denote $\bar{\theta} = K^{-1} \sum_{k=1}^{K} \theta_k$ and $\bar{\sigma^2} = K^{-1} \sum_{k=1}^{K} \sigma^2_k$. Let $\mathfrak{T}_{N,K} = K^{-1} \sum_{k=1}^{K} T_{N,K}^{(k)}$ be the average of $\{T_{N,K}^{(k)}\}_{k=1}^{K}$. To estimate the distribution of $K^{1/2} (\mathfrak{T}_{N,K} - \bar{\theta})$, we randomly draw $B$ resamples $\mathfrak{T}_{N,K}^{(1)}, \ldots, \mathfrak{T}_{N,K}^{(B)}$, $b = 1, \ldots, B$ from $F_{K,\mathfrak{T}}$, where $F_{K,\mathfrak{T}}$ is the empirical distribution of $\{\mathfrak{T}_{N,K}^{(k)}\}_{k=1}^{K}$. Denote $\mathfrak{T}_{N,K}^{(b)} = K^{-1} \sum_{k=1}^{K} \mathfrak{T}_{N,K}^{(b,k)}$ for $b = 1, \ldots, B$, then the empirical distribution of $\{K^{1/2} (\mathfrak{T}_{N,K}^{(b)} - \mathfrak{T}_{N,K})\}_{b=1}^{B}$ is used to estimate the one of $K^{1/2} (\mathfrak{T}_{N,K} - \bar{\theta})$. We call this algorithm the pseudo-distributed bootstrap and its procedure is summarized in Table 2.2.

The pseudo-distributed bootstrap is the conventional bootstrap procedure carried out on the statistics $\{\mathfrak{T}_{N,K}^{(k)}\}_{k=1}^{K}$, which are independent but not necessary identically distributed. The bootstrap procedures under non-i.i.d. models has been studied in Liu (1988). The key idea is that even the statistics $\{\mathfrak{T}_{N,K}^{(k)}\}_{k=1}^{K}$ are not i.i.d., the bootstrap still draws i.i.d. samples. The following proposition establishes the asymptotic properties of the pseudo-distributed bootstrap, which is an immediate result of Theorem 1 in Liu (1988).
Table 2.2 Algorithm for the pseudo-distributed bootstrap.

<table>
<thead>
<tr>
<th>Input:</th>
<th>data $\mathbf{X}^{(1)}<em>{N,K}, \ldots, \mathbf{X}^{(K)}</em>{N,K}$; $K$, number of subsets; $B$, number of Monte Carlo iterations; $\mathcal{T}$, function deriving statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output:</td>
<td>an estimate of the distribution of $K^{1/2}(\mathcal{T}_{N,K} - \bar{\theta})$</td>
</tr>
<tr>
<td>For $k \leftarrow 1$ to $K$ do</td>
<td></td>
</tr>
<tr>
<td>compute $\mathcal{T}^{(k)}<em>{N,K} = \mathcal{T}(\mathbf{X}^{(k)}</em>{N,K})$</td>
<td></td>
</tr>
<tr>
<td>End</td>
<td></td>
</tr>
<tr>
<td>Compute $\mathcal{T}<em>{N,K} = K^{-1} \sum</em>{k=1}^{K} \mathcal{T}^{(k)}_{N,K}$</td>
<td></td>
</tr>
<tr>
<td>For $b \leftarrow 1$ to $B$ do</td>
<td></td>
</tr>
<tr>
<td>generate resample $\mathcal{T}^{*b(1)}<em>{N,K}, \ldots, \mathcal{T}^{*b(K)}</em>{N,K}$ from $\mathcal{T}^{(1)}<em>{N,K}, \ldots, \mathcal{T}^{(K)}</em>{N,K}$</td>
<td></td>
</tr>
<tr>
<td>compute $\mathcal{T}^{*b}<em>{N,K} = K^{-1} \sum</em>{k=1}^{K} \mathcal{T}^{*b(k)}_{N,K}$</td>
<td></td>
</tr>
<tr>
<td>End</td>
<td></td>
</tr>
<tr>
<td>For $b \leftarrow 1$ to $B$ do</td>
<td></td>
</tr>
<tr>
<td>compute $K^{1/2}(\mathcal{T}^{*b}<em>{N,K} - \mathcal{T}</em>{N,K})$</td>
<td></td>
</tr>
<tr>
<td>End</td>
<td></td>
</tr>
</tbody>
</table>

Proposition 2.2. Let $\mathcal{T}^{(1)}_{N,K}, \ldots, \mathcal{T}^{(K)}_{N,K}$ be a sequence of statistics derived from each subset $\mathbf{X}^{(k)}_{N,K}$, with $E(\mathcal{T}^{(k)}_{N,K}) = \theta_k$ and $\text{Var}(\mathcal{T}^{(k)}_{N,K}) = \sigma_k^2$, $k = 1, \ldots, K$. Denote $\mathcal{T}_{N,K} = K^{-1} \sum_{k=1}^{K} \mathcal{T}^{(k)}_{N,K}$ and $\bar{\theta} = K^{-1} \sum_{k=1}^{K} \theta_k$. Let $\mathcal{T}^{*(1)}_{N,K}, \ldots, \mathcal{T}^{*(K)}_{N,K}$ be an i.i.d. sample from $F_{K,\mathcal{T}}$, where $F_{K,\mathcal{T}}$ is the empirical distribution of $\{\mathcal{T}^{(k)}_{N,K}\}_{k=1}^{K}$. Denote $\mathcal{T}^{*}_{N,K} = K^{-1} \sum_{k=1}^{K} \mathcal{T}^{*(k)}_{N,K}$. Suppose $\inf_k \sigma_k^2 > 0$ and $\sup_k E|\mathcal{T}^{(k)}_{N,K}|^{2+\delta} < \infty$ for some positive $\delta$. Moreover, assume $K^{-1} \sum_{k=1}^{K} (\theta_k - \bar{\theta})^2 \to 0$ as $K \to \infty$. Then with probability 1, as $K \to \infty$,

$$
\sup_{x \in \mathbb{R}} \left| P \left\{ K^{1/2}(\mathcal{T}^{*}_{N,K} - \mathcal{T}_{N,K}) \leq x \right| F_{K,\mathcal{T}} \right\} - P \left\{ K^{1/2}(\mathcal{T}_{N,K} - \bar{\theta}) \leq x \right\} \right| = o(1) \quad (2.11)
$$

This proposition ensures that under the conditions on the moments of $\{\mathcal{T}^{(k)}_{N,K}\}_{k=1}^{K}$, the pseudo-distributed bootstrap is asymptotic consistent. This result is quite general because we do not impose any other conditions on $\mathcal{T}^{(k)}_{N,K}$ rather than its moments. That is, $\mathcal{T}^{(k)}_{N,K}$ is not required to be in the form of (2.1). For specific statistics, we only need to check the moment conditions in Proposition 2.2 to get the consistency result. The following corollaries indicate that the pseudo-distributed bootstrap works for the distributed statistics in the form (2.3) under mild conditions.
Corollary 2.1. For $T_{N,K}$ in (2.3) with $\sigma^2 > 0$, under the conditions assumed in Theorem 2.4, let $T_{N,K}^{(k)} = N^{-1/2}K^{1/2}/n_k T_{N,K}'$ for $k = 1, \ldots, K$ and $F_{K,T}$ be the empirical distribution of $\{T_{N,K}^{(k)}\}^K_{k=1}$. Suppose $T_{N,K}^*(1), \ldots, T_{N,K}^{(K)}$ is an i.i.d. sample from $F_{K,T}$ and denote $T_{N,K}^* = K^{-1} \sum_{k=1}^K T_{N,K}^*(k)$. Assume $\mathbb{E} \{ \beta(X_1, X_2; F) \}^{2+\delta} < \infty$ and $\sup_k \mathbb{E} \{ n_k^{1/2} R_{N,K}^{(k)} \}^{2+\delta} < \infty$, in addition, $\sup_k N^{-1/2} K^{1/2} |n_k - NK^{-1}| \to 0$ as $N \to \infty$, then with probability 1, as $K \to \infty$,

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left\{ K^{1/2} (T_{N,K}^* - N^{1/2} K^{-1/2} T_{N,K}) \leq x \right| F_{K,T} \right\} - \mathbb{P} \left\{ N^{1/2} (T_{N,K} - \theta) \leq x \right\} \right| = o(1).$$

Corollary 2.1 shows that under moderate conditions on $n_k$, $K$ and the moments of $T_{N,K}^{(k)}$, the pseudo-distributed bootstrap is asymptotic consistent when $\sigma^2 > 0$. The resampling procedure is carried out on the scaled statistics $\{T_{N,K}^{(k)}\}^K_{k=1}$ as we have different sample size for each subset. The condition $\sup_k N^{-1/2} K^{1/2} |n_k - NK^{-1}| \to 0$ requires that the sample size of each subset should not be too far away from each other and this associates with the condition $K^{-1} \sum_{k=1}^K (\theta_k - \bar{\theta})^2 \to 0$ in Proposition 2.2.

Comparing with the distributed bootstrap, besides the gain in computing, an appealing property of the pseudo-distributed bootstrap is that it does not require the bootstrap statistic has certain form as assumed in Theorem 2.7 and 2.8. This makes the pseudo-distributed bootstrap more versatile and easier to verify. In addition, we would like to claim that the pseudo-distributed bootstrap has appropriate convergence rate. First, we assume that the remainder terms $\{R_{N,K}^{(k)}\}^K_{k=1}$ are small enough such that they can be dominated and thus omitted in the following discussion. Furthermore, suppose $n_k = NK^{-1}$ for $k = 1, \ldots, K$. Under the conditions $\mathbb{E} |\alpha(X_1; F)|^4 < \infty$ and $\mathbb{E} |\beta(X_1, X_2; F)|^4 < \infty$, by Theorem 5.18 of Petrov (1998),

$$\mathbb{P} \left\{ K^{1/2} s^{-1} (T_{N,K}^* - N^{1/2} K^{-1/2} T_{N,K}) \leq x \right| F_{K,T} \right\} \approx \Phi(x) - 6^{-1} K^{-1/2} s^{-3} \mathbb{E} \left\{ \left( T_{N,K}^* - N^{1/2} K^{-1/2} T_{N,K} \right)^3 \right\} F_{K,T} \left( x^2 - 1 \right) + O_p(K^{-1})$$

uniformly in $x \in \mathbb{R}$ where $s^2 = K^{-1} \sum_{k=1}^K \left( T_{N,K}^* - T_{N,K} \right)^2$. Mentioned that

$$\mathbb{E} \left\{ \left( T_{N,K}^* - N^{1/2} K^{-1/2} T_{N,K} \right)^3 \right\} F_{K,T} = O_p(n_k^{-1/2} + K^{-1/2}),$$
this clearly implies
\[
\sup_{x \in \mathbb{R}} \left| P \left\{ K^{1/2} \frac{1}{\bar{s}} (T_{N,K}^* - N^{1/2} K^{-1/2} T_{N,K}) \leq x \right| F_{K,T} \right| - \Phi(x) \right| = O_p(N^{-1/2} + K^{-1}).
\]
According to the proof of Theorem 2.4,
\[
\sup_{x \in \mathbb{R}} \left| P \left\{ N^{1/2} \sigma_\alpha^{-1} (T_{N,K} - \theta) \leq x \right| - \Phi(x) \right| = O(N^{-1/2}),
\]
it follows that
\[
\sup_{x \in \mathbb{R}} \left| P \left\{ K^{1/2} \frac{1}{\bar{s}} (T_{N,K}^* - N^{1/2} K^{-1/2} T_{N,K}) \leq x \right| F_{K,T} \right| - P \left\{ N^{1/2} \sigma_\alpha^{-1} (T_{N,K} - \theta) \leq x \right\}
\]
\[= O_p(N^{-1/2} + K^{-1}).\]
This result indicates that the convergence rate of the pseudo-distributed bootstrap is at the order of \(O_p(N^{-1/2} + K^{-1})\).

The next corollary gives theoretical support for the pseudo-distributed bootstrap when applied to the distributed statistics in the form (2.3) under degeneracy.

**Corollary 2.2.** For \(T_{N,K}\) in (2.3) with \(\sigma_\alpha^2 = 0\) but \(\sigma_\beta^2 > 0\), under Conditions 2.1, 2.3 where \(E(R_N) = b_1 N^{-\tau_1} + o(N^{-\tau_1})\) and \(Var(R_N) = o(N^{-\tau_2})\) for some \(\tau_1 > 1\), \(\tau_2 \geq 2\), assume \(n_k = NK^{-1}\) for \(k = 1, \ldots, K\) and \(K = O(N^{\tau'})\) for some positive constant \(\tau' < 1 - 1/(2\tau_1 - 1)\). Let \(T_{N,K}^{(k)} = n_k T_{N,K}^{(k)}\) for \(k = 1, \ldots, K\) and \(F_{K,T}\) be the empirical distribution function of \(\{T_{N,K}^{(k)}\}_{k=1}^K\).

Suppose \(T_{N,K}^{(1)}, \ldots, T_{N,K}^{(K)}\) is an i.i.d. sample from \(F_{K,T}\) and denote \(T_{N,K}^* = K^{-1} \sum_{k=1}^K T_{N,K}^{(k)}\).

Assume \(E|\beta(X_1, X_2; F)|^{2+\delta'} < \infty\) and \(\sup_k E \left| n_k R_{N,K}^{(k)} \right|^{2+\delta'} < \infty\) for some positive constant \(\delta'\), then with probability 1, as \(K \to \infty\),
\[
\sup_{x \in \mathbb{R}} \left| P \left\{ K^{1/2} (T_{N,K}^* - NK^{-1} T_{N,K}) \leq x \right| F_{K,T} \right| - P \left\{ NK^{-1/2} (T_{N,K} - \theta) \leq x \right\} \right| = o(1).
\]

Comparing with the case of \(\sigma_\alpha^2 > 0\), stronger conditions are needed for the degenerate statistics. First, \(\tau_1\) should be strictly larger than 1 such that \(R_{N,K}\) can be dominated by the quadratic term involving \(\beta\). Second, \(n_k\) is assumed to be the same for all subsets and \(K\) is required to have a smaller order. Last, stronger moment condition is needed for \(R_{N,K}^{(k)}\). These two corollaries reveal the versatility of the pseudo-distributed bootstrap. It works for both non-degenerate and degenerate statistics, the procedures are the same except that different scalers are used.
2.9 Comparison of the distributed bootstrap, the pseudo-distributed bootstrap, the BLB and the SDB

In this section, we compare the distributed bootstrap (DB), the pseudo-distributed bootstrap (PDB), the bag of little bootstrap (BLB) and the subsampled double bootstrap (SDB) in approximating the distribution function of $T_{N,K}$. First, we compare them regarding their computational complexity. In order to make our discussion more straightforward, we assume that the entire data are divided evenly into $K$ blocks, that is, $n_1 = \cdots = n_K = K^{-1}N$.

Suppose that the time cost for calculating the statistic in (2.1) on a sample of size $m$ is $t(m)$, then the computing time of the BLB with $B$ resamples for each of the $S$ subsets is $S \left\{ BKt(K^{-1}N) + t(n) \right\}$, where $n$ is the size of the subsets $\{X_{s,n}, s = 1, \ldots, S\}$. For the SDB with $S$ random subsets of size $n$, its computational complexity is $S \left\{ Kt(K^{-1}N) + t(n) \right\}$. If we generate $B$ resamples in the distributed bootstrap, the computational cost would be $(B+1)Kt(K^{-1}N)$ for the distributed bootstrap. In addition, the pseudo-distributed bootstrap requires $Kt(K^{-1}N) + BK$ time to fulfill its implementation with $B$ pseudo resamples. Thus the pseudo-distributed bootstrap is the fastest among these four resampling strategies, which is due to that the PDB benefits from avoiding recalculating the distributed statistic for each iteration. The distributed bootstrap and the SDB have similar computational cost if $S$ is close to $B$. However, the SDB can not be implemented distributively. For the BLB, its computational complexity depends on both $S$ and $B$. Moreover, if given a fixed time budget, it is not clear how to select $S$ and $B$ to achieve the optimal statistical accuracy, see more discussions in Sengupta, Volgushev and Shao (2016).

For massive dataset that stored in different locations, the advantage of the SDB is limited because it requires random sampling from the entire dataset, while the other three methods can be implemented distributively within each data block. Furthermore, only the pseudo-distributed bootstrap requires $K$ diverges, the other three approaches work for both cases when $K$ is finite or diverging. Comparison of these four resampling methods is summarized in Table 2.3.
Table 2.3 Comparison of the distributed bootstrap (DB), the pseudo-distributed bootstrap (PDB), the BLB and the SDB.

<table>
<thead>
<tr>
<th>Method</th>
<th>Computing Time</th>
<th>Distributed or Not</th>
<th>Requirement on $K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB</td>
<td>$(B + 1)Kt(K^{-1}N)$</td>
<td>distributed</td>
<td>finite or diverging</td>
</tr>
<tr>
<td>PDB</td>
<td>$Kt(K^{-1}N) + BK$</td>
<td>distributed</td>
<td>diverging</td>
</tr>
<tr>
<td>BLB</td>
<td>$S { BKt(K^{-1}N) + t(n) } $</td>
<td>distributed</td>
<td>finite or diverging</td>
</tr>
<tr>
<td>SDB</td>
<td>$S { Kt(K^{-1}N) + t(n) } $</td>
<td>not distributed</td>
<td>finite or diverging</td>
</tr>
</tbody>
</table>

2.10 Applications

Distance covariance and distance correlation, introduced in Székely, Rizzo and Bakirov (2007), are distance based methods that measuring and testing dependence between two random vectors. The distance covariance and correlation of two random vectors are equal to zero if and only if these two random vectors are independent. This attractive property makes the distance variance and correlation more versatile in detecting dependence than the classic Pearson correlation, especially for non-linear dependence. Distance covariance and correlation have been used in many literatures, for example, feature screening (Li, Zhong and Zhu, 2012), sufficient dimension reduction diagnostic (Chen, Cook and Zou, 2015) and testing mutual independence under high-dimensionality (Yao, Zhang and Shao, 2016). In this section, we focus on the distance covariance and its usage in testing independence for massive data.

Suppose $Y$ and $Z$ are two random vectors having finite first moments and taking values in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. Let $\phi_Y(t)$ and $\phi_Z(s)$ be the characteristic functions of $Y$ and $Z$, and $\phi_{Y,Z}(t,s)$ be the joint characteristic function of $Y$ and $Z$. The distance covariance between $Y$ and $Z$ is defined as

$$ dcov^2(Y, Z) = \int_{\mathbb{R}^{p \times q}} \|\phi_{Y,Z}(t,s) - \phi_Y(t)\phi_Z(s)\|^2 \omega(t,s) dt ds, $$

(2.12)

where $\omega(t,s) = (c_p c_q \|t\|_p^{1+p} \|s\|_q^{1+q})^{-1}$ is a weight function with $c_d = \pi^{(1+d)/2}/\Gamma\{(1 + d)/2\}$ and $\|a\|_d$ is the Euclidean norm of $a \in \mathbb{R}^d$. For a complex-valued function $\phi$, $\|\phi\|^2 = \phi \bar{\phi}$ where $\bar{\phi}$ is the conjugate of $\phi$. Denote $X = (Y^T, Z^T)^T$, under the condition that $E\|Y\|_p + E\|Z\|_q < \infty$, 

$$ dcov^2(Y, Z) = E\|Y - Y'\|_p\|Z - Z'\|_q - 2E\|Y - Y'\|_p\|Z - Z''\|_q + E\|Y - Y'\|_p E\|Z - Z'\|_q, $$

where $X' = (Y'^T, Z'^T)^T$ and $X'' = (Y''^T, Z''^T)^T$ are independent copies of $X$. 
Suppose that $X_N = \{X_1, \ldots, X_N\}$ is an observed sample from $F$, the joint distribution of $X$, and $X_i = (Y_i^T, Z_i^T)^T$ for $i = 1, \ldots, N$. Define $A_{ij} = \|X_i - X_j\|_p$ and $B_{ij} = \|Y_i - Y_j\|_q$ for $i, j = 1, \ldots, N$. In this paper, we consider the $U$-centered version of empirical distance covariance (Székely and Rizzo, 2014; Yao, Zhang and Shao, 2016) which is defined by

$$
dcov^2_N(Y, Z) = \{N(N-3)\}^{-1} \sum_{i \neq j} \tilde{A}_{ij} \tilde{B}_{ij},
$$

(2.13)

where

$$
\tilde{A}_{ij} = A_{ij} - (N - 2)^{-1} \sum_{l_1=1}^N A_{l_1 j} - (N - 2)^{-1} \sum_{l_2=1}^N A_{il_2} + \{(N - 1)(N - 2)\}^{-2} \sum_{l_1=1}^N \sum_{l_2=1}^N A_{l_1 l_2},
$$

and

$$
\tilde{B}_{ij} = B_{ij} - (N - 2)^{-1} \sum_{l_1=1}^N B_{l_1 j} - (N - 2)^{-1} \sum_{l_2=1}^N B_{il_2} + \{(N - 1)(N - 2)\}^{-2} \sum_{l_1=1}^N \sum_{l_2=1}^N B_{l_1 l_2},
$$

for $i \neq j$. By simple algebra,

$$
dcov^2_N(Y, Z) = \{N(N-1)\}^{-1} \sum_{i \neq j} A_{ij} B_{ij} + \{N(N-1)(N-2)\}^{-1} \sum_{i \neq j \neq l_1} A_{ij} B_{l_1 l_2},
$$

$$
+ \{N(N-1)(N-2)(N-3)\}^{-1} \sum_{i \neq j \neq l_1 \neq l_2} A_{ij} B_{l_1 l_2},
$$

from which we can see that $dcov^2_N(Y, Z)$ is an unbiased estimator of $dcov^2(Y, Z)$. Moreover, $dcov^2_N(Y, Z)$ is a U-statistic of degree 4.

When the sample size $N$ is large, there are two issues when computing $dcov^2_N(Y, Z)$. First, the computational complexity of $dcov^2_N(Y, Z)$ is at the order of $O(N^2)$, which is expensive when $N$ is large. Second, the memory requirement in calculating $dcov^2_N(Y, Z)$ is of order $O(N^2)$, which makes the computing less feasible for a larger $N$. For example, when $N = 10^5$, it requires more than 70 Gigabyte storage in memory in order to complete the calculation of $dcov^2_N(Y, Z)$. Thus we employ a distributed version of the distance covariance under the massive data scenario.

Suppose the entire data $X_N$ are divided into $K$ sub-samples with the $k$-th subset $X_{N,K}^{(k)} = \{X_{k,1}, \ldots, X_{k,n_k}\}$ of size $n_k$ for $k = 1, \ldots, K$. Denote $dcov^2_{N,K}(Y, Z)$ as the empirical distance covariance based on $X_{N,K}^{(k)}$. Then the distributed distance covariance is defined as

$$
dcov^2_{N,K}(Y, Z) = N^{-1} \sum_{k=1}^K n_k ddcov^2_{N,K}(Y, Z).
$$

(2.14)
It is easy to see that $dcov_{N,K}^2(Y, Z)$ is also an unbiased estimator of the population distance covariance $dcov^2(Y, Z)$. The distributed distance covariance enjoys two advantages in computing. Assume that $n_k, k = 1, \ldots, K$ have the same order, then the computing time of $dcov_{N,K}^2(Y, Z)$ is of order $O(K^{-1}N^2)$, which is much less than $O(N^2)$, the computational complexity of $dcov_N^2(Y, Z)$. Moreover, the memory requirement of computing $dcov_{N,K}^2(Y, Z)$ is only $O(K^{-2}N^2)$, which makes $dcov_{N,K}^2(Y, Z)$ more feasible when the size of the data is large. Besides that, computing $dcov_{N,K}^2(Y, Z)$ can be easily parallelized which is suited to modern parallel and distributed computing architectures.

Now we consider testing independence between $Y$ and $Z$ using the distance covariance. That is to test the hypothesis

$$H_0 : Y \text{ and } Z \text{ are independent} \quad \text{versus} \quad H_1 : Y \text{ and } Z \text{ are dependent.} \quad (2.15)$$

Under the null hypothesis that $Y$ and $Z$ are independent, if $E\|y\|_p^2 + E\|z\|_q^2 < \infty$, then the empirical distance covariance $dcov_N^2(Y, Z)$ is a degenerate U-statistic that has the following representation (Yao, Zhang and Shao, 2016):

$$dcov_N^2(Y, Z) = N^{-2} \sum_{1 \leq i < j \leq N} \beta(X_i, X_j; F) + R_N,$$

where $\beta(X_i, X_j; F) = 2U(Y_i, Y_j)V(Z_i, Z_j)$ with $U(y, y') = \|y - y'\|_p - E\|y - y'\|_p - E\|y - y' - Z\|_q + E\|Z - Z'\|_q$, and $V(z, z') = \|z - z'\|_q - E\|z - z'\|_q - E\|z - z'\|_q + E\|Z - Z'\|_q$, and $R_N = R(X_N; F)$ is the reminder term satisfies $E(R_N) = 0$ and $\text{Var}(R_N) = O(N^{-3})$. It is easy to verify that under the null hypothesis, $E\{\beta(X_1, X_2; F)|X_1\} = 0$ and $E\{\beta(X_1, X_2; F)\}^2 = 4E\{U(Y_1, Y_2)\}^2 E\{V(Z_1, Z_2)\}^2 \equiv \sigma_3^2 < \infty$.

If the empirical distance covariance $dcov_N^2(Y, Z)$ in (2.13) is used as the test statistic, we need to obtain a reference distribution for $dcov_N^2(Y, Z)$ using the bootstrap on the entire dataset (Arcones and Giné, 1992) or random permutation (Székely, Rizzo and Bakirov, 2007). For massive dataset, computing $dcov_N^2(Y, Z)$ itself is a big issue as discussed before. Moreover, the bootstrap and random permutation on the entire dataset are both computationally expensive as we need to recalculate the distance covariance for each resample or permutation. Thus we consider using the distributed distance covariance $dcov_{N,K}^2(Y, Z)$ as the test statistic. Under
the null hypothesis, if $E\|Y\|_p^2 + E\|Z\|_q^2 < \infty$, $dcov^2_{N,K}(Y, Z)$ can be rewritten as

$$dcov^2_{N,K}(Y, Z) = N^{-1} \sum_{k=1}^{K} n_k^{-1} \sum_{1 \leq i < j \leq n_k} \beta(X_{k,i}, X_{k,j}; F) + R_{N,K},$$

(2.16)

where $R_{N,K} = N^{-1} \sum_{k=1}^{K} n_k R(x^{(k)}_{N,K}; F)$ is the remainder term that satisfies $E(R_{N,K}) = 0$ and $Var(R_{N,K}) = O(K^2N^{-3})$. Thus according to Theorem 2.6, we have the following asymptotic result concerning $dcov^2_{N,K}(Y, Z)$.

**Corollary 2.3.** Assume there exists a constant $\delta' > 0$ such that $E\|Y\|_p^{2+\delta'} + E\|Z\|_q^{2+\delta'} < \infty$. There exist positive constants $c_1$ and $c_2$ such that $c_1 \leq \inf_{k_1,k_2} n_{k_1}/n_{k_2} \leq \sup_{k_1,k_2} n_{k_1}/n_{k_2} \leq c_2$. In addition, if $K \to \infty$ and $K = o(N)$, then under the null hypothesis that $Y$ and $Z$ are independent,

$$2^{1/2} K^{-1/2} N \sigma_{\beta}^{-1} dcov^2_{N,K}(Y, Z) \xrightarrow{d} N(0, 1)$$

as $N \to \infty$, where $\sigma_{\beta}^2 = 4E \{U(Y_1, Y_2)\}^2 E \{V(Z_1, Z_2)\}^2$.

In order to carry out the test using this asymptotic normality result, we need to get a consistent estimator of $\sigma_{\beta}^2$, or equivalently, estimators of $E \{U(Y_1, Y_2)\}^2$ and $E \{V(Z_1, Z_2)\}^2$. Mentioned that

$$dcov^2(Y, Y) = E \{U(Y_1, Y_2)\}^2 \quad \text{and} \quad dcov^2(Z, Z) = E \{V(Z_1, Z_2)\}^2$$

are the distance variances (Székely, Rizzo and Bakirov, 2007) of $Y$ and $Z$, respectively, it is straightforward to use the distributed version of the empirical distance variances to estimate $dcov^2(Y, Y)$ and $dcov^2(Z, Z)$. Let

$$dcov^2_{N,K}^{(k)}(Y, Y) = \{n_k(n_k - 3)\}^{-1} \sum_{1 \leq i \neq j \leq n_k} \tilde{A}^2_{(k,i)(k,j)},$$

$$dcov^2_{N,K}^{(k)}(Z, Z) = \{n_k(n_k - 3)\}^{-1} \sum_{1 \leq i \neq j \leq n_k} \tilde{B}^2_{(k,i)(k,j)},$$

then

$$\hat{\sigma}_{\beta,N,K}^2 = 4 \left\{ N^{-1} \sum_{k=1}^{K} n_k dcov^2_{N,K}^{(k)}(Y, Y) \right\} \left\{ N^{-1} \sum_{k=1}^{K} n_k dcov^2_{N,K}^{(k)}(Z, Z) \right\}$$

is an unbiased estimator of $\sigma_{\beta}^2$ under the null hypothesis. The consistency of this variance estimator is established in the next theorem.
Theorem 2.9. Under the conditions in Corollary 2.3, if \(E\|Y\|^4 + E\|Z\|^4 < \infty\), then as \(N \to \infty\),

\[
\hat{\sigma}^2_{\beta,N,K} / \sigma^2_{\beta} \overset{p}{\to} 1.
\]

Combine Corollary 2.3 and Theorem 2.9 with Slutsky’s theorem, we have

\[
2^{1/2} K^{-1/2} N \bar{\sigma}^{-1}_{\beta,N,K} d\text{cov}_{N,K}^2(Y, Z) \overset{d}{\to} N(0, 1)
\]

under the null hypothesis. Therefore, a test with nominal significant level \(\tau\) reject \(H_0\) if

\[
2^{1/2} K^{-1/2} N \bar{\sigma}^{-1}_{\beta,N,K} d\text{cov}_{N,K}^2(Y, Z) > z_{\tau},
\]

where \(z_{\tau}\) is the \(\tau\) upper-quantile of the standard normal distribution, and the test procedure is ensured to have the nominal level of significance asymptotically. Furthermore, the implementation of the test includes calculating \(d\text{cov}_{N,K}^2(Y, Z)\) and \(\hat{\sigma}^2_{\beta,N,K}\), which both require \(O(K^{-1}N^2)\) order computing time. This feature makes the testing procedure based on the distributed distance covariance more friendly to massive datasets. For each subset \(X^{(k)}_{N,K}\), after getting \(\hat{A}_{(k,i)(k,j)}\) and \(\hat{B}_{(k,i)(k,j)}\) for different \(i\) and \(j\), we can obtain \(d\text{cov}_{N,K}^2(Y, Z)\), \(d\text{cov}_{N,K}^2(Y, Y)\) and \(d\text{cov}_{N,K}^2(Z, Z)\) for \(k = 1, \ldots, K\), which are sufficient to fulfill the test.

Next we consider using the pseudo-distributed bootstrap to approximate the distribution of \(d\text{cov}_{N,K}^2(Y, Z)\) under the null hypothesis, and as a consequence, to implement the independence test. Assume the sizes of all subsets are equal, that is \(n_k = NK^{-1}\) for \(k = 1, \ldots, K\). Let \(F_{K,d\text{cov}}\) be the empirical distribution of \(\{d\text{cov}_{N,K}^2(Y, Z), \ldots, d\text{cov}_{N,K}^2(Y, Z)\}\). For \(b = 1, \ldots, B\), draw resample \(d\text{cov}_{N,K}^{2b(1)}(Y, Z), \ldots, d\text{cov}_{N,K}^{2b(K)}(Y, Z)\) from \(F_{K,d\text{cov}}\). Then

\[
d\text{cov}_{N,K}^{2b}(Y, Z) = K^{-1} \sum_{k=1}^{K} d\text{cov}_{N,K}^{2b(k)}(Y, Z), \quad b = 1, \ldots, B
\]

can be used to approximate the critical value of the test. According to Corollary 2.2, we have the following theoretical result supporting the pseudo-distributed bootstrap approach.

Corollary 2.4. Under the conditions in Corollary 2.3, assume \(n_k = NK^{-1}\) for \(k = 1, \ldots, K\). Let \(F_{K,d\text{cov}}\) be the empirical distribution of \(d\text{cov}_{N,K}^{2(1)}(Y, Z), \ldots, d\text{cov}_{N,K}^{2(K)}(Y, Z)\). Suppose that \(d\text{cov}_{N,K}^{2(k)}(Y, Z), \quad k = 1, \ldots, K\) are i.i.d. from \(F_{K,d\text{cov}}\) and denote

\[
d\text{cov}_{N,K}^{2*}(Y, Z) = K^{-1} \sum_{k=1}^{K} d\text{cov}_{N,K}^{2*(k)}(Y, Z).
\]
Then with probability 1, as \( K \to \infty \),
\[
\sup_{x \in \mathbb{R}} \left| P \left\{ 2^{1/2} K^{-1/2} N dcov_{N,K}^2(Y, Z) \leq x \right| F_{K,T} \right\} - P \left\{ 2^{1/2} K^{-1/2} N dcov_{N,K}^2(Y, Z) \leq x \right\} \right| = o(1).
\]

This corollary ensures that the test based on the pseudo-distributed bootstrap is consistent. Comparing with Theorem 2.9, fourth finite moment condition is not needed for the pseudo-distributed bootstrap approach. In addition, the procedure based on the pseudo-distributed bootstrap is computational competitive. We will compare these two different approaches in the simulation studies.

2.11 Numerical studies

In this section, we report numerical studies, using both simulated and real datasets, to evaluate the performance of our proposed distributed approach and bootstrap algorithms. We also compare our approaches with the BLB (Kleiner, et al., 2014) and the SDB (Sengupta, Volgushev and Shao, 2016) to reveal the advantages of our methods. We focus on two different setups, one is the Gini’s mean difference which measures the variability of a distribution, the other one is testing independence using the distance covariance. We fix the sample size as \( N = 100000 \) and vary the number of subsets \( K \) for different scenarios in the simulation studies. All the experiments executed in this section are ran in R with a single Intel(R) Core(TM) i7 4790K @4.0 GHz processor.

2.11.1 Gini’s mean difference

The Gini’s mean difference is defined as
\[
U_N = \{N(N - 1)\}^{-1} \sum_{1 \leq i \neq j \leq N} |X_i - X_j|,
\]
which is a U-statistic of degree two, and it is an unbiased estimator of the dispersion parameter \( \theta = E|X_i - X_j| \), which measures the variability of a distribution. Then \( U_N \) has the formation as in (2.10). For massive data, we introduce the distributed version of \( U_N \). Suppose the entire data are divided into \( K \) data blocks, with each of size \( n_k \). For each block, denote \( U_{N,K}^{(k)} \) as the Gini’s mean difference obtained from the \( k \)-th data block, then the distributed Gini’s mean
difference is $U_{N,K} = N^{-1} \sum_{k=1}^{K} n_k U_{N,K}^{(k)}$. As $U_N$ is an unbiased estimator of $\theta$, as long as $K = o(N)$, $U_{N,K}$ attains the same estimation efficiency as $U_N$.

For $U_N$, we can use the jackknife (Efron and Stein, 1981) to estimate its variance, denoted as $N^{-1} S^2_{U_N}$, where

$$S^2_{U_N} = 4(N - 1)(N - 2)^2 \sum_{i=1}^{N} \left\{ (N - 1)^{-1} \sum_{j \neq i} |X_i - X_j| - U_N \right\}^2,$$

and $S^2_{U_N}$ can be expanded as

$$S^2_{U_N} = \sigma^2_{\alpha,U} + N^{-1} \sum_{i=1}^{N} \gamma_U(X_i, F) + H_{U_N},$$

with $\gamma_U(X_i, F) = \{ \alpha_U(X_i, F) \}^2 - \sigma^2_{\alpha,U} + 2g(X_i)$ and $g(x) = E\{ \alpha_U(X_1, F)\beta_U(x, X_1; F) \}$, the remainder term $H_{U_N}$ satisfies $E(H_{U_N}) = O(N^{-2})$ and $E(H^2_{U_N}) = O(N^{-2})$ (Callaert and Veraverbeke, 1981). Then $S^2_{U_N}$ is in the form of (2.1) with vanished quadratic term. As an estimator of $\text{Var}(U_N)$, $N^{-1} S^2_{U_N}$ has a bias of order $N^{-2}$. Because the variance of $N^{-1} S^2_{U_N}$ is of order $O(N^{-3})$, the mean square error of $N^{-1} S^2_{U_N}$ is of order $O(N^{-3})$.

For the distributed statistic $U_{N,K}$, we propose a distributed jackknife variance estimator. Denote $n_k^{-1} S^2_{U_{N,K}}^{(k)}$ as the jackknife variance estimator of $U_{N,K}^{(k)}$ for the $k$-th block, then the distributed variance estimator for $U_{N,K}$ is defined as $N^{-1} S^2_{U_{N,K}} = N^{-1} \sum_{k=1}^{K} n_k S^2_{U_{N,K}}^{(k)}$. According to Theorem 2.1, the mean square error of $N^{-1} S^2_{U_{N,K}}$ as an estimator of $\text{Var}(U_{N,K})$ is $O(N^{-3} + K^2 N^{-4})$, which is still of order $N^{-3}$ if $K = o(N^{1/2})$. Also the mean square error of $N^{-1} S^2_{U_{N,K}}$ increases at the order of $K^2$ when $KN^{-1/2}$ diverges.

In the simulations, we consider $X_1, \ldots, X_N$ from three different distributions, thus we consider the following three scenarios.

(I). $\mathcal{N}(1, 1)$: Normal distribution with mean 1 and unit variance;

(II). Gamma$(3, 1)$: Gamma distribution with shape parameter 3 and scale parameter 1;


The first two distributions are continuous while the third one is discrete, and the true values of $\theta$ and $\text{Var}(U_{N,K})$ can be calculated algebraically for the first two distributions. For Poisson distribution, the corresponding $\theta$ and $\text{Var}(U_{N,K})$ can be approximated to a certain accuracy due to its discreteness. See Lomnicki (1952) for reference.
First, we want to study the performance of $U_{N,K}$ as an estimator of $\theta$ by comparing it with $U_N$, which is a special case of $U_{N,K}$ when $K = 1$. In addition, we examine how $N^{-1}S^2_{U_{N,K}}$ works as the estimator of the variance of $U_{N,K}$ in terms of mean square errors (MSE). Let $\mathcal{K} = \{5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000\}$ be the set of values for $K$ we considered here. For the sake of convenience, we divide the whole dataset into blocks with equal size $N/K$ in the simulations. And we use the notation MSE(·) to denote the mean square error of an estimator.

With the distributions and parameters set up above, for each simulation, we generate $\{X_i, i = 1, \ldots, N\}$ i.i.d. from the three distributions respectively. For each dataset, we calculate $U_N$ and its variance estimator $N^{-1}S^2_{U_N}$. Then for each $K \in \mathcal{K}$, we estimate $\theta$ using $U_{N,K}$ and estimate the variance of $U_{N,K}$ by $N^{-1}S^2_{U_{N,K}}$. Thus by carrying out a number of simulations, we can get MSE($U_N$), MSE($N^{-1}S^2_{U_N}$) and MSE($U_{N,K}$), MSE($N^{-1}S^2_{U_{N,K}}$) for each $K \in \mathcal{K}$ under three scenarios.

We summarize the average mean square errors of $U_{N,K}$ and $N^{-1}S^2_{U_{N,K}}$ for all three scenarios in Figure 2.2, with each based on 2000 simulations. In order to compare the results of $U_{N,K}$ with $U_N$ more intuitively, we plot the mean square error of $U_{N,K}$ and $N^{-1}S^2_{U_{N,K}}$ divided by those of $U_N$ and $N^{-1}S^2_{U_N}$, respectively, that is, each point in Figure 2.2 is the average of MSE($U_{N,K}$)/MSE($U_N$) (Figure 2.2(a)) or MSE($S^2_{U_{N,K}}$)/MSE($S^2_{U_N}$) (Figure 2.2(b)) for $K \in \mathcal{K}$. Figure 2.2(a) shows that, the mean square error of $U_{N,K}$ keeps increasing slowly as $K$ goes bigger when $K$ is not too large. Thus the distributed estimator $U_{N,K}$ is almost as good as $U_N$ for relatively small $K$, this coincides with our theoretical results that $U_{N,K}$ is unbiased and the variance increment occurs only in the second order term, so as long as $O(N^{-1})$ is still the dominate order of the variance of $U_{N,K}$ when $K = o(N)$, the increment in mean square error of $U_{N,K}$ is negligible. From the plot we can see that even when $K$ is as large as 5000, which means that each block only contains 20 data points, the relative increase of mean square error is less than 10 percent for all three distributions. Thus $U_{N,K}$ is still an reasonable substitute of $U_N$ as an estimator of $\theta$ in this situation.

For $N^{-1}S^2_{U_{N,K}}$, Figure 2.2(b) shows that $N^{-1}S^2_{U_{N,K}}$ is still comparable to $N^{-1}S^2_{U_N}$ in the sense of mean square error for relatively small $K$. But the mean square error for $N^{-1}S^2_{U_{N,K}}$ increases very fast for $K$ larger than 200, and the outcomes vary for different underlying distribu-
Figure 2.2 MSE($U_{N,K}$)/MSE($U_N$) and MSE($N^{-1}S^2_{UN,K}$)/MSE($N^{-1}S^2_{UN}$) against the number of blocks $K$ for each scenario.

As we discussed before, the mean square error of $N^{-1}S^2_{UN,K}$ is of order $O(N^{-3} + K^2 N^{-4})$ where the second order term increases at the order $K^2$. So for large $K$ such that the $O(K^2 N^{-4})$ order term dominates the term of order $O(N^{-3})$, the mean square error of $N^{-1}S^2_{UN,K}$ would increase at the order of $K^2$, as demonstrated in Figure 2.2(b).

In addition, Figure 2.3 depicts the variation of the absolute bias and variance of $N^{-1}S^2_{UN,K}$ that contributes to its mean square error. Clearly, by comparing Figure 2.3(a) with Figure 2.3(b), we see that the relative increment in the mean square error of $N^{-1}S^2_{UN,K}$ is mainly due to its bias increase, which agrees with the investigation concerning MSE($N^{-1}S^2_{UN,K}$) before.

Now we turn to study the performance of the distributed bootstrap and the pseudo-distributed bootstrap, and compare them with the BLB and the SDB. In this simulation, we consider constructing the 95% confidence intervals for $\theta$ based on $U_{N,K}$ for the number of subsets $K \in \{20, 50, 100, 200, 500, 1000\}$ using the distributed bootstrap, the pseudo-distributed bootstrap, the BLB and the SDB, respectively. For each simulated dataset and $K$, we use these four methods to estimate the width of confidence intervals and each method is allowed to run for 10 seconds in order to take the computing time into consideration. Under this setup, we can compare the performances of these methods within certain fixed time budget. For the BLB,
we fix the number of resamples $B = 100$ as in Kleiner, et al. (2014) and Sengupta, Volgushev and Shao (2016), the subset $X_{s,n}$ are chosen as the $s$-th data block, thus the size of the subset $X_{s,n}$ is equal to $N/K$. For the SDB, the size of the random subset $X_{s,n}^*$ is also chosen as $N/K$. Table 2.4 gives the summary of the number of iterations completed for each method within 10 seconds for different $K$. From this table, we can see that the pseudo-distributed bootstrap is the fastest one that has the most iterations among all these methods. The bag of little bootstrap is the slowest method, when $K = 20$, the BLB can not even finish one iteration in 10 seconds. The distributed bootstrap and the subsampled double bootstrap have similar performance. However, it is worth to mention that these results do not take the time expenditure of data communication between each data block into account, for the SDB which can not be implemented distributively, the communication between each block can be costly that slows down the SDB hugely.

In order to evaluate these four methods, we conduct 500 simulation replications and report the coverage rates and widths of the 95% confidence intervals for different methods, the results for Gaussian scenario is given in Table 2.5 and the results for Gamma and Poisson data can be found in Table 2.10 and Table 2.11 in the appendix. Table 2.5 shows that the coverage rates
Table 2.4  Number of iterations completed in 10 seconds for each method. DB: the distributed bootstrap, PDB: the pseudo-distributed bootstrap, BLB: the bag of little bootstrap, SDB: the subsampled double bootstrap.

<table>
<thead>
<tr>
<th></th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB</td>
<td>47</td>
<td>131</td>
<td>265</td>
<td>513</td>
<td>1067</td>
<td>1569</td>
</tr>
<tr>
<td>PDB</td>
<td>5000+</td>
<td>5000+</td>
<td>5000+</td>
<td>5000+</td>
<td>5000+</td>
<td>5000+</td>
</tr>
<tr>
<td>BLB</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>SDB</td>
<td>51</td>
<td>128</td>
<td>249</td>
<td>472</td>
<td>1015</td>
<td>1436</td>
</tr>
</tbody>
</table>

are a little bit far away from the nominal confidence level 95% for relatively small $K$. When $K = 20$, the reason why the distributed bootstrap and the subsampled double bootstrap do not have good performances is that they can not finish enough iterations to ensure the convergence of their algorithms. As the pseudo-distributed bootstrap is based on the assumption that $k$ diverges, the coverage rate is still not good when $K = 20$ even it can complete enough iterations in 10 seconds. For relatively large $K$, the performances of the DB, the PDB and the SDB are reasonable and comparable to each other. For the bag of little bootstrap, its performance depends on the choices of the number of resamples $B$ and the subset size $n$.

Table 2.5  Coverage rates and confidence interval widths (in parentheses) of the 95% confidence intervals with 10 seconds time budget for Gaussian data.

<table>
<thead>
<tr>
<th></th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB</td>
<td>0.918</td>
<td>0.934</td>
<td>0.940</td>
<td>0.926</td>
<td>0.938</td>
<td>0.936</td>
</tr>
<tr>
<td></td>
<td>(0.00960)</td>
<td>(0.00981)</td>
<td>(0.00991)</td>
<td>(0.00993)</td>
<td>(0.00994)</td>
<td>(0.00993)</td>
</tr>
<tr>
<td>PDB</td>
<td>0.918</td>
<td>0.940</td>
<td>0.944</td>
<td>0.936</td>
<td>0.940</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>(0.00962)</td>
<td>(0.00989)</td>
<td>(0.00994)</td>
<td>(0.00998)</td>
<td>(0.01000)</td>
<td>(0.01004)</td>
</tr>
<tr>
<td>BLB</td>
<td>NA</td>
<td>0.928</td>
<td>0.932</td>
<td>0.934</td>
<td>0.934</td>
<td>0.930</td>
</tr>
<tr>
<td></td>
<td>(NA)</td>
<td>(0.00958)</td>
<td>(0.00967)</td>
<td>(0.00973)</td>
<td>(0.00982)</td>
<td>(0.00978)</td>
</tr>
<tr>
<td>SDB</td>
<td>0.916</td>
<td>0.932</td>
<td>0.942</td>
<td>0.938</td>
<td>0.940</td>
<td>0.940</td>
</tr>
<tr>
<td></td>
<td>(0.00965)</td>
<td>(0.00984)</td>
<td>(0.00990)</td>
<td>(0.00996)</td>
<td>(0.00998)</td>
<td>(0.00998)</td>
</tr>
</tbody>
</table>

Next we calculate the relative errors of the confidence interval widths for each method. If $d$ is the true width and $\hat{d}$ is an estimator of the confidence interval width, the relative error is defined as $|\hat{d} - d|/d$. We approximate the true width by 5000 Monte Carlo simulations and the relative errors are averaged over 500 simulations. Follow the comparison strategy in
Sengupta, Volgushev and Shao (2016), these four methods are compared with respect to the
time evolution of relative errors. This is a way to explore which method can produce more
precise result given a fixed time budget. For each method, we calculate its relative error for
each iteration along with the computing time. For the BLB and the SDB, one iteration means
the completion of estimation procedure for one subset. The relative error is assigned to be 1
before the first iteration is completed.

Figure 2.4 shows the evolution of relative errors across time for the distributed bootstrap,
the pseudo-distributed bootstrap, the BLB and the SDB with different number of blocks, the
data are generated from $\mathcal{N}(1, 1)$. First, we can see that the pseudo-distributed bootstrap is
the fastest method to converge for all $K$. It takes time for the BLB to complete one iteration,
especially when $K$ is small, for example, when $K = 20$, the BLB can not even finish one iteration
in 10 seconds. Also when $K = 20$, the relative error of the pseudo-distributed bootstrap is
not reliable which means it can not produce a reasonable estimate of the confidence interval
width in this case. This is expectable because the convergence rate of the pseudo-distributed
bootstrap relies on the order of $K$. In contrast, as $K$ increases, the relative errors of the pseudo-
distributed bootstrap will decrease to an acceptable rate. For a sufficient large $K$ ($K \geq 200$),
its relative error is comparable with those of the distributed bootstrap, the BLB and the SDB.
The distributed bootstrap and the SDB have performance close to each other. However, the
SDB is not a distributed resampling approach, which would limit its usage when the entire
data are stored in different locations. Results for Gamma and Poisson scenarios are shown in
Figure 2.6 and Figure 2.7.

In conclusion, with a small time budget, the other three methods have advantages over
the BLB, especially when $K$ is small. When $K$ is relatively large, the pseudo-distributed
bootstrap has advantage in computing and can produce reasonable estimators. The SDB has
limitation that it can not be implemented distributively. The distributed bootstrap is designed
distributively and can work for a wide range of $K$. 
<table>
<thead>
<tr>
<th>Time (sec)</th>
<th>Relative Error</th>
<th>PDB</th>
<th>DB</th>
<th>BLB</th>
<th>SDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.4  Time evolution of relative errors under Gaussian scenario.
2.11.2 Distance covariance

The distributed version of distance covariance and its usage in testing independence between two multivariate random vectors has been studied in Section 2.10. By formulating the distributed distance covariance, we can reduce the requirements on computing time and memory space, which makes it meet the demands for handling massive datasets. In this section, we investigate the performance of the distributed distance covariance \( \text{dcov}^2_{N,K}(Y, Z) \) in testing independence by numerical studies. It is worth noting that for the conventional distance covariance \( \text{dcov}^2_N(Y, Z) \), random permutations are needed to implement the test (Székely, Rizzo and Bakirov, 2007), which could increase the computing cost dramatically. While for distributed distance covariance \( \text{dcov}^2_{N,K}(Y, Z) \), we can use the result in (2.17) based on the variance estimator \( \hat{\sigma}^2_{\beta,N,K} \) or the pseudo-distributed bootstrap (Corollary 2.4) to implement the test, where permutations are not needed. Moreover, the testing procedures based on \( \text{dcov}^2_{N,K}(Y, Z) \) only need \( O(K^{-2}N^2) \) memory space, which is much less than that of \( \text{dcov}^2_N(Y, Z) \), which is at the order of \( O(N^2) \). All simulation results in this section were based on 1000 iterations with the nominal significant level at 5%. The sample size is fixed at \( N = 100000 \) and the number of data blocks \( K \) is selected in the set \{20, 50, 100, 200, 500, 1000, 2000, 5000\}.

For the null hypothesis, we generate \( Y \) and \( Z \) independently from distribution \( \mathcal{G}_1 \) and \( \mathcal{G}_2 \), respectively. Three different combinations of \( \mathcal{G}_1 \) and \( \mathcal{G}_2 \) are considered: (I) \( \mathcal{G}_1 \) and \( \mathcal{G}_2 \) are both \( \mathcal{N}(0_p, I_p) \), where \( I_p \) is the \( p \)-dimensional identity matrix.; (II) \( \mathcal{G}_1 \) is \( \mathcal{N}(0_p, I_p) \), for \( \mathcal{G}_2 \), its \( p \) components are i.i.d. from student-\( t \) distribution with 3 degrees of freedom, that is, for \( Z = (Z_1, \ldots, Z_p)^T \), \( Z_i \sim t(3) \) and \( Z_i \) and \( Z_j \) are independent for \( i, j = 1, \ldots, p \); (III) For both \( \mathcal{G}_1 \) and \( \mathcal{G}_2 \), their marginal components are i.i.d. from student-\( t \) distribution with 3 degrees of freedom. Scenario (I) and (III) are also employed in Székely, Rizzo and Bakirov (2007). The dimension \( p \) is chosen as 5, 10, 20 and 40 for each scenario.

Table 2.6 reports the empirical sizes of the tests based on the distributed distance covariance \( \text{dcov}^2_{N,K}(Y, Z) \). \( T_{\text{Var}} \) stands for the testing procedure based on the variance estimator \( \hat{\sigma}^2_{\beta,N,K} \) in (2.17) and \( T_{\text{PDB}} \) represents the test using the pseudo-distributed bootstrap. Table 2.6 shows that the empirical sizes of both methods are close to the nominal level 5% for all combinations.
of $p$ and $K$ under all three scenarios. Thus, these two tests for independence based on the
distributed distance covariance both have good control of Type-I error for a wide range of $K$.
In addition, the performance of these two methods is very comparable to each other. However,
for relatively large $K$, $T_{\text{PDB}}$ has slightly better control of the empirical sizes than $T_{\text{Var}}$.

To compare the powers of these two tests, we generate $p$-dimensional random vectors $Y \sim G_1$ and $Z \sim G_2$, and the same three different combinations of $G_1$ and $G_2$ are considered as in the simulation setups for the null hypothesis. For $Y = (Y_1, \ldots, Y_p)^T$ and $Z = (Z_1, \ldots, Z_p)^T$ under each scenario, we simulate $\text{cor}(Y_i, Z_j) = \varrho|i-j-p|$ for $i, j = 1, \ldots, p$ and $\varrho = 0.05, 0.1$ are considered. The dimension parameter $p$ is taken to be 5, 10, 20 and 40. Under these setups, $Y$ and $Z$ are dependent with each other. Two testing procedures based on the distributed distance covariance $\text{dcov}_{N,K}^2(Y, Z)$ are carried out to testing the independence between $Y$ and $Z$. Table 2.7 and 2.8 give the empirical powers of $T_{\text{Var}}$ and $T_{\text{PDB}}$.

From Table 2.7 and 2.8, it is clear that the empirical powers of these two distributed distance
covariance based tests decrease as the dimension $p$ increases. In addition, as the number of
data blocks $K$ increases, the empirical powers of the tests also decrease. This is due to the
increase in the variance of $\text{dcov}_{N,K}^2(Y, Z)$ when $K$ increases. This is the price we need to pay
for using the distributed distance covariance. The computing time and memory requirement
can be reduced by increasing the number of data blocks, however, this will result in the power
loss of the tests for independence. This coincides with the discussions on degenerate $T_N$ and
the asymptotic efficiency of $T_{N,K}$.

2.11.3 Real data analysis

In this section, we use the 2016 airline on-time performance data as a case study to illustrate
how our proposed distributed inference work for massive data. The data is publicly available
from the Bureau of Transportation Statistics website (https://www.bts.gov/). The dataset
consists of the arrival and departure details for total 5,617,658 commercial flights within the
USA in 2016. We are interested in the arrival delay variable which presents the difference in
minutes between scheduled and actual arrival time for each flight, and negative values stand for
early arrivals. After removing missing and extreme values, we arrive at 5,536,426 data points.
Table 2.6  Sizes of independence tests based on $d_{cov}^2_{N,K}(Y,Z)$. $T_{Var}$: test using variance estimation in (2.17); $T_{PDB}$: test using the pseudo-distributed bootstrap.

<table>
<thead>
<tr>
<th>$K$</th>
<th>$p = 5$</th>
<th>$p = 10$</th>
<th>$p = 20$</th>
<th>$p = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T_{Var}$</td>
<td>$T_{PDB}$</td>
<td>$T_{Var}$</td>
<td>$T_{PDB}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scenario I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.051</td>
<td>0.050</td>
<td>0.053</td>
<td>0.058</td>
</tr>
<tr>
<td>50</td>
<td>0.055</td>
<td>0.047</td>
<td>0.046</td>
<td>0.049</td>
</tr>
<tr>
<td>100</td>
<td>0.051</td>
<td>0.048</td>
<td>0.041</td>
<td>0.043</td>
</tr>
<tr>
<td>200</td>
<td>0.045</td>
<td>0.041</td>
<td>0.035</td>
<td>0.035</td>
</tr>
<tr>
<td>500</td>
<td>0.051</td>
<td>0.048</td>
<td>0.049</td>
<td>0.044</td>
</tr>
<tr>
<td>1000</td>
<td>0.060</td>
<td>0.052</td>
<td>0.046</td>
<td>0.042</td>
</tr>
<tr>
<td>2000</td>
<td>0.067</td>
<td>0.062</td>
<td>0.059</td>
<td>0.053</td>
</tr>
<tr>
<td>5000</td>
<td>0.063</td>
<td>0.050</td>
<td>0.070</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scenario II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.058</td>
<td>0.050</td>
<td>0.055</td>
<td>0.057</td>
</tr>
<tr>
<td>50</td>
<td>0.057</td>
<td>0.056</td>
<td>0.052</td>
<td>0.047</td>
</tr>
<tr>
<td>100</td>
<td>0.051</td>
<td>0.045</td>
<td>0.058</td>
<td>0.056</td>
</tr>
<tr>
<td>200</td>
<td>0.055</td>
<td>0.051</td>
<td>0.048</td>
<td>0.049</td>
</tr>
<tr>
<td>500</td>
<td>0.053</td>
<td>0.046</td>
<td>0.040</td>
<td>0.038</td>
</tr>
<tr>
<td>1000</td>
<td>0.046</td>
<td>0.044</td>
<td>0.043</td>
<td>0.043</td>
</tr>
<tr>
<td>2000</td>
<td>0.057</td>
<td>0.050</td>
<td>0.052</td>
<td>0.045</td>
</tr>
<tr>
<td>5000</td>
<td>0.068</td>
<td>0.057</td>
<td>0.075</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scenario III</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.056</td>
<td>0.056</td>
<td>0.051</td>
<td>0.061</td>
</tr>
<tr>
<td>50</td>
<td>0.056</td>
<td>0.053</td>
<td>0.051</td>
<td>0.055</td>
</tr>
<tr>
<td>100</td>
<td>0.047</td>
<td>0.042</td>
<td>0.051</td>
<td>0.045</td>
</tr>
<tr>
<td>200</td>
<td>0.055</td>
<td>0.051</td>
<td>0.058</td>
<td>0.056</td>
</tr>
<tr>
<td>500</td>
<td>0.052</td>
<td>0.052</td>
<td>0.061</td>
<td>0.060</td>
</tr>
<tr>
<td>1000</td>
<td>0.056</td>
<td>0.053</td>
<td>0.061</td>
<td>0.058</td>
</tr>
<tr>
<td>2000</td>
<td>0.053</td>
<td>0.044</td>
<td>0.069</td>
<td>0.062</td>
</tr>
<tr>
<td>5000</td>
<td>0.069</td>
<td>0.054</td>
<td>0.070</td>
<td>0.055</td>
</tr>
</tbody>
</table>
Table 2.7  Powers of independence tests based on \( dcoV_{N,K}^2(Y,Z) \) for \( \rho = 0.05 \).

<table>
<thead>
<tr>
<th>( K )</th>
<th>( T_{\text{var}} )</th>
<th>( T_{PDB} )</th>
<th>( T_{\text{var}} )</th>
<th>( T_{PDB} )</th>
<th>( T_{\text{var}} )</th>
<th>( T_{PDB} )</th>
<th>( T_{\text{var}} )</th>
<th>( T_{PDB} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p = 5 )</td>
<td>( p = 10 )</td>
<td>( p = 20 )</td>
<td>( p = 40 )</td>
<td>( p = 5 )</td>
<td>( p = 10 )</td>
<td>( p = 20 )</td>
<td>( p = 40 )</td>
<td>( p = 5 )</td>
</tr>
<tr>
<td>scenario I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1.000</td>
<td>1.000</td>
<td>0.973</td>
<td>0.955</td>
<td>0.607</td>
<td>0.586</td>
<td>0.238</td>
<td>0.225</td>
</tr>
<tr>
<td>50</td>
<td>0.996</td>
<td>0.995</td>
<td>0.790</td>
<td>0.765</td>
<td>0.349</td>
<td>0.345</td>
<td>0.161</td>
<td>0.176</td>
</tr>
<tr>
<td>100</td>
<td>0.939</td>
<td>0.941</td>
<td>0.557</td>
<td>0.549</td>
<td>0.203</td>
<td>0.199</td>
<td>0.125</td>
<td>0.109</td>
</tr>
<tr>
<td>200</td>
<td>0.764</td>
<td>0.756</td>
<td>0.371</td>
<td>0.389</td>
<td>0.126</td>
<td>0.121</td>
<td>0.089</td>
<td>0.084</td>
</tr>
<tr>
<td>500</td>
<td>0.413</td>
<td>0.404</td>
<td>0.193</td>
<td>0.189</td>
<td>0.091</td>
<td>0.092</td>
<td>0.081</td>
<td>0.081</td>
</tr>
<tr>
<td>1000</td>
<td>0.243</td>
<td>0.238</td>
<td>0.164</td>
<td>0.167</td>
<td>0.080</td>
<td>0.077</td>
<td>0.066</td>
<td>0.065</td>
</tr>
<tr>
<td>2000</td>
<td>0.137</td>
<td>0.136</td>
<td>0.095</td>
<td>0.104</td>
<td>0.075</td>
<td>0.074</td>
<td>0.065</td>
<td>0.063</td>
</tr>
<tr>
<td>5000</td>
<td>0.109</td>
<td>0.107</td>
<td>0.085</td>
<td>0.087</td>
<td>0.063</td>
<td>0.064</td>
<td>0.050</td>
<td>0.047</td>
</tr>
<tr>
<td>scenario II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1.000</td>
<td>0.999</td>
<td>0.870</td>
<td>0.872</td>
<td>0.429</td>
<td>0.409</td>
<td>0.157</td>
<td>0.160</td>
</tr>
<tr>
<td>50</td>
<td>0.977</td>
<td>0.968</td>
<td>0.558</td>
<td>0.565</td>
<td>0.239</td>
<td>0.236</td>
<td>0.121</td>
<td>0.128</td>
</tr>
<tr>
<td>100</td>
<td>0.856</td>
<td>0.855</td>
<td>0.338</td>
<td>0.346</td>
<td>0.185</td>
<td>0.185</td>
<td>0.114</td>
<td>0.120</td>
</tr>
<tr>
<td>200</td>
<td>0.624</td>
<td>0.593</td>
<td>0.241</td>
<td>0.233</td>
<td>0.157</td>
<td>0.153</td>
<td>0.091</td>
<td>0.091</td>
</tr>
<tr>
<td>500</td>
<td>0.341</td>
<td>0.343</td>
<td>0.187</td>
<td>0.179</td>
<td>0.097</td>
<td>0.093</td>
<td>0.072</td>
<td>0.069</td>
</tr>
<tr>
<td>1000</td>
<td>0.245</td>
<td>0.250</td>
<td>0.130</td>
<td>0.124</td>
<td>0.081</td>
<td>0.081</td>
<td>0.059</td>
<td>0.054</td>
</tr>
<tr>
<td>2000</td>
<td>0.155</td>
<td>0.158</td>
<td>0.089</td>
<td>0.091</td>
<td>0.067</td>
<td>0.067</td>
<td>0.053</td>
<td>0.054</td>
</tr>
<tr>
<td>5000</td>
<td>0.092</td>
<td>0.095</td>
<td>0.068</td>
<td>0.069</td>
<td>0.059</td>
<td>0.060</td>
<td>0.046</td>
<td>0.051</td>
</tr>
<tr>
<td>scenario III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.996</td>
<td>0.995</td>
<td>0.719</td>
<td>0.681</td>
<td>0.359</td>
<td>0.309</td>
<td>0.130</td>
<td>0.147</td>
</tr>
<tr>
<td>50</td>
<td>0.908</td>
<td>0.893</td>
<td>0.404</td>
<td>0.376</td>
<td>0.200</td>
<td>0.205</td>
<td>0.088</td>
<td>0.089</td>
</tr>
<tr>
<td>100</td>
<td>0.713</td>
<td>0.699</td>
<td>0.264</td>
<td>0.285</td>
<td>0.152</td>
<td>0.140</td>
<td>0.081</td>
<td>0.078</td>
</tr>
<tr>
<td>200</td>
<td>0.452</td>
<td>0.448</td>
<td>0.155</td>
<td>0.163</td>
<td>0.132</td>
<td>0.136</td>
<td>0.064</td>
<td>0.063</td>
</tr>
<tr>
<td>500</td>
<td>0.262</td>
<td>0.254</td>
<td>0.115</td>
<td>0.122</td>
<td>0.103</td>
<td>0.106</td>
<td>0.062</td>
<td>0.063</td>
</tr>
<tr>
<td>1000</td>
<td>0.167</td>
<td>0.163</td>
<td>0.104</td>
<td>0.104</td>
<td>0.071</td>
<td>0.076</td>
<td>0.062</td>
<td>0.062</td>
</tr>
<tr>
<td>2000</td>
<td>0.139</td>
<td>0.143</td>
<td>0.072</td>
<td>0.071</td>
<td>0.064</td>
<td>0.065</td>
<td>0.051</td>
<td>0.052</td>
</tr>
<tr>
<td>5000</td>
<td>0.092</td>
<td>0.098</td>
<td>0.057</td>
<td>0.055</td>
<td>0.047</td>
<td>0.046</td>
<td>0.046</td>
<td>0.047</td>
</tr>
</tbody>
</table>
Table 2.8 Powers of independence tests based on $d cov_{N,K}^2(Y,Z)$ for $\varrho = 0.1$.

<table>
<thead>
<tr>
<th>$K$</th>
<th>$p = 5$</th>
<th>$p = 10$</th>
<th>$p = 20$</th>
<th>$p = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T_{Var}$</td>
<td>$T_{PDB}$</td>
<td>$T_{Var}$</td>
<td>$T_{PDB}$</td>
</tr>
<tr>
<td>scenario I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>50</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>100</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>200</td>
<td>1.000</td>
<td>1.000</td>
<td>0.998</td>
<td>0.999</td>
</tr>
<tr>
<td>500</td>
<td>1.000</td>
<td>1.000</td>
<td>0.931</td>
<td>0.933</td>
</tr>
<tr>
<td>1000</td>
<td>0.994</td>
<td>0.993</td>
<td>0.710</td>
<td>0.704</td>
</tr>
<tr>
<td>2000</td>
<td>0.871</td>
<td>0.873</td>
<td>0.427</td>
<td>0.435</td>
</tr>
<tr>
<td>5000</td>
<td>0.535</td>
<td>0.527</td>
<td>0.256</td>
<td>0.249</td>
</tr>
<tr>
<td>scenario II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>50</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>100</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>200</td>
<td>1.000</td>
<td>1.000</td>
<td>0.982</td>
<td>0.983</td>
</tr>
<tr>
<td>500</td>
<td>1.000</td>
<td>1.000</td>
<td>0.823</td>
<td>0.814</td>
</tr>
<tr>
<td>1000</td>
<td>0.964</td>
<td>0.965</td>
<td>0.575</td>
<td>0.564</td>
</tr>
<tr>
<td>2000</td>
<td>0.750</td>
<td>0.756</td>
<td>0.341</td>
<td>0.341</td>
</tr>
<tr>
<td>5000</td>
<td>0.406</td>
<td>0.408</td>
<td>0.164</td>
<td>0.158</td>
</tr>
<tr>
<td>scenario III</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>50</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>100</td>
<td>1.000</td>
<td>1.000</td>
<td>0.987</td>
<td>0.988</td>
</tr>
<tr>
<td>200</td>
<td>1.000</td>
<td>1.000</td>
<td>0.891</td>
<td>0.897</td>
</tr>
<tr>
<td>500</td>
<td>0.987</td>
<td>0.985</td>
<td>0.551</td>
<td>0.548</td>
</tr>
<tr>
<td>1000</td>
<td>0.840</td>
<td>0.835</td>
<td>0.289</td>
<td>0.302</td>
</tr>
<tr>
<td>2000</td>
<td>0.599</td>
<td>0.606</td>
<td>0.173</td>
<td>0.170</td>
</tr>
<tr>
<td>5000</td>
<td>0.273</td>
<td>0.273</td>
<td>0.126</td>
<td>0.127</td>
</tr>
</tbody>
</table>
In this case study, we treat the dataset as an univariate i.i.d. sample of size \( N = 5,536,426 \). The parameter of interest is the dispersion parameter \( \theta = \text{E}|X_i - X_j| \), which can be estimated by the Gini’s mean difference

\[
U_N = \left\{ N(N - 1) \right\}^{-1} \sum_{1 \leq i \neq j \leq N} |X_i - X_j|.
\]

Due to the massive size of the dataset, we consider using the distributed version of \( U_N \) to do inferences on \( \theta \). Denote \( U_{N,K} \) as the distributed version of the Gini’s mean difference with the entire data randomly divided evenly into \( K \) blocks. Table 2.9 presents the values of \( U_{N,K} \) and its jackknife variance estimator \( N^{-1}S^2_{U_{N,K}} \) for different \( K \) along with their computing time.

According to Table 2.9, it is clear that the computing time of both \( T_{N,K} \) and \( S^2_{U_{N,K}} \) is linearly decreasing in \( K \), and the value of \( S^2_{U_{N,K}} \) is increasing in \( K \). The computing time for \( S^2_{U_{N,K}} \) when \( K = 50000 \) is about a factor of 20000 less than the case when \( K = 1 \), however, the estimated variance only has 0.3% increase. This result indicates that for the Gini’s mean difference, the distributed statistic \( T_{N,K} \) can maintain sufficient estimation efficiency while decrease computing time massively.

Next we applied the distributed bootstrap, the pseudo-distributed bootstrap, the BLB and the SDB on this dataset to estimate the 95% confidence interval width of \( \theta \) based on \( T_{N,K} \) with different number of blocks \( K \in \{50, 200, 500, 2000, 5000, 20000\} \). Same as in the simulation studies, we chose the subset size \( n = N/K \) for the BLB and the SDB, and the number of resamples \( B \) was chosen as 100 for the BLB. Each method was allowed to run for 60 minutes and the time evolution of the estimated confidence interval width is shown in Figure 2.5. The BLB is the slowest to converge and the pseudo-distributed bootstrap is the fastest to stabilized. Thus, the pseudo-distributed bootstrap has advantage for small time budgets. For relatively large \( K \), these four methods tend to obtain similar estimates of the confidence interval width.

### 2.12 Conclusion

We considered distributed inferences for massive data with a concentration on a general type of statistics \( T_N \). To conquer computational issues, we formulated the distributed statistics \( T_{N,K} \) and studied its statistical efficiency under a general approach. We also focused on
Table 2.9  $U_{N,K}$ and $S_{UN,K}^2$ for airline on-time dataset along with computing time (in seconds) given in the parentheses.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_{N,K}$</td>
<td>29.94390</td>
<td>29.94389</td>
<td>29.94389</td>
<td>29.94393</td>
<td>29.94394</td>
<td>29.94391</td>
<td>29.94394</td>
<td>29.94398</td>
</tr>
<tr>
<td></td>
<td>(18014.36)</td>
<td>(8887.95)</td>
<td>(3419.52)</td>
<td>(1724.94)</td>
<td>(857.74)</td>
<td>(328.38)</td>
<td>(165.78)</td>
<td>(82.99)</td>
</tr>
<tr>
<td>$S_{UN,K}^2$</td>
<td>3625.759</td>
<td>3625.759</td>
<td>3625.761</td>
<td>3625.771</td>
<td>3625.774</td>
<td>3625.810</td>
<td>3625.839</td>
<td>3625.893</td>
</tr>
<tr>
<td></td>
<td>(101674.42)</td>
<td>(51434.09)</td>
<td>(21183.67)</td>
<td>(9634.81)</td>
<td>(4969.88)</td>
<td>(1894.72)</td>
<td>(947.36)</td>
<td>(477.83)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>5000</th>
<th>10000</th>
<th>20000</th>
<th>50000</th>
<th>100000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_{N,K}$</td>
<td>29.94370</td>
<td>29.94377</td>
<td>29.94385</td>
<td>29.94396</td>
<td>29.94376</td>
<td>29.94320</td>
<td>29.94563</td>
<td>29.94439</td>
</tr>
<tr>
<td></td>
<td>(32.96)</td>
<td>(16.17)</td>
<td>(8.17)</td>
<td>(3.31)</td>
<td>(1.67)</td>
<td>(0.88)</td>
<td>(0.37)</td>
<td>(0.22)</td>
</tr>
<tr>
<td>$S_{UN,K}^2$</td>
<td>3625.866</td>
<td>3625.933</td>
<td>3626.034</td>
<td>3626.704</td>
<td>3627.682</td>
<td>3629.835</td>
<td>3636.678</td>
<td>3646.08637</td>
</tr>
<tr>
<td></td>
<td>(189.43)</td>
<td>(94.95)</td>
<td>(47.97)</td>
<td>(19.31)</td>
<td>(10.03)</td>
<td>(5.92)</td>
<td>(4.73)</td>
<td>(6.40)</td>
</tr>
</tbody>
</table>
Figure 2.5  Time evolution of estimated confidence interval width.
the degenerate case when the linear term in $T_N$ vanishes. Furthermore, regarding the computational advantages of $T_{N,K}$ against $T_N$, we proposed bootstrap algorithms to approximate the distribution function of $T_{N,K}$. Theoretical properties of the proposed methods were established and we applied them on the independence test using the distance covariance regarding massive datasets. Finally, we used numerical studies to illustrate our theoretical results.

In practice, an important issue of the distributed approaches is the choice of the number of blocks $K$. Increasing $K$ would decrease the computational cost, however, this leads to a loss in statistical efficiency. This topic has been touched in Section 2.4, where storage and memory requirement are considered. However, it is still an issue on how to select $K$ in practice that balances the computing time and statistical efficiency. Instead of giving fixed time budget, another problem is to minimize computing time subject to attaining certain statistical efficiency. This is similar to a sample size determination problem. We leave these for future study. We focused our development of the distributed approaches on i.i.d. data. It is still unclear how these methods work for dependent data. Furthermore, in Kleiner, et al. (2014), higher order accuracy of BLB has been studied. Thus, it is of interest to investigate higher order correctness and convergence rates of our proposed distributed approaches, and we plan to investigate this in the future.

### 2.13 Appendix

In this section, we present the proofs of main theorems in this paper. Before stating the proofs of the main theorems, we give several lemmas that will be frequently used later. The first lemma is a generalization of Esseen’s inequality which we may refer to Theorem 5.7 in Petrov (1998).

**Lemma 2.1.** Let $Y_1, \ldots, Y_N$ be independent random variables. For $i = 1, \ldots, N$, $E(Y_i) = 0$ and $E|Y_i|^{2+\delta} < \infty$ for some positive $\delta \leq 1$. Then

$$
\sup_{x \in \mathbb{R}} \left| P \left( B_N^{-1/2} \sum_{i=1}^N Y_i \leq x \right) - \Phi(x) \right| \leq C_1 B_N^{-1-\delta/2} \sum_{i=1}^N E|Y_i|^{2+\delta},
$$

where $B_N = \sum_{i=1}^N \text{Var}(Y_i)$ and $C_1$ is a positive constant.
The second lemma is a modified version of Marcinkiewicz-Zygmund strong law of larger numbers and its proof can be found in Liu (1988).

**Lemma 2.2.** Let $Y_1, \ldots, Y_N$ be independent random variables with $E|Y_i|^\eta + \varepsilon < \infty$ for some positive $\eta$, $0 < \eta < 2$ and $\varepsilon > 0$. Then as $N \to \infty$,

$$N^{-1/\eta} \sum_{i=1}^{N} (Y_i - a_i) \to 0$$

almost surely, where $a_i = EY_i$ if $\eta \geq 1$ and $a_i = 0$ otherwise.

2.13.1 Proof of Theorem 2.4

**Proof.** According to Lemma 2.1, we know that

$$\sup_{x \in \mathbb{R}} |P(V_N \leq x) - \Phi(x)| \leq N^{-\delta/2} C_1\sigma^{-2-\delta}\alpha(X_1; F)^{1+\delta},$$

where $V_N = N^{-1/2}\sigma^{-1}\alpha(X_i; F)$. Denote $\Delta_N = N^{-3/2}\sigma^{-1}\sum_{1 \leq i < j \leq N} \beta(X_i, X_j; F)$ and $\Delta_{N,K} = N^{-1/2}\sigma^{-1}\sum_{k=1}^{K} n_k^{-1}\sum_{1 \leq i < j \leq n_k} \beta(X_{k,i}, X_{k,j}; F)$, then by Cauchy-Schwarz inequality,

$$P(\Delta_N \geq (\ln N)^{-1}) = O(N^{-1}(\ln N)^2),$$

$$P(\Delta_{N,K} \geq (\ln N)^{-1}) = O(KN^{-1}(\ln N)^2).$$

If $E(R_N) = b_1N^{-\tau_1} + o(N^{-\tau_1})$ and $\text{Var}(R_N) = o(N^{-\tau_2})$, under Condition 2.3 and the assumption $K = O(N^{r'})$, $E(R_{N,K}) = O(N^{\tau_1 r' - \tau_1})$ and $\text{Var}(R_{N,K}) = o(N^{\tau_2 r' - \tau'_1 - \tau_2})$, again by Cauchy-Schwarz inequality,

$$P(|R_N| \geq N^{-1/2}(\ln N)^{-1}) = o(N^{1-\tau_2}(\ln N)^2),$$

$$P(|R_{N,K}| \geq N^{-1/2}(\ln N)^{-1}) = o(N^{\tau_2 r' - \tau' - \tau_2 + 1}(\ln N)^2),$$

under the condition that $\tau' < 1 - 1/(2\tau_1)$. Finally, by the fact that

$$\sup_{x \in \mathbb{R}, \ |t|<(\ln N)^{-1}} |\Phi(x+t) - \Phi(x)| = o(1),$$

we complete the proof of Theorem 2.4.
2.13.2 Proof of Theorem 2.6

Proof. When $M$ is finite, the result can be easily obtained from independence between each subset and the classic limit theorem for degenerate U-statistic (Serfling, 1980).

Now we focus on the case when $K \to \infty$. Without loss of generality, assume $\theta = 0$. Denote $T_{N,K}^{(k)} = \Delta_{N,K}^{(k)} + R_{N,K}^{(k)}$ where $\Delta_{N,K}^{(k)} = n_k^{-2} \sum_{1 \leq i < j \leq n_k} \beta(X_{k,i}, X_{k,j}; F)$. Under the condition that $R_{N,K} = o_p(K^{1/2}N^{-1})$, it is sufficient to prove that $2^{1/2}K^{-1/2}N\sigma_{\beta}^{-1}\Delta_{N,K}$ convergence to standard normal distribution as $K, N \to \infty$, here

$$
\Delta_{N,K} = N^{-1} \sum_{k=1}^{K} n_k \Delta_{N,K}^{(k)} = N^{-1} \sum_{k=1}^{K} n_k^{-1} \sum_{1 \leq i < j \leq n_k} \beta(X_{k,i}, X_{k,j}; F).
$$

Rewrite

$$
2^{1/2}K^{-1/2}N\sigma_{\beta}^{-1}\Delta_{N,K} = \sum_{k=1}^{K} 2^{1/2}K^{-1/2}n_k\sigma_{\beta}^{-1}\Delta_{N,K}^{(k)} = \sum_{k=1}^{K} \Delta_{N,K}^{(k)},
$$

where $\Delta_{N,K}^{(k)} = 2^{1/2}K^{-1/2}n_k\sigma_{\beta}^{-1}\Delta_{N,K}^{(k)}$ satisfies $E(\Delta_{N,K}^{(k)}) = 0$ and $\text{Var}(\Delta_{N,K}^{(k)}) = K^{-1}(1-n_k^{-1})$ for $k = 1, \ldots, K$. As $\Delta_{N,K}^{(k)}$, $k = 1, \ldots, K$ are independent, it is enough to check the Lindeberg condition for $\Delta_{N,K}^{(k)}$. Note that $s_{N,K}^2 \equiv \sum_{k=1}^{K} \text{Var}(\Delta_{N,K}^{(k)}) = 1 - K^{-1} \sum_{k=1}^{K} n_k^{-1}$, then for every $\varepsilon > 0$,

$$
\begin{align*}
\sum_{k=1}^{K} & E \left[ \left( \Delta_{N,K}^{(k)} \right)^2 \mathbf{1} \left\{ \left| \Delta_{N,K}^{(k)} \right| > \varepsilon s_{N,K} \right\} \right] \\
= & s_{N,K}^{-2} \sum_{k=1}^{K} K^{-1} \left( 1 - n_k^{-1} \right) E \left[ \left| K^{1/2} \Delta_{N,K}^{(k)} \right|^2 \mathbf{1} \left\{ \left| K^{1/2} \Delta_{N,K}^{(k)} \right| > \varepsilon s_{N,K} K^{1/2} \right\} \right] \\
\leq & s_{N,K}^{-2} \sum_{k=1}^{K} K^{-1} \left( 1 - n_k^{-1} \right) \left( \varepsilon s_{N,K} K^{1/2} \right)^{-\delta'} E \left[ \left| K^{1/2} \Delta_{N,K}^{(k)} \right|^{2+\delta'} \right] \\
\leq & C_2 s_{N,K}^{-2} \sum_{k=1}^{K} K^{-1} \left( 1 - n_k^{-1} \right) \left( \varepsilon s_{N,K} K^{1/2} \right)^{-\delta'} \\
\to & 0, \quad \text{as} \quad K \to \infty.
\end{align*}
$$

The next-to-last line is from the moment inequalities of U-statistics (Koroljuk and Borovskich, 1994) and $C_2$ is a positive constant. Thus we finish the proof of Theorem 2.6. \qed
2.13.3 Proof of Theorem 2.7

Proof. According to the proof of Theorem 2.4, under the conditions assumed in Theorem 2.4, we have

\[ \sup_{x \in \mathbb{R}} \left| P \left\{ N^{1/2} \sigma^{-1}_\alpha (T_{N,K} - \theta) \leq x \right\} - \Phi(x) \right| = o(1), \]

thus, it is sufficient to show that under the conditions assumed in the theorem

\[ \sup_{x \in \mathbb{R}} \left| P \left\{ N^{1/2} \hat{\sigma}^{-1}_{\alpha,N} (T_{N,K}^* - \theta_N^*) \leq x | F_N \right\} - \Phi(x) \right| = o_p(1). \]  

(2.18)

We proceed to show that \( E \{ |\alpha(X_1^*; F_N)|^{2+\delta} | F_N \} \) is bounded in probability. Note that

\[ E \left\{ |\alpha(X_1^*; F_N)|^{2+\delta} \right\} \]

\[ = N^{-1} \sum_{i=1}^{N} |\alpha(X_i; F_N)|^{2+\delta} \]

\[ \leq N^{-1} \sum_{i=1}^{N} C_\delta \left\{ |\alpha(X_i; F_N) - \alpha(X_i; F)|^{2+\delta} + |\alpha(X_i; F)|^{2+\delta} \right\}, \]

By the condition that \( \sup_{x \in \mathbb{R}} |\alpha(x; F_N) - \alpha(x; F)| = o_p(1) \) and the strong law of large numbers, we have \( E \{ |\alpha(X_1^*; F_N)|^{2+\delta} | F_N \} = o_p(1) \). Similarly, we can show that \( E \{ |\beta(X_1^*; X_2^*; F_N)|^{2} | F_N \} < \infty \) in probability. Thus, under the condition \( P \left\{ |R_{N,K}^*| \geq N^{-1/2} (\ln N)^{-1} | F_N \right\} = o_p(1) \), by carrying out similar procedures in the proof of Theorem 2.4, we can prove (2.18) which leads to the result that

\[ \sup_{x \in \mathbb{R}} \left| P \left\{ N^{1/2} \hat{\sigma}^{-1}_{\alpha,N} (T_{N,K}^* - \theta_N^*) \leq x | F_N \right\} - P \left\{ N^{1/2} \sigma^{-1}_\alpha (T_{N,K} - \theta) \leq x \right\} \right| = o_p(1). \]

In addition,

\[ \hat{\sigma}^2_{\alpha,N} - \sigma^2_\alpha = E \left\{ \alpha^2(X_1^*; F_N) | F_N \right\} - \sigma^2_\alpha \]

\[ = N^{-1} \sum_{i=1}^{N} \alpha^2(X_i; F_N) - \sigma^2_\alpha \]

\[ = N^{-1} \sum_{i=1}^{N} \left\{ \alpha(X_i; F_N) - \alpha(X_i; F) \right\}^2 + N^{-1} \sum_{i=1}^{N} 2\alpha(X_i; F) \left\{ \alpha(X_i; F_N) - \alpha(X_i; F) \right\} \]

\[ + N^{-1} \sum_{i=1}^{N} \left\{ \alpha^2(X_i) - \sigma^2_\alpha \right\} \]

\[ = o_p(1). \]
2.13.4 Proof of Theorem 2.8

Proof. The proof of Theorem 2.8 is similar with the one of Theorem 2.7 and thus is omitted here.

2.13.5 Proof of Corollary 2.1

Proof. Let \( \mathcal{T}^{(k)}_{N,K} = \mathcal{T}^{(k)}_{N,K} - N^{1/2}K^{-1/2}\theta \). Denote

\[
\theta_k = \mathbb{E}\left( \mathcal{T}^{(k)}_{N,K} \right) = N^{-1/2}K^{1/2}(n_k - NK^{-1})\theta + N^{-1/2}K^{1/2}n_k\mathbb{E}\left( R^{(k)}_{N,K} \right)
\]

and \( \sigma_k^2 = \text{Var}\left( \mathcal{T}^{(k)}_{N,K} \right) = N^{-1}K n_k\sigma_n^2 \{1 + o(1)\} \) for \( k = 1, \ldots, K \), then

\[
\bar{\theta} = K^{-1} \sum_{k=1}^{K} \theta_k = N^{-1/2}K^{-1/2} \sum_{k=1}^{K} n_k\mathbb{E}\left( R^{(k)}_{N,K} \right).
\]

Denote \( \mathcal{T}^{*(k)}_{N,K} = \mathcal{T}^{(k)}_{N,K} - N^{1/2}K^{-1/2}\theta \), then \( \mathcal{T}^{*(k)}_{N,K}, k = 1, \ldots, K \) are i.i.d. from \( F_{K,T}(x + N^{1/2}K^{-1/2}\theta) \). Since \( \mathbb{E}|\alpha(X_1; F)|^{2+\delta} < \infty, \mathbb{E}|\beta(X_1, X_2; F)|^{2+\delta} < \infty \) and \( \sup\mathbb{E}|n_k^{1/2} R^{(k)}_{N,K}|^{2+\delta} < \infty \), we have \( \sup\mathbb{E}\left( \mathcal{T}^{(k)}_{N,K} \right)^{2+\delta} < \infty \). This immediately results in

\[
\mathbb{E}\left\{ \left| \mathcal{T}^{*(k)}_{N,K} \right|^{2+\delta_1} \right| F_{K,T} \} = K^{-1} \sum_{k=1}^{K} \left| \mathcal{T}^{(k)}_{N,K} \right|^{2+\delta_1} < \infty
\]

with probability 1 for some \( \delta_1 \) satisfies \( 0 < \delta_1 < \delta \) (Lemma 2.2). Define

\[
\mathcal{S}^{*}_{N,K} = K^{-1} \sum_{k=1}^{K} \mathcal{T}^{*(k)}_{N,K} = \mathcal{T}^{*}_{N,K} - N^{1/2}K^{-1/2}\theta,
\]

then \( \mathcal{S}^{*}_{N,K} - \mathcal{S}_{N,K} = \mathcal{T}^{*}_{N,K} - N^{1/2}K^{-1/2}\mathcal{T}_{N,K} \) where \( \mathcal{S}_{N,K} = K^{-1} \sum_{k=1}^{K} \mathcal{T}^{(k)}_{N,K} \). Denote \( \bar{s}^2 = K^{-1} \sum_{k=1}^{K} \sigma_k^2 \) and

\[
\bar{s}^2 = K^{-1} \sum_{k=1}^{K} \left( \mathcal{S}^{(k)}_{N,K} - \mathcal{S}_{N,K} \right)^2,
\]

then \( \sup n^{-1/2} K^{-1/2} |n_k - NK^{-1}| \to 0 \) implies that

\[
s^2 - \bar{s}^2 = K^{-1} \sum_{k=1}^{K} \left\{ \left( \mathcal{S}^{(k)}_{N,K} - \theta_k + \bar{\theta} - \bar{\theta} - \mathcal{S}_{N,K} \right)^2 - \sigma_k^2 \right\} \to 0
\]
almost surely. According to Lemma 2.1,

\[
\sup_{x \in \mathbb{R}} \left| P \left\{ K^{1/2} \bar{s}^{-1} \left( T^{(k)}_{N,K} - N^{1/2} K^{-1/2} T_{N,K} \right) \leq x \right| F_{K,T} \right\} - \Phi(x) \right| \\
\leq C_1 \bar{s}^{-2-\delta_1} K^{-\delta_1/2} E \left\{ \left| \bar{\xi}_{N,K} - \bar{\xi}_{N,K} \right|^{2+\delta_1} |F_{K,T}| \right\} \\
\leq C_2 \bar{s}^{-2-\delta_1} K^{-\delta_1/2} \left[ E \left\{ \left| \bar{\xi}_{N,K} \right|^{2+\delta_1} \right| F_{K,T} \right\} + \left| \bar{\xi}_{N,K} \right|^{2+\delta_1} \right] \\
\to 0
\]

almost surely as \( K \to \infty \). It has been shown in the proof of Theorem 2.4 that

\[
\sup_{x \in \mathbb{R}} \left| P \left\{ N^{1/2} \sigma_{\alpha}^{-1} (T_{N,K} - \theta) \leq x \right| - \Phi(x) \right| = o(1).
\]

Combine this with the fact that \( \bar{s}^2 - \sigma_{\alpha}^2 \to 0 \) with probability 1, we complete the proof of this corollary.

\[ \Box \]

### 2.13.6 Proof of Corollary 2.2

**Proof.** Use similar techniques as in the proof of Corollary 2.1, let \( \bar{\xi}^{(k)}_{N,K} = T^{(k)}_{N,K} - N K^{-1} \theta = n_k(T^{(k)}_{N,K} - \theta) \), denote \( \theta_k = E(\bar{\xi}^{(k)}_{N,K}) = n_k E(R^{(k)}_{N,K}) \) and \( \sigma_k^2 = Var(\bar{\xi}^{(k)}_{N,K}) = 2^{-1} \sigma^{2} \{ 1 + o(1) \} \) for \( k = 1, \ldots, K \), then \( \bar{\xi}_{N,K} = K^{-1} \sum_{k=1}^{K} \xi^{(k)}_{N,K} = N K^{-1} (T_{N,K} - \theta) \) and \( \bar{\theta} = K^{-1} \sum_{k=1}^{K} \theta_k = K^{-1} \sum_{k=1}^{K} n_k E(R^{(k)}_{N,K}) \).

Under the conditions that \( E|\beta(X_1, X_2; F)|^{2+\delta'} < \infty \) and \( \sup_k E \left| n_k R^{(k)}_{N,K} \right|^{2+\delta'} < \infty \), we obtain \( \sup_k E \left| n_k (T^{(k)}_{N,K} - \theta) \right|^{2+\delta'} < \infty \), which in turn implies that

\[
\sup_{x \in \mathbb{R}} \left| P \left\{ N K^{-1/2} \bar{\sigma}^{-1} (T_{N,K} - \theta) \leq x \right| - \Phi(x) \right| = o(1),
\]

where \( \bar{\sigma}^2 = K^{-1} \sum_{k=1}^{K} \sigma_k^2 \).

Denote \( \bar{\xi}^{(k)}_{N,K} = T^{(k)}_{N,K} - N K^{-1} \theta \), then \( \bar{\xi}^{(k)}_{N,K}, k = 1, \ldots, K \) are i.i.d. from \( F_{K,T}(x + N K^{-1} \theta) \).

As \( \sup_k E \left( \left| \bar{\xi}^{(k)}_{N,K} \right|^{2+\delta_2} \right) < \infty \), this immediately results in

\[
E \left\{ \left| \bar{\xi}^{(k)}_{N,K} \right|^{2+\delta_2} \left| F_{K,T} \right\} = K^{-1} \sum_{k=1}^{K} \left| \bar{\xi}^{(k)}_{N,K} \right|^{2+\delta_2} < \infty
\]

because \( K \to \infty \).
with probability 1 and \( \delta_2 \) satisfies \( 0 < \delta_2 < \delta' \) (Lemma 2.2). Define

\[
\bar{T}_{N,K} = K^{-1} \sum_{k=1}^{K} \bar{T}_{N,K}^{(k)} = T_{N,K} - NK^{-1} \theta,
\]

then \( \bar{T}_{N,K} - \bar{T}_{N,K} = T_{N,K} - NK^{-1} \bar{T}_{N,K} \). Denote \( s^2 = K^{-1} \sum_{k=1}^{K} \left( \bar{T}_{N,K}^{(k)} - \bar{T}_{N,K} \right)^2 \), then \( s^2 - \sigma^2 = K^{-1} \sum_{k=1}^{K} \left( \bar{T}_{N,K}^{(k)} - \bar{T}_{N,K} - \theta \bar{k} + \theta_{k} + \bar{\theta} - \bar{T}_{N,K} \right)^2 - \sigma_k^2 \),

which convergence to 0 almost surely. According to Lemma 2.1,

\[
\sup_{x \in \mathbb{R}} \left| P \left\{ K^{1/2} s^{-1} (T_{N,K}^* - NK^{-1} T_{N,K}) \leq x \right| F_{K,T} \right) - \Phi(x) \right| \leq C_1 s^{-2 - \delta_2} K^{-\delta_2/2} E \left\{ \left| \bar{T}_{N,K}^* - \bar{T}_{N,K} \right|^{2+\delta_2} |F_{K,T}| \right) \rightarrow 0
\]

almost surely as \( K \rightarrow \infty \), thus we finish the proof of Corollary 2.2.

\[\square\]

### 2.13.7 Proof of Theorem 2.9

**Proof.** Under the null hypothesis, \( E \left( \hat{\sigma}_{\beta,N,K}^2 \right) = \sigma_{\beta}^2 \), thus it is sufficient to show that \( \text{Var} \left( \hat{\sigma}_{\beta,N,K}^2 \right) = o(1) \) as \( N \rightarrow \infty \).

Mentioned that

\[
\text{Var} \left( \hat{\sigma}_{\beta,N,K}^2 \right) = \text{Var} \left[ 4 \left\{ N^{-1} \sum_{k=1}^{K} n_k \text{dcov}^2_{N,K}(Y,Y) \right\} \left\{ N^{-1} \sum_{k=1}^{K} n_k \text{dcov}^2_{N,K}(Z,Z) \right\} \right]
\]

\[
= 16 \text{E} \left\{ N^{-1} \sum_{k=1}^{K} n_k \text{dcov}^2_{N,K}(Y,Y) \right\}^2 \text{E} \left\{ N^{-1} \sum_{k=1}^{K} n_k \text{dcov}^2_{N,K}(Z,Z) \right\}^2
\]

\[
- 16 \text{dcov}^4(Y,Y) \text{dcov}^4(Z,Z)
\]

\[
= 16 N^{-2} \sum_{k=1}^{K} n_k^2 \text{Var} \left\{ \text{dcov}^2_{N,K}(Y,Y) \right\} \text{dcov}^4(Z,Z)
\]

\[
+ 16 N^{-2} \sum_{k=1}^{K} n_k^2 \text{Var} \left\{ \text{dcov}^2_{N,K}(Z,Z) \right\} \text{dcov}^4(Y,Y)
\]

\[
+ 16 N^{-4} \sum_{k=1}^{K} n_k^2 \text{Var} \left\{ \text{dcov}^2_{N,K}(Y,Y) \right\} \sum_{k=1}^{K} n_k^2 \text{Var} \left\{ \text{dcov}^2_{N,K}(Z,Z) \right\}.
\]
Since $\text{dcov}^2_{N,K}(Y,Y)$ and $\text{dcov}^2_{N,K}(Z,Z)$ are U-statistics of degree 4, under the condition that $E\|Y\|^4_p + E\|Z\|^4_q < \infty$, $\text{Var}\left\{\text{dcov}^2_{N,K}(Y,Y)\right\} = O(n_k^{-1})$ and $\text{Var}\left\{\text{dcov}^2_{N,K}(Z,Z)\right\} = O(n_k^{-1})$, these two results immediately leads to $\text{Var}\left(\hat{\sigma}^2_{\beta,N,K}\right) = o(1)$ and the proof is complete.
Table 2.10  Coverage rates and confidence interval widths (in parentheses) of the 95% confidence intervals with 10 seconds time budget for Gamma data.

<table>
<thead>
<tr>
<th></th>
<th>K</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>500</td>
<td>1000</td>
</tr>
<tr>
<td>PDB</td>
<td>0.924</td>
<td>0.936</td>
<td>0.940</td>
<td>0.942</td>
<td>0.942</td>
<td>0.944</td>
</tr>
<tr>
<td></td>
<td>(0.01940)</td>
<td>(0.01998)</td>
<td>(0.02023)</td>
<td>(0.02031)</td>
<td>(0.02033)</td>
<td>(0.02041)</td>
</tr>
<tr>
<td>DB</td>
<td>0.928</td>
<td>0.938</td>
<td>0.942</td>
<td>0.946</td>
<td>0.942</td>
<td>0.946</td>
</tr>
<tr>
<td></td>
<td>(0.01963)</td>
<td>(0.02002)</td>
<td>(0.02024)</td>
<td>(0.02018)</td>
<td>(0.02022)</td>
<td>(0.02011)</td>
</tr>
<tr>
<td>BLB</td>
<td>NA</td>
<td>0.934</td>
<td>0.936</td>
<td>0.926</td>
<td>0.930</td>
<td>0.930</td>
</tr>
<tr>
<td></td>
<td>(NA)</td>
<td>(0.01951)</td>
<td>(0.01948)</td>
<td>(0.01938)</td>
<td>(0.01925)</td>
<td>(0.01899)</td>
</tr>
<tr>
<td>SDB</td>
<td>0.926</td>
<td>0.938</td>
<td>0.940</td>
<td>0.948</td>
<td>0.944</td>
<td>0.948</td>
</tr>
<tr>
<td></td>
<td>(0.01964)</td>
<td>(0.02003)</td>
<td>(0.02024)</td>
<td>(0.02024)</td>
<td>(0.02040)</td>
<td>(0.02036)</td>
</tr>
</tbody>
</table>

Table 2.11  Coverage rates and confidence interval widths (in parentheses) of the 95% confidence intervals with 10 seconds time budget for Poisson data.

<table>
<thead>
<tr>
<th></th>
<th>K</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>500</td>
<td>1000</td>
</tr>
<tr>
<td>PDB</td>
<td>0.921</td>
<td>0.938</td>
<td>0.941</td>
<td>0.945</td>
<td>0.949</td>
<td>0.949</td>
</tr>
<tr>
<td></td>
<td>(0.01986)</td>
<td>(0.02032)</td>
<td>(0.02046)</td>
<td>(0.02053)</td>
<td>(0.02059)</td>
<td>(0.02066)</td>
</tr>
<tr>
<td>DB</td>
<td>0.944</td>
<td>0.932</td>
<td>0.946</td>
<td>0.944</td>
<td>0.948</td>
<td>0.952</td>
</tr>
<tr>
<td></td>
<td>(0.01987)</td>
<td>(0.02017)</td>
<td>(0.02035)</td>
<td>(0.02052)</td>
<td>(0.02046)</td>
<td>(0.02042)</td>
</tr>
<tr>
<td>BLB</td>
<td>NA</td>
<td>0.936</td>
<td>0.940</td>
<td>0.938</td>
<td>0.932</td>
<td>0.932</td>
</tr>
<tr>
<td></td>
<td>(NA)</td>
<td>(0.01970)</td>
<td>(0.01968)</td>
<td>(0.01962)</td>
<td>(0.01953)</td>
<td>(0.01940)</td>
</tr>
<tr>
<td>SDB</td>
<td>0.922</td>
<td>0.936</td>
<td>0.952</td>
<td>0.946</td>
<td>0.950</td>
<td>0.948</td>
</tr>
<tr>
<td></td>
<td>(0.01976)</td>
<td>(0.02017)</td>
<td>(0.02041)</td>
<td>(0.02049)</td>
<td>(0.02054)</td>
<td>(0.02052)</td>
</tr>
</tbody>
</table>
Figure 2.6 Time evolution of relative errors under Gamma scenario.
Figure 2.7  Time evolution of relative errors under Poisson scenario.
CHAPTER 3. MORE POWERFUL TESTS FOR SPARSE HIGH DIMENSIONAL COVARIANCES MATRICES

A paper accepted by *Journal of Multivariate Analysis*

Liuhua Peng, Song Xi Chen, Wen Zhou

Abstract

This paper considers improving the power of tests for the identity and sphericity hypotheses regarding high dimensional covariance matrices. The power improvement is achieved by employing the banding estimator for the covariance matrices, which leads to significant reduction in the variance of the test statistics in high dimension. Theoretical justification and simulation experiments are provided to ensure the validity of the proposed tests. The tests are used to analyze a dataset from an acute lymphoblastic leukemia gene expression study for an illustration.

3.1 Introduction

This paper is interested in testing hypothesis for high-dimensional covariance matrices, $\Sigma$, of a $p$-dimensional random vector $X$. In practice, it is often of scientific interest to test whether or not a prescribed dependence structure is supported by data, for instance

$$H_0 : \Sigma = \Sigma_0 \quad \text{vs.} \quad H_1 : \Sigma \neq \Sigma_0$$

and

$$H_0 : \Sigma = \sigma^2 \Sigma_0 \quad \text{vs.} \quad H_1 : \Sigma \neq \sigma^2 \Sigma_0 \quad \text{for some unknown} \ \sigma^2 > 0,$$
for a known non-degenerate covariance matrix $\Sigma_0$. Among many practical applications, genomic studies usually motivate (3.1) or (3.2): it is not uncommon to postulate a grouping structure among genes of interest such that genes are not correlated across groups (Katsani et al., 2014), i.e. $\Sigma$ is presumed in a diagonal block shape upon permutations. Additionally, in the fields of image segmentation, epidemiology and ecology, large numbers of pixels or population abundances are collected across the spatial domain. Certain spatial autocorrelations are usually prescribed for fitting data to a parametric or semiparametric model for predictions (Bolker, 2008; Cressie, 1993). It is important to verify whether or not these hypothetical dependence structures are supported by data.

For identically and independently distributed data $X_1, \ldots, X_n$ with unknown common mean $\mu$ and covariance $\Sigma_0$, linear transform $\Sigma_0^{-1/2}X_i$ reduces (3.1) and (3.2) to

$$H_0 : \Sigma = I_p \quad \text{vs.} \quad H_1 : \Sigma \neq I_p,$$

(3.3)

and

$$H_0 : \Sigma = \sigma^2 I_p \quad \text{vs.} \quad H_1 : \Sigma \neq \sigma^2 I_p,$$

(3.4)

where $I_p$ is the $p$-dimensional identity matrix. Hypotheses (3.3) and (3.4) are called the identity and sphericity hypothesis, respectively. For fixed $p$, likelihood ratio test has been developed and widely applied. We refer to Anderson (2003) for more details. Let $\hat{\Sigma}$ be the sample covariance matrix. John (1971, 1972) and Nagao (1973) showed that for a fixed $p$, test statistics

$$V_n = p^{-1}\text{tr}\left\{ \left( \hat{\Sigma} - I_p \right)^2 \right\} \quad \text{and} \quad U_n = p^{-1}\text{tr}\left[ \left\{ p\hat{\Sigma}/\text{tr}(\hat{\Sigma}) - I_p \right\}^2 \right]$$

(3.5)

provide the most powerful invariant tests for both the identity and sphericity hypotheses against the local alternatives. Traditional tests, however, are not applicable to the large $p$, small $n$ paradigm since the sample covariance matrix is singular with probability one whenever $p > n$ and is no longer a consistent estimator if $p$ is not a smaller order of $n$ (Bai and Yin, 1993; Bai, Silverstein and Yin, 1988).

Tests for covariance matrices suited for the high dimensionality have been developed over the recent years. Ledoit and Wolf (2002) established the asymptotic properties of statistics in
(3.5) for $p/n \to c \in (0, +\infty)$ and proposed tests for identity (3.3) and sphericity (3.4) under the Gaussian assumption. Jiang (2004) developed a sphericity test based on the max-type statistic $L_n = \max_{1 \leq i < j \leq p} |\hat{\rho}_{ij}|$, where $\hat{\rho}_{ij}$ denotes the sample correlation coefficient between the $i$-th and $j$-th components of $X$. With the aid of the random matrix theory, Bai et al. (2009) proposed a modified likelihood ratio statistic for testing (3.3) for $p/n \to y \in (0, 1)$. To avoid the issue of inconsistency of $\hat{\Sigma}$ when $p > n$, Chen, Zhang and Zhong (2010) proposed U-statistic based testing procedures for both the identity and sphericity hypotheses. Their tests require much relaxed assumptions on the data distribution, and allow $p$ diverges in $n$ in any rates. See, for example, Cai and Jiang (2011); Hallin and Paindaveine (2006); Srivastava and Yanagihara (2010); Srivastava and Reid (2012); Srivastava, Yanagihara and Kubokawa (2014); Zou et al., (2014) for alternative test formulations, and Bai et al. (2009); Schott (2005); Zheng et al., (2014); Qiu and Chen (2012) for related works. One limitation of these high dimensional tests is a loss of power under sparse high dimension situations, largely due to a rapid increase in the variance of the test statistic as the $p$ gets larger. For instance, in the formulation of the identity test, estimation of the discrepancy measure $p^{-1}\text{tr}\{(\hat{\Sigma} - I_p)^2\}$ involves all the entries of the sample covariance. As a result, the test statistics incurs larger variation as the dimension gets larger. The increased variance dilutes the signal $p^{-1}\text{tr}\{(\hat{\Sigma} - I_p)^2\}$ of the test and hence brings down its power.

While we are gathering more dimensions in the data as more features are recorded, the information content of the data is not necessarily increasing at the same rate as the dimension. Indeed, it is commonly acknowledged that parameters associated with high dimensional data can be sparse in the sense of that many of the parameters are either zero or taking small values. This was the rationale behind the proposal of LASSO in Tibshirani (1996) as well as other regularization-based estimations in regression and covariance matrices; see Bickel and Levina (2008b); Cai, Zhang and Zhou (2010); Fan and Li (2001); Rothman, Levina and Zhu (2009). We consider in this paper tests for covariance matrices by utilizing the regularization-based estimation constructed for a specific class of sparse covariance matrices, the so-called bandable covariances, introduced by Bickel and Levina (2008a). The bandable class is naturally suited as alternative hypotheses to the null identity and the sphericity hypotheses. Specifically, we
formulate the test statistics by employing the banded covariance estimator proposed in Bickel and Levina (2008a). This allows us to take advantage of the knowledge of sparsity in the $\Sigma$. We demonstrate in this paper that the new test formulations have a remarkable power enhancement over the existing high dimensional tests for the covariance which do not utilize the sparsity information.

The rest of the paper is organized as follows. We introduce our motivations in Section 3.2 and present the testing procedures in Section 3.3. The theoretical properties of the proposed tests are also investigated in Section 3.3. Section 3.4 is devoted to a discussion on the selection of $k$ for the proposed tests. Numerical results are displayed in Section 3.5 to investigate the performance of the tests in practice. Both simulation studies and applications of the proposed tests to an acute lymphoblastic leukemia gene expression dataset are reported. Section 3.6 concludes the article with a brief discussion. Technical proofs and supplementary material contains more details on the numerical studies are given in Section 3.7.

### 3.2 Motivations and preliminaries

Our investigation is motivated by the notion of the bandable covariance class introduced by Bickel and Levina (2008a), which is defined as

$$
U(\varepsilon_0, C, \alpha) = \left\{ \Sigma = (\sigma_{ij})_{1 \leq i,j \leq p} : \max_j \sum_{|i-j| > k_0} |\sigma_{ij}| \leq C k_0^{-\alpha} \text{ for all } k_0 \geq 0, \right.
$$

$$
0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\varepsilon_0 \right\}, \quad (3.6)
$$

where $\varepsilon_0$, $C$ and $\alpha$ are positive constants. The bandable covariance prescribes a general decaying pair-wise covariances $\sigma_{ij}$ for large $|i - j|$. The increasing sparsity as the pair-wise covariance moves away from the main diagonal is ideally suited as alternative models for the identity and sphericity hypotheses. It is noted that the $\Sigma$ of the original random vector $X$ may not be bandable. We assume there is a permutation of $X$ such that the corresponding covariance is bandable. There are algorithms, for instance the angle-based ordering algorithm (Friendly, 2002) or the Isoband algorithm (Wagaman and Levina, 2009), which may be used to permute the data so that $\Sigma$ under the permutation is more bandable.
For the above bandable covariance class, Bickel and Levina (2008a) proposes the banding covariance estimator. Let \( k < p \) be a positive integer and write the sample covariance matrix \( \hat{\Sigma} = (\hat{\sigma}_{ij})_{1 \leq i, j \leq p} \). The banding estimator of \( \Sigma \) with bandwidth \( k \) is \( \hat{\Sigma}_{k,p} = \hat{\Sigma}_k = B_k(\hat{\Sigma}) = (\hat{\sigma}_{ij}I(|i - j| \leq k))_{1 \leq i, j \leq p} \). For \( \Sigma \in \mathcal{U}(\varepsilon_0, C, \alpha) \), Bickel and Levina (2008a) established the consistency of \( B_k(\hat{\Sigma}) \) to \( \Sigma \) under the spectral norm by letting \( k \) divergence at the rate \( (n^{-1}\ln p)^{-1/(2(\alpha + 1))} \) for sub-Gaussian distributed data when \( \ln p/n \to 0 \).

Encouraged by this important advance in high dimensional covariance estimation, we consider replacing \( \hat{\Sigma} \) in (3.5) by \( B_k(\hat{\Sigma}) \) leads to the following test statistics

\[
P^{-1}\text{tr}\left[ \left( B_k(\hat{\Sigma}) - I_p \right)^2 \right] \quad \text{and} \quad P^{-1}\text{tr}\left[ \left( pB_k(\hat{\Sigma})/\text{tr}(\hat{\Sigma}) - I_p \right)^2 \right].
\]

Comparing with the statistics \( V_n \) and \( U_n \) given in (3.5), the above formulations based on the banding estimator \( B_k(\hat{\Sigma}) \) are expected to be less variable since those pair-wise sample covariances \( \hat{\sigma}_{ij} \) located further away from the \( k \)-th superdiagonals (subdiagonals) are excluded due to the banding operation. Indeed, for \( \Sigma \) in the bandable class, most of the signals (those larger \( \sigma_{ij} \)) are located closer to the main diagonals. This form of sparsity suggests us to discount covariances which are far away from the main diagonals. It is obvious that the formulation is critically dependent on the choice of the banding width \( k \). Recently, Qiu and Chen (2015) have proposed a data driven method for choosing \( k \) by minimizing an empirical version of \( ||B_k(\Sigma) - \Sigma||_F \), where \( || \cdot ||_F \) denotes the Frobenius norm.

The seemingly natural constructions given in (3.7) need to be refined in order to be suitable for more relaxed relationship between \( p \) and \( n \) and without the sub-Gaussian assumption in Bickel and Levina (2008a). Our aim here is to obtain unbiased estimators for \( \text{tr}\{B_k(\Sigma)\} \) and \( \text{tr}\{[B_k(\Sigma)]^2\} \) for \( \Sigma \in \mathcal{U}(\varepsilon_0, C, \alpha) \). Denote \( \mathbf{X}_i = (x_{i1}, \ldots, x_{ip})^\top \) for each \( i \), and let

\[
L_{n_1}(i, j) = \frac{1}{P_n^{2}} \sum_{l_1 \neq l_2} x_{i1}x_{i1,l_2}x_{i2}x_{i2,l_2}, \quad L_{n_2}(i, j) = \frac{1}{P_n^{3}} \sum_{l_1, l_2, l_3} x_{i1}x_{i2,l_3}x_{i3}x_{i3,l_3},
\]

\[
L_{n_3}(i, j) = \frac{1}{P_n^{4}} \sum_{l_1, l_2, l_3, l_4} x_{i1}x_{i2,l_4}x_{i3,l_4}x_{i4}, \quad L_{n_4}(i) = \frac{1}{n} \sum_{l=1}^{n} x_{il}^2, \quad L_{n_5}(i) = \frac{1}{P_n^{2}} \sum_{l_1 \neq l_2} x_{i1}x_{i2,l_2},
\]

where \( P_n^r = n!/(n - r)! \) for \( r = 2, 3, 4 \) and \( \sum \) denotes the summation over mutually different indices. Notice that \( \sum_{i=1}^{p} \{L_{n_1}(i) - L_{n_5}(i)\} \) and \( \sum_{|i - j| \leq k} \{L_{n_1}(i, j) - 2L_{n_2}(i, j) + L_{n_3}(i, j)\} \) are unbiased estimators of \( \text{tr}(\Sigma) \) and \( \text{tr}\{[B_k(\Sigma)]^2\} \), respectively.
We consider two discrepancy measures for the identity and sphericity hypotheses:

\[
p^{-1}\text{tr}[(B_k(\Sigma) - I_p)^2] = p^{-1}\text{tr}[(B_k(\Sigma))^2] - 2p^{-1}\text{tr}(\Sigma) + 1,
\]

and

\[
\frac{1}{p} \text{tr} \left[ \left( \frac{B_k(\Sigma)}{(1/p)\text{tr}(B_k(\Sigma))} - I_p \right)^2 \right] = \frac{p\text{tr}[(B_k(\Sigma))^2]}{(\text{tr}(\Sigma))^2} - 1,
\]

upon which, we propose the following two unbiased estimators to these two measures

\[
V_{n,k} = p^{-1} \sum_{|i-j| \leq k} \{L_{n1}(i,j) - 2L_{n2}(i,j) + L_{n3}(i,j)\} - 2p^{-1} \sum_{i=1}^{p} \{L_{n4}(i) - L_{n5}(i)\} + 1, \quad (3.8)
\]

and

\[
U_{n,k} = p \left[ \sum_{|i-j| \leq k} \{L_{n1}(i,j) - 2L_{n2}(i,j) + L_{n3}(i,j)\} \right] - 1. \quad (3.9)
\]

It is noted that \(V_{n,k}\) and \(U_{n,k}\) reduce to the statistics in Chen, Zhang and Zhong (2010) when the banding width \(k = p - 1\) which involves all the sample covariances. Thus, \(V_{n,k}\) and \(U_{n,k}\) are regularized versions of those proposed by Chen, Zhang and Zhong (2010). By utilizing the sparse information that \(\Sigma \in \mathcal{U}(\varepsilon_0, C, \alpha)\), those \(\hat{\sigma}_{ij}\) beyond the \(k\)-th superdiagonal are avoided which makes \(V_{n,k}\) and \(U_{n,k}\) have less variations and hence more powerful tests as we will demonstrate later.

It is easy to see that \(V_{n,k}\) and \(U_{n,k}\) are invariant under the location shift. Hence, without loss of generality, we assume data has been centered such that \(\mu = 0\).

**Assumption 3.1.** \(\Sigma \in \mathcal{U}(\varepsilon_0, C, \alpha)\) for some constants \(\varepsilon_0, C\) and \(\alpha\) which are unrelated to \(p\).

Similarly to Bai and Saranadasa (1996) and Chen, Zhang and Zhong (2010), we assume the following multivariate model for \(X_i\).

**Assumption 3.2.** Data \(X_1, \ldots, X_n\) are independent and identically distributed \(p\)-dimensional random vectors such that

\[
X_i = \Gamma Z_i \quad \text{for} \quad i = 1, \ldots, n, \quad (3.10)
\]

where \(\Gamma = (\Gamma_{ij})_{1 \leq i \leq p, 1 \leq j \leq m}\) is a constant loading matrix with \(p \leq m\) and \(\Gamma\Gamma^\top = \Sigma\), \(Z_i = (z_{i1}, \ldots, z_{im})^\top\)'s are independent and identically \(p\)-dimensional random vectors with zero mean.
and identity covariance. Furthermore, we assume \( \sup_j E(z_{1j}^8) < C_1 \) for some constant \( C_1 > 0 \) and there exists a constant \( \Delta < \infty \) such that \( E(z_{1j}^4) = 3 + \Delta \) for each \( j \). For any integer \( \ell_v \geq 0 \) with \( \sum_{v=1}^q \ell_v \leq 8 \),

\[
E(z_{1i_1}^{\ell_1} \ldots z_{1i_q}^{\ell_q}) = E(z_{1i_1}^{\ell_1}) \ldots E(z_{1i_q}^{\ell_q}) \tag{3.11}
\]

whenever \( i_1, \ldots, i_q \) are distinct.

This model, first employed by Bai and Saranadasa (1996) for testing high-dimensional mean vectors, ensures that the observations \( X_i \) are linearly generated by \( m \)-variate \( Z_i \) consisted of white noises. The dimension \( m \) of \( Z_i \) is finite but diverge to infinity as \( p \) and \( n \) both go to infinity. So the dimension of \( Z_i \) is arbitrary as long as \( m \geq p \) that equips the model flexibility in generating \( X_i \) with covariance \( \Sigma \). The distribution of \( Z_i \) is not restricted to particular families, and is therefore nonparametric. Assumption 3.2 has been extensively employed in high dimensional multivariate testing problems, for example see Chen, Zhang and Zhong (2010); Touloumis, Tavaré and Marioni (2015).

### 3.3 Testing procedures

#### 3.3.1 Identity test

We first derive the mean and variance of the test statistic \( V_{n,k} \) for the identity hypothesis. Derivations given in Lemma 3.1 in the appendix show that under Assumptions 3.1 and 3.2, as \( n, p \to \infty \) and \( k \to \infty \), if \( k = o(\min(n^{1/2}, p^{1/2})) \),

\[
E(V_{n,k}) = p^{-1} \text{tr} [\{B_k(\Sigma) - I_p\}^2] \quad \text{and} \quad \text{Var}(V_{n,k}) = p^{-2} \sigma_{V_{n,k}}^2 \{1 + o(1)\} \tag{3.12}
\]

where

\[
\sigma_{V_{n,k}}^2 = \frac{1}{n} \text{tr} [\Sigma \{B_k(\Sigma) - I_p\}]^2 
+ 4 \Delta n^{-1} \text{tr} \left[ \left\{ \Gamma^T(B_k(\Sigma) - I_p) \Gamma \right\} \circ \left\{ \Gamma^T(B_k(\Sigma) - I_p) \Gamma \right\} \right]. \tag{3.13}
\]
\[ \tau_{n,k}^2(\Sigma) = \frac{4}{n(n-1)} \sum_{|i-j| \leq k, |i_2-j_2| \leq k} \sigma_{i_1,i_2}^2 \sigma_{j_1,j_2}^2 + \frac{4}{n(n-1)} \sum_{|i-j| \leq k, |i_2-j_2| \leq k} \sigma_{i_1,i_2} \sigma_{j_1,j_2} \sigma_{i_1,j_2} \sigma_{j_1,i_2} 
\]
\[ + \frac{8\Delta}{n(n-1)} \sum_{|i-j| \leq k, |i_2-j_2| \leq k} \sigma_{i_1,i_2} \sigma_{j_1,j_2} f_{i_1,j_1} f_{i_2,j_2}(\Sigma) + \frac{2\Delta^2}{n(n-1)} \sum_{|i-j| \leq k, |i_2-j_2| \leq k} f_{i_1,j_1} f_{i_2,j_2}(\Sigma), \]

(3.14)

and \( \odot \) denotes the Hadamard product of matrices. In (3.14), \( f_{i_1,j_1} f_{i_2,j_2}(\Sigma) = \sum_{r=1}^m \Gamma_{i_1,r} \Gamma_{j_1,r} \Gamma_{i_2,r} \Gamma_{j_2,r} \) for the loading matrix \( \Gamma = (\Gamma_{jl})_{p \times m} \) in Assumption 3.2.

**Theorem 3.1.** Under Assumptions 3.1 and 3.2, as \( n \to \infty, p \to \infty \) and \( k \to \infty \), if \( k = o(\min(n^{1/2},p^{1/2})) \),

\[ \sigma_{V_{n,k}}^{-1} \{ p V_{n,k} - \text{tr}[\{ B_k(\Sigma) - I_p \}^2] \} \to \mathcal{N}(0,1) \]  

(3.15)

in distribution.

In Theorem 3.1, no explicit restrictions on the growth rates of \( p \) and \( n \) are imposed. We note that the banding width prescribed in Bickel and Levina (2008a) that allows consistent estimation was \( k = (n^{-1} \ln p)^{-1/(2(\alpha+1))} \) for the sub-Gaussian distributed data and \( k = (n^{-1/2} p^{2/\beta})^{-1/(1+\alpha+2/\beta)} \) for data with bounded \( \beta \)-th moment, where \( \alpha \) is the sparsity index in the bandable class. For both cases, the condition \( k = o(\min(n^{1/2},p^{1/2})) \) assumed in Theorem 3.1 allows wider range of the banding width than that in Bickel and Levina (2008a). This is because testing hypotheses usually requires less stringent assumptions than the estimation. It is also noted that the asymptotic normality holds even for a fixed \( k \), but \( \sigma_{V_{n,k}}^2 \) will be in a more involved form with more terms than those in (3.13).

Under the null identity hypothesis \( H_0 \) in (3.3), \( \text{E}(V_{n,k}) = 0 \) and \( \text{Var}(V_{n,k}) = p^{-2} \sigma_{V_{n,k}}^2 + o(p^{-2} \sigma_{V_{n,k}}^2) \) where

\[ \sigma_{V_{n,k}}^2 = \tau_{n,k}^2(I_p) = 4\{n(n-1)\}^{-1}(2pk + 2p - k^2 - k) + 8\Delta\{n(n-1)\}^{-1} \sum_{|i-j| \leq k} f_{i_1,j_1} f_{i_2,j_2}(I_p) 
\]
\[ + 2\Delta^2\{n(n-1)\}^{-1} \sum_{|i_1-j_1| \leq k, |i_2-j_2| \leq k} f_{i_1,j_1} f_{i_2,j_2}(I_p). \]

From Theorem 3.1, the asymptotic null distribution of \( V_{n,k} \) is

\[ p \sigma_{V_{n,k}}^{-1} V_{n,k} \overset{d}{\to} \mathcal{N}(0,1). \]  

(3.16)
To facilitate testing, we need to estimate \( \sigma_{\hat{\nu}, n, k}^2 \). Given that the loading matrix \( \Gamma \) is not observable, it is difficult to estimate \( f_{i_1j_1} (I_p) \) and \( f_{i_1j_1i_2j_2} (I_p) \) directly from data. However, we note that under the \( H_0 \)

\[
\sigma_{\hat{\nu}, n, k}^2 = 2\{n(n-1)\}^{-1} \sum_{|i_1-j_1| \leq k} \sum_{|i_2-j_2| \leq k} \{ \text{E}(x_{i_1j_1} x_{i_1j_2} x_{i_2j_2}) - \sigma_{i_1j_1} \sigma_{i_2j_2} \}^2 ,
\]

which suggests that \( \sigma_{\hat{\nu}, n, k}^2 \) can be estimated by

\[
\hat{\sigma}_{\hat{\nu}, n, k}^2 = 2\{n(n-1)\}^{-1} \sum_{|i_1-j_1| \leq k} \sum_{|i_2-j_2| \leq k} \left\{ \frac{1}{P^2} \sum_{l_1 \neq l_2} x_{i_1i_1} x_{i_1j_1} x_{i_1j_2} x_{i_2j_2} \right. \\
- \frac{2}{P^3} \sum_{l_1, l_2, l_3} x_{i_1i_1} x_{i_1j_1} x_{i_2j_2} x_{l_3i_1} x_{i_1j_1} x_{l_3j_2} \\
+ \frac{1}{P^4} \sum_{l_1, l_2, l_3, l_4} x_{i_1i_1} x_{i_1j_1} x_{l_2j_2} x_{l_3i_1} x_{l_3j_1} x_{l_4j_2} \left\} .
\]

(3.17)

The consistency of \( \hat{\sigma}_{\hat{\nu}, n, k}^2 \) is implied from the following proposition.

**Proposition 3.1.** Under Assumptions 3.1 and 3.2, and if \( \mu = 0 \), then

\[
\text{E} \left( \hat{\sigma}_{\hat{\nu}, n, k}^2 \right) = \sigma_{\hat{\nu}, n, k}^2 \quad \text{and} \quad \text{Var} \left( \hat{\sigma}_{\hat{\nu}, n, k}^2 / \sigma_{\hat{\nu}, n, k}^2 \right) = O(k^2 n^{-1} + n^{-1}).
\]

In practice, to cater for the case of \( \mu \neq 0 \), we can replace \( x_{i\ell} \) by \( x_{i\ell} - \bar{x}_{\ell} \) where \( \bar{x}_{\ell} = n^{-1} \sum_{\ell=1}^{n} x_{i\ell} \) for each \( \ell \) to center the data so that \( \mu = 0 \) is satisfied. As Proposition 3.1 implies \( \hat{\sigma}_{\hat{\nu}, n, k}^2 / \sigma_{\hat{\nu}, n, k}^2 \to 1 \) in probability under the \( H_0 \). We have

\[
p \hat{\sigma}_{\hat{\nu}, n, k}^{-1} V_{n, k} \to \mathcal{N}(0, 1)
\]

(3.18)

in distribution under the \( H_0 \). Therefore, a regularized identity test with a nominal significant level \( \alpha \) rejects \( H_0 : \Sigma = I_p \) if

\[
p \hat{\sigma}_{\hat{\nu}, n, k}^{-1} V_{n, k} > z_\alpha
\]

where \( z_\alpha \) is the \( \alpha \) upper-quantile of \( \mathcal{N}(0, 1) \). As long as \( k = o(\min(n^{1/2}, p^{1/2})) \), the asymptotic normality given in (3.18) ensures the nominal level of significance asymptotically.
3.3.2 Power of the identity test

To evaluate the power of the test for the identity hypothesis, let $\delta_{V_{n,k}} = \text{tr}\{B_k(\Sigma) - I_p\}^2$, which can be viewed as the signal of the test problem under the alternative. The power of the regularized identity test

$$
\beta_{V_{n,k}}(\alpha) = \Pr\left(p\sigma_{V_{n,k}0}^{-1} V_{n,k} \geq z_\alpha\right)
$$

$$
= \Pr\left(\sigma_{V_{n,k}}^{-1} (p V_{n,k} - \text{tr}\{B_k(\Sigma) - I_p\}^2) \geq \sigma_{V_{n,k}0}^{-1} \sigma_{V_{n,k}0} z_\alpha - \sigma_{V_{n,k}}^{-1} \delta_{V_{n,k}}\right)
$$

$$
= 1 - \Phi\left(\sigma_{V_{n,k}}^{-1} \sigma_{V_{n,k}0} z_\alpha - \sigma_{V_{n,k}}^{-1} \delta_{V_{n,k}}\right),
$$

where $\Phi$ is the standard normal distribution. It can be shown that $\sigma_{V_{n,k}}^{-1} \sigma_{V_{n,k}0}$ is bounded. Hence, the power of the proposed identity test is largely determined by $\sigma_{V_{n,k}}^{-1} \delta_{V_{n,k}}$, which may be regarded as the signal to noise ratio of the test.

We now discuss the role of the banding width on the power of the test. We note that both $\delta_{V_{n,k}}$ and $\sigma_{V_{n,k}}^2$ are increasing with respect to $k$. For two banding widths $k_2 > k_1$, suppose that $\delta_{V_{n,k_1}} > 0$, it may be shown that $\sigma_{V_{n,k_2}}^{-1} \delta_{V_{n,k_2}} \geq \sigma_{V_{n,k_1}}^{-1} \delta_{V_{n,k_1}}$ if and only if

$$
(\delta_{V_{n,k_2}} - \delta_{V_{n,k_1}})/\delta_{V_{n,k_1}} + 1 \geq \left\{(\sigma_{V_{n,k_2}}^2 - \sigma_{V_{n,k_1}}^2)/\sigma_{V_{n,k_1}}^2 + 1\right\}^{1/2}.
$$

Therefore, if the relative signal increment $(\delta_{V_{n,k_2}} - \delta_{V_{n,k_1}})/\delta_{V_{n,k_1}}$ can off-set the relative increase in the noise as specified above, the test with the larger $k_2$ is more powerful than that with the smaller $k_1$. On the contrary, if the relative increase in the signal cannot off-set the relative increase in the noise, using the large $k$ leads to a loss of power.

To answer the question that when will the increase of the banding width not bring in more power, we note that

$$
\delta_{V_{n,k_2}} - \delta_{V_{n,k_1}} = \sum_{k_1 < |i-j| \leq k_2} \sigma_{ij}^2
$$

and for a positive constant $c$,

$$
\sigma_{V_{n,k_2}}^2 - \sigma_{V_{n,k_1}}^2 \geq 4n(n - 1)^{-1} \sum_{k_1 < |i-j| \leq k_2} \sigma_{ii}^2 \sigma_{jj}^2 \geq c(k_2 - k_1)pn^{-2}.
$$

This implies that a necessary condition for $\sigma_{V_{n,k_2}}^{-1} \delta_{V_{n,k_2}} \geq \sigma_{V_{n,k_1}}^{-1} \delta_{V_{n,k_1}}$ is

$$
(\delta_{V_{n,k_2}} - \delta_{V_{n,k_1}})/\delta_{V_{n,k_1}} \geq c_1(k_2 - k_1)n^{-1}
$$
for a positive constant $c_1$. Hence, if the relative signal increment is smaller than $c_1(k_2 - k_1)n^{-1}$, using the larger $k_2$ will result in a loss in the power.

To gain further insight, let us consider the case of banded covariance in which $\Sigma = B_{\tilde{k}}(\Sigma)$ for a $\tilde{k}$. In this case, $\delta_{V_{n,k}} = \delta_{V_{n,k}}$ while $\sigma_{V_{n,k}}^2$ keeps increasing for all $k \geq \tilde{k}$. This means that $\sigma_{V_{n,k}}^{-1}\delta_{V_{n,k}}$ gets smaller and a loss of power occurs as $k$ gets larger beyond $\tilde{k}$. Since our proposed test is identical to the one in Chen, Zhang and Zhong (2010) if $k = p - 1$, the proposed identity test is asymptotically more powerful for a properly selected $k$ under the banded scenario.

The following theorem establishes the consistency of the proposed identity test.

**Theorem 3.2.** Under Assumptions 3.1 and 3.2, $p \to \infty$, $k \to \infty$ and $k = o(\min(n^{1/2}, p^{1/2}))$, as $n \to \infty$, $\beta_{V_{n,k}} \to 1$ provided $\{\tau_{n,k}(\Sigma)\}^{-1}\delta_{V_{n,k}} \to \infty$.

We note here that $\tau_{n,k}(\Sigma)$ defined in (3.14) is the first term of $\sigma_{V_{n,k}}^2$. Theorem 3.2 implies that the proposed identity test is able to detect alternatives with power tending to 1 when $\{\tau_{n,k}(\Sigma)\}^{-1}\delta_{V_{n,k}} \to \infty$. We note that when $\{\tau_{n,k}(\Sigma)\}^{-1}\delta_{V_{n,k}} \to \infty$, $\tau_{n,k}(\Sigma)$ dominates the other two terms of $\sigma_{V_{n,k}}^2$ in (3.13), which means $\sigma_{V_{n,k}}^{-1}\delta_{V_{n,k}} \to \infty$. It may be shown that $\tau_{n,k}(\Sigma)$ is at most $O(k^2pn^{-2})$ under Assumptions 3.1 and 3.2. Hence, the test has power approaching to one as long as $p^{-1}\delta_{V_{n,k}}$ is a larger order of $kp^{-1/2}n^{-1}$, which is much weaker than $n^{-1}$, the corresponding lower limit for the test of Chen, Zhang and Zhong (2010).

### 3.3.3 Sphericity test

We firstly establish the asymptotic properties of $U_{n,k}$. Let

$$
\sigma_{U_{n,k}}^2 = \frac{\tau_{n,k}^2(\Sigma)}{\text{tr}^2\{B_k(\Sigma)\Sigma\}} + \frac{2}{n} \text{tr} \left[ \left\{ \frac{B_k(\Sigma)\Sigma}{\text{tr}(B_k(\Sigma)\Sigma)} - \frac{\Sigma}{\text{tr}(\Sigma)} \right\}^2 \right] + \frac{\Delta}{n} \text{tr} \left( \left[ \Gamma^\top \left\{ \frac{B_k(\Sigma)}{\text{tr}(B_k(\Sigma)\Sigma)} - \frac{I_p}{\text{tr}(\Sigma)} \right\} \Gamma \right] \circ \left[ \Gamma^\top \left\{ \frac{B_k(\Sigma)}{\text{tr}(B_k(\Sigma)\Sigma)} - \frac{I_p}{\text{tr}(\Sigma)} \right\} \Gamma \right] \right).
$$

**Theorem 3.3.** Under Assumptions 3.1 and 3.2, as $n \to \infty$, $p \to \infty$ and $k \to \infty$, if $k = o(\min(n^{1/2}, p^{1/2}))$,

$$
\sigma_{U_{n,k}}^{-1} \left[ \frac{\text{tr}^2(\Sigma)}{\text{tr}(B_k(\Sigma)\Sigma)} \left( \frac{U_{n,k} + 1}{p} \right) - 1 \right] \to \mathcal{N}(0,1)
$$

in distribution.
The asymptotic variance under the null hypothesis is
\[
\sigma_{U_{n,k}0}^2 = \sigma^{-2} p^{-2} \left[ 4\{n(n-1)^{-1}(2pk - k^2 - k + 2p) + 8\Delta \{n(n-1)\}^{-1} \sum_{|i-j| \leq k} f_{ijij}(I_p) \\
+ 2\Delta^2 \{n(n-1)\}^{-1} \sum_{|i_1-j_1| \leq k} \sum_{|i_2-j_2| \leq k} f_{i_1j_1i_2j_2}(I_p) \right].
\]

Then, Theorem 3.3 implies that under the null sphericity hypothesis
\[
\sigma_{U_{n,k}0}^{-1} U_{n,k} \to \mathcal{N}(0, 1) \quad (3.20)
\]
in distribution. Using a similar approach to estimating the null variance \(\sigma_{V_{n,k}0}^2\) in Section 3.3.1, an estimator of the null variance \(\sigma_{U_{n,k}0}^2\) is
\[
\hat{\sigma}_{U_{n,k}0}^2 = \sigma_{V_{n,k}0}^2 \left\{ \sum_{|i-j| \leq k} (L_{1,ij} - 2L_{2,ij} + L_{3,ij}) \right\}^{-2}.
\]

It may be shown that \(\hat{\sigma}_{U_{n,k}0}^2 \to \sigma_{U_{n,k}0}^2\) in probability and \(\hat{\sigma}_{U_{n,k}0}^{-1} U_{n,k}\) converges to \(\mathcal{N}(0, 1)\) in distribution. These lead to a sphericity test with a nominal significance level \(\alpha\) that rejects \(H_0 : \Sigma = \sigma^2 I_p\) if
\[
\hat{\sigma}_{U_{n,k}0}^{-1} U_{n,k} > z_{\alpha}.
\]

Let \(\delta_{U_{n,k}} = 1 - \{\text{tr}(\Sigma)/[\text{ptr}(B_k(\Sigma)\Sigma)]\}.\) The power of the sphericity test is
\[
\beta_{U_{n,k}}(\alpha) = \Pr(\sigma_{U_{n,k}0}^{-1} U_{n,k} \geq z_{\alpha})
= 1 - \Phi \left[ \left\{ \frac{\text{tr}^2(\Sigma)}{\text{ptr}(B_k(\Sigma)\Sigma)} \right\} \left( \frac{\sigma_{U_{n,k}0}}{\sigma_{U_{n,k}}} \right) z_{\alpha} - \frac{\delta_{U_{n,k}}}{\sigma_{U_{n,k}}} \right].
\]

As \(\text{tr}^2(\Sigma)/[\text{ptr}(B_k(\Sigma)\Sigma)]\) and \(\sigma_{U_{n,k}0}/\sigma_{U_{n,k}}\) are both bounded, the power is largely influenced by \(\delta_{U_{n,k}}/\sigma_{U_{n,k}}\), which can be viewed as the signal to noise ratio of the testing problem.

To gain insight on the power, we study the signal to noise ratio as what we did for the identity test in section 3.3.1. It can be shown that for \(k_2 > k_1\), \(\sigma_{U_{n,k_2}}^{-1} \delta_{U_{n,k_2}} \geq \sigma_{U_{n,k_1}}^{-1} \delta_{U_{n,k_1}}\) if and only if
\[
(\delta_{U_{n,k_2}} - \delta_{U_{n,k_1}})/\delta_{U_{n,k_1}} + 1 \geq \{(\sigma_{U_{n,k_2}}^2 - \sigma_{U_{n,k_1}}^2)/\sigma_{U_{n,k_1}}^2 + 1\}^{1/2}.
\]

We also note that, for \(k_2 > k_1\)
\[
\delta_{U_{n,k_2}} - \delta_{U_{n,k_1}} = p^{-1} \text{tr}^2(\Sigma) \frac{\sum_{k_1 < |i-j| \leq k_2} \sigma_{ij}^2}{\text{tr}\{B_{k_1}^2(\Sigma)\}\text{tr}\{B_{k_2}^2(\Sigma)\}}.
\]
which indicates that the increase in the signal is largely driven by \( \sum_{k_1 < |i-j| \leq k_2} \sigma^2_{ij} \), those \( \sigma_{ij} \) between the \( k_1 \) and \( k_2 \)-th superdiagonals. At the same time, it can be shown that \( \sigma^2_{U_{n,k}} \) is increasing with respect to \( k \) at a rate at least \( p^{-1}n^{-2} \). Hence, a power-enhancing strategy is to use the smallest \( k \) that captures the most signals.

The following theorem establishes the consistency of the proposed sphericity test (3.21).

**Theorem 3.4.** Under Assumptions 3.1 and 3.2, \( p \rightarrow \infty, k \rightarrow \infty \) and \( k = o(\min(n^{1/2}, p^{1/2})) \) as \( n \rightarrow \infty \), if \( \text{tr}\{B_k(\Sigma)\Sigma\}\{\tau_{n,k}(\Sigma)\}^{-1}\delta_{U_{n,k}} \rightarrow \infty \) then \( \beta_{U_{n,k}} \rightarrow 1 \).

It can be shown that \( \tau_{n,k}(\Sigma)/\text{tr}\{B_k(\Sigma)\Sigma\} \) is at most \( O(kp^{-1/2}n^{-1}) \). Hence, the proposed sphericity test is consistent as long as \( \delta_{U_{n,k}} \) is a larger order of \( kp^{-1/2}n^{-1} \), which is much lower than the corresponding lower limit of the test of Chen, Zhang and Zhong (2010) since \( k = o(p^{1/2}) \). The latter test requires \( \delta_{U_{n,p-1}} \) is a larger order of \( n^{-1} \).

### 3.4 Selection of \( k \)

Given the beneficial power property of the tests based on the banding operation, we report numerical results of the proposed tests with respect to \( k \) in this section.

We start with evaluating the impacts of \( k \) on the size of the tests. Clearly, under the null hypotheses for both (3.3) and (3.4), \( \Sigma \in \mathcal{U}(\varepsilon_0, C, \alpha) \). Given the established asymptotic normality for the two test statistics, the size of the proposed tests are expected to be close to the nominal significance level for a wider range of \( k \) as long as \( k = o(\min(n^{1/2}, p^{1/2})) \). To confirm this, we ran simulations for the sphericity test. We generated independent and identically distributed random vectors \( X_i \) from \( \mathcal{N}(0, \Sigma) \) where \( \Sigma = 2I_p \) and considered \( k = \lceil a_in^{1/3} \rceil \) for \( \{a_i\}_{i=1}^{10} = \{0, 0.1, 0.3, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0\} \). The sample size and the dimensions considered were \( n = 20, 40 \) and \( 60 \), and \( p = 38, 89 \) and \( 181 \), respectively. Figure 3.1 displays the empirical size of the proposed sphericity test with respect to \( k \)’s for the combinations of \( n \) and \( p \) based on 1000 simulations. The figure shows that the empirical size was largely close to the nominal 5% level for the wider choices of \( k \)’s.

To gain information on the role of \( k \) on the power of the tests, we considered the identity test at 5% level of significance under the alternative where \( \Sigma_1 = I_p + 0.6^2 \Omega, \Omega = (\omega_{ij})_{1 \leq i,j \leq p} \)
with $\omega_{ij} = I(|i - j| = \ell)$ for a fixed integer $\ell$. According to Theorem 3.2, the proposed identity test is powerful if $\{\tau_{n,k}(\Sigma_1)\}^{-1}\delta_{V_{n,k}} \rightarrow \infty$. For the given covariance structure, $\delta_{V_{n,k}} = 0$ for $k < \ell$ and $\delta_{V_{n,k}} = 0.2592(p - \ell)$ for $k \geq \ell$. In the meanwhile, $\tau_{n,k}^2(\Sigma_1)$ is strictly increasing along with $k$. Thus, $\{\tau_{n,k}(\Sigma_1)\}^{-1}\delta_{V_{n,k}}$ is maximized at $k = \ell$ for any given $n$ and $p$, namely the power would be maximized if $k$ agreed with the underlying bandwidth $\ell$.

We carried out a simulation experiment for $\mathcal{N}(0, \Sigma_1)$ distributed data with $\Sigma_1$ defined as above and $\ell = 2$, and $n = 20$ and $p = 20$. Table 3.1 reports the empirical power of the proposed identity test with banding widths ranging from 0 to 7 based on 1000 replications. It shows that the test gained powers as $k$ was increased to $\ell = 2$, peaked at $k = \ell = 2$, and then declined afterward. This power profile was highly consistent with the discussion made toward the end of Section 3.3.1 regarding the signals and the noise of the test statistic.

<table>
<thead>
<tr>
<th>$k$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>0.072</td>
<td>0.077</td>
<td>0.942</td>
<td>0.902</td>
<td>0.856</td>
<td>0.831</td>
<td>0.798</td>
<td>0.766</td>
</tr>
</tbody>
</table>

While the above results were assuring, we need a practical way to select the banding width $k$ in order to carry out the tests. Qiu and Chen (2015) proposed a selection method by minimizing an empirical estimate of $\mathbb{E}\|B_k(S_n) - \Sigma\|_F^2$. They demonstrated that the approach has superior performance than the cross-validation approach based on random sample splitting proposed in

---

Figure 3.1  Empirical size of the 5% sphericity test with respect to the banding width $k$ for various sample sizes and dimensions.
Bickel and Levina (2008a). The cross-validation was formulated based on a score function of \( k \) that measured the discrepancy between \( B_k(S_n) \) based on one part of the split sample, and the sample covariance based on the remaining sample. The issue with the cross-validation was that the use of the inconsistent sample covariance makes the approach unreliable in high dimension.

We carried out 1000 simulation experiments to investigate the performance of the banding width denoted as \( k_{qc} \) prescribed by Qiu and Chen (2015) for the same Gaussian model with covariance matrix \( \Sigma_1 \) in the setting for Table 3.1 for the identity test. We considered \( n = 20, 40, 80 \) and \( p = 20, 60, 100, 160, 200, 300 \) in the simulations. Figure 3.2 reports the empirical proportion that the selected banding width \( k_{qc} \) agreed with the true value, which was \( \ell = 2 \), and otherwise. Red solid lines with squares report the proportion of times that the selected \( k \) agrees with the true value \( \ell = 2 \), the green dashed lines with circles report the proportion that the selected \( k < 2 \), and the blue dot-dash lines with triangles report the proportion that the \( k > 2 \).

The performance of the banding width selection algorithm was satisfactory. Figure 3.2 shows that even when \( n \) was small at 20, the algorithm could still identify the true banding width with sufficient probability, and the precision improved as the dimension \( p \) was increased. When \( n \) was 40, the proportion of correct selection started to be close to 100%. The corresponding empirical powers of the proposed identity test using \( k_{qc} \) were close to one for most combinations of \( n \) and \( p \). Hence, this numerical study provided support to using \( k_{qc} \) for the regularized tests for high dimensional covariance matrices.

![Figure 3.2](image.jpg) Proportions of the banding width selected by the method of Qiu and Chen (2015).
3.5 Numerical results

3.5.1 Simulation studies

We report results of simulation experiments which were designed to evaluate the performance of the proposed tests for identity and sphericity of high dimensional $\Sigma$. To demonstrate the improvement in power of the proposed tests, we also experimented two existing high dimensional tests: the one by Chen, Zhang and Zhong (2010) (CZZ hereafter) and the test by Ledoit and Wolf (2002) (LW hereafter). The LW test is applicable for Gaussian data only. The test statistics of CZZ and LW tests include all the components of the sample covariance matrix in their formulation and were expected to have lower power as the test statistics can bear larger variation.

We considered two types of innovations for generating data according to Assumption 3.2: the Gaussian innovation where $Z_i$ were $\mathcal{N}(0_m, I_m)$ distributed; and the Gamma innovation where $Z_i$ had $m$ independent and identically distributed components, and each component was the standardized Gamma random variable with parameters 4 and 0.5 such that it had zero mean and unit variance. In the simulations for the sphericity test, the null hypothesis was $H_0: \Sigma = \sqrt{2}I_p$.

When evaluating the powers of the tests for both identity and sphericity hypotheses, we considered the following three forms of $\Sigma$.

- **Diagonal Form:** Set $\Sigma = \text{diag}(4\times1_{[vp]}, 2\times1_{p-[vp]})$ where $[x]$ denotes the integer truncation of $x$ and $v$ characterizes the sparsity of the signals. We chose $v = 0.05$ and 0.1.

- **Banded Form:** $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p}$ with $\sigma_{ij} = \rho^{i-j}I(\{i-j\} \leq 1)$ and $\rho = 0.10$.

- **Bandable Form:** Take $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p}$ with $\sigma_{ij} = I(i = j) + \theta (i-j)^{-\rho}I(i \neq j)$ with $\theta = 1.2$ and $\rho = 0.16$, which prescribed a polynomial decay as considered by Qiu and Chen (2015).

To mimic the large $p$, small $n$ scenario, we set $n = 20, 40, 60$ and 80, and for each $n$ let $p = 38, 55, 89, 159, 181, 331, 343, 642$, also set $m = p$ when generating the data according to Assumption 3.2. All the simulation results were based on 1000 iterations with nominal
significance level at 5%. The parameter $k$ was determined by $k = k_{qc} + 1$ where $k_{qc}$ was the banding width by the method of Qiu and Chen (2015). We add 1 to $k_{qc}$ was to ensure the banding estimator $B_k(\Sigma)$ contains enough signals.

Table 3.2 displays the empirical sizes of the proposed identity test, the CZZ and LW tests for testing (3.3). It shows that for the Gaussian data the LW test maintained the size better than the other two tests as the LW test was designed for Gaussian data. It is observed that as the sample size was increased, both the proposed identity test and the CZZ test had empirical sizes approaching to the nominal significance level. For the Gamma distributed innovation, as displayed in Table 3.2, the proposed identity test and the CZZ test had the empirical sizes close to the nominal level, while the LW test failed to control the size. The empirical sizes of the three tests for testing the sphericity hypothesis are reported in the appendix, which were quite similar to those of the identity tests in Table 3.2.

To compare the powers of the tests, we considered $n = 40, 60, 80$ for $p$ set as above. For the Gamma distributed data, only the proposed identity and the CZZ tests were considered since the LW test was no longer applicable. Figures 3.3–3.5 display the empirical powers of the of the proposed identity test, the CZZ and LW tests. It was very clear that the proposed identity test outperformed the other two tests for both data generating distributions and the covariance models considered in the simulation. For the alternative $\Sigma$ with the diagonal form (in Figures 3.3 and 3.4), the powers of all three tests were improved for all tests when $v$ is large as expected; the LW and CZZ tests have comparable powers for the Gaussian data (Figure 3.3). Figure 3.5 reports the empirical power for the bandable and banded alternative. It shows that the proposed identity test outperformed the other two tests. For the alternative in the bandable form, neither the CZZ test nor the LW test gained extra powers as $p$ was increased, while the proposed identity test had its power increased as $p$ was increased as shown in panels (a) and (b) of Figure 3.5). As $n$ was increased, all three tests gained powers as expected.

Given the slow decay rate of off-diagonal entries, the alternative in the banded form had a relatively larger banding width than other two types of alternatives. For the banded alternative, all tests gain powers in growing $p$ or $n$ and the CZZ test has power approaching to the proposed identity test as $n$ increasing (panels (c) and (d) in Figure 3.5).
Table 3.2  Empirical sizes of the proposed test (IT) for the identity hypothesis, along with those of the tests by Chen, Zhang and Zhong (2010) (CZZ) and Ledoit and Wolf (2002) (LW).

<table>
<thead>
<tr>
<th></th>
<th>IT</th>
<th>CZZ</th>
<th>LW</th>
<th>IT</th>
<th>CZZ</th>
<th>LW</th>
<th>IT</th>
<th>CZZ</th>
<th>LW</th>
<th>IT</th>
<th>CZZ</th>
<th>LW</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>0.077</td>
<td>0.075</td>
<td>0.059</td>
<td>0.072</td>
<td>0.057</td>
<td>0.064</td>
<td>0.057</td>
<td>0.056</td>
<td>0.051</td>
<td>0.062</td>
<td>0.064</td>
<td>0.053</td>
</tr>
<tr>
<td>55</td>
<td>0.080</td>
<td>0.058</td>
<td>0.059</td>
<td>0.056</td>
<td>0.053</td>
<td>0.047</td>
<td>0.052</td>
<td>0.046</td>
<td>0.058</td>
<td>0.061</td>
<td>0.066</td>
<td>0.063</td>
</tr>
<tr>
<td>89</td>
<td>0.076</td>
<td>0.084</td>
<td>0.057</td>
<td>0.072</td>
<td>0.057</td>
<td>0.045</td>
<td>0.063</td>
<td>0.071</td>
<td>0.045</td>
<td>0.072</td>
<td>0.071</td>
<td>0.048</td>
</tr>
<tr>
<td>159</td>
<td>0.073</td>
<td>0.065</td>
<td>0.057</td>
<td>0.064</td>
<td>0.062</td>
<td>0.049</td>
<td>0.060</td>
<td>0.057</td>
<td>0.049</td>
<td>0.068</td>
<td>0.056</td>
<td>0.050</td>
</tr>
<tr>
<td>181</td>
<td>0.077</td>
<td>0.087</td>
<td>0.049</td>
<td>0.057</td>
<td>0.057</td>
<td>0.050</td>
<td>0.046</td>
<td>0.056</td>
<td>0.056</td>
<td>0.053</td>
<td>0.051</td>
<td>0.053</td>
</tr>
<tr>
<td>331</td>
<td>0.081</td>
<td>0.071</td>
<td>0.063</td>
<td>0.066</td>
<td>0.055</td>
<td>0.061</td>
<td>0.060</td>
<td>0.051</td>
<td>0.057</td>
<td>0.060</td>
<td>0.054</td>
<td>0.062</td>
</tr>
<tr>
<td>343</td>
<td>0.069</td>
<td>0.072</td>
<td>0.054</td>
<td>0.062</td>
<td>0.065</td>
<td>0.053</td>
<td>0.056</td>
<td>0.056</td>
<td>0.051</td>
<td>0.060</td>
<td>0.042</td>
<td>0.044</td>
</tr>
<tr>
<td>642</td>
<td>0.074</td>
<td>0.071</td>
<td>0.054</td>
<td>0.056</td>
<td>0.060</td>
<td>0.055</td>
<td>0.056</td>
<td>0.059</td>
<td>0.053</td>
<td>0.057</td>
<td>0.045</td>
<td>0.045</td>
</tr>
</tbody>
</table>

Gaussian distributed innovation

<table>
<thead>
<tr>
<th></th>
<th>IT</th>
<th>CZZ</th>
<th>LW</th>
<th>IT</th>
<th>CZZ</th>
<th>LW</th>
<th>IT</th>
<th>CZZ</th>
<th>LW</th>
<th>IT</th>
<th>CZZ</th>
<th>LW</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 40$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>0.084</td>
<td>0.076</td>
<td>0.198</td>
<td>0.078</td>
<td>0.071</td>
<td>0.209</td>
<td>0.055</td>
<td>0.067</td>
<td>0.221</td>
<td>0.052</td>
<td>0.059</td>
<td>0.199</td>
</tr>
<tr>
<td>55</td>
<td>0.081</td>
<td>0.074</td>
<td>0.182</td>
<td>0.067</td>
<td>0.072</td>
<td>0.212</td>
<td>0.065</td>
<td>0.060</td>
<td>0.192</td>
<td>0.058</td>
<td>0.062</td>
<td>0.191</td>
</tr>
<tr>
<td>89</td>
<td>0.085</td>
<td>0.080</td>
<td>0.213</td>
<td>0.066</td>
<td>0.055</td>
<td>0.195</td>
<td>0.059</td>
<td>0.051</td>
<td>0.195</td>
<td>0.066</td>
<td>0.065</td>
<td>0.185</td>
</tr>
<tr>
<td>159</td>
<td>0.080</td>
<td>0.073</td>
<td>0.168</td>
<td>0.075</td>
<td>0.061</td>
<td>0.178</td>
<td>0.067</td>
<td>0.064</td>
<td>0.207</td>
<td>0.056</td>
<td>0.046</td>
<td>0.189</td>
</tr>
<tr>
<td>181</td>
<td>0.087</td>
<td>0.081</td>
<td>0.176</td>
<td>0.071</td>
<td>0.059</td>
<td>0.200</td>
<td>0.059</td>
<td>0.056</td>
<td>0.212</td>
<td>0.062</td>
<td>0.052</td>
<td>0.228</td>
</tr>
<tr>
<td>331</td>
<td>0.088</td>
<td>0.071</td>
<td>0.198</td>
<td>0.067</td>
<td>0.074</td>
<td>0.193</td>
<td>0.059</td>
<td>0.073</td>
<td>0.182</td>
<td>0.060</td>
<td>0.049</td>
<td>0.188</td>
</tr>
<tr>
<td>343</td>
<td>0.083</td>
<td>0.079</td>
<td>0.182</td>
<td>0.063</td>
<td>0.064</td>
<td>0.198</td>
<td>0.073</td>
<td>0.065</td>
<td>0.172</td>
<td>0.062</td>
<td>0.063</td>
<td>0.189</td>
</tr>
<tr>
<td>642</td>
<td>0.076</td>
<td>0.065</td>
<td>0.183</td>
<td>0.074</td>
<td>0.056</td>
<td>0.173</td>
<td>0.061</td>
<td>0.060</td>
<td>0.184</td>
<td>0.056</td>
<td>0.065</td>
<td>0.178</td>
</tr>
</tbody>
</table>

Gamma distributed innovation
Figure 3.3  Empirical powers for the identity tests against the alternative in the diagonal form with the Gaussian distributed innovation.

Figure 3.4  Empirical powers for the identity tests against the alternative in the diagonal form with the Gamma distributed innovation.
(a) Bandable alternative with Gaussian innovations.

(b) Bandable alternative with Gaussian innovations.

(c) Banded alternative with Gaussian innovations.

(d) Banded alternative with Gamma innovations.

Figure 3.5 Empirical powers for the identity tests against the alternatives in either the bandable or banded forms of the proposed test with either the Gaussian or Gamma distributed innovation.
Figure 3.6  Empirical powers for the sphericity tests against the alternative in the diagonal form with the Gaussian distributed innovation.

Figure 3.7  Empirical powers for the sphericity tests against the alternative in the diagonal form with the Gamma distributed innovation.
Figure 3.8  Empirical powers for the sphericity tests against the alternatives in either the bandable or banded forms of the proposed test with either the Gaussian or Gamma distributed innovation.
The power performance of the proposed test for the sphericity along with the CZZ and LW tests are reported in Figures 3.6–3.8, which suggest that the proposed sphericity test was much more powerful than the other two tests under the three forms of the alternatives $\Sigma$. More simulation results are reported in the appendix.

3.5.2 Empirical study

We analyzed an acute lymphoblastic leukemia (ALL) data reported in Chiaretti et al. (2004) to demonstrate the proposed regularized tests for identity and sphericity. The data contain microarray expressions for patients having acute lymphoblastic leukemia of either T-lymphocyte type or B-lymphocyte type. We focused on the sub-sample of B-lymphocyte type leukemia in this analysis. The 78 patients of B-lymphocyte type leukemia were classified into two groups: the BCR/ABL fusion (36 patients) and cytogenetically normal NEG (42 patients). The original dataset has been analyzed by Chen and Qin (2010); Chiaretti et al. (2004), and Dudoit, Keles and van der Laan (2011), using different methodologies.

Our analysis is to study the covariance structures for sets of genes defined within the gene ontology (GO) framework. It is known that genes tend to work collectively to achieve certain biological tasks, which gave rise to the identification of gene-sets (also called GO terms) with respect to three broader categories of biological functions: biological processes (BP), cellular components (CC) and molecular functions (MF). The gene-sets are technically defined in the gene ontology (GO) system via structured vocabularies which produce unique name for a gene-set. After a preliminary screening with the gene-filtering approach advocated in Gentleman et al. (2005), there left 2694 unique gene-sets in the BP category, 352 in the CC category and 419 in the MF category for the ALL data. The largest gene-set had 3048, 3140 and 303 genes in BP, CC and MF, respectively.

Our aim was to study the dependence structures in the expression levels of gene-sets between the BCR/ABL and NEG groups for each of the three functional categories by testing hypotheses (3.3) and (3.4) for appropriately transformed data. The procedure is described as following. To attain bandable covariance structure so that we can apply the proposed tests, we employed the re-ordering algorithm in Friendly (2002) to each gene-set in both the NEG and the BCR/ABL
fusion groups to obtain a permutation of the genes in that gene-set so that the covariance was more bandable. As a demonstration of this data re-ordering algorithm, we compare the heat maps of the correlation matrices before and after the re-ordering based on samples from the NEG group for the gene-set GO:0000086 (G2/M transition of mitotic cell cycle), which has 63 genes, in Figure 3.9 in the appendix.

For a gene-set of a functional category with the NEG sample, say the $g$-th gene-set, its covariance $\Sigma_{\text{neg},g}$ was estimated by using the banding estimator, denoted as $\hat{\Sigma}_{\text{neg},g}$ with the banding width determined by the method in Qiu and Chen (2015). And $\hat{\Sigma}_{\text{neg},g}^{-1/2}$ was used to transform the same gene-set in the BCR/ABL group. For each transformed gene-set in the BCR/ABL group, we tested the hypotheses (3.3) and (3.4) using the proposed regularized identity and sphericity tests when $p > 10$ and use the tests by John (1971, 1972) for those gene-sets with smaller dimensions. For the proposed tests, $k$ was chosen as $k_{qc} + 1$. We essentially tested hypotheses $H_0 : \Sigma_{\text{neg},g} = \Sigma_{\text{BCR/ABL},g}$ or $\Sigma_{\text{neg},g} = \sigma_g^2 \Sigma_{\text{BCR/ABL},g}$ for some $\sigma_g^2 > 0$ for each $g = 1, \ldots, \mathcal{G}$ where $\mathcal{G}$ is the total number of gene-sets in the category. The test of Chen, Zhang and Zhong (2010) was also performed to serve as a comparison.

By controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995) at 0.001, we have identified gene-sets that have significantly different covariance structures between the NEG and BCR/ABL groups. Table 3.3 provides a broad classification for the gene-sets identified by the proposed test and the test of Chen, Zhang and Zhong (2010). The table shows that the dependence structure between NEG and BCR/ABL were largely different with quite a large number of significantly differential expressed gene-sets, which was due to a large number of very small $p$-values (Figure 3.10 in the appendix). Biologically speaking, the NEG and BCR/ABL cases have different genetic mechanisms that cytogenetically normal leukemia is not associated with large chromosomal abnormalities while the BCR/ABL leukemia involves fusion of BCR and ABL genes in Philadelphia chromosome (Pakakasama et al., 2008).

Table 3.3 reveals that the proposed tests identified more gene-sets than the CZZ test. It is interesting to notice that the proposed sphericity test has identified GO:0004527 and GO:0004869 as diseases-associated gene-sets in the MF category while they were missed by the CZZ test, and biologically these two gene-sets correspond to exonuclease activity and endopeptidase in-
Table 3.3 Number of identified gene-sets against null hypotheses. ST and IT stand for the proposed sphericity and identity tests, respectively. The last column reports the means and standard deviations for $k_{qc}$.

<table>
<thead>
<tr>
<th>GO Category</th>
<th>Total</th>
<th>Sphericity hypothesis (3.4)</th>
<th>Identity hypothesis (3.3)</th>
<th>$k$ after re-ordering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ST only</td>
<td>Both</td>
<td>CZZ only</td>
</tr>
<tr>
<td>BP</td>
<td>2694</td>
<td>21</td>
<td>2338</td>
<td>0</td>
</tr>
<tr>
<td>CC</td>
<td>352</td>
<td>0</td>
<td>317</td>
<td>0</td>
</tr>
<tr>
<td>MF</td>
<td>418</td>
<td>7</td>
<td>363</td>
<td>1</td>
</tr>
</tbody>
</table>

Inhibitor activity, which have been recognized associated to the disease development of different types of leukemia recently (Shi et al., 2015; Tsakou et al., 2012).

### 3.6 Discussion

In this paper, we introduced two powerful tests for the identity and sphericity hypotheses of large covariance matrices and showed that the proposed testing procedures perform particularly well against sparse alternatives from some particular classes. The proposed tests leverage the sparsity information of the alternatives in high dimensional settings that leads to significant reduction in the variance of the test statistics. The theoretical properties of the proposed tests were established. We also explored how the proposed tests improve the powers comparing to the test by Chen, Zhang and Zhong (2010). Furthermore, we discussed the selection of $k$ for the proposed tests in practice. Finally, we examined the proposed tests by numerical studies and illustrated its applications in real data analysis.

The forms of the identity hypothesis that $\Sigma = I_p$ and the sphericity hypothesis that $\Sigma = \sigma^2 I_p$ are two idealized hypotheses. Despite being idealized, they play central roles in either the classical multivariate analysis where the dimension of data $p$ is fixed or the high dimensional multivariate analysis where $p$ is diverging and can be larger than the sample size as treated in this paper. The literature of the classical multivariate analysis include those of Anderson (2003), John (1971, 1972) and Nagao (1973), while the contemporary literature includes Chen, Zhang and Zhong (2010) and Ledoit and Wolf (2002) among others.

The identity hypothesis in (3.3) actually covers the hypothesis $H_0 : \Sigma = \Sigma_0$ for a known invertible covariance matrix $\Sigma_0$. By transforming the data via left multiplying $\Sigma_0^{-1/2}$, the
identity hypothesis can be carried out for the transformed data. The sphericity hypothesis can be treated similarly. In practice, the hypotheses $\Sigma_0$ can be postulated based on the empirical estimates of $\Sigma$. This was what we have done in the case study by first permuting the data to rearrange the data components so that those with high correlations are grouped closer than those with low correlation. After the permutation, we employed the banding estimator of the high dimensional covariance matrix of Bickel and Levina (2008a) to attain a banded form for $\Sigma$, which are used to standardize the data. We are aware that using the estimated $\Sigma_0$ would introduce issues of inference as the estimation error may affect the asymptotic distribution. While this would not be a big issue in the classical fixed dimensional context, it would be an issue when $p$ diverges. We would consider this issue in a future study.

Another issue is a concern on the computational costs required in carrying out the proposed test procedures. The main cost of computation is in computing the raw test statistics and in estimating their variance. To gain information on the computation time needed, we report in Table 3.4 the time needed to accomplish a single sphericity test under the alternative bandable form $\Sigma$ in the simulation study for different $n$ and $p$ on a PC with Intel(R) Core(TM) i7-4790K processor and CPU speed of 4.0GHz. The computing times in the table suggest that the computation burden for carrying out the test procedure is manageable even for relatively larger values of $n$ and $p$ with a quite standard computational capacity.

Table 3.4  Computational time (in seconds) for a single sphericity test under the alternative $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p}$ with $\sigma_{ij} = I(i = j) + \theta|i - j|^\rho I(i \neq j)$ with different $n$ and $p$. 

<table>
<thead>
<tr>
<th></th>
<th>38</th>
<th>55</th>
<th>89</th>
<th>159</th>
<th>181</th>
<th>331</th>
<th>343</th>
<th>641</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n=20$</td>
<td>0.172</td>
<td>0.354</td>
<td>0.932</td>
<td>3.051</td>
<td>3.961</td>
<td>13.408</td>
<td>14.830</td>
<td>51.652</td>
</tr>
<tr>
<td>$n=40$</td>
<td>0.299</td>
<td>0.633</td>
<td>1.659</td>
<td>5.403</td>
<td>7.023</td>
<td>23.863</td>
<td>25.926</td>
<td>91.187</td>
</tr>
<tr>
<td>$n=60$</td>
<td>0.425</td>
<td>0.901</td>
<td>2.399</td>
<td>7.747</td>
<td>10.109</td>
<td>34.104</td>
<td>37.088</td>
<td>131.226</td>
</tr>
<tr>
<td>$n=80$</td>
<td>0.550</td>
<td>1.182</td>
<td>3.133</td>
<td>10.119</td>
<td>13.158</td>
<td>44.805</td>
<td>48.658</td>
<td>186.654</td>
</tr>
</tbody>
</table>

3.7 Appendix

We begin this appendix by presenting some notation and technical preliminaries that will be used in the proof of the main results. For a matrix $\mathbf{M} = (m_{ij})_{1 \leq i,j \leq p}$, we denote $\lambda(\mathbf{M})$ the
eigenvalues of $\mathbf{M}$ with $\lambda_{\text{min}}(\mathbf{M})$ and $\lambda_{\text{max}}(\mathbf{M})$ the smallest and largest eigenvalues of $\mathbf{M}$, respectively, and denote the matrix norms by $\|\mathbf{M}\|_1 \equiv \max_j \sum_i |m_{ij}|$, $\|\mathbf{M}\| \equiv \left\{ \lambda_{\text{max}}(\mathbf{M}^T \mathbf{M}) \right\}^{1/2}$, and $\|\mathbf{M}\|_\infty \equiv \max_i \sum_j |m_{ij}|$. For symmetric $\Sigma \in \mathcal{U}(\varepsilon_0, C, \alpha)$, we have the following properties

1. $\|\Sigma\| \leq 1/\varepsilon_0$ and $\|B_k(\Sigma) - \Sigma\|_1 \leq Ck^{-\alpha}$;
2. $\|B_k(\Sigma)\| \leq 1/\varepsilon_0 + Ck^{-\alpha}$;
3. $|\lambda_{\text{min}}\{B_k(\Sigma)\} - \lambda_{\text{min}}(\Sigma)| \leq Ck^{-\alpha}$; and
4. there exist $C_1, C_2 > 0$ such that $C_1 \leq \min_i \sigma_{ii} \leq \max_i \sigma_{ii} \leq C_2$ for $\Sigma = (\sigma_{ij})_{1 \leq i, j \leq p}$.

Properties (2) and (3) imply that for sufficiently large $k$, there exits a positive constant $\delta_0$ such that

$$0 < \delta_0 \leq \lambda_{\text{min}}\{B_k(\Sigma)\} \leq \lambda_{\text{max}}\{B_k(\Sigma)\} \leq 1/\delta_0,$$

which means $\text{tr}(\Sigma) = O(p)$ and $\text{tr}\{B_k(\Sigma)\} = O(p)$ for sufficiently large $k$. In addition, some algebraic computations yield following useful results in remaining derivations:

(i) $\sum_{|i_1 - j_1| \leq k} \sum_{i_2=1}^p \sigma_{i_1 j_1} f_{i_1 j_1 i_2 j_2} (\Sigma) = \text{tr} \left[ \{\Gamma^T B_k(\Sigma) \Gamma\} \circ \{\Gamma^T \Sigma \} \right] \leq \text{tr}\{B_k(\Sigma)\Sigma^2\}$;

(ii) $\sum_{|i_1 - j_1| \leq k} \sum_{|i_2 - j_2| \leq k} \sigma_{i_1 j_1} \sigma_{i_2 j_2} f_{i_1 j_1 i_2 j_2} (\Sigma) = \text{tr} \left[ \{\Gamma^T B_k(\Sigma) \Gamma\} \circ \{\Gamma^T B_k(\Sigma) \Gamma\} \right] \\
\leq \text{tr} \left[ \{B_k(\Sigma)\Sigma^2\}^2 \right] $;

(iii) $\sum_{|i_1 - j_1| \leq k} \sum_{|i_2 - j_2| \leq k} \sigma_{i_1 i_2}^2 \sigma_{j_1 j_2}^2 \leq (2k + 1)^2 \text{tr}\{(\Sigma \circ \Sigma)^2\}$,

$\sum_{|i_1 - j_1| \leq k} \sum_{|i_2 - j_2| \leq k} \sigma_{i_1 i_2} \sigma_{j_1 j_2} \sigma_{i_1 j_2} \sigma_{j_1 i_2} \leq (2k + 1)^2 \text{tr}\{(\Sigma \circ \Sigma)^2\}$;

(iv) $\sum_{|i_1 - j_1| \leq k} \sum_{|i_2 - j_2| \leq k} \sigma_{i_1 i_2} \sigma_{j_1 j_2} f_{i_1 j_1 i_2 j_2} (\Sigma) \leq (2k + 1)^2 \text{tr}\{(\Gamma \circ \Gamma)(\Gamma \circ \Gamma)^T(\Sigma \circ \Sigma)\}$; and,

(v) $\sum_{|i_1 - j_1| \leq k} \sum_{|i_2 - j_2| \leq k} f_{i_1 j_1 i_2 j_2}^2 (\Sigma) \leq (2k + 1)^2 \text{tr}\left[ (\Gamma \circ \Gamma)(\Gamma \circ \Gamma)^T \right]^2$.

3.7.1 Critical lemmas

In this section, we collect some technical lemmas that will be used in the proofs. The first lemma provides algebraic representations of the variances and covariances of $L_{m_1}$, $L_{m_2}$, $L_{m_3}$, $L_{m_4}$ and $L_{m_5}$.
Lemma 3.1. Under Assumptions 3.1 and 3.2, for \( \tau_{n,k}^2(\Sigma) \) defined in (3.14) and \( n^* = n(n - 1)(n - 2)(n - 3) \)

\[
\text{Var} \left\{ \sum_{|i-j| \leq k} L_{n_1}(i, j) \right\} = \tau_{n,k}^2(\Sigma) + 8n^{-1}\text{tr} \left\{ B_k(\Sigma)\Sigma \right\}^2 + 4\Delta n^{-1}\text{tr} \left\{ \left( \Sigma^\top B_k(\Sigma)\Sigma \right) \circ \left( \Sigma^\top B_k(\Sigma)\Sigma \right) \right\},
\]

\[
\text{Var} \left\{ \sum_{|i-j| \leq k} L_{n_2}(i, j) \right\} = \frac{2}{n(n-1)}\text{tr} \{ B_k(\Sigma)\Sigma B_k(\Sigma)\Sigma \} + \frac{1}{n(n-1)(n-2)} \sum_{|i_1-j_1| \leq k} \sum_{|i_2-j_2| \leq k} \left\{ (\sigma_{i_1i_2}\sigma_{j_1j_2} + \sigma_{i_1j_2}\sigma_{j_1i_2})^2 + \Delta(\sigma_{i_1i_2}\sigma_{j_1j_2} + \sigma_{i_1j_2}\sigma_{j_1i_2})f_{i_1j_1i_2j_2}(\Sigma) \right\},
\]

\[
\text{Var} \left\{ \sum_{|i-j| \leq k} L_{n_3}(i, j) \right\} = \frac{8}{n^*} \sum_{|i_1-j_1| \leq k} \sum_{|i_2-j_2| \leq k} \sigma_{i_1i_2}^2\sigma_{j_1j_2}^2 + \frac{16}{n^*} \sum_{|i_1-j_1| \leq k} \sum_{|i_2-j_2| \leq k} \sigma_{i_1i_2}\sigma_{i_1j_2}\sigma_{j_1i_2}\sigma_{j_1j_2},
\]

\[
\text{Var} \left\{ \sum_{i=1}^p L_{n_4}(i) \right\} = 2n^{-1}\text{tr}(\Sigma^2) + \Delta n^{-1}\text{tr} \left\{ \left( \Sigma^\top \Sigma \right) \circ \left( \Sigma^\top \Sigma \right) \right\},
\]

\[
\text{Var} \left\{ \sum_{i=1}^p L_{n_5}(i) \right\} = 2(n(n-1))^{-1}\text{tr}(\Sigma^2),
\]

and furthermore,

\[
\text{Cov} \left\{ \sum_{|i-j| \leq k} L_{n_1}(i, j), \sum_{i=1}^p L_{n_4}(i) \right\} = 4n^{-1}\text{tr} \{ B_k(\Sigma)\Sigma^2 \} + 2\Delta n^{-1}\text{tr} \left\{ \left( \Sigma^\top B_k(\Sigma)\Sigma \right) \circ \left( \Sigma^\top \Sigma \right) \right\},
\]

\[
\text{Cov} \left\{ \sum_{i=1}^p L_{n_4}(i), \sum_{i=1}^p L_{n_5}(i) \right\} = 0.
\]

The proof of Lemma 3.1 is based on standard yet tedious computations that we omit here. Lemma 3.1 implies that \( \text{Var}\{ \sum_{|i-j| \leq k} L_{n_3}(i, j) \} = O(n^{-2}\text{Var}\{ \sum_{|i-j| \leq k} L_{n_1}(i, j) \}) \), and under the assumption that \( k \to \infty \) and \( k = o(\min(n^{1/2}, p^{1/2})) \), it yields

\[
\text{Var} \left\{ \sum_{|i-j| \leq k} L_{n_2}(i, j) \right\} = o \left( \text{Var} \left\{ \sum_{|i-j| \leq k} L_{n_1}(i, j) \right\} \right)
\]

and

\[
\text{Var} \left\{ \sum_{i=1}^p L_{n_5}(i) \right\} = o \left( \text{Var} \left\{ \sum_{|i-j| \leq k} L_{n_1}(i, j) \right\} \right).
\]

The next lemma is on the asymptotic normality of statistics \( T_{n,k} \).
Lemma 3.2. Under Assumptions 3.1 and 3.2, for any real sequences \( \{a_n\}_{n=1}^{\infty} \) and \( \{b_n\}_{n=1}^{\infty} \),

\[
T_{n,k} = a_n \sum_{|i-j| \leq k} L_{n_1}(i, j) + b_n \sum_{i=1}^{p} L_{n_4}(i)
\]

satisfies

\[
\{ \text{Var}(T_{n,k}) \}^{-1/2} \{ T_{n,k} - \mathbb{E}(T_{n,k}) \} \to \mathcal{N}(0, 1) \tag{3.22}
\]

in distribution provided \( k = o(\min(n^{1/2}, p^{1/2})) \).

Proof. Let \( \mathcal{F}_0 = \{\emptyset, \Omega\} \) and \( \mathcal{F}_t = \sigma\{X_1, \ldots, X_t\} \) for \( t = 1, \ldots, n \) be the sequence of \( \sigma \)-fields generated by data, and denote \( \mathbb{E}_t(\cdot) \equiv \mathbb{E}(\cdot|\mathcal{F}_t) \) and \( \mathbb{E}(\cdot) \equiv \mathbb{E}_0(\cdot) \). Write \( T_{n,k} - \mathbb{E}(T_{n,k}) = \sum_{t=1}^{n} D_{t,k} \), where \( D_{t,k} = \mathbb{E}_t(T_{n,k}) - \mathbb{E}_{t-1}(T_{n,k}) \). It is easy to see that \( D_{t,k} \) is \( \mathcal{F}_t \) measurable and \( \mathbb{E}_{t-1}(D_{t,k}) = 0 \) for each \( t \geq 1 \) so that for every \( n \), \( \{D_{t,k}, \mathcal{F}_t\}_{1 \leq t \leq n} \) is a martingale difference array. By the martingale central limit theorem, (3.22) is straightforward once one can show that, as \( n \to \infty \),

\[
\frac{\sum_{t=1}^{n} \sigma^2_{t,k}}{\text{Var}(T_{n,k})} \xrightarrow{p} 1 \quad \text{and} \quad \frac{\sum_{t=1}^{n} \mathbb{E}(D^4_{t,k})}{\text{Var}^2(T_{n,k})} \to 0, \tag{3.23}
\]

with \( \sigma^2_{t,k} = \mathbb{E}_{t-1}(D^2_{t,k}) \).

As \( \mathbb{E}(\sum_{t=1}^{n} \sigma^2_{t,k}) = \text{Var}(T_{n,k}) \), it suffices to show that \( \text{Var}(\sum_{t=1}^{n} \sigma^2_{t,k}) = o(\text{Var}^2(T_{n,k})) \) to derive the first part of (3.23). By Lemma 3.1,

\[
\text{Var}(T_{n,k}) = a_n^2 r^2_{n,k}(\Sigma) + 2n^{-1} \text{tr}\left\{ [2a_n B_k(\Sigma) + b_n I_p] \Sigma \right\}^2 + \Delta n^{-1} \text{tr}\left\{ \left[ \Gamma^\top (2a_n B_k(\Sigma) + b_n I_p) \Gamma \right] \right\} .
\]

Also, notice that

\[
D_{t,k} = 2a_n \{n(n-1)\}^{-1} \left[ X_t^\top B_k(Q_{t-1}) X_t - \text{tr}\{B_k(Q_{t-1}) \Sigma \} \right] + 2a_n n^{-1} \left[ X_t^\top B_k(\Sigma) X_t - \text{tr}\{B_k(\Sigma) \Sigma \} \right] + b_n n^{-1} \left\{ X_t^\top X_t - \text{tr}(\Sigma) \right\},
\]

where \( Q_{t-1} = (Q_{i,j}^{i,j})_{1 \leq i,j \leq p} = \sum_{s=1}^{t-1} (X_s X_s^\top - \Sigma) \) with \( Q_{i,j}^{i,j} = \sum_{s=1}^{t-1} (x_{si} x_{sj} - \sigma_{ij}) \), so

\[
\sum_{t=1}^{n} \sigma^2_{t,k} = R_{1,n} + R_{2,n} + R_{3,n} + R_{4,n} + nC,
\]
for some constant $C$ and

$$R_{1,n} = 8a_n(n^2(n-1))^{-1} \sum_{t=1}^{n} \text{tr} \left[ B_k(Q_{t-1}) \Sigma \left( 2a_nB_k(\Sigma) + b_nI_p \right) \Sigma \right],$$

$$R_{2,n} = 4\Delta a_n(n^2(n-1))^{-1} \sum_{t=1}^{n} \text{tr} \left[ \left( B_k(Q_{t-1}) \Gamma \right) \circ \left( \Gamma^T (2a_nB_k(\Sigma) + b_nI_p) \Gamma \right) \right],$$

$$R_{3,n} = 8a_n^2(n^2(n-1))^2 \sum_{t=1}^{n} \text{tr} \left[ \{ B_k(Q_{t-1}) \Sigma \}^2 \right],$$

$$R_{4,n} = 4\Delta a_n^2(n^2(n-1))^2 \sum_{t=1}^{n} \text{tr} \left[ \left( B_k(Q_{t-1}) \Gamma \right) \circ \left( \Gamma^T B_k(Q_{t-1}) \Gamma \right) \right].$$

It is sufficient to show $\text{Var}(R_{i,n}) = o(\text{Var}^2(T_{n,k}))$ for $i = 1, 2, 3, 4$ to obtain $\text{Var}(\sum_{t=1}^{n} \sigma_{t,k}^2) = o(\text{Var}^2(T_{n,k})).$

We first study $R_{1,n}$. Denote $\Omega_k = \Sigma \{ 2a_nB_k(\Sigma) + b_nI_p \} \Sigma$. Recall the fact that $\text{tr} \{ B_k(A)C \} = \text{tr} \{ AB_k(C) \}$ for symmetric matrices $A$ and $C$ with conformable sizes, we have

$$\text{tr} \{ B_k(Q_{t-1}) \Omega_k \} = \sum_{s=1}^{t-1} \left[ X_s^T B_k(\Omega_k) X_s - \text{tr} \{ B_k(\Sigma) \Omega_k \} \right].$$

Therefore, for each $s \geq 1$

$$\text{Var} \left[ X_s^T B_k(\Omega_k) X_s - \text{tr} \{ B_k(\Sigma) \Omega_k \} \right] = 2\text{tr} \{ \Sigma B_k(\Omega_k) \}^2 + \Delta \text{tr} \left[ \left( \Gamma^T B_k(\Omega) \Gamma \right) \circ \left( \Gamma^T B_k(\Omega) \Gamma \right) \right].$$

By the algebraic properties summarized before, it holds for some constant $\gamma > 0$ that

$$\text{tr} \{ \Sigma B_k(\Omega_k) \}^2 \leq \gamma \text{tr} \{ B_k(\Sigma) \Sigma \}^2 \text{tr} \{ \{ 2a_nB_k(\Sigma) + b_nI_p \} \Sigma \}^2 = o \left( n^3 \text{Var}^2(T_{n,k}) \right),$$

and also $\text{tr} \left[ \{ \Gamma^T B_k(\Omega_k) \Gamma \} \circ \{ \Gamma^T B_k(\Omega_k) \Gamma \} \right] \leq \text{tr} \{ \Sigma B_k(\Omega_k) \}^2$. We therefore conclude that there exist constant $C > 0$ such that

$$\text{Var}(R_{1,n}) \leq Cn^{-3} \text{Var} \left[ \text{tr} \{ B_k(Q_{t-1}) \Omega_k \} \right] = o(\text{Var}^2(T_{n,k})).$$

For $R_{3,n}$, denote for any $1 \leq s_1, s_2 \leq n$,

$$Y_{s_1s_2} = \sum_{|i_1-j_1| \leq k} \sum_{|i_2-j_2| \leq k} \sigma_{s_1i_2} \sigma_{s_1j_2} (x_{s_1i_1} x_{s_1j_1} - \sigma_{s_1j_1}) (x_{s_2i_2} x_{s_2j_2} - \sigma_{s_2j_2}).$$

Then

$$\text{tr} \left[ \{ B_k(Q_{t-1}) \Sigma \}^2 \right] = \sum_{s=1}^{t-1} Y_{ss} + \sum_{s_1 \neq s_2} Y_{s_1s_2}.$$
Notice that $E(Y_{s_1s_2}) = 0$ for any $s_1 \neq s_2$ and $E(Y_{s_1s_2}Y_{s_3s_4}) = 0$ for any \((s_1, s_2, s_3, s_4)\) except \(s_1 = s_2 = s_3 = s_4, s_1 = s_3\) and \(s_2 = s_4\) or \(s_1 = s_4\) and \(s_2 = s_3\). Hence, for any \(t \leq l\)

\[
\text{Cov}(\text{tr}\{(B_k(Q_{l-1})\Sigma)^2\}, \text{tr}\{(B_k(Q_{l-1})\Sigma)^2\}) = (t-1)\text{Var}(Y_{11}) + 2(t-1)(t-2)\text{Var}(Y_{12}).
\]

It is therefore sufficient to show \(\text{Var}(Y_{11}) = o(n^5\text{Var}^2(T_{n,k}))\) and \(\text{Var}(Y_{12}) = o(n^4\text{Var}^2(T_{n,k}))\).

As \(k = o(n^{1/2})\), then for constant \(C_1 > 0\)

\[
E(Y_{11}^2) = \sum_{|i_1 - j_1| \leq k} \sum_{|i_2 - j_2| \leq k} \sum_{|i_3 - j_3| \leq k} \sum_{|i_4 - j_4| \leq k} \left[ \sigma_{i_1i_2}\sigma_{i_1j_1}\sigma_{i_3i_4}\sigma_{i_3j_4} \times E\{(x_{i_1i_2}x_{i_1j_1} - \sigma_{i_1j_1})(x_{i_1i_2}x_{i_3j_3} - \sigma_{i_3j_3})(x_{i_3i_4}x_{i_3j_4} - \sigma_{i_3j_4})\} \right] \leq C_1(2k+1)^4\text{tr}^2(\Sigma^4) = o(n^5\text{Var}^2(T_{n,k}))
\]

so that \(\text{Var}(Y_{11}) = o\{n^5\text{Var}^2(T_{n,k})\}\). Similarly, for some constant \(C_2 > 0\)

\[
\text{Var}(Y_{12}) = \sum_{|i_1 - j_1| \leq k} \sum_{|i_2 - j_2| \leq k} \sum_{|i_3 - j_3| \leq k} \sum_{|i_4 - j_4| \leq k} \left[ \sigma_{i_1i_2}\sigma_{i_1j_1}\sigma_{i_3i_4}\sigma_{i_3j_4} \times \{\sigma_{i_1i_3}\sigma_{j_1j_3} + \sigma_{i_1j_3}\sigma_{j_1i_3} + \Delta f_{i_1j_1i_3j_3}(\Sigma)\}\{\sigma_{i_2i_4}\sigma_{j_2j_4} + \sigma_{i_2j_4}\sigma_{j_2i_4} + \Delta f_{i_2j_2i_4j_4}(\Sigma)\} \right] \leq C_2(2k+1)^4\text{tr}(\Sigma^8) = o(n^4\text{Var}^2(T_{n,k}))
\]

Thus,

\[
\text{Var}\{\text{tr}\{(B_k(Q_{l-1})\Sigma)^2\}\} = t^2o(n^4\text{Var}^2(T_{n,k})),
\]

which implies that for some positive constants \(\gamma, \gamma_1\) and \(\gamma_2\)

\[
\text{Var}(R_{3,n}) \leq \gamma n^{-8}\text{Var}\left[ \sum_{t=1}^{n} \text{tr}\{(B_k(Q_{l-1})\Sigma)^2\}\right] \leq \gamma_1 n^{-5}\text{Var}(Y_{11}) + \gamma_2 n^{-4}\text{Var}(Y_{12})
\]

\[
= o(\text{Var}^2(T_{n,k})).
\]

Likewise, we can show that \(\text{Var}(R_{i,n}) = o(\text{Var}^2(T_{n,k}))\) for \(i = 2, 4\), by which we obtain the first part of (3.23) using the concentration property.

It remains to show the second part of (3.23). By standard algebraic computations, one can show that \(D_{t,k} = M_{t,1} + M_{t,2}\) where

\[
M_{t,1} = 2a_n(n(n-1))^{-1}\left[ X_t^\top B_k(Q_{l-1})X_t - \text{tr}\{B_k(Q_{l-1})\Sigma\} \right],
\]

\[
M_{t,2} = n^{-1}\left[ X_t^\top \{2a_nB_k(\Sigma) + b_nI_p\}X_t - \text{tr}\{(2a_nB_k(\Sigma) + b_nI_p)\Sigma\} \right].
\]
3.7.2 Proof of Proposition 3.1

Since \( \sum_{t=1}^{n} E(D_{t,k}^4) \leq C \left\{ \sum_{t=1}^{n} E(M_{t,1}^4) + \sum_{t=1}^{n} E(M_{t,2}^4) \right\} \) for constant \( C > 0 \), it suffices to show \( E(M_{t,i}^4) = o(\text{Var}^2(T_{n,k})) \) for \( i = 1, 2 \). By Cauchy-Schwartz inequality, for some constant \( c > 0 \),

\[
E \left[ X_t^\top B_k(Q_{t-1})X_t - tr \{ B_k(Q_{t-1})\Sigma \} \right]^4 \leq cE \left[ \text{tr}^2 \{ (B_k(Q_{t-1})\Sigma)^2 \} \right].
\]

On the other hand,

\[
E \left[ \text{tr} \{ (B_k(Q_{t-1})\Sigma)^2 \} \right] = (t-1) \sum_{|i_1-j_1| \leq k} \sum_{|i_2-j_2| \leq k} \{ \sigma_{i_1i_2}\sigma_{j_1j_2} + \sigma_{i_1j_2}\sigma_{i_2j_1} + \Delta f_{i_1j_1i_2j_2}(\Sigma) \} \sigma_{i_1i_2}\sigma_{j_1j_2},
\]

and

\[
\text{Var} \left[ \text{tr} \{ (B_k(Q_{t-1})\Sigma)^2 \} \right] = t^2o \left(n^2\text{Var}(T_{n,k}) \right)^2,
\]

we have \( E \left[ \text{tr}^2 \{ (B_k(Q_{t-1})\Sigma)^2 \} \right] = t^2O \left(n^2\text{Var}(T_{n,k}) \right)^2 \). Therefore, it yields

\[
\sum_{t=1}^{n} E(M_{t,1}^4) = 16a_n^4n^{-8} \sum_{t=1}^{n} E \left[ X_t^\top B_k(Q_{t-1})X_t - tr \{ B_k(Q_{t-1})\Sigma \} \right]^4
\]

\[
\leq n^4 O \left(n^2\text{Var}(T_{n,k}) \right)^2 = o(\text{Var}(T_{n,k})).
\]

Similarly, for some constant \( C > 0 \),

\[
\sum_{t=1}^{n} E(M_{t,2}^4) = n^{-4} \sum_{t=1}^{n} E \left[ X_t^\top \{ 2a_nB_k(\Sigma) + b_nI_p \}X_t - tr \{ (2a_nB_k(\Sigma) + b_nI_p)\Sigma \} \right]^4
\]

\[
\leq Cn^{-3}t^2 \left[ \{ 2a_nB_k(\Sigma) + b_nI_p \}\Sigma \right]^2,
\]

which implies \( \sum_{t=1}^{n} E(M_{t,2}^4) = o \left( \text{Var}^2(T_{n,k}) \right) \). We therefore obtain the second part in (3.23), and derive the assertion of Lemma 3.2.

\[\square\]

3.7.2 Proof of Proposition 3.1

Simple algebraic computations yield

\[
\hat{\sigma}_{n,k}^2 = 2\{n(n-1)\}^{-1} \sum_{|i_1-j_1| \leq k} \sum_{|i_2-j_2| \leq k} \{ \Pi_1 - 2\Pi_2 + \Pi_3 \},
\]

where

\[
\Pi_1 = \frac{1}{P^3} \sum_{l_1 \neq l_2} \{ (x_{1i_1}x_{1j_1} - \sigma_{i_1j_1})(x_{1i_2}x_{1j_2} - \sigma_{i_2j_2})(x_{1i_1}x_{1j_1} - \sigma_{i_1j_1})(x_{1i_2}x_{1j_2} - \sigma_{i_2j_2}) \},
\]

\[
\Pi_2 = \frac{1}{P^3} \sum_{l_1,l_2,l_3} \{ (x_{1i_1}x_{1j_1} - \sigma_{i_1j_1})(x_{1i_2}x_{1j_2} - \sigma_{i_2j_2})(x_{1i_3}x_{1j_3} - \sigma_{i_3j_3})(x_{1i_2}x_{1j_2} - \sigma_{i_2j_2}) \},
\]

\[
\Pi_3 = \frac{1}{P^3} \sum_{l_1,l_2,l_3} \{ (x_{1i_1}x_{1j_1} - \sigma_{i_1j_1})(x_{1i_2}x_{1j_2} - \sigma_{i_2j_2})(x_{1i_3}x_{1j_3} - \sigma_{i_3j_3})(x_{1i_2}x_{1j_2} - \sigma_{i_2j_2}) \}.
\]
Thus, $\text{E}(\hat{\sigma}^2_{n,k0}) = \sigma^2_{Vn,k0}$. Similar to Lemma 3.1,

$$
\text{Var} \left( \sum_{|i-i_j| \leq k} \sum_{|i_2-j_2| \leq k} \Pi_1 \right) = \left( -4n^{-1} \sum_{|i-i_j| \leq k} \sum_{|i_2-j_2| \leq k} \{ \text{E}(x_{i_1}x_{j_1}x_{i_2}x_{j_2}) - \sigma_{i_1} \sigma_{i_2} \sigma_{j_1} \sigma_{j_2} \}^2 \right) + 4n^{-1} \sum_{|i-i_j| \leq k} \sum_{|i_2-j_2| \leq k} \sum_{|i_3-j_3| \leq k} \sum_{|i_4-j_4| \leq k} \left[ \{ \text{E}(x_{i_1}x_{i_2}x_{i_3}x_{i_4}) - \sigma_{i_1} \sigma_{i_2} \} \{ \text{E}(x_{j_1}x_{j_2}x_{j_3}x_{j_4}) - \sigma_{j_1} \sigma_{j_2} \} \} \times \text{E}\{ (x_{i_1}x_{i_2}x_{i_3}x_{i_4}) (x_{j_1}x_{j_2}x_{j_3}x_{j_4}) \} \right] \{ 1 + O(n^{-1}) \}
$$

$$
= O \left( n^3 + k^2 n^3 \right) \sigma^2_{Vn,k0}.
$$

The variances for the remaining terms of $\hat{\sigma}^2_{n,k0}$ can be estimated similarly and the proposition follows.

### 3.7.3 Proof of Theorem 3.1

Denote

$$
\tilde{V}_{n,k} = \sum_{|i-j| \leq k} L_{n_1}(i,j) - 2 \sum_{i=1}^p L_{n_4}(i) + p.
$$

Under Assumptions 3.1 and 3.2, $\text{E}(\tilde{V}_{n,k}) = \text{tr}\{ (B_k(\Sigma) - I_p)^2 \}$. From Lemma 3.1, we have $\text{Var}(\tilde{V}_{n,k}) = \sigma^2_{Vn,k} + o(\sigma^2_{Vn,k})$, $\text{Var}\{ \sum_{|i-j| \leq k} L_{n_2}(i,j) \} = o(\sigma^2_{Vn,k})$, $\text{Var}\{ \sum_{|i-j| \leq k} L_{n_3}(i,j) \} = o(\sigma^2_{Vn,k})$, and $\text{Var}\{ \sum_{i=1}^p L_{n_5}(i) \} = o(\sigma^2_{Vn,k})$. Theorem 3.1 therefore follows Lemma 3.2 and Slutsky’s Theorem.

### 3.7.4 Proof of Theorem 3.2

It follows Theorems 3.1 and (3.16) that

$$
\beta_{V_{n,k}} = 1 - \Phi \left( \sigma_{V_{n,k}}^{-1} \sigma_{V_{n,k} 0} 2z_0 - \sigma_{V_{n,k}}^{-1} \text{tr}\{ (B_k(\Sigma) - I_p)^2 \} \right).
$$

As discussed before, it can be seen that

$$
\sigma_{V_{n,k}} \leq \text{tr}\{ (B_k(\Sigma) - I_p)^2 \} \left( \frac{\tau_{n,k}(\Sigma)}{\text{tr}^2\{ (B_k(\Sigma) - I_p)^2 \}} + \frac{8 + 4\Delta \text{tr}\{ (B_k(\Sigma) - I_p)^2 \}^2}{n \text{tr}^2\{ (B_k(\Sigma) - I_p)^2 \}} \right)^{1/2},
$$
so that under the condition \( \{\tau_{n,k}(\Sigma)\}^{-1}\tr\left[\{B_k(\Sigma) - I_p\}^2\right] \to \infty \), it suffices to show
\[
\frac{\tr[\Sigma\{B_k(\Sigma) - I_p\}]^2}{n\tr^2\left[\{B_k(\Sigma) - I_p\}^2\right]} \to 0,
\]
(3.24)
provided \( \sigma_{\tr,n,k}^{-1}\sigma_{\tr,n,k,0} \) is bounded.

Denote \( \lambda_1 \leq \cdots \leq \lambda_p \) the eigenvalues of \( B_k(\Sigma) \) that for some \( \delta_0, 0 < \delta_0 \leq \lambda_1 \leq \lambda_p \leq \delta_0^{-1} \) for sufficiently large \( k \). Standard algebraic computations give
\[
\tr\left[\Sigma\{B_k(\Sigma) - I_p\}\right]^2 = \tr(B_k(\Sigma)\{B_k(\Sigma) - I_p\})^2 + \tr(\{\Sigma - B_k(\Sigma)\}\{B_k(\Sigma) - I_p\})^2 + 2\tr(B_k(\Sigma)\{B_k(\Sigma) - I_p\}\{\Sigma - B_k(\Sigma)\}\{B_k(\Sigma) - I_p\})
\]
\[
\equiv I_1 + I_2 + I_3.
\]
By the fact that \( ||\Sigma - B_k(\Sigma)|| \leq Ck^{-\alpha}, I_i = o\left(n\tr^2\left[\{B_k(\Sigma) - I_p\}^2\right]\right) \) for \( i = 2, 3 \). Notice that
\[
\tr(B_k(\Sigma)\{B_k(\Sigma) - I_p\})^2 = \sum_{i=1}^p \lambda_i^2(\lambda_i - 1)^2 \leq \delta_0^{-2}\sum_{i=1}^p (\lambda_i - 1)^2 = \delta_0^{-2}\tr\left[\{B_k(\Sigma) - I_p\}^2\right],
\]
from which (3.24) follows. We therefore obtain Theorem 3.2.

### 3.7.5 Proof of Theorem 3.3

Denote
\[
\tilde{U}_{n,k} = \sum_{|i-j| \leq k} \frac{L_{m_1}(i,j)}{\tr(B_k(\Sigma)\Sigma)} - \frac{2\sum_{i=1}^p L_{m_4}(i)}{\tr(\Sigma)} + 1.
\]
By Lemmas 3.1 and 3.2, we have \( \text{Var}(\tilde{U}_{n,k}) = \sigma_{\tilde{U}_{n,k}}^2\{1 + o(1)\} \) and \( \sigma_{\tilde{U}_{n,k}}^{-1}\tilde{U}_{n,k} \to \mathcal{N}(0, 1) \) in distribution. Together with results that \( \text{Var}\left[\sum_{|i-j| \leq k} \frac{L_{m_2}(i,j)}{\tr(B_k(\Sigma)\Sigma)}\right] = o(\sigma_{\tilde{U}_{n,k}}^2), \)
\( \text{Var}\left(\sum_{i=1}^p L_{m_5}(i)/\tr(\Sigma)\right) = o(\sigma_{\tilde{U}_{n,k}}^2) \) and \( \text{Var}\left[\sum_{|i-j| \leq k} \frac{L_{m_3}(i,j)}{\tr(B_k(\Sigma)\Sigma)}\right] = o(\sigma_{\tilde{U}_{n,k}}^2) \), we have
\[
\sigma_{\tilde{U}_{n,k}}^{-1}\tilde{U}_{n,k} \to \mathcal{N}(0, 1)
\]
in distribution, where
\[
\tilde{U}_{n,k} = \frac{\sum_{|i-j| \leq k}\{L_{m_1}(i,j) - 2L_{m_2}(i,j) + L_{m_3}(i,j)\}}{\tr(B_k(\Sigma)\Sigma)} - \frac{2\sum_{i=1}^p \{L_{m_4}(i) - L_{m_5}(i)\}}{\tr(\Sigma)} + 1.
\]
Denote \( \epsilon_n = \left[\sum_{i=1}^p \{L_{m_4}(i) - L_{m_5}(i)\}\right] / \tr(\Sigma) \) with \( \mathbb{E}(\epsilon_n) = 0 \), then
\[
\text{Var}(\epsilon_n) = \tr^{-2}(\Sigma)\left[2n^{-1}\tr(\Sigma^2) + \Delta n^{-1}\tr\left\{\left(\Gamma^\top\Gamma\right)\circ\left(\Gamma^\top\Gamma\right)\right\} + 2\{n(n-1)\}^{-1}\tr(\Sigma^2)\right]
\]
\[
\leq [(2 + \Delta)n^{-1} + 2\{n(n-1)\}^{-1}]\tr^{-2}(\Sigma)\tr(\Sigma^2) = o(\sigma_{\tilde{U}_n}).
\]
Thus, Theorem 3.3 follows from
\[
\left( \frac{\text{tr}^2(\Sigma)}{\text{tr}\{B_k(\Sigma)\Sigma\}} \right) \left( \frac{U_{n,k} + 1}{p} \right) - 1 = \frac{\widehat{U}_{n,k} - \epsilon_n^2}{(1 + \epsilon_n^2)^2}.
\]

### 3.7.6 Proof of Theorem 3.4

It follows Theorems 3.3 and (3.21)
\[
\beta_{U_{n,k}} = 1 - \Phi \left( \left[ \frac{\text{tr}^2(\Sigma)}{\text{trace}(B_k(\Sigma)\Sigma)} \right] \left( \frac{\sigma_{U_{n,k}0}}{\sigma_{U_{n,k}}} \right) z - \sigma_{U_{n,k}}^{-1} \left[ 1 - \frac{\text{tr}^2(\Sigma)}{\text{trace}(B_k(\Sigma)\Sigma)} \right] \right).
\]

Since \(\sigma_{U_{n,k}0}/\sigma_{U_{n,k}}\) and \(p^{-1}\text{tr}^2(\Sigma)/\text{trace}\{B_k(\Sigma)\Sigma\}\) are bounded and
\[
\sigma_{U_{n,k}} \leq \left[ \frac{\tau_{n,k}^2(\Sigma)}{\text{trace}\{B_k(\Sigma)\Sigma\}\Sigma^2} + \frac{8 + 4\Delta}{n} \text{tr} \left\{ \left( \frac{B_k(\Sigma)\Sigma - \Sigma}{\text{trace}(B_k(\Sigma)\Sigma)} \right)^2 \right\} \right]^{1/2},
\]
similar to Theorem 3.2, under the condition \(\{\tau_{n,k}(\Sigma)\}^{-1}[\text{trace}\{B_k(\Sigma)\Sigma - \text{tr}^2(\Sigma)/p\}] \to \infty\), it is sufficient to show
\[
n^{-1}\text{tr} \left\{ \left( \frac{\Sigma^2}{\text{trace}(\Sigma^2)} - \frac{\Sigma}{\text{trace}(\Sigma)} \right)^2 \right\} \left\{ 1 - \frac{\text{tr}^2(\Sigma)}{p\text{trace}(\Sigma^2)} \right\}^{-2} \to 0. \tag{3.25}
\]

Standard algebra and definition in (3.6) implies that
\[
\text{tr} \left\{ \{ \Sigma^2 - \text{tr}^{-1}(\Sigma)\text{trace}(\Sigma^2)\Sigma \}^2 \right\}
\leq \epsilon_0^{-2} \sum_{i=1}^{p} \left( \lambda_i(\Sigma) - \text{tr}^{-1}(\Sigma)\text{trace}(\Sigma^2) \right)^2
= \epsilon_0^{-2}\text{tr}^2(\Sigma) \sum_{i=1}^{p} \left( \text{tr}^{-1}(\Sigma)\lambda_i(\Sigma) - p^{-1} + p^{-1} - \text{tr}^{-2}(\Sigma)\text{trace}(\Sigma^2) \right)^2
= \epsilon_0^{-2}\text{tr}^2(\Sigma) \left[ \text{tr} \left\{ (\text{tr}^{-1}(\Sigma)\Sigma - p^{-1} I_p)^2 \right\} + p \left\{ p^{-1} - \text{tr}^{-2}(\Sigma)\text{trace}(\Sigma^2) \right\}^2 \right],
\]
and \(\{\text{tr}(\Sigma^2) - p^{-1}\text{tr}^2(\Sigma)\}^2 = \text{tr}(\Sigma)\text{trace} \left\{ (\text{tr}^{-1}(\Sigma)\Sigma - p^{-1} I_p)^2 \right\}, \) which implies (3.25) and Theorem 3.4 follows.

### 3.7.7 More numerical results
Table 3.5 Empirical sizes of the proposed test (ST) for the sphericity hypothesis, along with those of the tests by Chen, Zhang and Zhong (2010) (CZZ) and Ledoit and Wolf (2002) (LW).

<table>
<thead>
<tr>
<th></th>
<th>IT</th>
<th>CZZ</th>
<th>LW</th>
<th>IT</th>
<th>CZZ</th>
<th>LW</th>
<th>IT</th>
<th>CZZ</th>
<th>LW</th>
<th>IT</th>
<th>CZZ</th>
<th>LW</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>159</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>181</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>331</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>343</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>642</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>p</strong></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gaussian distributed innovation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>0.063</td>
<td>0.061</td>
<td>0.060</td>
<td>0.052</td>
<td>0.060</td>
<td>0.044</td>
<td>0.054</td>
<td>0.056</td>
<td>0.048</td>
<td>0.049</td>
<td>0.055</td>
<td>0.050</td>
</tr>
<tr>
<td>55</td>
<td>0.070</td>
<td>0.056</td>
<td>0.055</td>
<td>0.054</td>
<td>0.056</td>
<td>0.055</td>
<td>0.045</td>
<td>0.058</td>
<td>0.050</td>
<td>0.055</td>
<td>0.056</td>
<td>0.064</td>
</tr>
<tr>
<td>89</td>
<td>0.061</td>
<td>0.058</td>
<td>0.052</td>
<td>0.056</td>
<td>0.055</td>
<td>0.051</td>
<td>0.062</td>
<td>0.053</td>
<td>0.047</td>
<td>0.061</td>
<td>0.055</td>
<td>0.053</td>
</tr>
<tr>
<td>159</td>
<td>0.068</td>
<td>0.062</td>
<td>0.058</td>
<td>0.055</td>
<td>0.058</td>
<td>0.045</td>
<td>0.052</td>
<td>0.049</td>
<td>0.046</td>
<td>0.058</td>
<td>0.052</td>
<td>0.041</td>
</tr>
<tr>
<td>181</td>
<td>0.069</td>
<td>0.055</td>
<td>0.054</td>
<td>0.051</td>
<td>0.054</td>
<td>0.056</td>
<td>0.062</td>
<td>0.054</td>
<td>0.051</td>
<td>0.057</td>
<td>0.056</td>
<td>0.045</td>
</tr>
<tr>
<td>331</td>
<td>0.077</td>
<td>0.047</td>
<td>0.068</td>
<td>0.062</td>
<td>0.049</td>
<td>0.059</td>
<td>0.045</td>
<td>0.057</td>
<td>0.053</td>
<td>0.058</td>
<td>0.051</td>
<td></td>
</tr>
<tr>
<td>343</td>
<td>0.084</td>
<td>0.061</td>
<td>0.061</td>
<td>0.057</td>
<td>0.053</td>
<td>0.059</td>
<td>0.063</td>
<td>0.058</td>
<td>0.053</td>
<td>0.068</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>642</td>
<td>0.059</td>
<td>0.051</td>
<td>0.064</td>
<td>0.041</td>
<td>0.047</td>
<td>0.065</td>
<td>0.069</td>
<td>0.062</td>
<td>0.039</td>
<td>0.060</td>
<td>0.051</td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gamma distributed innovation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>0.089</td>
<td>0.083</td>
<td>0.187</td>
<td>0.063</td>
<td>0.078</td>
<td>0.166</td>
<td>0.062</td>
<td>0.056</td>
<td>0.210</td>
<td>0.053</td>
<td>0.054</td>
<td>0.196</td>
</tr>
<tr>
<td>55</td>
<td>0.071</td>
<td>0.183</td>
<td>0.071</td>
<td>0.069</td>
<td>0.184</td>
<td>0.047</td>
<td>0.070</td>
<td>0.181</td>
<td>0.056</td>
<td>0.064</td>
<td>0.184</td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>0.074</td>
<td>0.193</td>
<td>0.065</td>
<td>0.056</td>
<td>0.185</td>
<td>0.064</td>
<td>0.058</td>
<td>0.198</td>
<td>0.056</td>
<td>0.046</td>
<td>0.205</td>
<td></td>
</tr>
<tr>
<td>159</td>
<td>0.057</td>
<td>0.183</td>
<td>0.067</td>
<td>0.066</td>
<td>0.197</td>
<td>0.062</td>
<td>0.055</td>
<td>0.191</td>
<td>0.063</td>
<td>0.051</td>
<td>0.178</td>
<td></td>
</tr>
<tr>
<td>181</td>
<td>0.081</td>
<td>0.190</td>
<td>0.065</td>
<td>0.058</td>
<td>0.181</td>
<td>0.071</td>
<td>0.053</td>
<td>0.183</td>
<td>0.052</td>
<td>0.055</td>
<td>0.181</td>
<td></td>
</tr>
<tr>
<td>331</td>
<td>0.084</td>
<td>0.191</td>
<td>0.063</td>
<td>0.059</td>
<td>0.204</td>
<td>0.061</td>
<td>0.043</td>
<td>0.173</td>
<td>0.057</td>
<td>0.056</td>
<td>0.208</td>
<td></td>
</tr>
<tr>
<td>343</td>
<td>0.078</td>
<td>0.201</td>
<td>0.070</td>
<td>0.059</td>
<td>0.199</td>
<td>0.065</td>
<td>0.059</td>
<td>0.175</td>
<td>0.061</td>
<td>0.063</td>
<td>0.172</td>
<td></td>
</tr>
<tr>
<td>642</td>
<td>0.069</td>
<td>0.199</td>
<td>0.074</td>
<td>0.057</td>
<td>0.171</td>
<td>0.064</td>
<td>0.046</td>
<td>0.172</td>
<td>0.069</td>
<td>0.071</td>
<td>0.163</td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.9  Comparison of heat maps of correlation matrices before and after reordering of a gene set, GO:0000086 from BP category in NEG group.
Figure 3.10  Histograms of $p$-values for identity hypotheses (right panels) and sphericity hypotheses (left panels) for the three gene ontologies by the proposed tests.
CHAPTER 4. VARIABLE IMPORTANCE ASSESSMENT AND BACKWARD VARIABLE SELECTION FOR HIGH-DIMENSIONAL DATA

A paper to be submitted to *Bioinformatics*

Liuhua Peng, Long Qu, Dan Nettleton

Abstract

Variable selection in high-dimensional scenarios is of great interest in statistics. One application involves identifying differentially expressed genes in genomic analysis. Existing methods for addressing this problem have some limits or disadvantages. Many rely on parametric assumptions that may not be satisfied in practice. Many focus only on univariate marginal distributions and ignore information contained in multivariate distributions. Other approaches fail when the number of variables or the number of important variables is large relative to the sample size. In this paper, we propose distance based variable importance measures to deal with these problems, which are inspired by the Multi-Response Permutation Procedure (MRPP) and closely related energy distance between distributions. These variable importance assessments can effectively measure the importance of an individual dimension by quantifying its influence on the differences between two multivariate distributions. A backward selection algorithm is developed that can be used in high-dimensional variable selection to discover important variables and to focus multivariate tests for distributional differences on the most relevant dimensions, which improves the power of the original MRPP. Both simulations and real data applications demonstrate that our proposed method enjoys good properties and has advantages over other methods.
4.1 Introduction

With the explosive and continued advancement of high-throughput biotechnologies, simultaneous measurement of more and more biological variables from any single experimental subject has become increasingly affordable and is frequently used in biomedical research. Typical examples of such experiments include microarray-based gene expression profiling, next-generation sequencing-based gene expression profiling (mRNA-Seq), genome-wide association (GWA) and genomic selection (GS) studies based on high-throughput genotyping, genome-wide scanning of copy number variations (CNV), expression quantitative trait loci (eQTL) mapping, identification of genomic locations interacting with proteins through chromatin immunoprecipitation (ChIP) followed by microarray (ChIP-Chip) or sequencing (ChIPSeq) technologies, identification of brain regions associated with activity or disease by modern imaging technologies (e.g., functional magnetic resonance imaging, or fMRI), and the emerging metagenome-wide association studies (meta-GWAS) where the composition of microbial genomes are targeted for association with complex diseases.

High-dimensional variable selection is in great need in multiple scientific disciplines, particularly in modern genomics and personalized medicine. Identifying genes that differ in expression across two or more treatments or conditions is of great interest in genomic analysis. Identification of differentially expressed (DE) genes not only gives information about gene functionality, but also provides insight into the molecular genetic mechanisms underlying biological processes. Microarray and RNA-seq technologies enable researchers to simultaneously measure thousands of potentially interesting variables. A distinguishing feature of these applications is that only a very limited number of experimental units (subjects) can be measured due to expense, leading to the so-called “small n, large p” problem. Furthermore, the response variables are expected to have very complex dependence structures governed by underlying biological processes that are not well characterized. In such a complex setting, only a small number of the most interesting and biologically meaningful variables or groups of variables are the primary targets for in-depth investigation.
Although variable selection is not a new problem in statistics, existing statistical tools for variable selection in such high-dimensional contexts are still limited in capability. Current methods usually suffer from one or more of the following shortcomings: the number of selected variables cannot exceed the sample size; variable importance is evaluated based on a comparison of univariate marginal distributions; variables are selected in a forward manner; variable importance is based on distorted dependence structures that are not evident in the data; strong model assumptions (typically on the mean structure) are imposed during variable selection; and the selected variables can only capture certain aspects of the dependency. Because of these limitations, there is a great gap between the needs of biological researchers and the capability of existing statistical tools for variable selection. This paper aims at avoiding or alleviating these aforementioned shortcomings.

The Multi-Response Permutation Procedure (MRPP) described by Mielke and Berry (2007) is a powerful tool that can detect differences between multivariate distributions. The test statistic is based on a weighted average of within-treatment pairwise distance and the testing procedure is carried out by permuting the observations. Moreover, under some mild conditions, the MRPP test is equivalent to the test proposed in Székely and Rizzo (2004), which was inspired by the properties of the so called “energy distance” (Székely and Rizzo, 2013). Distance-based methods such as the MRPP and energy statistics, have good features when dealing with multivariate and even high-dimensional problems.

Inspired by these former works, in this paper, we introduce distance-based variable importance measures for high-dimensional contexts that automatically take covariance structures into consideration. We also propose a backward selection algorithm that can be used to select variables and improve multivariate tests for distributional differences by concentrating the power of tests on signal-bearing dimensions. Examples in both real data and simulation studies show that our method has good performance when detecting differentially expressed genes in genomic analysis.

The paper is organized as follows. Two ways of calculating variable importance are described in Section 4.2. Section 4.3 introduces the proposed backward selection algorithm. Section 4.4 presents a test for differences in multivariate distributions that begins with the proposed
backward selection procedure to concentrate the power of the test on important dimensions. Section 4.5 gives an example of applying our method on a real data set. Simulation studies are shown in Section 4.6.

### 4.2 Variable importance

#### 4.2.1 The multi-response permutation procedure

The MRPP is described as a “distance function approach” by Mielke and Berry (2007). Briefly, the MRPP chooses a distance measure and uses a weighted average of within-treatment pairwise distances as the test statistic. This strategy is motivated by using the error sums of squares as a test statistic in the analysis of variance, which is known to be permutationally equivalent to using the \( F \)-statistic.

Consider a \( K \)-sample comparison experiment with an \( R \)-dimensional response vector. Let \( Y_i \) be the \( i \)-th observation of the \( R \)-dimensional response vector, with \( r \)-th element \( Y_{i,r} \). The MRPP distance measure between observations \( i \) and \( j \) is usually chosen to be Euclidean distance,

\[
\Delta(i, j) = \sqrt{\sum_{r=1}^{R} (Y_{i,r} - Y_{j,r})^2} = ||Y_i - Y_j||. \tag{4.1}
\]

Suppose we have a total of \( N \) independent observations, each of which comes from exactly one of the \( K \) treatments. Let \( M_b(i) \) be the treatment label of observation \( i \) under the \( b \)-th permutation of the \( N \) observations, where \( b \) is the factoradic number that indexes all \( B = N! \) permutations of the observations, and \( b = 0 \) indicates the original assignment of treatment labels to observations. Further let \( n_k \) be the sample size in the \( k \)-th treatment such that \( N = \sum_{k=1}^{K} n_k \). The MRPP test statistic is

\[
z_b(\Delta) = \sum_{k=1}^{K} C_k \left\{ \frac{2}{n_k(n_k - 1)} \sum_{(i,j) \in T_b(k)} \Delta(i, j) \right\}, \tag{4.2}
\]

where \( C_k \) is the group weight usually chosen to be proportional to \( n_k/N \) or \( (n_k - 1)/(N - K) \), and \( T_b(k) = \{(i, j) : M_b(i) = M_b(j) = k, \ i < j, \ i = 1, 2, \ldots, N, \ j = 1, 2, \ldots, N\} \). The final permutation \( p \)-value for testing the null hypothesis of no distributional difference across the \( K \)
treatments is

\[ p(\Delta) = \frac{1}{B} \sum_{b=0}^{B-1} I \{ z_0(\Delta) \geq z_b(\Delta) \}. \] (4.3)

Let \( M \) be the sorted non-redundant set of sample sizes \( \{ n_i : i = 1, 2, \ldots, K \} \) with the \( j \)-th element \( m_j \), for \( j = 1, 2, \ldots, |M| \). Because all within treatment permutations are equivalent and exchanging treatment labels between treatment \( i \) and \( i' \) when \( n_i = n_{i'} \) also results in the same test statistic, the support of the \( p \)-value is the discrete set \( \{ b'/B' : b' = 1, 2, \ldots, B' \} \) where \( B' = \frac{N!}{\prod_{i=1}^{K} n_i! \prod_{j=1}^{m_j} m_j!} \). When \( B' \) is large, we may randomly sample \( \tilde{B} \ll B' \) permutations from \( B' \) non-equivalent permutations to save computing time. Whether using all \( B' \) permutations or using a random subset of permutations, the type I error rate is bounded above by \( \alpha \) when the null hypothesis of distributional equality is rejected if and only if \( p \leq \alpha \).

The MRPP has the advantage of recognizing, accounting for, and utilizing dependence information among the \( R \)-dimensions and capturing information about the joint distribution rather than only each marginal distribution. It has demonstrated good performance in the context of gene set testing (Nettleton, Recknor and Reecy, 2008). The associated multiple testing problem (Liang and Nettleton, 2010) and variations of the test for more targeted hypotheses on variances (Qu et al., 2008) have been addressed.

### 4.2.2 An MRPP-based importance measure

The focus of the MRPP is on detecting differences between multivariate distributions. The procedure does not provide a measure of the importance of any one variable with respect to the information it contains about distributional differences. In this section, we introduce a method that ranks the importance of the \( R \) response variables and performs response variable selection for the MRPP. Briefly, the ranking procedure consists of a hypothetical perturbation method that tilts each of the \( R \) dimensions and a scheme for assessing the effects of such perturbations. Intuitively, the dimensions that lead to large differences in results under a small perturbation will be more influential and potentially more important than other dimensions.

Because the MRPP procedure relies on the distance measure \( \Delta \), it is natural to consider a weighted Euclidean distance as an extension of the Euclidean distance for use in the MRPP.
Let \( \omega \) be an \( R \)-vector with \( r \)-th element \( \omega_r \geq 0 \) being the weight for the \( r \)-th dimension. Then the weighted Euclidean distance between observations \( i \) and \( j \) is

\[
\Delta_\omega(i,j) = \sqrt{\sum_{r=1}^{R} \omega_r(Y_{i,r} - Y_{j,r})^2}.
\] (4.4)

When \( \omega_r = 1 \) for all \( r = 1, 2, \ldots, R \), (4.4) is equivalent to (4.1). Note that introducing the weights is only conceptual and in practice \( \omega_r \) can be always set to 1. The advantage of using weights is that we can hypothetically increase or decrease some \( \omega_r \) as a means of data perturbation for the purpose of evaluating variable importance. For example, setting \( \omega_r = 0 \) is equivalent to omitting dimension \( r \) from analysis, which is similar to dropping a regressor in regression variable selection.

Given our method of perturbation, we now seek a measure of the effect of the perturbation. Because the end result for a permutation test is a permutation \( p \)-value, it is reasonable to consider how much the permutation \( p \)-value is changed by perturbation. However, because of the discreteness of the support of the permutation \( p \)-values defined in (4.3), if the perturbation in \( \omega_r \) is too small, the permutation \( p \)-value \( p(\Delta_\omega) \) may not change. On the other hand, we also want the perturbation to be as small as possible to faithfully reflect the original data set. To solve this conflict, we consider an approximation to the discrete permutation \( p \)-value in (4.3) by a continuous \( p \)-value. Our choice is to treat the \( B \) permutation test statistics as a random sample of size \( B \) from an infinite population with a cumulative distribution function (CDF) \( F \) and density \( f \), and to apply the kernel method to estimate \( F \) and \( f \). Using the Gaussian kernel with bandwidth \( h \), the kernel estimate of \( F \) is

\[
\hat{F}(z) = \frac{1}{B} \sum_{b=0}^{B-1} \Phi \left\{ \frac{z - z_b(\Delta_\omega)}{h} \right\},
\]

where \( \Phi \) is the CDF for the standard normal distribution. The continuous approximation to the discrete \( p \)-value in (4.3) is then given by \( \hat{p}(\Delta_\omega) = \hat{F}(z_0(\Delta_\omega)) \), evaluated at \( \omega_r = 1 \) for all \( r = 1, 2, \ldots, R \).

The choice of bandwidth \( h \) is well known to be crucial for the performance of kernel density estimation. We choose to select \( h \) to minimize the mean squared error of \( \hat{f} = \hat{F}' \) at the fixed point \( z_0(\Delta) \). The asymptotically optimal \( h \) is known (Parzen, 1962) and can be estimated
by plugging in initial kernel estimates of $f$ and $f''$ (Wand and Jones, 1995) using an ad hoc bandwidth determined by, e.g., Scott’s rule (Scott, 1992).

We can now compute the importance, $\iota_r$, of variable $r$ as the partial derivative of the continuous $p$-value $\tilde{p}$ with respect to weight $\omega_r$, evaluated at $\omega_r = 1$ for $r = 1, 2, \cdots, R$. If the derivate is negative, then increasing the weight will decrease the $p$-value, i.e., the $r$-th dimension is important. On the other hand, if the derivative is positive, then increasing the weight will increase the $p$-value, and focusing more on the $r$-th dimension diminishes the significance of the MRPP.

Specifically, the importance (or the influence) of the $r$-th dimension is computed as

$$\iota_r = \frac{\partial \tilde{p}}{\partial \omega_r} \bigg|_{\omega = 1} = \frac{1}{Bh} \sum_{b=0}^{B-1} \phi \left\{ \frac{z_0(\Delta) - z_b(\Delta)}{h} \right\} \{ z_0(\nabla_r) - z_b(\nabla_r) \},$$

(4.5)

where $\phi$ is the standard normal density and

$$\nabla_r(i,j) = \frac{\partial \Delta(\omega)}{\partial \omega_r} (i,j) \bigg|_{\omega = 1} = \frac{(Y_{i,r} - Y_{j,r})^2}{2\Delta(i,j)}$$

for all $b = 0, 1, \cdots, B - 1$. Notice that $\iota_r$ does not depend on $\iota_s$ for $s \neq r$. Hence, even though the computational load is the same as a separate permutation for each dimension, simple parallel algorithms are readily available for computing the importance for different dimensions by different processors.

Other than taking account of the dependency and robustness to normality, another important advantage of MRPP – allowing the dimensionality to exceed the sample size – is also inherited by our variable ranking procedure. When we measure the importance of the $r$-th dimension, the remaining $R - 1$ dimensions have not been excluded from the data, even if $R - 1 > N$. This allows backward variable selection procedures to be possible in high dimensions and is particularly advantageous compared to marginal screening procedures, for example, the marginal Pearson correlation screening (Fan and Lv, 2008), the marginal distance correlation screening (Li, Zhong and Zhu, 2012), the marginal maximal information coefficient (MIC) (Reshef et al., 2011; Speed, 2011), or the marginal empirical likelihood screening (Chang, Tang and Wu, 2013, 2016).

Compared to similar permutation methods that permute each dimension separately to assess variable importance, e.g., as in the random forest procedure (Breiman, 2001), our method
does not distort the inter-relationship between the variable under consideration and the remaining variables. Thus, our method is more faithful to the observed data and reflects the true importance of a variable in the joint distribution of all response variables, rather than in the distribution where the variable under consideration and the remaining variables are artificially decorrelated through permutation. This is very important in terms of biological interpretations. In molecular biology, it is well known that intracellular environment is crucial and genes interact with each other in a complex manner. The same gene may have different functions depending on how related genes are expressed. Therefore, a statistical procedure that assesses whether a gene is important must account for the expression levels of other genes. Because our method does this accounting, it provides a potentially more meaningful solution to molecular biology researchers.

4.2.3 An alternative importance measure

In Section 4.2.2, the variable importance measure \( \iota_r \) was derived by using a kernel-smoothing approximation to the permutation distribution of the MRPP test statistic. As argued in Section 4.2.2, we believe \( \iota_r \) is an intuitively appealing measure of importance that avoids drawbacks of other approaches. However, \( \iota_r \) has its own drawbacks. First, the reliance on the permutation distribution of the MRPP statistic involves nontrivial computational expense in high-dimensional problems. Second, the need to specify a bandwidth parameter \( h \) is an inconvenience. In this section, we propose an alternative variable importance measure that takes \( \iota_r \) as a starting point and attempts to eliminate its drawbacks while maintaining its appealing features. For ease of exposition, we focus on the two-sample problem here; the extension to \( K \) groups is straightforward.

Let \( Y_1, \ldots, Y_{n_1} \sim F_1 \) and \( Y_{n_1+1}, \ldots, Y_N \sim F_2 \) where \( N = n_1 + n_2 \) and all \( N \) vector-valued observations are independent. According to the definition given in Section 4.2.2, the importance of the \( r \)-th dimension is

\[
\iota_r = \frac{1}{Bh} \sum_{b=0}^{B-1} \phi \left( \frac{z_0(\Delta) - z_b(\Delta)}{h} \right) \left\{ z_0(\nabla_r) - z_b(\nabla_r) \right\},
\]

where \( z_0(\cdot) \) and \( z_b(\cdot) \) are cumulative distribution functions, and \( \phi \) is a kernel function.
where
\[ \nabla_r(i, j) = \frac{(Y_{i,r} - Y_{j,r})^2}{2 \Delta(i, j)} = \frac{(Y_{i,r} - Y_{j,r})^2}{2 \sqrt{\sum_{s=1}^{R}(Y_{i,s} - Y_{j,s})^2}}. \]

Note that \( \nabla_r(i, j) \) is very similar to
\[ A_r(i, j) \equiv \frac{(Y_{i,r} - Y_{j,r})^2}{\sqrt{\sum_{s=1}^{R}(Y_{i,s} - Y_{j,s})^2} + \sqrt{\sum_{s \neq r}(Y_{i,s} - Y_{j,s})^2}}. \]

Furthermore, with \( Y^{(r)}_i \equiv (Y_{i,1}, \ldots, Y_{i,r-1}, Y_{i,r+1}, \ldots, Y_{i,R})^T \) for all \( i = 1, \ldots, N \), it is straightforward to show that
\[ A_r(i, j) = ||Y_i - Y_j|| - ||Y^{(r)}_i - Y^{(r)}_j||, \]
the Euclidean distance between \( Y_i \) and \( Y_j \) less the Euclidean distance between \( Y_i \) and \( Y_j \) when the \( r \)-th dimension is dropped from both vectors. As we subsequently show, there are advantages to replacing \( \nabla_r(i, j) \) in \( \iota_r \) with \( A_r(i, j) \) and dropping the kernel function term in \( \iota_r \) to obtain the following alternative measure of the importance of the \( r \)-th variable:
\[ \tau_r = \frac{1}{B} \sum_{b=0}^{B-1} \{ z_0(A_r) - z_b(A_r) \} = z_0(A_r) - \frac{1}{B} \sum_{b=0}^{B-1} z_b(A_r). \]

Although this alternative measure of variable importance seems to require computation for \( B \) data permutations, simple algebra shows that
\[ \frac{1}{B} \sum_{b=0}^{B-1} z_b(A_r) = \frac{1}{N(N-1)} \sum_{i \neq j} A_r(i, j) \]
\[ = \frac{1}{N(N-1)} \sum_{i \neq j} \left( ||Y_i - Y_j|| - ||Y^{(r)}_i - Y^{(r)}_j|| \right). \]

If with take \( C_k \) in (4.2) to be \( n_k/N \) for \( k = 1, 2 \), we have
\[ \tau_r = \frac{2}{N(n_1 - 1)} \sum_{(i,j) \in T_0(1)} \left( ||Y_i - Y_j|| - ||Y^{(r)}_i - Y^{(r)}_j|| \right) \]
\[ + \frac{2}{N(n_2 - 1)} \sum_{(i,j) \in T_0(2)} \left( ||Y_i - Y_j|| - ||Y^{(r)}_i - Y^{(r)}_j|| \right) \]
\[ - \frac{2}{N(N - 1)} \sum_{1 \leq i < j \leq N} \left( ||Y_i - Y_j|| - ||Y^{(r)}_i - Y^{(r)}_j|| \right). \]

Thus, computation of \( \tau_r \) requires calculation of \( ||Y_i - Y_j|| - ||Y^{(r)}_i - Y^{(r)}_j|| \) for all \( i \neq j \) but avoids the need to do calculations for multiple permutations. Also \( \tau_r \) is free of any redundant parameters.
Another advantage of $\tau_r$ over $\iota_r$ is its relationship to the concept of energy distance (Székely and Rizzo, 2004, 2005, 2013). If the $R$-dimensional random vectors $X$ and $Y$ are independent with $E\|X\| + E\|Y\| < \infty$, their energy distance is defined as

$$
\varepsilon(X, Y) = 2E\|X - Y\| - E\|X - X'\| - E\|Y - Y'\|,
$$

where $X'$ is an i.i.d copy of $X$ and $Y'$ is an i.i.d copy of $Y$. The work of Székely and Rizzo shows that $\varepsilon(X, Y) \geq 0$ with equality to zero if and only if $X$ and $Y$ are identically distributed.

If we let $Z_1, Z'_1 \sim F_1$ and $Z_2, Z'_2 \sim F_2$ with independence among all these vectors, we have

$$
E(\tau_r) = -\frac{n_1n_2}{N(N-1)} \left(2E\|Z_1 - Z_2\| - E\|Z_1 - Z'_1\| - E\|Z_2 - Z'_2\|\right)
+ \frac{n_1n_2}{N(N-1)} \left(2E\|Z_1^{(r)} - Z_2^{(r)}\| - E\|Z_1^{(r)} - Z'_1^{(r)}\| - E\|Z_2^{(r)} - Z'_2^{(r)}\|\right)
= -\frac{n_1n_2}{N(N-1)} \left\{\varepsilon(Z_1, Z_2) - \varepsilon\left(Z^{(r)}, Z_2^{(r)}\right)\right\}.
$$

Thus, $\tau_r$ is proportional to an unbiased estimator of the change in energy distance between distributions when the $r$-th dimension is deleted. If the $r$-th dimension is important, we should expect a decrease in energy distance between distributions when the $r$-th dimension is deleted and, therefore, a negative value for $\tau_r$. Thus, as for $\iota_r$, negative values for $\tau_r$ indicate importance of the $r$-th variable. From empirical studies not detailed here, $\{\iota_r : r = 1, 2, \cdots, R\}$ and $\{\tau_r : r = 1, 2, \cdots, R\}$ tend to be highly correlated. Because $\tau_r$ is computationally less expensive and does not require specification of a bandwidth parameter $h$, we use $\tau_r$ as our measure of variable importance throughout the subsequent sections of this paper.

### 4.3 Backward selection

Our variable importance measure quantifies how one variable plays a role in the difference between two distributions. When dealing with high-dimensional distributions, the importance of individual dimensions can be obscured by irrelevant dimensions whose joint distribution is identical across treatment groups. To eliminate such dimensions and focus attention on the most important variables, we propose a backward variable selection algorithm that can trim away irrelevant variables in a stepwise manner.
Our backward selection algorithm is defined as follows. Let $S(\ell)$ be the indices of the selected variables at iteration $\ell$. Let $D(\ell)$ be the indices of the deleted variables at iteration $\ell$. Initialize $S(0) = \{1, \ldots, R\}$ and $D(0) = \emptyset$. For $\ell \geq 1$, perform the following steps:

1. For each $r \in S(\ell - 1)$, let $\tau_{\ell,r}$ be the measure of variable importance for dimension $r$ when data vectors consist only of variables indexed by $S(\ell - 1)$. Set $\Gamma(\ell) = \{\tau_{\ell,r} : r \in S(\ell - 1)\}$.

2. Let $s_{\ell,r} = 1 \ast \text{sign}(\tau_{\ell,r})$ for all $r \in S(\ell - 1)$ and $s_{\ell,r} = 1$ for all $r \in D(\ell - 1)$. Let $\text{Sign}(\ell) = \{s_{\ell,r} : r = 1, \ldots, R\}$.

3. For all $r \in S(\ell - 1)$, let $\gamma_{\ell,r}$ be the rank of $\tau_{\ell,r}$ in $\Gamma(\ell)$. For all $r \in D(\ell - 1)$, let $\gamma_{\ell,r} = R - \tilde{\ell} + 1$, where $\tilde{\ell}$ is the number of the iteration when the $r$-th variable was moved from the selected set to the deleted set (see step 4 below). Let $\text{Rank}(\ell) = \{\gamma_{\ell,r} : r = 1, \ldots, R\}$.

4. Find $\max \Gamma(\ell)$, and let $d(\ell)$ be the index corresponding to the maximum element of $\Gamma(\ell)$.

   (a) If $\max \Gamma(\ell) \geq 0$, compute the $p$-value of the MRPP test of distributional equality between treatment groups based only on the variables whose indices are in the set $D(\ell - 1) \cup \{d(\ell)\}$. If the $p$-value is less than a user-chosen threshold for significance, set $S(\ell) = S(\ell - 1)$, $D(\ell) = D(\ell - 1)$, and $L = \ell$ and stop iterating. Otherwise, set $S(\ell) = S(\ell - 1) \setminus d(\ell)$ and $D(\ell) = D(\ell - 1) \cup \{d(\ell)\}$ and continue iterating.

   (b) If $\max \Gamma(\ell) < 0$, set $S(\ell) = S(\ell - 1)$, $D(\ell) = D(\ell - 1)$, and $L = \ell$ and stop iterating.

By using the results obtained from the backward selection, there are several approaches that can be used to make decisions about importance of variables. The first issue is to assess which variables are important. We can deal with this problem in two ways. The first intuitive approach is to declare all variables with indices in the set $S(L)$ to be important and all variables indexed by $D(L)$ to be unimportant. A second method is based on the signs recorded in $\text{Sign}(\ell)$ for $\ell = 1, \ldots, L$. Ideally, each important variable will have a negative sign for each iteration, but random variation in the importance measures can lead to positive signs in some iterations for some important variables. Thus, it may make sense to consider variable $r$ important if $s_{\ell,r} < 0$ for some large proportion (for example, 80%) of iterations $\ell = 1, \ldots, L$. We denote this set of
important variables determined by the sign vectors as $\mathbf{S}_{\text{Sign}}(L, \delta)$, where $\delta$ is a threshold that specifies the percentage of negative signs needed for one variable to be classified as important. Furthermore, by calculating the average rank of variables according to the ranks contained in \{Rank($\ell$), $\ell = 1, \cdots, L$\}, we can compare the relative importance of variables based on their average ranks.

One motivation of doing backward selection is that, by iteratively deleting variables that are not important, we can reduce the dimensions of the data to relieve the effect of “The Curse of Dimensionality”. This method tends to work well, especially when the covariance structure of the data sets are complicated. Examples that illustrate this point will be given in Section 4.6.

### 4.4 A modified MRPP

The backward selection algorithm proposed in the previous section can be used as a follow-up procedure to identify important variables when the original MRPP test detects a difference in multivariate distributions among treatment groups. In this section, we explain how our backward selection procedure can alternatively be used prior to the original MRPP test to concentrate attention on the most important subset of the variables that contains information about potentially lower-dimensional multivariate distribution differences embedded within high-dimensional data vectors. The procedure is defined as follows.

1. Starting with the original dataset, perform backward selection to obtain the $R_0$ variables judged to be most important. $R_0$ can be determined by the cardinality of $\mathbf{S}(L)$ or $\mathbf{S}_{\text{Sign}}(L, \delta)$ described in Section 4.3. Alternatively, $R_0$ can be pre-selected. Compute the MRPP test statistic given in (4.2) using only the $R_0$ variables chosen by backward selection. Use $z_{0,bs}(\Delta_{R_0})$ to represent the value of the test statistic.

2. For the $b$-th permutation of the original dataset, where $b$ is the factoradic number that indexes all $B = N!$ permutations of the observations, do backward selection on the permuted data to select $R_0$ variables. After backward selection, calculate the MRPP test
statistic only with the $R_0$ variables selected from the permuted data. Use $z_{b,bs}(\Delta_{R_0})$ to represent the value of the test statistic for permutation $b$.

3. The modified MRPP $p$-value is defined as

$$p_{bs}(\Delta) = \frac{1}{B} \sum_{b=0}^{B-1} I \{ z_{0,bs}(\Delta_{R_0}) \geq z_{b,bs}(\Delta_{R_0}) \}. \quad (4.6)$$

This testing procedure is similar to the original MRPP, but instead of using all $R$ dimensions, we impose backward selection to focus on the variables that carry the strongest signal for distributional differences. For each permuted data, we select the same number of variables ($R_0$) as were selected for the original data so that the $p$-value $p_{bs}(\Delta)$ is derived by comparing the average of within-group pairwise distances based on data vector of constant dimensionality. Moreover, because the permutation $p$-value involves comparing MRPP test statistics computed from varying subsets of the original $R$ variables, it is important to standardize each variable prior to conducting this modified MRPP test.

It has been discovered that with fixed number of signal-bearing dimensions, the power of the original MRPP and the test based on energy distance (Szekely and Rizzo, 2004) decreases with increasing dimension (Ramdas et al., 2015). The modified MRPP can alleviate the power drop by focusing on the subset of important variables. Simulation studies presented in Section 4.6 show that the size of modified MRPP can be well controlled and that the power of the modified procedure can exceed that of the original MRPP test on all $R$ dimensions, especially when there are a relatively small number of variables responsible for multivariate distributional differences.

### 4.5 Real data analysis

The MRPP has been used to detect differentially expressed gene sets in the analysis of microarray gene expression data in Nettleton, Recknor and Reecy (2008). To demonstrate the usefulness of our proposed variable selection method, we performed backward selection based on our variable importance measure on a real microarray data set and compared it with other methods.
We use the ALL dataset which consists of transcript abundance measurements on 12625 genes for 128 different individuals with acute lymphoblastic leukemia (ALL). As described in Hahne et al. (2008), two subsets of interest in the data are 40 individuals with B-cell tumors that carry the BCR/ABL mutation and 35 individuals with B-cell tumors that have no observed cytogenetic abnormalities. We restrict our attention to an analysis of these 75 samples. As suggested in Hahne et al. (2008), we consider a subset of 2149 genes that shows the greatest variation in transcript abundance levels across the 75 samples. Among these 2149 genes is a set of 196 genes associated with the gene ontology (GO) term “positive regulation of transcription from RNA polymerase II promoter” that we use as an example to illustrate our approaches.

For ease of reference, we will refer to the 40 samples with the BCR/ABL mutation as group 1 and the 35 without cytogenic abnormalities as group 2. The $p$-value of the MRPP test for a difference between groups 1 and 2 with respect to the 196-dimensional multivariate gene expression distribution is 0.002 based on 1000 permutations. This small $p$-value provides significant evidence of a distributional difference between groups 1 and 2 but provides no information about which of the 196 genes may be primarily responsible for the difference.

We conducted backward selection on the 196 genes. The algorithm terminates after 182 iterations ($L = 182$) at which point the importance measure of each of the remaining genes is negative. The final inclusion set $S(L)$ contains 14 genes. When we apply the MRPP to the 14 selected genes, the MRPP $p$-value is less than 0.001. On the other hand, the MRPP $p$-value on the 182 excluded genes is 0.173. Hence, our proposed backward selection algorithm seems to be able to remove a majority of genes whose joint distribution does not appear to differ across groups. This allows researchers to focus follow-up efforts on the subset of 14 genes judged to be important by our procedure. Figure 4.1 gives the MRPP $p$-values for the sets of selected and deleted genes in each iteration. We can see from the plot that the MRPP $p$-values on the selected genes across all iterations remains significantly small while the MRPP $p$-values for the set of deleted genes is high for the first 100 iterations and then steadily decreases until the backward selection procedure terminates.

Form each iteration, we can collect the sign of the importance measure for each remaining gene. Averaging the signs for each gene across all iterations, we obtain the proportion of
Figure 4.1 The MRPP $p$-values on the sets of selected and deleted genes in each iteration.

iterations that each particular gene is declared as important. First, all 14 genes in the inclusion set $S(L)$, have negative importance measures for all 182 iterations. Second, if we choose $\delta = 0.9$, 0.95 and 0.99, then the corresponding sets of important genes determined by $S_{\text{Sign}}(L, \delta)$ have 20, 18 and 14 genes, respectively. Thus, $S_{\text{Sign}}(L, 0.99)$ is exactly the same as $S(L)$. Similarly, we can use the rank vectors $\{\text{Rank}(\ell), \ell = 1, \ldots, L\}$ to rank all genes in terms of their importance in differentially expression. From the average rank of each gene, one gene not included in $S(L)$ emerges as the ninth most important gene, but overall importance rankings based on $S(L)$, $\{\text{Sign}(\ell), \ell = 1, \ldots, L\}$ and $\{\text{Rank}(\ell), \ell = 1, \ldots, L\}$ were similar.

There are other analysis options for identifying which of the 196 transcription factor activity genes are most relevant to the difference between groups 1 and 2. Perhaps the most obvious approach would be to conduct a two-sample $t$-test separately for each gene. When controlling the false discovery rate at approximately 0.05 using the method of Benjamini and Hochberg (1995), 11 out of the 196 genes are identified as differentially expressed. A similar approach involves conducting a moderated two-sample $t$-test as implemented in the R package limma
introduced in Smyth (2004) and again controlling false discovery rate at approximately the 0.05 level. This approach, which borrows information across genes to estimate the error variance for each gene, yields 12 genes that include the 11 identified by the traditional two-sample \( t \)-test approach. The 14 genes identified by our backward selection algorithm include the 12 genes identified by the moderated \( t \)-test approach. Thus, in this example our backward selection procedure provides some additional discoveries but overall performs similarly to the conventional approaches.

To gain further insight into the performance of the backward selection algorithm, we compare the differences between the correlation matrices of the selected genes for each treatment. Genes are usually regulated together to carry out their functions, and the proposed importance measure and backward selection method can take differences between the covariance matrices into consideration. Hence, a good subset of selected differential expressed genes may include genes whose correlation matrices differ between two groups. For the data set we analyzed, the average absolute differences between the sample correlation matrices was 0.187 for the original 196 genes, 0.241 for the 14 selected genes, and 0.182 for the 182 excluded genes. Thus, it seems that our procedure succeeded in focusing attention on a subset of transcription factor activity genes whose correlation structure differs across groups to a greater extent than the correlation structure of transcription factor activity genes in general.

4.6 Simulations

4.6.1 Backward selection

The results from Section 4.5 indicate that our proposed backward selection algorithm performs similarly to conventional approaches but may also possess some advantages for detecting differentially expressed genes. Because the true differential expression status of genes is unknown in applications, in this section, we examine the performance of backward selection relative to the \( t \)-test approaches when applied to data simulated from the ALL dataset in such a way that true differential expression status is known.
We continue to focus on the 40 samples with BCR/ABL mutation (group 1) contrasted against the 35 individuals without BCR/ABL mutation (group 2). Among the set of 2149 genes that remained after applying the filtering criteria described in Section 4.5, there are 14697 different GO terms associated with at least one of these genes. Thus, we can define 14697 different gene sets corresponding to these 14697 GO terms. We focus on the subset of these 14697 gene sets with cardinality no smaller than 40 and MRPP $p$-value less than 0.05 for testing equality of joint expression distributions between groups 1 and 2. This results in 859 gene sets selected for further study. The number of genes in these sets range from 40 to 2103. The histogram of the number of genes and MRPP test $p$-values for each gene set is shown in Figure 4.2. For each of the 859 sets, the following procedures was used to simulate datasets.

1. Use the two-sample $t$-test with FDR control at 0.05, limma with FDR control at 0.05, and our proposed backward selection algorithm with inclusion set $S(L)$ to obtain three sets of genes that are detected as differentially expressed. Find the union of these three sets, denoted as $\Theta$, and let $p_1$ be the number of genes in $\Theta$.

2. Randomly select $2 \times 15$ individuals without replacement from group 1, and randomly divide the selected samples into two groups, each with 15 samples. Denote these two groups as group $1'$ and group $2'$.
3. Randomly select 15 samples without replacement from group 2 and extract the subset of \( p_1 \) genes in \( \Theta \).

4. For each individual in group 2', replace the data for the \( p_1 \) genes in \( \Theta \) with the data extracted from a group 2 sample in step 3.

These four steps produce two samples, each of size 15, drawn from multivariate distributions that differ only for the subset of dimensions corresponding to \( \Theta \). Figure 4.3 summarizes the number of genes selected by each method along with the size of \( \Theta \) for each gene set. It is clear that our proposed backward selection algorithm tends to select more genes than the competing approaches for larger \( R \).

![Figure 4.3 Number of genes selected by each method for each gene set.](image)

Following steps 1 through 4, we simulate 1000 data sets for each gene set. For each simulated dataset, we find the \( p_1 \) most significant genes using the two-sample \( t \)-test, \( \text{limma} \), and our backward selection procedure. For the backward selection algorithm, the most significant genes
are collected according to the average of rank vectors \( \{ \text{Rank}(\ell), \ell = 1, \ldots, L \} \). We then calculate the percentage of the selected genes not in \( \Theta \) for each of the three methods and average the results over 1000 simulation replications to get the average false positive rate for the three methods. Let \( \rho_t, \rho_l \) and \( \rho_{bs} \) represent those average false positive rates for the two-sample \( t \)-test, limma, and backward selection, respectively. The results for comparing the three methods are given in Figure 4.4.

By looking at plots (a) through (c) in Figure 4.4, we see that our backward selection approach tends to produce the lowest false positive rates (and thus the highest discovery rates), especially for high-dimensional gene sets. The two-sample \( t \)-test has worst performance against the other two methods. Plot (d) in Figure 4.4 shows that backward selection is usually the best of the three methods, and its advantage over the other methods increases as the dimension increases.

If, as in the data analysis example of Section 4.5, the \( t \)-test-based approaches tend to identify a proper subset of the genes identified by the backward selection procedure when applied to the 859 gene sets, the preceding simulation studies could be biased in favor of our backward selection procedure because the union of the sets of genes identified in step 1 may tend to coincide with the genes selected by the backward selection procedure. To avoid this potential source of bias, we focus on the subset of gene sets in which the number of genes selected was the same (but not necessarily identity) for all three methods, which results in 61 gene sets. The number of genes in those 61 gene sets range from 40 to 391 and the histogram is given in Figure 4.5(a). The results for the 61 gene sets satisfying this additional criterion are shown in Figure 4.5(b)–4.5(d).

From Figure 4.5, we see that for most of the gene sets, the backward selection algorithm performs best and that limma performs better than the traditional two-sample \( t \)-test. Table 4.1 gives the number of gene sets for which each method ranks first, second, or third among the three methods with respect to false positive rate. For 45 out of the 61 sets, our backward selection algorithm has the lowest false positive rates. The limma method has the fewest worst false positive rates, among these 61 gene sets. Based on our simulation results, we conclude that, overall, the proposed backward selection method performs best among the three methods.
Figure 4.4 Comparison of average false positive rate for three methods, $\rho_t$ for two-sample t-test, $\rho_l$ for limma and $\rho_{bs}$ for backward selection. (a) $\rho_t - \rho_{bs}$; (b) $\rho_l - \rho_{bs}$; (c) $\rho_l - \rho_t$; (d) the proportion of 1000 simulation replications in which each method has the lowest false positive rate across gene sets with of varying size.
Figure 4.5  (a): Histogram of number of genes in the 61 gene sets for which all three methods initially selected the same number of genes. (b) to (d): Comparison of average false positive rate for three methods when applied to the 61 gene sets.
Table 4.1 Number of gene sets for which each method ranks 1, 2, or 3 with respect to false positive rate. (Rank 1 is best and 3 is worst).

<table>
<thead>
<tr>
<th>Method</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Two-sample t-test</td>
<td>2</td>
</tr>
<tr>
<td>\textit{limma}</td>
<td>14</td>
</tr>
<tr>
<td>Backward selection</td>
<td>45</td>
</tr>
</tbody>
</table>

4.6.2 Modified MRPP

A modified MRPP test procedure based on our proposed backward selection algorithm is introduced in Section 4.4. In this section, we compare the performance of our proposed modified MRPP with the original MRPP for testing differences between two multivariate distributions by conducting Monte Carlo simulations. We consider six different sample size pairs \((n_1, n_2)\) in combination five different choices for the data vector dimension \(R\). We focus primarily on combinations where the dimension exceeds the sample size (see Table 4.2). All our size and power estimates are based on 1000 Monte Carlo simulations at the nominal level \(\alpha = 0.05\).

Table 4.2 Combinations of \(n_1\) and \(n_2\) considered in the simulations.

<table>
<thead>
<tr>
<th>(n_1)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n_2)</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>40</td>
<td>40</td>
<td>80</td>
</tr>
<tr>
<td>(N)</td>
<td>40</td>
<td>60</td>
<td>100</td>
<td>80</td>
<td>120</td>
<td>160</td>
</tr>
</tbody>
</table>

First, we evaluate the size of the proposed test. The data \(\{Y_1, \ldots, Y_{n_1}; Y_{n_1+1}, \ldots, Y_N\}\) are generated from the \(R\)-dimensional multivariate normal distribution with mean vector \(0_R\) and covariance matrix \(\Sigma = (0.5|i-j|)_{1 \leq i,j \leq R}\). For each simulated data set, we carry out the original MRPP along with our proposed modified MRPP for testing distributional differences between two groups \(\{Y_1, \ldots, Y_{n_1}\}\) and \(\{Y_{n_1+1}, \ldots, Y_N\}\). The number of permutations is set at 1000 for both testing procedures. To implement the modified MRPP, we choose the number of variables selected \((R_0)\) in several different ways. First, we consider setting \(R_0\) to be the number of variables in the inclusion set \(S(L)\) obtained from backward selection. Alternative choices are obtained by prespecifying \(R_0\) as 2, 4, 8, 16, or \(\sqrt{R}\).
The sizes for both original and our modified MRPP tests summarized in Table 4.3 show that the original MRPP test maintains the size well around the nominal significant level 0.05. When the number of variables $R_0$ is pre-specified before the testing procedure, the sizes of the modified MRPP can also be well controlled. However, when $R_0$ is chosen as the cardinality of the inclusion set $S(L)$ after the backward selection procedure, the sizes of the modified MRPP are slightly larger than the nominal level. Overall, the modified MRPP has good control of Type-I error under the null hypothesis that the distributions of two groups are the same.

Table 4.3 Empirical sizes of original MRPP (MRPP$_{Org}$) and modified MRPP with different choices of $R_0$. For Mod$_{S(L)}$, $R_0$ is chosen as the cardinality of $S(L)$. For Mod$_2$, Mod$_4$, Mod$_8$, Mod$_{16}$ and Mod$_{\sqrt{R}}$, $R_0$ is pre-specified as 2, 4, 8, 16 and $\sqrt{R}$, respectively.

<table>
<thead>
<tr>
<th>$R$</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRPP$_{Org}$</td>
<td>0.042</td>
<td>0.047</td>
<td>0.047</td>
<td>0.056</td>
<td>0.044</td>
<td>0.052</td>
<td>0.055</td>
<td>0.052</td>
<td>0.044</td>
<td>0.059</td>
</tr>
<tr>
<td>Mod$_{S(L)}$</td>
<td>0.064</td>
<td>0.074</td>
<td>0.063</td>
<td>0.063</td>
<td>0.050</td>
<td>0.075</td>
<td>0.068</td>
<td>0.073</td>
<td>0.056</td>
<td>0.055</td>
</tr>
<tr>
<td>Mod$_2$</td>
<td>0.046</td>
<td>0.046</td>
<td>0.051</td>
<td>0.047</td>
<td>0.046</td>
<td>0.046</td>
<td>0.041</td>
<td>0.058</td>
<td>0.046</td>
<td>0.046</td>
</tr>
<tr>
<td>Mod$_4$</td>
<td>0.038</td>
<td>0.047</td>
<td>0.045</td>
<td>0.046</td>
<td>0.044</td>
<td>0.050</td>
<td>0.044</td>
<td>0.052</td>
<td>0.043</td>
<td>0.041</td>
</tr>
<tr>
<td>Mod$_8$</td>
<td>0.038</td>
<td>0.042</td>
<td>0.044</td>
<td>0.053</td>
<td>0.042</td>
<td>0.055</td>
<td>0.044</td>
<td>0.053</td>
<td>0.042</td>
<td>0.042</td>
</tr>
<tr>
<td>Mod$_{16}$</td>
<td>0.042</td>
<td>0.040</td>
<td>0.042</td>
<td>0.045</td>
<td>0.040</td>
<td>0.053</td>
<td>0.056</td>
<td>0.052</td>
<td>0.035</td>
<td>0.043</td>
</tr>
<tr>
<td>Mod$_{\sqrt{R}}$</td>
<td>0.043</td>
<td>0.047</td>
<td>0.042</td>
<td>0.047</td>
<td>0.042</td>
<td>0.049</td>
<td>0.042</td>
<td>0.052</td>
<td>0.036</td>
<td>0.044</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$R$</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRPP$_{Org}$</td>
<td>0.047</td>
<td>0.045</td>
<td>0.047</td>
<td>0.049</td>
<td>0.044</td>
<td>0.049</td>
<td>0.048</td>
<td>0.052</td>
<td>0.041</td>
<td>0.046</td>
</tr>
<tr>
<td>Mod$_{S(L)}$</td>
<td>0.063</td>
<td>0.063</td>
<td>0.055</td>
<td>0.065</td>
<td>0.056</td>
<td>0.070</td>
<td>0.065</td>
<td>0.061</td>
<td>0.063</td>
<td>0.044</td>
</tr>
<tr>
<td>Mod$_2$</td>
<td>0.044</td>
<td>0.046</td>
<td>0.040</td>
<td>0.046</td>
<td>0.054</td>
<td>0.043</td>
<td>0.044</td>
<td>0.034</td>
<td>0.051</td>
<td>0.049</td>
</tr>
<tr>
<td>Mod$_4$</td>
<td>0.044</td>
<td>0.037</td>
<td>0.048</td>
<td>0.040</td>
<td>0.049</td>
<td>0.048</td>
<td>0.039</td>
<td>0.043</td>
<td>0.050</td>
<td>0.050</td>
</tr>
<tr>
<td>Mod$_8$</td>
<td>0.041</td>
<td>0.035</td>
<td>0.038</td>
<td>0.043</td>
<td>0.053</td>
<td>0.045</td>
<td>0.041</td>
<td>0.046</td>
<td>0.054</td>
<td>0.044</td>
</tr>
<tr>
<td>Mod$_{16}$</td>
<td>0.046</td>
<td>0.039</td>
<td>0.039</td>
<td>0.047</td>
<td>0.056</td>
<td>0.048</td>
<td>0.041</td>
<td>0.043</td>
<td>0.055</td>
<td>0.035</td>
</tr>
<tr>
<td>Mod$_{\sqrt{R}}$</td>
<td>0.041</td>
<td>0.038</td>
<td>0.037</td>
<td>0.044</td>
<td>0.056</td>
<td>0.039</td>
<td>0.041</td>
<td>0.043</td>
<td>0.056</td>
<td>0.037</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$R$</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRPP$_{Org}$</td>
<td>0.038</td>
<td>0.035</td>
<td>0.042</td>
<td>0.056</td>
<td>0.046</td>
<td>0.051</td>
<td>0.041</td>
<td>0.057</td>
<td>0.047</td>
<td>0.058</td>
</tr>
<tr>
<td>Mod$_{S(L)}$</td>
<td>0.064</td>
<td>0.061</td>
<td>0.060</td>
<td>0.072</td>
<td>0.056</td>
<td>0.071</td>
<td>0.066</td>
<td>0.071</td>
<td>0.071</td>
<td>0.057</td>
</tr>
<tr>
<td>Mod$_2$</td>
<td>0.044</td>
<td>0.042</td>
<td>0.042</td>
<td>0.044</td>
<td>0.054</td>
<td>0.047</td>
<td>0.044</td>
<td>0.052</td>
<td>0.048</td>
<td>0.053</td>
</tr>
<tr>
<td>Mod$_4$</td>
<td>0.044</td>
<td>0.043</td>
<td>0.041</td>
<td>0.054</td>
<td>0.052</td>
<td>0.043</td>
<td>0.044</td>
<td>0.047</td>
<td>0.055</td>
<td>0.056</td>
</tr>
<tr>
<td>Mod$_8$</td>
<td>0.040</td>
<td>0.042</td>
<td>0.041</td>
<td>0.054</td>
<td>0.044</td>
<td>0.045</td>
<td>0.039</td>
<td>0.046</td>
<td>0.051</td>
<td>0.050</td>
</tr>
<tr>
<td>Mod$_{16}$</td>
<td>0.039</td>
<td>0.041</td>
<td>0.041</td>
<td>0.059</td>
<td>0.049</td>
<td>0.048</td>
<td>0.039</td>
<td>0.046</td>
<td>0.058</td>
<td>0.047</td>
</tr>
<tr>
<td>Mod$_{\sqrt{R}}$</td>
<td>0.039</td>
<td>0.040</td>
<td>0.042</td>
<td>0.056</td>
<td>0.052</td>
<td>0.043</td>
<td>0.040</td>
<td>0.046</td>
<td>0.059</td>
<td>0.046</td>
</tr>
</tbody>
</table>
To investigate the power improvement of the modified MRPP test, for each combination of sample sizes \(n_1, n_2\) and dimension \(R\), we simulate \(\{Y_1, \ldots, Y_{n_1}\}\) from the \(R\)-dimensional multivariate normal distribution with mean vector \(0_R\) and covariance matrix \(\Sigma = (0.5|i-j|)_{1 \leq i,j \leq R}\), while \(\{Y_{n_1+1}, \ldots, Y_N\}\) are drawn from the same distribution but with a location shift in the first four dimensions of its mean vector; that is, \(\mu = (\nu 1_4^T, 0_{R-4}^T)^T\) for the second group. The magnitude of location shift \(\nu\) is chosen as 0.5, 1.0 and 1.5, respectively. Again, for the modified MRPP, the number of variables \(R_0\) used for testing is chosen in the same fashion as for our investigation of the test size. Figure 4.6–4.8 displays the empirical powers of the original MRPP and the modified MRPP tests.

It is clear that the powers increase as \(\nu\) increases for all testing approaches. In addition, as the sample sizes grow larger, all tests gain extra power as expected. The empirical powers of the original MRPP decrease as the dimension \(R\) increases, especially for the case when \(\nu = 1\). The empirical powers of the modified MRPP also decrease when \(R\) grows, but the decrease is much slower than that of the original MRPP. Because only a subset of important variables are used for testing the differences between two distributions based on our backward selection algorithm, the modified MRPP exhibits noticeable gains in power relative to the original MRPP for the largest \(R\) settings. The empirical results suggests that the choice of \(R_0\) does not have too much impact on the performance of the modified MRPP for the simulation scenarios we considered.

### 4.7 Conclusion

In this paper, we introduced importance measures based on MRPP and energy distance. The importance measures quantify the contribution of each variable in the difference between multivariate distributions. We developed a backward selection algorithm to address the variable selection problem for high-dimensional data. We examined the proposed backward selection approach by numerical studies and illustrated its applications in real data analysis. Furthermore, we modified the original MRPP using our proposed backward selection algorithm. Empirical evidence shows that the modified MRPP can not only preserve the nominal significance level, but also improve the power of the original MRPP by concentrating on the subset of most important variables when many variables are unimportant.
Figure 4.6  Empirical powers of the original MRPP and modified MRPP with different choices of $R_0$ for $\nu = 0.5$
Figure 4.7  Empirical powers of the original MRPP and modified MRPP with different choices of $R_0$ for $\nu = 1$

(a) $n_1 = 20, n_2 = 20$

(b) $n_1 = 20, n_2 = 40$

(c) $n_1 = 20, n_2 = 80$

(d) $n_1 = 40, n_2 = 40$

(e) $n_1 = 40, n_2 = 80$

(f) $n_1 = 80, n_2 = 80$
Figure 4.8  Empirical powers of the original MRPP and modified MRPP with different choices of $R_0$ for $\nu = 1.5$
CHAPTER 5. GENERAL CONCLUSIONS

In this dissertation, we studied several topics related with statistical inference for massive data and high-dimensional data.

In Chapter 2, for a general class of statistics, we studied the statistical efficiency of the distributed statistics and proposed bootstrap algorithms to approximate the distribution of the distributed statistics. An advantage of the distributed statistic is that it can save computing time and make statistical inference more feasible under the massive data setting. It was shown that for non-degenerate statistics, the distributed statistic can preserve the same estimation efficiency as the full sample statistic with a properly selected number of data blocks. In addition, it has the same asymptotic distribution as the full sample statistic. It is of interest to investigate whether the distributed statistic enjoys high-order correctness.

The proposed bootstrap algorithms can be implemented distributively, which is a crucial guideline for statistical methods when analyzing massive data, especially when the data are stored in different locations. The consistency of the bootstrap algorithms are established and it would be interesting to study the convergence rate of the proposed distributed bootstrap methods.

In Chapter 3, we developed new testing procedures for the identity and sphericity hypotheses regarding high-dimensional covariance matrices. We demonstrated that our proposed tests improve the power of existing methods when the underlying covariance matrix belongs to the bandable covariance class. The covariance matrix of the original random vector may not be bandable. It is assumed that there is a permutation of the random vector such that the corresponding covariance matrix is bandable. However, it is not clear whether permuting the data would distort the size of the proposed tests or not, and this is deserved to study in the future.
In Chapter 4, the importance measures inspired by the MRPP and the energy distance are proposed to quantify the importance of each dimension for high-dimensional data. Simulations studies showed that the backward selection based on the importance measure has advantages over other methods. Besides the MRPP and the energy distance, the maximum mean discrepancy (Gretton et al., 2007, 2012) is a kernel-based measure that quantifies the disparity between two multivariate distributions, which is closely related to the energy distance. It is of interest to consider constructing a kernel-based importance measure by utilizing the idea of weighting and perturbation on the maximum mean discrepancy.


