January 2009

Statistical Aggregation: Theory and Applications

Ruibin Xi
Washington University in St. Louis

WASHINGTON UNIVERSITY

Department of Mathematics

Dissertation Examination Committee:

Nan Lin, Chair

Yixin Chen

Jimin Ding

Jeff Gill

Stanley Sawyer

Mladen Victor Wickerhauser

STATISTICAL AGGREGATION: THEORY AND APPLICATIONS

by

Ruibin Xi

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2009

St. Louis, Missouri

ACKNOWLEDGMENTS

I would like to express my deep gratitude to all professors in my defense committee. I am full of gratefulness to my advisor, Prof. Nan Lin, who gives me enormous amount of insightful suggestions and advices during the last four years. This thesis could not be finished or even could not be started without his help. I am indebted to Prof. Stanley Sawyer, from whom I probably learned much more than I realized about mathematics and statistics. I could not grow up as a statistician from a PhD student without his help. I sincerely thank Prof. Yixin Chen who gave me many valuable suggestions and advices that made my research more applicable to real applications. I could not be thankful enough to Prof. Jeff Gill for his patience, kindness and great help to me. I am obligated to Prof. Jimin Ding whose great help is invaluable to me. I sincerely thank Prof. Mladen Victor Wickerhauser. He is very nice and willing to provide help at any time.

I am full of gratitude to the Department of Mathematics at Washington University in St. Louis. I could not obtain my Ph.D degree without the support of the department. I am indebted to all faculties in the department. I gratefully thank my first two year advisor Prof. Rachel Robert for her support and encouragement. I sincerely thank Prof. Xiang Tang. I benefit a lot from discussions with him. I am full of gratefulness to all staffs in the department for their help and support. Especially, I

would like to express my gratitude to Ms. Marry Ann Stenner. She gave me so much help from the very beginning of my PhD study to my graduation. The helps from all my friends are also greatly appreciated.

Last but not least, I must thank my parents and my fiancée Jing zhang. It is only with their support and understanding that I could finish my PhD degree.

TABLE OF CONTENTS

Page

Page

ABSTRACT

Statistical Aggregation: Theory and Applications

by

Xi, Ruibin

Doctor of Philosophy in Mathematics,

Washington University in St. Louis, August, 2009.

Professor Nan Lin, Chairperson

Due to their size and complexity, massive data sets bring many computational challenges for statistical analysis, such as overcoming the memory limitation and improving computational efficiency of traditional statistical methods. In the dissertation, I propose the statistical aggregation strategy to conquer such challenges posed by massive data sets. Statistical aggregation partitions the entire data set into smaller subsets, compresses each subset into certain low-dimensional summary statistics and aggregates the summary statistics to approximate the desired computation based on the entire data. Results from statistical aggregation are required to be asymptotically equivalent.

Statistical aggregation processes the entire data set part by part, and hence overcomes memory limitation. Moreover, statistical aggregation can also improve the computational efficiency of statistical algorithms with computational complexity at

the order of $O(N^m)$ $(m > 1)$ or even higher, where N is the size of the data. Statistical aggregation is particularly useful for online analytical processing (OLAP) in data cubes and stream data, where fast response to queries is the top priority. The "partition-compression-aggregation" strategy in statistical aggregation actually has been considered previously for OLAP computing in data cubes. But existing research in this area tends to overlook the statistical property of the analysis and aims to obtain identical results from aggregation, which has limited the application of this strategy to very simple analyses. Statistical aggregation instead can support OLAP in more sophisticated statistical analyses.

In this dissertation, I apply statistical aggregation to two large families of statistical methods, estimating equation (EE) estimation and U-statistics, develop proper compression-aggregation schemes and show that the statistical aggregation tremendously reduces their computational burden while maintaining their efficiency. I further apply statistical aggregation to U-statistic based estimating equations and propose new estimating equations that need much less computational time but give asymptotically equivalent estimators.

1. Introduction

Nowadays, many statistical analyses need be performed on massive data sets, such as Internet traffic data, business transaction records and satellite feeds. These data sets can be too large to fit in a computer's internal memory and bring a series of special computational challenges. Even when the massive data sets can fit in a computer's memory, oftentimes, the analysis may not be finished within an acceptable amount of time when fast analysis is desired.

For a massive static data set that does not evolve over time, e.g. transaction history of a company, a simple solution is to obtain a reduced data set by sub-sampling the massive data set, which makes the relevant statistical computation tractable [1]. However, this method could be "sub-optimal" due to the sub-sampling variability. For time-evolving data, sub-sampling methods are usually not applicable, as only the most recent raw data are stored in the memory, and therefore it's very expensive or impossible to sub-sample from the historical raw data. Furthermore, applications in massive data sets often need on-line analytical processing (OLAP) computing and fast response to queries is the top priority for any OLAP tool. The response time should be in the order of seconds, minutes at most, even if complex statistical analyses are involved. Queries are usually interested in different parts of the massive data set. Sub-sampling for each query is then computationally inefficient and cannot support

fast OLAP computing. In this thesis, I propose the statistical aggregation strategy to conquer the difficulties posed by massive data sets.

Next, I will briefly review the stream data [2, 3] and data cubes [4, 5, 6]. Analyses in both environments require to perform the same analyses for different subsets while the raw data often can not be saved permanently. This makes statistical aggregation particularly useful.

1.1 Stream Data

Stream data are data records coming rapidly along time. Examples include phone records in large call centers, web search activities, and network traffic. Formally, stream data are a sequence of data items $z_1, \cdots, z_t, \cdots, z_N$ such that the items are read once in increasing order of the indices t [3]. These data sets increase explosively over time and are typically stored in secondary storage devices, making access, particularly random access, very expensive. Meanwhile, analysis needs to be repeated from time to time when more data are available. This demands algorithms that process the raw data only once and then compress them into low-dimensional statistics based on which the desired analysis can be performed exactly or approximately. Some recent research on stream data include clustering [7, 8] and classification [9]. Statistical aggregation provides a general solution to fast statistical analysis for stream data.

1.2 Data Cubes

Data cube is a popular OLAP tool in data warehousing. It models the massive data set as a multidimensional hyper-rectangle. *Dimensional attributes* in data cubes are the perspectives or entities with respect to which an organization wants to keep records. Usually each dimension attribute has multiple levels of abstraction formed by conceptual hierarchies. For example, country, state, city, and street are four levels of abstraction in a dimension for location. Attributes other than dimensional attributes in data cubes are *measure attributes*. A *cell* is a tuple in a multi-dimensional data cube space that each dimensional attribute and measure attribute take specific value. Given two distinct cells c_1 and c_2, c_1 is an *ancestor* of c_2, or c_2 a *descendant* of c_1 if on every dimensional attribute, either c_1 and c_2 share the same value, or c_1's value is a generalized value of c_2's in the dimension's concept hierarchy. A cell c is called a *base cell* if it does not have any descendant. A cell c is an *aggregated cell* if it is an ancestor of some base cells.

Example 1: Suppose a chain supermarket records its sales with respect to location, time and product. We then can use a data cube with three dimensional attributes *location*, *time* and *product* and one measure attribute *sale* to model this data warehouse. Figure 1.1 (a) shows a part of this data cube, where c_1 is the cell with (*location, time, product*) being (*MO, 2009,* P_3). Figure 1.1 (b) shows some descendant cells of c_1, where *location* takes value among cities in Missouri, *time* among

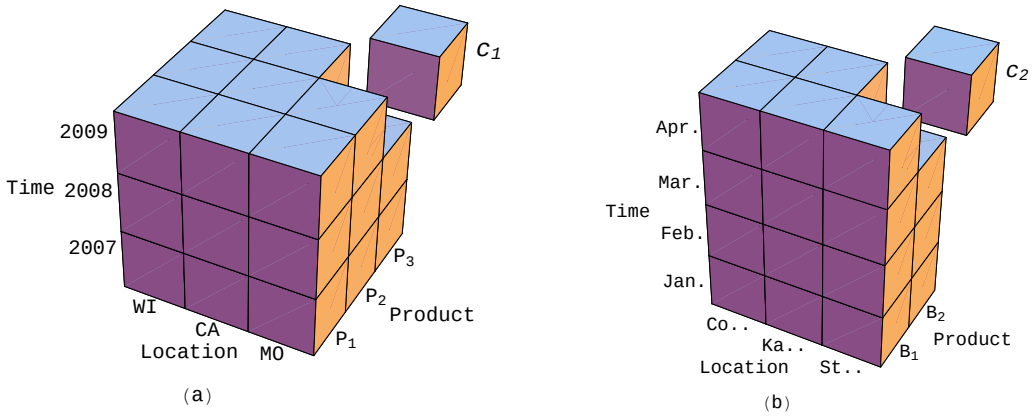

Figure 1.1. The data cube in Example 1. (a) c_1 is the cell with $location = MO$, $Time = 2009$ and $Product = P_3$. (b) Descendant cells of c_1, where c_2 is the descendant cell of c_1 with $location = St.\ Louis$, $time = Apr.\ 2009$ and $product = B_2$

months in 2009 and *product* as one of two brands of product P_3. In particular, c_2 is a descendant cell of c_1. ■

Computer scientists noticed that some simple summary statistics like *sum*, *count*, *average* can be first computed for the base cells of the data cube, and then these simple summary statistics for higher-level cells can be obtained by aggregating the compressed summary statistics in base cells without accessing the raw data. Thus, we can pre-compress all base cells into these summary statistics in one scan. Then, to answer a query about a specific cell c, we only need aggregate the compressed summary statistics of base cells inside the cell c together. Therefore, data cubes can support fast OLAP computing of these simple summary statistics by avoiding accessing raw data. Recently, some researchers developed compression-aggregation schemes for more advanced statistical analysis including parametric models such as linear re-

gression [10, 11], general multiple linear regression [12, 13] and predictive filters [12], as well as nonparametric statistical models such as näive Bayesian classifiers [14] and linear discriminant analysis [15]. Statistical aggregation introduced here provides a general solution to fast OLAP computation of more advanced statistical analyses.

1.3 Statistical Aggregation

Current data cube techniques usually view statistical analysis purely as an algorithm and pays little attention to its statistical properties. *Statistical aggregation* instead utilizes the statistical properties of statistical analyses and is a general strategy for statistical analyses on massive data sets. The basic idea of the statistical aggregation is as follows. The entire data set is first partitioned into K subsets, often determined by dimensional attributes in a data cube context, and the data in each subset are compressed to some summary statistics. At last, the summary statistics are aggregated to approximate the statistics of interest without accessing the raw data. Unlike in current data cube techniques, the resulted statistics given by the statistical aggregation are only required to be asymptotically equivalent to but not exactly equal to the statistics of interest. With this looser but statistically satisfactory requirement, statistical aggregation can support more sophisticated statistical analyses.

The statistical aggregation also serves as a general strategy for statistical analysis on massive data set. By partitioning the entire data set into small subsets and compressing each piece into some summary statistics, the statistical aggregation con-

quers the memory and storage problems raised by massive data sets. The statistical aggregation can be readily applied to support OLAP computing in data cubes. The base cells only need store the summary statistics from the statistical aggregation and we can approximate the statistics of interest for other cells by aggregating the corresponding base cells.

Another application of statistical aggregation is to expediate the computation of statistical analyses whose computational complexity is high. For example, the computational burden of a degree m U-statistic [16] is $O(N^m)$, where N is the size of the entire data set. In Chapter 3, I apply the statistical aggregation to U-statistics and propose the aggregated U-statistics (AU-statistics). The AU-statistics is asymptotically equivalent to the U-statistic but its computational burden is just $O(N^{(m+1)/2})$ if we partition the entire data data into $K = O(\sqrt{N})$ pieces.

When applying the statistical aggregation to the specific statistical analysis, one has to find appropriate summary statistics and the corresponding aggregation algorithm. The dimension of summary statistics should be low and independent of the size of the data set and the aggregation algorithm should be simple and easy computationally. The summary statistics used in statistical aggregation is closed related to *sufficient statistics* [17]. In fact, if the parameter estimation of the Gaussian distribution is under consideration, one can develop a compression-aggregation scheme using sufficient statistics as summary statistics of each subset. However, it is generally very difficult or impossible to find low-dimensional sufficient statistics since many statistical analyses are semi-parametric or even non-parametric. Thus, we generally

have to resort to the asymptotic properties of the estimator under consideration and develop its compression-aggregation scheme. In this dissertation, I apply the statistical aggregation strategy to two large families of estimators, estimating equation (EE) estimators and U-statistics. The compression-aggregation schemes for EE estimators and U-statistics are developed based on Taylor's expansion of the estimating equation and asymptotic normality of U-statistics, respectively.

The dissertation is organized as following. In Chapter 2, I apply the statistical aggregation strategy to EE estimators. I show in theory that the proposed aggregated EE (AEE) estimator is asymptotically equivalent to the EE estimator if K goes to infinity not too fast. Simulation studies validate the theory and show that the AEE estimator is computationally very efficient. I also apply the AEE estimator to the data cube context and show its remarkable performance in saving computational time. In Chapter 3, I apply the statistical aggregation strategy to U-statistics and show in theory that the AU-statistic is asymptotically equivalent to U-statistic and its computational complexity is much lower than that of U-statistic. In Chapter 4, I use the technique developed in Chapter 3 to functional regression models (FRM) [18] and propose a new estimating equation for the FRMs. The estimator from the new estimating equation is asymptotically equivalent to the original estimator presented in [18], but computationally more efficient. I then conclude my thesis and discuss other possible applications of this strategy and future researches in the last chapter.

2. Aggregation of Estimating Equation Estimations

Many parametric and semi-parametric statistical estimation techniques can be unified into the estimating equation framework, such as the OLS estimator, the quasi-likelihood estimator (QLE) [19] and robust M-estimators [20, 21, 22]. In this chapter, I will apply the statistical aggregation strategy to estimating equation (EE) estimations in massive data sets. I first partition the massive data sets into many subsets and then compress the raw data into the EE estimates and the first-order derivative of the estimating equation before discarding the raw data. The saved statistics allow to reconstruct an approximation to the original estimating equation in each subset, and hence an approximation to the equation for the entire data set after aggregating over all subsets. I will show that, when the number of subsets is bounded or goes to infinity not too fast, the solution to the approximated estimating equation, called the aggregated EE (AEE) estimator, is consistent and asymptotically equivalent to the original EE estimator under some mild regularity conditions. I will also show in theory and in simulation studies that the AEE estimator provides more accurate estimates than estimates from a subsample of the entire data set, which is commonly used for static massive data sets. The AEE estimator is not only an accurate approximation to the EE estimator, but also computationally more efficient shown by simulation studies.

2.1 Aggregation for Linear Regression

In this section, I review the regression cube technique [12] to illustrate the idea of aggregation for linear regression analysis.

Suppose that we have N independent observations $(y_1, \mathbf{x}_1), \cdots, (y_N, \mathbf{x}_N)$, where y_i is a scalar response, $\mathbf{x}_i$ is a $p \times 1$ covariate vector, $i = 1, \ldots, N$. Let $\mathbf{y} = (y_1, \ldots, y_N)^T$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$. A linear regression model assumes that $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. Suppose that $\mathbf{X}^T\mathbf{X}$ is invertible, the OLS estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_N = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Suppose that the entire data set is partitioned into K subsets with $\mathbf{y}_k$ and $\mathbf{X}_k$ being the values of the response and covariates, and $\hat{\boldsymbol{\beta}}_k = (\mathbf{X}_k^T\mathbf{X}_k)^{-1}\mathbf{X}_k^T\mathbf{y}_k$ is the OLS estimate in the kth subset, $k = 1, \ldots, K$. Then, we have $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_K^T)^T$ and $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_K^T)^T$. Since $\mathbf{X}^T\mathbf{X} = \sum_{k=1}^K \mathbf{X}_k^T\mathbf{X}_k$ and $\mathbf{X}^T\mathbf{y} = \sum_{k=1}^K \mathbf{X}_k^T\mathbf{y}_k$, the regression cube technique sees that

$$\hat{\boldsymbol{\beta}}_N = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \left(\sum_{k=1}^K \mathbf{X}_k^T\mathbf{X}_k\right)^{-1} \sum_{k=1}^K \mathbf{X}_k^T\mathbf{X}_k\hat{\boldsymbol{\beta}}_k, \tag{2.1}$$

which suggests that we can compute the OLS estimate for the entire data set without accessing the raw data after saving $(\mathbf{X}_k^T\mathbf{X}_k, \hat{\boldsymbol{\beta}}_k)$ for each subset. The size of $(X_k^T X_k, \hat{\boldsymbol{\beta}}_k)$ is $p^2 + p$, so we only need to save $Kp(p+1)$ numbers, which achieves very efficient compression since both K and p are far less than N in practice. The success of this technique largely depends on the linearity of the estimating equation in parameter $\boldsymbol{\beta}$ and the estimating equation of the entire data set is a simple summation of the equations in all subsets. That is, $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{k=1}^K \mathbf{X}_k^T(\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}) = 0$.

2.2 The AEE Estimator

In this section, I consider, more generally, estimating equation estimation in massive data sets and propose our AEE estimator to provide computationally tractable estimation by approximation and aggregation.

Given independent observations $\{\mathbf{z}_i, i = 1, \cdots, N\}$, suppose that there exists $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ such that $\sum_{i=1}^N E[\boldsymbol{\psi}(\mathbf{z}_i, \boldsymbol{\beta}_0)] = 0$ for some score function $\boldsymbol{\psi}$. The score function is a vector function of the same dimension p as the parameter $\boldsymbol{\beta}_0$ in general. The EE estimator $\hat{\boldsymbol{\beta}}_N$ of $\boldsymbol{\beta}_0$ is defined as the solution to the estimating equation $\sum_{i=1}^N \boldsymbol{\psi}(\mathbf{z}_i, \boldsymbol{\beta}) = 0$. In regression analyses, we have $\mathbf{z}_i = (y_i, \mathbf{x}_i^T)$ with response variable y and predictor $\mathbf{x}$ and the score function is usually given as $\boldsymbol{\psi}(\mathbf{z}, \boldsymbol{\beta}) = \phi(y - \mathbf{x}^T\boldsymbol{\beta})\mathbf{x}$ for some function ϕ. When ϕ is the identify function, it gives the OLS estimator and the estimating equation is linear in $\boldsymbol{\beta}$. However, the score function $\boldsymbol{\psi}$ is more often nonlinear, and this nonlinearity imposes difficulty to find low-dimensional summary statistics based on which the EE estimate for the entire data set can be obtained by aggregation as in (2.1). Therefore, I instead aim at finding an estimator that accurately approximates the EE estimator, and can still be computed by aggregation. Our basic idea is to approximate the nonlinear estimating equation by its first-order approximation, whose linearity then allows us to find representations similar to (2.1) and hence the proper low-dimensional summary statistics.

Again, consider partitioning the entire data set into K subsets. To simplify our notation, I assume that all subsets are of equal size n. This condition is not necessary

for the theory, though. Denote the observations in the kth subset by $\mathbf{z}_{k1}, \cdots, \mathbf{z}_{kn}$. The EE estimate $\hat{\boldsymbol{\beta}}_{nk}$ based on observations in the kth subset is then the solution to the following estimating equation,

$$\mathbf{M}_k(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{z}_{ki}, \boldsymbol{\beta}) = 0. \tag{2.2}$$

Let

$$\mathbf{A}_k = -\sum_{i=1}^{n} \frac{\partial \boldsymbol{\psi}(\mathbf{z}_{ki}, \hat{\boldsymbol{\beta}}_{nk})}{\partial \boldsymbol{\beta}}. \tag{2.3}$$

Since $\mathbf{M}_k(\hat{\boldsymbol{\beta}}_{nk}) = 0$, we have $\mathbf{M}_k(\boldsymbol{\beta}) = \mathbf{A}_k(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{nk}) + \mathbf{R}_2 = \mathbf{F}_k(\boldsymbol{\beta}) + \mathbf{R}_2$ from the Taylor expansion of $\mathbf{M}_k(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}_{nk}$, where $\mathbf{R}_2$ is the residual term in the Taylor expansion. The AEE estimator $\tilde{\boldsymbol{\beta}}_{NK}$ is then the solution to $\mathbf{F}(\boldsymbol{\beta}) = \sum_{k=1}^{K} \mathbf{F}_k(\boldsymbol{\beta}) = 0$, which leads to

$$\tilde{\boldsymbol{\beta}}_{NK} = \left(\sum_{k=1}^{K} \mathbf{A}_k \right)^{-1} \sum_{k=1}^{K} \mathbf{A}_k \hat{\boldsymbol{\beta}}_{nk}. \tag{2.4}$$

This representation suggests the following algorithm to compute the AEE estimator.

1. **Partition**. Partition the entire data set into K subsets with each containable in the computer's memory.

2. **Compression**. For the kth subset, save $(\hat{\boldsymbol{\beta}}_{nk}, \mathbf{A}_k)$ and discard the raw data. Repeat for $k = 1, \cdots, K$.

3. **Aggregation**. Calculate the AEE estimator $\tilde{\boldsymbol{\beta}}_{NK}$ using (2.4).

This implementation processes the data part by part and requires to store only $K(p^2+p)$ numbers after compressing the data, and therefore overcomes the computer's memory constraint.

2.3 Asymptotic Properties

In this section, I give the strong consistency of the AEE estimator and its asymptotic equivalence to the original EE estimator, which supports that the AEE estimator serves as a valid replacement of the original EE estimator. All proofs will be given in Chapter 2 Section 2.7

Let the score function be $\boldsymbol{\psi}(\mathbf{z}_i, \boldsymbol{\beta}) = (\psi_1(\mathbf{z}_i, \boldsymbol{\beta}), \cdots, \psi_p(\mathbf{z}_i, \boldsymbol{\beta}))^T$. I first specify some technical conditions.

(C1) The score function $\boldsymbol{\psi}$ is measurable for any fixed $\boldsymbol{\beta}$ and is twice continuously differentiable with respect to $\boldsymbol{\beta}$.

(C2) The matrix $-\frac{\partial \boldsymbol{\psi}(\mathbf{z}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ is semi-positive definite (s.p.d.), and $-\sum_{i=1}^{n} \frac{\partial \boldsymbol{\psi}(\mathbf{z}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ is positive definite (p.d.) in a neighborhood of $\boldsymbol{\beta}_0$ when n is large enough.

(C3) The EE estimator $\hat{\boldsymbol{\beta}}_n$ is strongly consistent, i.e. $\hat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}_0$ almost surely (a.s.) as $n \to \infty$.

(C4) There exists two p.d. matrices, $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ such that $\boldsymbol{\Lambda}_1 \leq n^{-1}\mathbf{A}_k \leq \boldsymbol{\Lambda}_2$ for all $k = 1, \cdots, K$, i.e. for any $\mathbf{v} \in \mathbb{R}^p$, $\mathbf{v}^T\boldsymbol{\Lambda}_1\mathbf{v} \leq n^{-1}\mathbf{v}^T\mathbf{A}_k\mathbf{v} \leq \mathbf{v}^T\boldsymbol{\Lambda}_2\mathbf{v}$, where $\mathbf{A}_k$ is given in (2.3).

(C5) In a neighborhood of $\boldsymbol{\beta}_0$, the norm of the second-order derivatives $\frac{\partial^2 \boldsymbol{\psi}_j(\mathbf{z}_i,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}$ is bounded uniformly, i.e. $\left\| \frac{\partial^2 \boldsymbol{\psi}_j(\mathbf{z}_i,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \right\| \leq C_2$ for all i, j, where C_2 is a constant.

(C6) There exits a real number $\alpha \in (1/4, 1/2)$ such that for any $\eta > 0$, the EE estimator $\hat{\boldsymbol{\beta}}_n$ satisfies $P(n^{\alpha}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| > \eta) \leq C_{\eta} n^{2\alpha-1}$, where $C_{\eta} > 0$ is a constant only depending on η.

Condition (C2) makes the AEE estimator $\tilde{\boldsymbol{\beta}}_{NK}$ well-defined. Condition (C3) is necessary for the strong consistency of the AEE estimator and is satisfied by almost all EE estimators in practice. Conditions (C4) and (C5) are required to prove the strong consistency of the AEE estimator, and are often true when each subset contains enough observations. Condition (C6) guarantees the consistency of the AEE estimator and the asymptotic equivalence of the AEE and EE estimators when the partition number K also goes to infinity as the number of observation goes to infinity. In Section 2.5, I will show that Condition (C6) is satisfied for the quasi-likelihood estimators considered in [23] under some regularity conditions.

Theorem 1 *Let* $k_0 = argmax_{1\leq k\leq K}\{\|\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0\|\}$. *Under Conditions (C1)-(C3), if the partition number* K *is bounded, we have* $\|\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0\| \leq K\|\hat{\boldsymbol{\beta}}_{nk_0} - \boldsymbol{\beta}_0\|$. *If Condition (C4) is also true, we have* $\|\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0\| \leq C\|\hat{\boldsymbol{\beta}}_{nk_0} - \boldsymbol{\beta}_0\|$ *for some constant* C *independent of* n *and* K. *Furthermore, if Condition (C5) is satisfied, we have* $\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N\| \leq C_1 \left(\|\hat{\boldsymbol{\beta}}_{nk_0} - \boldsymbol{\beta}_0\|^2 + \|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\|^2 \right)$ *for some constant* C_1 *independent of* n *and* K.

Theorem 1 shows that if the partition number K is bounded, then the AEE estimator is also strongly consistent. Usually, we have $\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\| = o(\|\hat{\boldsymbol{\beta}}_{nk_0} - \boldsymbol{\beta}_0\|)$. Therefore, the last part of Theorem 1 implies that $\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_0\| \leq 2C\|\hat{\boldsymbol{\beta}}_{nk_0} - \boldsymbol{\beta}_0\|^2 + \|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\|$.

Theorem 2 *Let $\hat{\boldsymbol{\beta}}_N$ be the EE estimator based on the entire data set. Then under Conditions (C1) - (C2), (C4)-(C6), if the partition number K satisfies $K = O(n^\gamma)$ for some $0 < \gamma < \min\{1 - 2\alpha, 4\alpha - 1\}$, we have $P(\sqrt{N}\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N\| > \delta) = o(1)$ for any $\delta > 0$.*

Theorem 2 says that if the EE estimator $\hat{\boldsymbol{\beta}}_N$ is a consistent estimator and the partition number K goes to infinity slowly, then the AEE estimator $\tilde{\boldsymbol{\beta}}_{NK}$ is also a consistent estimator. In general, one can easily use Theorem 2 to show the asymptotic normality of the AEE estimator if the EE estimator is asymptotically normally distributed, and further to prove the asymptotic equivalence of the two estimators. An application to QLE is given in the next section.

2.4 The Aggregated QLE

In this section, I demonstrate the applicability of the AEE technique to quasi-likelihood estimation and call the resulted estimator the aggregated quasi-likelihood estimator (AQLE). I consider a simplified version of QLE discussed in [23]. Suppose that we have N independent observations $(y_i, \mathbf{x}_i)$, $i = 1, \cdots, N$, where y is a scalar response and $\mathbf{x}$ is a p-dimensional vector of explanatory variables. Let μ be a contin-

uously differentiable function such that $\dot{\mu}(t) = d\mu/dt > 0$ for all t. Suppose that we have

$$E(y_i) = \mu(\boldsymbol{\beta}_0^T \mathbf{x}_i) \quad i = 1, \cdots, N. \tag{2.5}$$

for some $\boldsymbol{\beta}_0 \in \mathbb{R}^p$. Then the QLE of $\boldsymbol{\beta}_0$, $\hat{\boldsymbol{\beta}}_N$, is the solution to the estimating equation

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{N} [y_i - \mu(\boldsymbol{\beta}^T \mathbf{x}_i)]\mathbf{x}_i = 0, \tag{2.6}$$

Let $\varepsilon_i = y_i - \mu(\boldsymbol{\beta}_0^T \mathbf{x}_i)$ and $\sigma_i^2 = \mathrm{Var}(y_i)$. The following theorem shows that Condition (C6) is satisfied for the QLE under some regularity conditions.

Theorem 3 *Consider a generalized linear model specified by (2.5) with fixed design. Suppose that y_i's are independent and that λ_N is the minimum eigenvalue of $\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T$. If there are two positive constants C and M such that $\lambda_N/N > C$ and $\sup_i\{\|\mathbf{x}_i\|,\ \|\sigma_i^2\|\} \leq M$, then for any $\eta > 0$ and $\alpha \in (0, 1/2)$,*

$$P(N^\alpha \|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\| > \eta) \leq C_1 (m_\eta \eta)^{-2} N^{2\alpha - 1},$$

where $C_1 = pM^3C^{-3}$ is a constant, and $m_\eta > 0$ is a constant only depending on η.

Now suppose that the entire data set is partitioned into K subsets. Let $\{(y_{ki}, \mathbf{x}_{ki})\}_{i=1}^{n}$ be the observations in the kth subset with $n = N/K$.

(B1) The link function μ is twice continuously differentiable and the derivative of the link function is always positive, i.e. $\dot{\mu}(t) > 0$.

(B2) The vectors $\mathbf{x}_{ki}$ are fixed and uniformly bounded, and the minimum eigenvalue λ_k of $\sum_{j=1}^{n} \mathbf{x}_{kj}\mathbf{x}_{kj}^T$ satisfies $\lambda_k/n > C > 0$ for all k and n.

(B3) The variances of y_{ki}, σ_{ki}^2, are bounded uniformly.

Condition (B1) is needed for proving Conditions (C1) and (C5). Conditions (B1)-(B2) together guarantee Conditions (C2), (C4) and (C5). And it is easy to verify that all the conditions assumed in Theorem 1 of [23] are satisfied under Conditions (B1)-(B2). Hence, by Theorem 1 in [23] the QLEs $\hat{\boldsymbol{\beta}}_{nk}$ are strongly consistent. Theorem 3 implies that the QLEs $\hat{\boldsymbol{\beta}}_{nk}$ satisfy Condition (C6) under Conditions (B1)-(B3). Therefore, Theorem 1 and Theorem 2 hold for the AQLE under Conditions (B1)-(B3). Furthermore, the AQLE $\tilde{\boldsymbol{\beta}}_{NK}$ has the following asymptotic normality.

Theorem 4 *Let* $\boldsymbol{\Sigma}_N = \sum_{i=1}^{N} \sigma_i^2 \mathbf{x}_i \mathbf{x}_i^T$. *Suppose that there exist a constant* c_1 *such that* $\sigma_i^2 > c_1^2$ *for all* i *and* $\sup_i E(|\varepsilon_i|^r) < \infty$ *for some* $r > 2$. *Then under Conditions (B1)-(B3), if* $K = O(n^\gamma)$ *for some* $0 < \gamma < \min\{1 - 2\alpha, 4\alpha - 1\}$, *we have* $\boldsymbol{\Sigma}_N^{-1/2}\mathbf{D}_N(\boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I_p)$ *and* $\tilde{\boldsymbol{\beta}}_{NK}$ *is asymptotically equivalent to the QLE* $\hat{\boldsymbol{\beta}}_N$, *where* $\mathbf{D}_N(\boldsymbol{\beta}) = -\sum_{i=1}^{N} \dot{\mu}(\mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i^T\mathbf{x}_i$.

2.5 Simulation Studies

In this section, I illustrate the computational advantages of the AEE estimator by simulation studies. Consider computing the maximum likelihood estimator (MLE) of the regression coefficients in logistic regression with five predictors $x_1, \cdots, x_5$. Let y_i

be the binary response and $\mathbf{x}_i = (1, x_{i1}, \cdots, x_{i5})^T$. In a logistic regression model, we have

$$Pr(y_i = 1) = \mu(\mathbf{x}_i^T \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}, \quad i = 1, \cdots, N.$$

And the MLE of the regression coefficients $\boldsymbol{\beta}$ is a special case of the QLE as the solution to (2.6). Set the true regression coefficients as $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_5) = (1, 2, 3, 4, 5, 6)$ and the sample size as $N = 500,000$. The predictor values are drawn independently from the standard normal distribution.

Then, compute $\tilde{\boldsymbol{\beta}}_{NK}$, the AEE estimate of $\boldsymbol{\beta}$, with different partition numbers for $K = 1000, 950, \ldots, 100, 90, \cdots, 10$. In compressing the subsets, I use the Newton-Raphson method to calculate the MLE $\hat{\boldsymbol{\beta}}_{nk}$ in every subset k, $k = 1, \ldots, K$. For comparison, I also compute $\hat{\boldsymbol{\beta}}_N$, the MLE from the entire data set, which is equivalent to $\tilde{\boldsymbol{\beta}}_{NK}$ when $K = 1$. All programs are written in C and our computer has a 3.4GHz Pentium processor and 1.00GB memory.

Figure 2.1 plots the relative bias $\|\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0\|/\|\boldsymbol{\beta}_0\|$ against the number of partitions K. The linearly increasing trend can be well explained by the theory in Section 2.3 and 2.4. In Section 2.3, I argued that the magnitude of $\|\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0\|$ is close to $2C_1\|\hat{\boldsymbol{\beta}}_{nk_0} - \boldsymbol{\beta}_0\|^2 + \|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\|$. From Theorem 1 in [23], we have $\|\hat{\boldsymbol{\beta}}_{nk_0} - \boldsymbol{\beta}_0\|^2 = o([\log n]^{1+\delta}/n)$. Since $\log n \ll n$, $\|\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0\|$ is close to $o(1/n) = o(K/N)$, which increases linearly with K when N is held fixed. Since N is fixed in the simulation, $\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\|$ is fixed and so $\|\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0\|$ will roughly increase linearly with K.

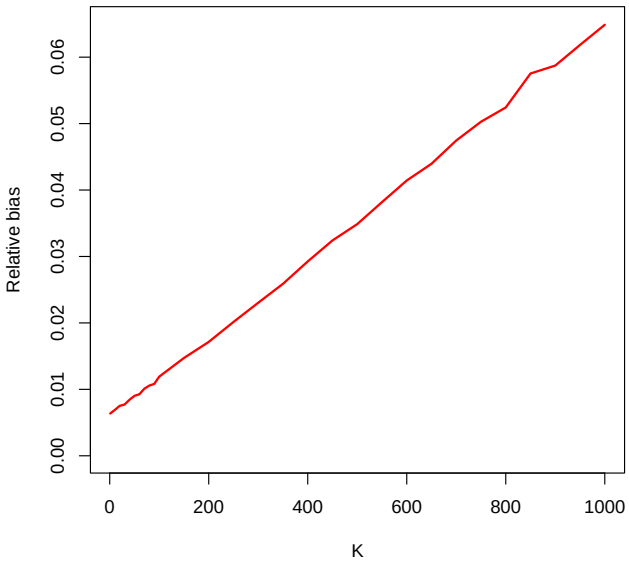

Figure 2.1. Relative bias against number of partitions

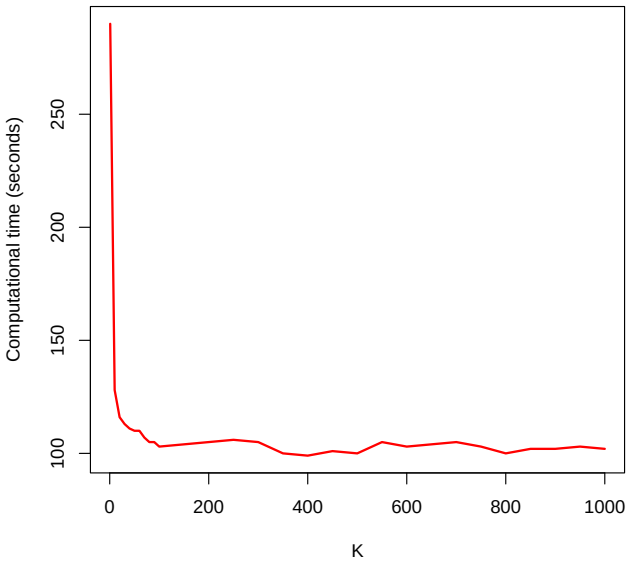

Figure 2.2. Computation time against number of partition K

Figure 2.2 plots the computational time against the number of partitions. It takes 290 seconds to compute the MLE ($K = 1$) and 128 seconds to compute the AEE estimator when $K = 10$, which shows a reduction of more than 50%. As K increases, the computational time soon stabilizes. This shows that we may choose a relatively small K as long as the size of each subset does not exceed the storage limit or memory constraint. On the other hand, we see that the AEE estimator provides not only an efficient storage solution, but also a viable way to achieve more efficient computation even when the EE estimate using all the raw data can be computed.

Next, I will show that the AEE estimator is more accurate than estimates based on sub-sampling. In our study, we can view $\hat{\boldsymbol{\beta}}_{nk}$ from each subset as estimates based on a sub-sample of the entire data set. Table 2.1 presents the percentages of $\hat{\boldsymbol{\beta}}_{nk}$ with relative bias $\|\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0\|/\|\boldsymbol{\beta}_0\|$ above that of the AEE estimator for different partition numbers. It is seen that that more than 90% of $\hat{\boldsymbol{\beta}}_{nk}$'s have relative bias larger than that of the $\tilde{\beta}_{NK}$, which clearly shows that the AEE estimator is generally more accurate than estimators based on sub-sampling.

Table 2.1
Performance of $\hat{\boldsymbol{\beta}}_{nk}$.

K	500	100	50	10
Percentage	94%	97%	94%	90%

2.6 Applications to Data Cubes and Data Streams

In this section, I discuss applications of the AEE estimator in two massive data environments: data cubes and data streams. Analyses in both environments require to perform the same analyses for different subsets while the raw data often can not be saved permanently. Efficient compression of the raw data by the AEE method enables remarkable computational reduction for estimating equation estimation in these two scenarios. In both cases, the size of the compressed data is independent of and far smaller than that of the raw data for most applications.

2.6.1 Methods

The AEE method can be applied to data cubes to support OLAP of EE estimation. Using the AEE method, I first compress the raw data in each base cell into the EE estimate $\hat{\boldsymbol{\beta}}_{nk}$ and A_k in (2.3). This only requires to scan the raw data once and then we can discard the raw data. And the EE estimate in any higher level cell can be approximated by computing the AEE estimate using the aggregation in (2.4). This aggregation is very fast since only simple operations are needed. Consequently, fast OLAP computation and efficient storage are both achieved when EE estimation is needed for many different cells.

The AEE method provides a natural solution to EE estimation for stream data. I first choose a sequence of integers $\{n_k\}$ such that $\sum_{k=1}^{K} n_k = N$. Choices of $\{n_k\}$ can be decided by the pyramidal time frame proposed by Aggarwal et al. [24] to

guarantee that the EE estimates for any time interval can be approximated well. Let $m_0 = 0$, $m_k = \sum_{l=1}^{k} n_l$ for $k = 1, \cdots, K$. At each time point m_k, I calculate and store the EE estimate $\hat{\boldsymbol{\beta}}_{nk}$ and A_k based on data items $z_{m_{k-1}}, \cdots, z_{m_k}$ in the time interval $[m_{k-1}, m_k]$. According to the property of the pyramidal time frame in [24], we can obtain a good approximation to the EE estimate in any time interval by computing the AEE estimator using (2.4).

2.6.2 Simulation Studies

Consider again maximum likelihood estimation in logistic regression to demonstrate the remarkable value of the AEE method. Since after the partitioning for the data streams is decided, each time interval can be viewed as a base cell in data cubes, our simulation focuses on data cubes only. In this simulation, I use the same simulated data as in Section 6 with two additional variables: location and time. Location has 20 levels and time has 50 levels, so we have $1000 = 50 \times 20$ base cells in total. In reality, this data set can be business transaction records in 50 months for 20 cities. Suppose that there are 500 records for each city in each month. Consider the situation where a business analyst is interested in computing the MLE in 100 different cubes. I simulate each of these 100 cubes by first randomly selecting D from $\{1, \cdots, 1000\}$ as the number of base cells contained in a cube, and then randomly choosing D base cells from the 1000 base cells.

Compare the computation time of the AEE estimates with that of computing the EE estimates directly from the raw data. Table 2.2 shows that the AEE method first

spent a moderate amount of time to compress all base cells and then finished the aggregation for all 100 queries almost timelessly, while it took about 70 times longer to compute the 100 EE estimates from the raw data. Obviously, we can expect even more significant time reduction when the calculation is needed for more cubes.

Table 2.2
Comparison of computational time.

	AEE estimate	EE estimate
Compression	97 seconds	NA
Aggregation	0.0 second	6771 seconds

2.7 Proof of Theorems

I first prove two lemmas that are needed for proofs of the theorems in this chapter.

Definition 2.7.1 *Let* $\mathbf{A}$ *be a* $d \times d$ *positive definite matrix. The* norm *of* $\mathbf{A}$*, is defined as* $\|\mathbf{A}\| = \sup_{\mathbf{v} \in \mathbb{R}^d, \mathbf{v} \neq 0} \frac{\|\mathbf{A}\mathbf{v}\|}{\|\mathbf{v}\|}$.

Lemma 1 *Suppose that* $\mathbf{A}$ *is a* $d \times d$ *positive definite matrix. Let* λ *be the smallest eigenvalue of* $\mathbf{A}$*, then we have* $\mathbf{v}^T\mathbf{A}\mathbf{v} \geq \lambda \mathbf{v}^T\mathbf{v} = \lambda \|\mathbf{v}\|^2$ *for any vector* $\mathbf{v} \in \mathbb{R}^d$*. On the contrary, if there exists a constant* $C > 0$ *such that* $\mathbf{v}^T\mathbf{A}\mathbf{v} \geq C\|\mathbf{v}\|^2$ *for any vector* $\mathbf{v} \in \mathbb{R}^d$*, then* $C \leq \lambda$*.*

Proof Since $\mathbf{A}$ is a positive definite matrix, there exits a $d \times d$ unitary matrix $\mathbf{U}$ and a $d \times d$ diagonal matrix $\mathbf{D}$ such that $\mathbf{A} = \mathbf{U}^T\mathbf{D}\mathbf{U}$ and $\mathbf{D}$'s diagonal elements are the eigenvalues of $\mathbf{A}$. Take any $\mathbf{v} \in \mathbb{R}^d$, and let $\mathbf{u} = \mathbf{U}\mathbf{v}$. Then $\mathbf{v}^T\mathbf{A}\mathbf{v} =$

$\mathbf{v}^T\mathbf{U}^T\mathbf{D}\mathbf{U}\mathbf{v} = \mathbf{u}^T\mathbf{D}\mathbf{u}$. Since $\mathbf{D}$ is diagonal and λ is the smallest element of $\mathbf{D}$'s main diagonal elements, we have $\mathbf{u}^T\mathbf{D}\mathbf{u} \geq \lambda\mathbf{u}^T\mathbf{u}$. Then since $\mathbf{U}$ is a unitary matrix, we have $\mathbf{u}^T\mathbf{u} = \mathbf{v}^T\mathbf{U}^T\mathbf{U}\mathbf{v} = \mathbf{v}^T\mathbf{v}$, and therefore, $\mathbf{v}^T\mathbf{A}\mathbf{v} \geq \lambda\|\mathbf{v}\|^2$.

Now, suppose that $\mathbf{v}^T\mathbf{A}\mathbf{v} \geq C\|\mathbf{v}\|^2$ for some $C > 0$. We know that there exists an eigenvector $\mathbf{v}_\lambda \neq 0$ corresponding to the eigenvalue λ such that $\mathbf{A}\mathbf{v}_\lambda = \lambda\mathbf{v}_\lambda$, and consequently, ${\mathbf{v}_\lambda}^T\mathbf{A}\mathbf{v}_\lambda \geq \lambda\|\mathbf{v}_\lambda\|^2$. Then from the assumption ${\mathbf{v}_\lambda}^T\mathbf{A}\mathbf{v}_\lambda \geq C\|\mathbf{v}_\lambda\|^2$, the second part of Lemma 1 follows. ■

Lemma 2 *Let* $\mathbf{A}$ *be a* $d \times d$ *positive definite matrix and* λ *is the smallest eigenvalue of* $\mathbf{A}$*. If* $\lambda \geq c > 0$ *for some constant* c*, one has* $\|\mathbf{A}^{-1}\| \leq c^{-1}$*.*

Proof Since $\mathbf{A}$ is a positive definite matrix, $\mathbf{A}^{-1}$ is also a positive definite matrix and the reciprocals of the eigenvalues of $\mathbf{A}$ are the eigenvalues of $\mathbf{A}^{-1}$. Thus λ^{-1} must be the largest eigenvalue of $\mathbf{A}^{-1}$. Hence, for any $\mathbf{v} \in \mathbb{R}^d$, we have $\|\mathbf{A}\mathbf{v}\| \leq \lambda^{-1}\|\mathbf{v}\|$. Therefore, $\|\mathbf{A}^{-1}\| \leq \lambda^{-1} \leq c^{-1}$. ■

In the following, I will give the proofs for all theorems in this chapter.

Proof [Proof of Theorem 1] From Conditions (C2) and (C5), we know that matrix $\mathbf{A}_k$ is positive definite for each $k = 1, \cdots, K$ when n is sufficiently large. Hence, $\sum_{k=1}^K \mathbf{A}_k$ is a positive definite matrix. In particular, $\left(\sum_{k=1}^K \mathbf{A}_k\right)^{-1}$ exists and Equation (2.4) is valid. Subtracting $\boldsymbol{\beta}_0$ from both sides of (2.4), we get

$$\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0 = \left(\sum_{k=1}^K \mathbf{A}_k\right)^{-1} \left[\sum_{k=1}^K \mathbf{A}_k(\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0)\right].$$

Thus,

$$\|\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0\| \leq \sum_{k=1}^{K} \left\| \left(\sum_{k=1}^{K} \mathbf{A}_k \right)^{-1} \mathbf{A}_k (\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0) \right\| \leq \sum_{k=1}^{K} \|\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0\|. \tag{2.7}$$

The second inequality comes from the fact $\|(\sum_{k=1}^{K} \mathbf{A}_k)^{-1}\mathbf{A}_k\| \leq 1$. Hence the first part of Theorem 1 follows.

Now suppose that Condition (C3) is also true. Let $\lambda_1 > 0$ be the smallest eigenvalue of the matrix $\boldsymbol{\Lambda}_1$ and λ_2 be the largest eigenvalue of the matrix $\boldsymbol{\Lambda}_2$. Then for any vector $\mathbf{v} \in \mathbb{R}^p$, we have $\mathbf{v}^T \frac{1}{n} \mathbf{A}_k \mathbf{v} \geq \mathbf{v}^T \boldsymbol{\Lambda}_1 \mathbf{v} \geq \lambda_1 \|\mathbf{v}\|^2$. Hence, $\mathbf{v}^T \frac{1}{nK} \sum_{k=1}^{K} \mathbf{A}_k \mathbf{v} \geq \lambda_1 \|\mathbf{v}\|^2$. Then from Lemmas 1 and 2, we have $\left\| (\frac{1}{nK} \sum_{i=1}^{K} \mathbf{A}_k)^{-1} \right\| \leq {\lambda_1}^{-1}$. Then since $\|n^{-1}\mathbf{A}_k\| \leq \|\boldsymbol{\Lambda}_2\| \leq \lambda_2$, it follows that

$$\left\| \left(\sum_{k=1}^{K} \mathbf{A}_k \right)^{-1} \mathbf{A}_k \right\| \leq \left\| \left(\frac{1}{nK} \sum_{k=1}^{K} \mathbf{A}_k \right)^{-1} \right\| \cdot \left\| \frac{1}{nK} \mathbf{A}_k \right\| \leq \frac{\lambda_2}{K\lambda_1}.$$

For $C = \lambda_2/\lambda_1$, we get

$$\|\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0\| \;\leq\; \sum_{k=1}^{K} \left\| \left(\sum_{k=1}^{K} \mathbf{A}_k \right)^{-1} \mathbf{A}_k (\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0) \right\| \leq C \|\hat{\boldsymbol{\beta}}_{nk_0} - \boldsymbol{\beta}_0\|.$$

Now suppose Condition (C5) is also satisfied.Let $\hat{\boldsymbol{\beta}}_N$ be the EE estimate based on the entire data set. Then we have $\mathbf{M}(\hat{\boldsymbol{\beta}}_N) = \sum_{k=1}^{K} \mathbf{M}_k(\hat{\boldsymbol{\beta}}_N) = 0$. By the Taylor expansion, we have

$$\mathbf{M}_k(\hat{\boldsymbol{\beta}}_N) = \mathbf{M}_k(\hat{\boldsymbol{\beta}}_{nk}) + \mathbf{A}_k(\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_{nk}) + \mathbf{R}_{nk}, \tag{2.8}$$

where the jth element of $\mathbf{R}_{nk}$ is

$$(\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_{nk})^T \sum_{i=1}^{n} \frac{\partial^2 \boldsymbol{\psi}_j(\mathbf{z}_{ki}, \boldsymbol{\beta}_k^*)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} (\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_{nk})$$

for some $\boldsymbol{\beta}_k^*$ between $\hat{\boldsymbol{\beta}}_N$ and $\hat{\boldsymbol{\beta}}_{nk}$. Therefore, we actually have $\|\mathbf{R}_{nk}\| \leq Cn\|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_{nk}\|^2 \leq 2Cn(\|\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0\|^2 + \|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\|^2)$ for some constant C. Since $\mathbf{M}_k(\hat{\boldsymbol{\beta}}_{nk}) = 0$ and $\mathbf{M}(\hat{\boldsymbol{\beta}}_N) = 0$, if we take summation over k on both side of Equation (2.8), we get $\sum_{k=1}^K \mathbf{A}_k(\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_{nk}) + \sum_{k=1}^K \mathbf{R}_{nk} = \sum_{k=1}^K \mathbf{A}_k(\hat{\boldsymbol{\beta}}_N - \tilde{\boldsymbol{\beta}}_{NK}) + \sum_{k=1}^K \mathbf{R}_{nk} = 0$, where the first equation comes from the definition of $\tilde{\boldsymbol{\beta}}_{NK}$. Hence, we have $\hat{\boldsymbol{\beta}}_N - \tilde{\boldsymbol{\beta}}_{NK} = \left(\sum_{k=1}^K \mathbf{A}_k\right)^{-1} \sum_{k=1}^K \mathbf{R}_{nk}$. Then similar to the first part of the proof, we get $\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N\| \leq C_1(\|\boldsymbol{\beta}_{nk_0} - \boldsymbol{\beta}_0\|^2 + \|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\|^2)$ for some constant C_1. ■

Proof [Proof of Theorem 2] Suppose that all the random variables are defined on a probability space $(\Omega, \mathcal{F}, P)$. Let $\Omega_{n,k,\eta} = \{\omega \in \Omega : \ n^\alpha \|\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0\| \leq \eta\}$, $\Omega_{N,\eta} = \{\omega \in \Omega : \ N^\alpha \|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\| \leq \eta\}$ and $\Gamma_{N,K,\eta} = \cap_{k=1}^K \Omega_{n,k,\eta} \cap \Omega_{N,\eta}$. From Condition (C7), for any $\eta > 0$, we have

$$P(\Gamma_{N,K,\eta}^c) \ \leq \ P(\Omega_{N,\eta}^c) + \sum_{k=1}^{K} P(\Omega_{n,k,\eta}^c) \leq C_\eta(N^{2\alpha-1} + Kn^{2\alpha-1}).$$

Since $K = O(n^\gamma)$ and $\gamma < 1 - 2\alpha$, we have $P(\Gamma_{N,K,\eta}^c) \to 0$ as $n \to \infty$.

Let $\mathbf{R}_{nk}$ be as in the proof of Theorem 1. For all $\omega \in \Gamma_{N,K,\eta}$, we have $\boldsymbol{\beta}_k^* \in B_\eta(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \eta\}$ since $B_\eta(\boldsymbol{\beta}_0)$ is a convex set and $\hat{\boldsymbol{\beta}}_N$, $\hat{\boldsymbol{\beta}}_{nk} \in B_\eta(\boldsymbol{\beta}_0)$. When η is small enough, the neighborhood in the Condition (C5) contains

$B_\eta(\boldsymbol{\beta}_0)$. Hence, we have $\|\mathbf{R}_{nk}\| \leq C_2 pn\|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_{nk}\|^2$ for all $\omega \in \Gamma_{N,K,\eta}$ when η is small enough. Therefore, for all $\omega \in \Gamma_{N,K,\eta}$, we have the following inequalities,

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}_N - \tilde{\boldsymbol{\beta}}_{NK}\| &\leq \left\| \left(\frac{1}{nK} \sum_{k=1}^{K} \mathbf{A}_k \right)^{-1} \right\| \left\| \frac{1}{nK} \sum_{k=1}^{K} \mathbf{R}_{nk} \right\| \\
&\leq \frac{\lambda_1^{-1} C_2 p}{K} \sum_{k=1}^{K} \|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_{nk}\|^2 \\
&\leq C n^{-2\alpha} \eta^2,
\end{aligned}$$

where $C = 4\lambda_1^{-1} C_2 p$ and λ_1 is the minimum eigenvalue of the matrix $\boldsymbol{\Lambda}_1$ as in the proof of Theorem 1. For any $\delta > 0$, take $\eta_\delta > 0$ such that $C\eta_\eta^2 < \delta$. Then for any $\omega \in \Gamma_{N,K,\eta_\delta}$ and $K = O(n^\gamma)$ for $\gamma < \min\{1 - 2\alpha, 4\alpha - 1\}$, we have

$$\sqrt{N}\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N\| \leq \sqrt{N} n^{-2\alpha} \delta = O(n^{(1+\gamma-4\alpha)/2})\delta.$$

Therefore, when n is large enough, we have $\Gamma_{N,K,\eta_\delta} \subset \{\omega \in \Omega : \sqrt{N}\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N\| \leq \delta\}$ and hence, $P(\sqrt{N}\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N\| > \delta) \leq P(\Gamma^c_{N,K,\eta_\delta}) \to 0$ as $n \to \infty$. ■

To prove Theorem 3, we need the following two lemmas. The proof of Lemma 3 can be found in [25] and the proof of Lemma 4 is in [23].

Lemma 3 *Suppose that A, B are two $p \times p$ positive definite matrices. Then*

(1) $A \geq B$ if and only if $A^{-1} \leq B^{-1}$

(2) If we have $AB = BA$, then $A \geq B$ implies $A^2 \geq B^2$.

Lemma 4 *Let H be a smooth injection from $\mathbb{R}^p$ to $\mathbb{R}^p$ with $H(\mathbf{x}_0) = \mathbf{y}_0$. Define $B_\delta(\mathbf{x}_0) = \{\mathbf{x} \in \mathbb{R}^p,\ \|\mathbf{x} - \mathbf{x}_0\| \le r\}$ and $S_\delta(\mathbf{x}_0) = \partial B_\delta(\mathbf{x}_0) = \{\mathbf{x} \in \mathbb{R}^p,\ \|\mathbf{x} - \mathbf{x}_0\| = \delta\}$. Then $\inf_{\mathbf{x} \in S_\delta(\mathbf{x}_0)} \|H(\mathbf{x}_0) - \mathbf{y}_0\| \ge r$ implies (1) $B_r(\mathbf{y}_0) = \{\mathbf{y} \in \mathbb{R}^p,\ \|\mathbf{y} - \mathbf{y}_0\| = \delta\} \subseteq H(B_\delta(\mathbf{x}_0))$; (2) $H^{-1}(B_r(\mathbf{y}_0)) \subseteq B_\delta(\mathbf{x}_0)$.*

Proof [Proof of Theorem 3] Suppose that all the random variables are defined on a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathbf{a}_N = (\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T)^{-1} \sum_{i=1}^N \mathbf{x}_i\varepsilon_i$ and $G_N(\boldsymbol{\beta}) = (\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T)^{-1} \sum_{i=1}^N [\mu(\boldsymbol{\beta}^T\mathbf{x}_i) - \mu(\boldsymbol{\beta}_0^T\mathbf{x}_i)]\mathbf{x}_i$, where $\varepsilon_i = y_i - \mu(\boldsymbol{\beta}_0^T\mathbf{x}_i)$. Then, the QLE $\hat{\boldsymbol{\beta}}_N$ is the solution of the equation $G_N(\hat{\boldsymbol{\beta}}_N) = \mathbf{a}_N$.

Take any $\eta > 0$, and let $m_\eta = \inf\{\dot{\mu}(\boldsymbol{\beta}^T\mathbf{x}) :\ \|\mathbf{x}\| \le M \text{ and } \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \eta\}$. Obviously, $m_\eta > 0$ only depends on η for the given M. Take any $\boldsymbol{\beta} \in \mathbb{R}^p$ with $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \eta$, we have by the mean-value theorem,

$$\begin{aligned} G_N(\boldsymbol{\beta}) &= (\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T)^{-1} \sum_{i=1}^N [\mu(\boldsymbol{\beta}^T\mathbf{x}_i) - \mu(\boldsymbol{\beta}_0^T\mathbf{x}_i)]\mathbf{x}_i \\ &= (\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T)^{-1} \sum_{i=1}^N \dot{\mu}(\boldsymbol{\beta}_i^T\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0), \end{aligned}$$

where $\boldsymbol{\beta}_i \in \mathbb{R}^p$ lies on the line segment between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$.

Since $\|\mathbf{x}_i\| \le M$, we have $\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T \le MNI_p$, where I_p is the $p \times p$ identity matrix, and hence by Lemma 3, $(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T)^{-2} \ge M^{-2}N^{-2}I_p$. On the other hand, since $\lambda_N/N > C$ and $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_0\| \le \eta$, we have

$$\sum_{i=1}^N \dot{\mu}(\boldsymbol{\beta}_i^T\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T \ge \sum_{i=1}^N m_\eta\mathbf{x}_i\mathbf{x}_i^T \ge m_\eta CNI_p.$$

Therefore, the following inequality holds

$$\begin{aligned}\|G_N(\boldsymbol{\beta})\|^2 &= (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\Big(\sum_{i=1}^N \dot{\mu}(\boldsymbol{\beta}_i^T\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T\Big)\Big(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T\Big)^{-2}\Big(\sum_{i=1}^N \dot{\mu}(\boldsymbol{\beta}_i^T\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T\Big)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\\ &\geq (MN)^{-2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\Big(\sum_{i=1}^N \dot{\mu}(\boldsymbol{\beta}_i^T\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T\Big)^2(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\\ &\geq (MN)^{-2}(m_\eta CN)^2\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 = \left(\frac{m_\eta C}{M}\right)^2\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2,\end{aligned}$$

i.e. $\|G_N(\boldsymbol{\beta})\| \geq m_\eta C\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|/M$ for $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \leq \eta$. In particular, $\|G_N(\boldsymbol{\beta})\| \geq m_\eta C\eta/M$ for all $\boldsymbol{\beta} \in S_\eta(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \ \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| = \eta\}$. Therefore, by Lemma 4, if $\|\mathbf{a}_N\| \leq m_\eta C\eta/M$, there exists an $\hat{\boldsymbol{\beta}}_N \in \mathbb{R}^p$, $\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\| \leq \eta$, such that $G_N(\hat{\boldsymbol{\beta}}_N) = \mathbf{a}_N$.

Let $\alpha \in (0, 1/2)$, define $W_{N,\eta} = \{\omega \in \Omega : \ N^\alpha\|\mathbf{a}_N\| \leq m_\eta C\eta/M\}$. Then by Chebyshev's inequality, we have

$$P(W_{N,\eta}^c) = P(N^\alpha\|\mathbf{a}_N\| > m_\eta C\eta/M) \leq M^2N^{2\alpha}E[\|\mathbf{a}_N\|^2]/(m_\eta C\eta)^2.$$

Furthermore,

$$E[\|\mathbf{a}_N\|^2] = tr[E(\mathbf{a}_N\mathbf{a}_N^T)] = tr[(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T)^{-1}(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T\sigma_i^2)(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T)^{-1}].$$

From $\sigma_i^2 \leq M$, we have $\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T\sigma_i^2 \leq M\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T$. Therefore,

$$tr[(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T)^{-1}(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T\sigma_i^2)(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T)^{-1}] \leq tr[M(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T)^{-1}] \leq pM(CN)^{-1}.$$

That is, $P(W_{N,\eta}^c) \leq pM^3C^{-3}(m_\eta\eta)^{-2}N^{2\alpha-1}$.

For $\omega \in W_{N,\eta}$, $\|\mathbf{a}_N\| \leq m_\eta C\eta/M$. By Lemma 4, there exists an $\hat{\boldsymbol{\beta}}_N \in \mathbb{R}^p$, $\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\| \leq \eta$, such that $G_N(\hat{\boldsymbol{\beta}}_N) = \mathbf{a}_N$. Furthermore, for $\omega \in W_{N,\eta}$ we have $N^\alpha\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\| \leq N^\alpha \left(\frac{m_\eta C}{M}\right)^{-1} \|\mathbf{a}_N\| \leq \eta$. Hence, $W_{N,\eta} \subseteq \Omega_{N,\eta} = \{\omega \in \Omega :\ N^\alpha\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\| \leq \eta\}$. At last we get

$$P(N^\alpha\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\| > \eta) = P(\Omega^c_{N,\eta}) \leq P(W^c_{N,\eta}) \leq pM^3C^{-3}(m_\eta\eta)^{-2}N^{2\alpha-1}.$$

■

Proof [Proof of Theorem 4] We first prove

$$\boldsymbol{\Sigma}_N^{-1/2}\mathbf{M}(\boldsymbol{\beta}_0) = \boldsymbol{\Sigma}_N^{-1/2}\sum_{i=1}^N \mathbf{x}_i[y_i - \mu(\mathbf{x}_i^T\boldsymbol{\beta}_0)] \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_p). \tag{2.9}$$

Let $\boldsymbol{\lambda}$ be any given unit p-dimensional vector. Put $\xi_{Ni} = \boldsymbol{\lambda}^T\boldsymbol{\Sigma}_N^{-1/2}\mathbf{x}_i\varepsilon_i$ and $\xi_N = \boldsymbol{\lambda}^T\boldsymbol{\Sigma}_N^{-1/2}\mathbf{M}(\boldsymbol{\beta}_0)$. Hence we have $\mathrm{E}(\xi_{ni}) = 0$, $i = 1, \cdots, N$, and $\mathrm{Var}(\xi_N) = 1$. From the Cramér-Wold theorem and the Linderberg central limit theorem, to prove (2.9), we only need to prove that, for any $\epsilon > 0$, $g_N(\epsilon) := \sum_{i=1}^N \mathrm{E}(|\xi_{Ni}|^2 I(|\xi_{Ni}| > \epsilon)) \longrightarrow 0$ as $N \longrightarrow \infty$. Let $a_{Ni} = \boldsymbol{\lambda}^T\boldsymbol{\Sigma}_N^{-1/2}\mathbf{x}_i$. Then we have

$$|\xi_{Ni}|^2 = \varepsilon_i^2\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_N^{-1/2}\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\Sigma}_N^{-1/2}\boldsymbol{\lambda} = \varepsilon_i^2 a_{Ni}^2$$

. By the assumption $\sigma_i^2 > c_1^2$, we have $\mathbf{\Sigma}_N > c_1^2 \sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T$, i.e. $\mathbf{\Sigma}_N - c_1^2 \sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T$ is a positive definite matrix, and hence,

$$\sum_{i=1}^N a_{Ni}^2 = \boldsymbol{\lambda}^T \mathbf{\Sigma}_N^{-1/2} \left(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T \right) \mathbf{\Sigma}_N^{-1/2} \boldsymbol{\lambda} \leq c_1^{-2}.$$

Then by the assumption $\sup_i \mathrm{E}(|\varepsilon_i|^r) < \infty$ for some $r > 2$, we have

$$\begin{aligned} g_N(\epsilon) &= \sum_{i=1}^N |a_{Ni}|^2 \mathrm{E}\left[|\varepsilon_i|^2 I(|\varepsilon_{Ni}| > \epsilon/|a_{Ni}|)\right] \leq \sum_{i=1}^N |a_{Ni}|^2 |a_{Ni}|^{r-2} \epsilon^{r-2} \mathrm{E}(|\varepsilon_i|^r) \\ &\leq c_1^{-2} \epsilon^{r-2} \sup_i \mathrm{E}(|\varepsilon_i|^r) \max_{1 \leq i \leq N} (|a_{Ni}|^{r-2}) \to 0 \quad \text{as} \quad n \to \infty. \end{aligned}$$

Therefore, we have proved (2.9). It is easy to check that all the conditions in Corollary 2.2 in [26] are satisfied here, the QLE $\hat{\boldsymbol{\beta}}_N$ has the following Badahur representation

$$\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0 = -\mathbf{D}_N^{-1}(\boldsymbol{\beta}_0) \sum_{i=1}^N \mathbf{x}_i [y_i - \mu(\mathbf{x}_i^T \boldsymbol{\beta}_0)] + O(N^{-3/4} (\log N)^3) \qquad \text{a.s.},$$

where $\mathbf{D}_N(\boldsymbol{\beta}) = -\sum_{i=1}^N \dot{\mu}(\mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i^T \mathbf{x}_i$. Then since $\mathbf{\Sigma}_N^{-1/2} = O(N^{-1/2})$ and $\mathbf{D}_N(\boldsymbol{\beta}_0) = O(N)$, we get

$$\begin{aligned} & -\mathbf{\Sigma}_N^{-1/2} \mathbf{D}_N(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) \\ = \; & \mathbf{\Sigma}_N^{-1/2} \sum_{i=1}^N \mathbf{x}_i [y_i - \mu(\mathbf{x}_i^T \boldsymbol{\beta}_0)] + \mathbf{\Sigma}_N^{-1/2} \mathbf{D}_N(\boldsymbol{\beta}_0) O(N^{-3/4} (\log N)^3) \\ = \; & \mathbf{\Sigma}_N^{-1/2} \sum_{i=1}^N \mathbf{x}_i [y_i - \mu(\mathbf{x}_i^T \boldsymbol{\beta}_0)] + O(N^{-1/4} (\log N)^3) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_p). \end{aligned}$$

For the AQLE, we have $-\boldsymbol{\Sigma}_N^{-1/2}\mathbf{D}_N(\boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}}_{NK}-\boldsymbol{\beta}_0) = -\boldsymbol{\Sigma}_N^{-1/2}\mathbf{D}_N(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_N-\boldsymbol{\beta}_0+\tilde{\boldsymbol{\beta}}_{NK}-\hat{\boldsymbol{\beta}}_N)$. Since $\|-\boldsymbol{\Sigma}_N^{-1/2}\mathbf{D}_N(\boldsymbol{\beta}_0)\| = O(N^{-1/2})$, Theorem 2 and Theorem 3 together implies that $\|\boldsymbol{\Sigma}_N^{-1/2}\mathbf{D}_N(\tilde{\boldsymbol{\beta}}_{NK}-\hat{\boldsymbol{\beta}}_N)\| = o_p(1)$ and hence

$$-\boldsymbol{\Sigma}_N^{-1/2}\mathbf{D}_N(\boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}}_{NK}-\boldsymbol{\beta}_0)\xrightarrow{d}\mathcal{N}(\mathbf{0},\mathbf{I}_p)$$

for $K = O(n^\gamma)$ with $\gamma < \min\{1-2\alpha, 4\alpha-1\}$. ■

3. Aggregation of U-statistics

Many commonly used statistics, especially, in many rank-based nonparametric procedure, can be put as U-statistics, such as the sample mean, the sample variance, the Mann-Whitney-Wilcoxon test statistic [27, 28], and Kendall's τ rank correlation [29]. U-statistics have long been known as a class of nonparametric estimators with good theoretical properties such as unbiasedness and asymptotic normality. However, in general, the time complexity of computing the U-statistic of degree m is $O(N^m)$, which is computationally costly for massive data sets when $m \geq 2$. For example, for a data set of $10,000$ observations, it takes about 4 hours to calculate the symmetry test statistic [30], a U-statistic of degree 3, using codes written in C on a computer with a 1.6 GHz Pentium processor and a 512 MB memory.

In this chapter, I will discuss how to apply statistical aggregation to U-statistics to reduce the computational complexity. I propose two unbiased nonparametric statistics, the aggregated U-statistic (AU-statistic) and the average aggregated U-statistic (AAU-statistic). The AU-statistic is obtained by first partitioning the entire data set into smaller subsets and then aggregating U-statistics from each subset by taking a weighted average. And the AAU-statistic is the average of AU-statistics computed from different random partitioning. Both statistics are shown to be asymptotically equivalent to the U-statistics under proper partitioning, while the AAU-statistic of-

fers a smaller finite sample variance than the AU-statistics at a price of some extra computational time. For a data set of size N, if we take the number of partitioning as $K = o(N)$, the computational complexity of both statistics is $O(K(N/K)^m)$, which means that they can be computed much faster when $m \geq 2$ as each subset is of a much smaller size.

3.1 Review of U-statistics

Let $X_1, \cdots, X_N$ be a random sample from an unknown distribution P in a non-parametric family $\mathcal{P}$. Suppose that $h(x_1, \cdots, x_m)$ is a measurable function defined on $\mathbb{R}^m$ that is symmetric in its arguments and satisfies $\vartheta = E[h(X_1, \cdots, X_m)] < \infty$. Then an unbiased estimator of ϑ is given by

$$U_N = \binom{N}{m}^{-1} \sum_{1 \leq i_1 < \cdots < i_m \leq N} h(X_{i_1}, \cdots, X_{i_m}), \tag{3.1}$$

where the summation is over the set of all $\binom{N}{m}$ combinations of m integers, $i_1 < i_2 < \ldots < i_m$ chosen from $\{1, 2, \ldots, N\}$. Here, U_N is called a U-statistic with kernel h and degree m.

The fundamental theory of U-statistics was first developed by [16], in which the asymptotic properties of U-statistics were derived using the projection method. Consider a U-statistic with kernel h and of degree m as in (3.1). For $k = 1, \cdots, m$,

define $h_k(x_1, \ldots, x_k) = E[h(x_1, \ldots, x_k, X_{k+1}, \ldots, X_m)]$. Then the projection of U_N on $(X_1, \ldots, X_N)$ is defined as

$$\breve{U}_N = E(U_N) + \sum_{i=1}^{N}[E(U_N|X_i) - E(U_N)] = \vartheta + \frac{m}{N}\sum_{i=1}^{N}\tilde{h}_1(X_i), \tag{3.2}$$

where $\tilde{h}_1(x) = h_1(x) - E[h(X_1, \ldots, X_m)]$. Based on the expansion in (3.2), one can obtain the following lemma [31, 32].

Lemma 5 *Let $\zeta_k = var(h_k(X_1, \cdots, X_k))$. Assuming that $E[h(X_1, \cdots, X_m)]^2 < \infty$, we have*

(i) if $\zeta_j = 0$ for $j < k$ and $\zeta_k > 0$ for some $k = 1, \cdots, m$, then

$$var(U_N) = \frac{k!}{N^k}\binom{m}{k}^2\zeta_k + O\left(\frac{1}{N^{k+1}}\right);$$

(ii) $E(U_N) = E(\breve{U}_N)$ and $E(U_N - \breve{U}_N)^2 = var(U_N) - var(\breve{U}_N)$.

Then, one can obtain the following asymptotic normality of U-statistics.

Theorem 5 *Assuming $E[h(X_1, \cdots, X_m)]^2 < \infty$, if $\zeta_1 > 0$, we have*

$$\sqrt{N}[U_N - \vartheta] \xrightarrow{d} \mathcal{N}(0, m^2\zeta_1) \quad as \quad N \to \infty.$$

For more detailed expositions of the general topic, see [31], [33] and [34].

3.2 The AU-statistics and the AAU-statistics

In this section, I propose the AU-statistic and the AAU-statistic and derive their asymptotic properties.

3.2.1 The AU-statistic and its Asymptotic Property

Let $\{X_1, \cdots, X_N\}$ be a random sample from an unknown distribution P. The AU-statistic is define as follows. First, partition the random sample into K subsets with observations in the kth subset denoted by $\{X_{k1}, \cdots, X_{kn_k}\}$ and the U-statistic based on them as U_{kn_k}. It is obvious that $\sum_{k=1}^{K} n_k = N$. Then, the AU-statistic is given by the following weighted average,

$$\tilde{U}_N = \frac{1}{N} \sum_{k=1}^{K} n_k U_{kn_k}. \tag{3.3}$$

We have the following asymptotic result about the AU-statistic.

Theorem 6 *Let $\tilde{U}_N$ be given by (3.3) with $E[h(X_1, \cdots, X_m)]^2 < \infty$. Then if $\zeta_1 > 0$ and $K = o(N)$, one has*

$$\sqrt{N}[\tilde{U}_N - \vartheta] \xrightarrow{d} \mathcal{N}(0, m^2\zeta_1) \quad as \quad N \to \infty.$$

Proof Let $\breve{U}_{kn_k}$ be the projection of the U-statistic U_{kn_k} on $X_{k1}, \cdots, X_{kn_k}$. It then follows from Lemma 5 that $E[(U_{kn_k} - \breve{U}_{kn_k})^2] = O(n_k^{-2})$. Therefore, we have

$$\begin{aligned}
\tilde{U}_N &= \frac{1}{N}\sum_{k=1}^{K} n_k \breve{U}_{kn_k} + \frac{1}{N}\sum_{k=1}^{K} n_k (U_{kn_k} - \breve{U}_{kn_k}) \\
&= \frac{1}{N}\sum_{k=1}^{K} n_k \left[\vartheta + \frac{m}{n_k}\sum_{i=1}^{n_k} \tilde{h}_1(X_{ki})\right] + \frac{1}{N}\sum_{k=1}^{K} n_k (U_{kn_k} - \breve{U}_{kn_k}) \\
&= \vartheta + \frac{m}{N}\sum_{i=1}^{N} \tilde{h}_1(X_i) + R_N, \qquad (3.4)
\end{aligned}$$

where $R_N = N^{-1}\sum_{k=1}^{K} n_k (U_{kn_k} - \breve{U}_{kn_k})$. Let $\Delta_k = U_{kn_k} - \breve{U}_{kn_k}$. Then, since Δ_k's are independent of each other, we have $E(R_N^2) = N^{-2}\sum_{k=1}^{K} n_k^2 E[(U_{kn_k} - \breve{U}_{kn_k})^2] = O(KN^{-2})$ and hence $E(NR_N^2) = O(K/N) \to 0$ as $N \to \infty$. By Chebyshev's inequality, $\sqrt{N}R_N = o_p(1)$. Finally, by the central limit theorem, we get

$$\sqrt{N}[\tilde{U}_N - \vartheta] \xrightarrow{d} \mathcal{N}(0, m^2\zeta_1) \quad \text{as} \quad N \to \infty.$$

■

Theorem 6 shows that, if the number of partitions is properly chosen, i.e., $K = o(N)$, the AU-statistics are asymptotically equivalent to the U-statistics. Meanwhile, the time complexity of the AU-statistics is much less than that of the U-statistics as it does not calculate the "pairs" across different subsets. For example, if we take $K = \sqrt{N}$ and let each partition have the same number of observations, then the time complexity of the AU-statistics would be $K \cdot O((N/K)^m) = O(N^{(m+1)/2})$ which is far less than $O(N^m)$ when $m \geq 2$ for moderately large N.

3.2.2 The AAU-statistic and its Asymptotic Property

While the AU-statistic is shown to be asymptotically equivalent to the U-statistic, however, it generally tends to have larger variance than the corresponding U-statistics in the finite sample case since AU-statistics use less "pairs" of observations. Notice that, unlike the U-statistics, the AU-statistics are not symmetric statistics and different partitions of the data set will result in different estimates of the parameter ϑ. Therefore, the average of the AU-statistics given by different partitions should be a more accurate estimator of the parameter ϑ than a single AU-statistic.

Let B be a fixed positive integer. For each $b = 1, \cdots, B$, we randomly partition the data set $\{X_1, \cdots, X_N\}$ into K subsets. Let $\tilde{U}_N^b$ be the AU-statistic for the bth partition. Then, I define the AAU-statistic as

$$\hat{U}_N = \frac{1}{B} \sum_{b=1}^{B} \tilde{U}_N^b. \tag{3.5}$$

Note that $\tilde{U}_N^b$, $b = 1, \cdots, B$, have the same asymptotic distribution but are not independent random variables. Hence, $\text{var}(\tilde{U}_N^b) = \text{var}(\tilde{U}_N)$ is a constant over $b = 1, \cdots, B$. We have $\text{var}(\hat{U}_N) \leq \text{var}(\tilde{U}_N)$, since $\text{cov}(X, Y) \leq \text{var}(X)^{1/2}\text{var}(Y)^{1/2}$ for any two random variables X, Y with finite second order moments. Therefore, the AAU-statistics have no larger variances than the AU-statistics. Using the representation of the AU-statistics in (3.4), it is straightforward to show the following asymptotic normality of the AAU-statistic.

Theorem 7 *For a given positive integer B, under the assumptions of Theorem 6, we have*

$$\sqrt{N}[\hat{U}_N - \vartheta] \xrightarrow{d} \mathcal{N}(0, m^2\zeta_1) \quad as \quad N \to \infty.$$

Therefore, the AAU-statistic is also asymptotically equivalent to the U-statistic. In the finite sample case, the AAU-statistic, being the average of the AU-statistics, is expected to have a smaller variance than an AU-statistic, which is also verified by simulation studies in Section 3.3. Even though a larger B seems to provide a statistic with a smaller variance, I do not recommend to use a large B, because it will make the AAU-statistic lose its computational advantage over the U-statistic. Furthermore, simulation studies in Section 3.3 show that small values of B ($B = 5$ or 10) already provide very good estimates, and a larger choice of B is unnecessary.

3.3 Simulation Studies

In this section, the two aggregation methods are applied to computing two U-statistics, symmetry test statistics [30] and Kendall's τ [29], and use simulations to show that the AU-statistics and the AAU-statistics are computationally much more efficient than the U-statistics and meanwhile well approximate it. To expediate the computation of U-statistics, all the programs are written in C. And the simulations were done on a computer with a 1.6 GHz Pentium processor and 512 MB memory.

3.3.1 Symmetry Test Statistics

Randles et al. [30] proposed a nonparametric method to test the symmetry of a data distribution. The test statistics is a U-statistic of order $m = 3$ with kernel function

$$h(x, y, z) = \frac{1}{3}[\text{sign}(x + y - 2z) + \text{sign}(x + z - 2y) + \text{sign}(y + z - 2x)],$$

where $\text{sign}(u) = -1,\ 0$, or 1 as $u <, =$, or > 0. In the simulation, 200 data sets are generated from the standard normal distribution, each of which is of size 2,000. Since the standard normal distribution is symmetric about 0, the symmetry test statistic is expected to be a good estimate of $\vartheta = 0$. For each simulated data set, the symmetry test statistic is computed in four different ways: U-statistics as in (3.1), AU-statistics as in (3.3) and the AAU-statistics as in (3.5) with $B = 5$ and $B = 10$, respectively. When computing AU-statistics and AAU-statistics, I also try different partition number $K = 20, 60$, and 100 to assess its impact, and all K subsets are kept of equal size.

Figure 3.1 shows box plots of the biases of the symmetry test statistics computed using different methods for $K = 20, 60$ and 100. It shows that the AU-statistics spread a little bit wider than the AAU-statistics and the U-statistics, especially for larger K's. But overall, both the AU-statistics and the AAU-statistics perform very well for all $K = 20, 60, 100$. Table 3.1 provides a numerical comparison of different methods on their biases, variances and average computational time. We also see

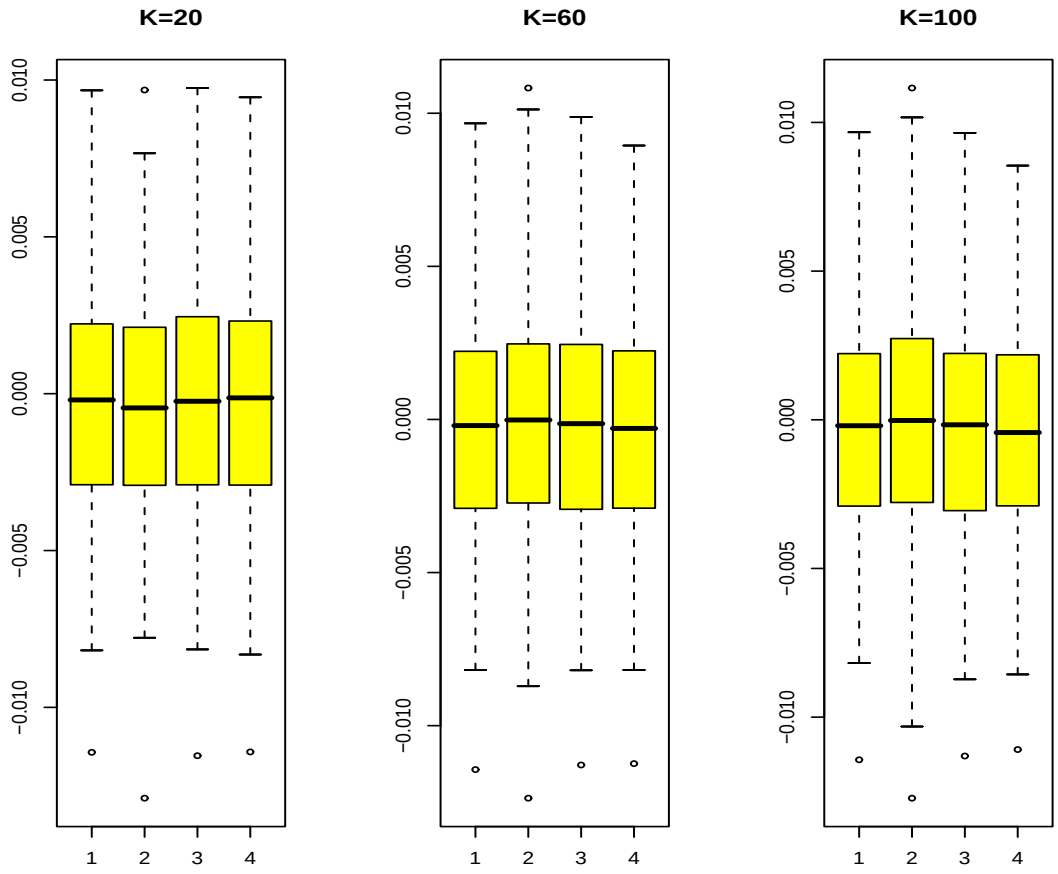

Figure 3.1. Boxplots of the biases of the symmetry test statistics. 1: U-statistics; 2: AU-statistics; 3: AAU-statistics ($B = 5$) and 4: AAU-statistics ($B = 10$).

that the variances of the AU-statistics are slightly larger than that of the U-statistics and the AAU-statistics especially when K is larger, whereas the latter two have similar variances. It is also shown that the AU-statistics and the AAU-statistics are computationally much more efficient than the U-statistics. The AU-statistics and AAU-statistics only take less than 1% of the computational time of the U-statistics

in most cases. With a larger K, the AU-statistics tend to have larger variances, but this is no longer seen for the AAU-statistics even with $B = 5$. A larger K also reduces the computational time dramatically for both AU-statistics and AAU-statistics. So, in general, I recommend to use the AAU-statistics with a relatively large K and a small B.

Table 3.1
Comparison of U-statistics, AU-statistics and AAU-statistics on computing the symmetry test statistic.

	K	Bias($\times 10^{-4}$)	Variance($\times 10^{-5}$)	Time (seconds)
U	1	-3.3	1.31	148
AU	20	-3.2	1.39	0.38
	60	-1.7	1.66	0.04
	100	-1.6	1.91	0.01
AAU ($B = 5$)	20	-3.8	1.32	1.82
	60	-3.9	1.31	0.20
	100	-4.1	1.34	0.07
AAU ($B = 10$)	20	-3.7	1.31	3.96
	60	-3.9	1.28	0.41
	100	-4.1	1.30	0.14

3.3.2 Kendall's τ

Now, I consider computing Kendall's τ, which is popularly used for quantifying the association of two random variables nonparametrically. Let $\mathbf{Z}_1 = (X_1, Y_1)^T$, ..., $\mathbf{Z}_N = (X_N, Y_N)^T$ be a series of independently and identically distributed (i.i.d.)

random vectors in $\mathbb{R}^2$. Kendall's τ is then $\tau_N = 1-2U_N$ with U_N being a U-statistic of order 2 with the kernel function $h(\mathbf{z}_1, \mathbf{z}_2) = I(x_1 < x_2, y_1 > y_2) + I(x_1 > x_2, y_1 < y_2)$ for $\mathbf{z}_1 = (x_1, y_1) \in \mathbb{R}^2$ and $\mathbf{z}_2 = (x_2, y_2) \in \mathbb{R}^2$, where I is the indicator function. When two variables are independent, we have $E(\tau_N) = 0$.

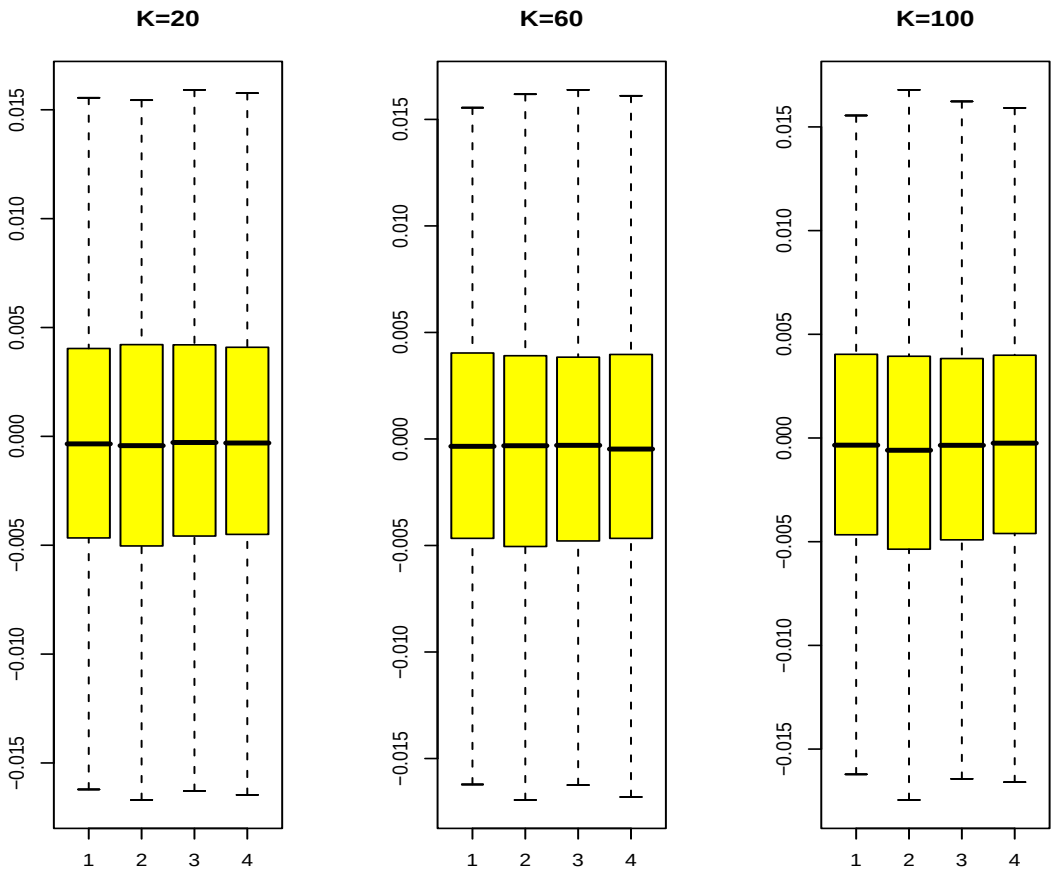

Figure 3.2. Boxplots of the biases of Kendall's τ. 1: U-statistics; 2: AU-statistics; 3: AAU-statistics ($B = 5$) and 4: AAU-statistics ($B = 10$).

Table 3.2
Comparison of U-statistics, AU-statistics and AAU-statistics on computing Kendall's τ.

	K	Mean($\times 10^{-4}$)	Variance($\times 10^{-5}$)	Time (seconds)
U	1	-3.0	4.27	9.14
AU	20	-3.4	4.36	0.46
	60	-3.9	4.51	0.15
	100	-4.8	4.60	0.09
AAU ($B = 5$)	20	-3.1	4.28	2.42
	60	-3.2	4.30	0.82
	100	-3.3	4.31	0.50
AAU ($B = 10$)	20	-3.1	4.27	4.85
	60	-3.1	4.31	1.65
	100	-2.8	4.29	1.00

In the simulation, I generate 200 data sets with 10,000 observations each from the bivariate standard normal distribution. Due to the independence between the two variables, we should expect Kendall's τ to be a good estimate of 0. Again, I compute Kendall's τ in four different ways for each simulated data set: U-statistics as in (3.1), AU-statistics as in (3.3) and the AAU-statistics as in (3.5) with $B = 5$ and $B = 10$, respectively. Comparison of the four methods is given in Figure 3.2 and Table 3.2. Similar results are seen as in simulation studies for computing the symmetry test statistic in Section 3.3.1. All methods perform well with little bias and the resulted estimators have similar distributions. Again, we see that the AAU-statistic with a relatively large K and a small B ($K = 100$, $B = 5$) seems to be the best choice when balancing the performance between the variance and the computational time.

3.4 An Application to Testing Serial Dependence

Ferguson et al. [35] proposed to use Kendall's τ to test against serial dependence in a univariate time series context. Here, I consider to apply the AU-statistics and AAU-statistics to compute Kendall's τ and test the serial independence against the nonzero first order correlation on both simulated data and real stock data. Results show that tests based on AU-statistics and AAU-statistics perform equally well as the original test in [35].

Suppose that we have a univariate time series $X_1,\cdots,X_{N+1}$. Let τ_N be Kendall's τ based on bivariate random vectors $(X_1, X_2)^T,\ldots,(X_N, X_{N+1})^T$. Then, $3\sqrt{N}\tau_N/2$ is asymptotically standard normal when assuming zero first order autocorrelation [35]. Therefore, one can test against the nonzero first order autocorrelation for the time series by rejecting the independence null hypothesis if $|\tau_N| > 2z_{\alpha/2}/3\sqrt{N}$ at significance level $\alpha > 0$, where $z_{\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution. Denote by $\tilde{\tau}_N$ and $\hat{\tau}_N$ Kendall's τ given by the AU-statistic and the AAU-statistic, respectively. As they have the same asymptotic distribution as τ_N, we can establish tests based on them using the same rejection rule.

I first use simulated data to compare the three tests based on τ_N, $\tilde{\tau}_N$ and $\hat{\tau}_N$, respectively. I generate 200 data sets of size 10,000 from an AR(1) model with autocorrelation $\rho = 0$, 0.02 and 0.05 respectively. Table 3.3 shows the computational time and Type I error rates (level) or powers for the three tests with $\alpha = 0.05$. We

Table 3.3
Testing serial dependence: simulated data.

		$\rho = 0$		$\rho = 0.02$		$\rho = 0.05$	
Statistics	K	Time (Sec)	Level	Time (Sec)	Power	Time (Sec)	Power
τ_N	1	9.23	0.055	9.16	0.485	9.24	1.0
$\tilde{\tau}_N$	20	0.46	0.055	0.47	0.485	0.46	1.0
	60	0.16	0.055	0.16	0.485	0.15	1.0
	100	0.10	0.055	0.093	0.485	0.090	1.0
$\hat{\tau}_N$ $(B = 5)$	20	2.30	0.055	2.35	0.485	2.32	1.0
	60	0.78	0.055	0.79	0.485	0.80	1.0
	100	0.47	0.055	0.48	0.485	0.48	1.0
$\hat{\tau}_N$ $(B = 10)$	20	4.64	0.055	4.63	0.485	4.62	1.0
	60	1.55	0.055	1.56	0.485	1.56	1.0
	100	0.96	0.055	0.95	0.485	0.99	1.0

see that all tests have the same Type I error rates or powers, while tests based on the AU-statistics and AAU-statistics require far less computational time.

Second, I apply the three tests to historical stock prices (close price) of Ford Motor Co. and General Electric (GE) Co. downloaded from finance.yahoo.com. The data contain 7,883 observations for the Ford stock and 11,639 observations for the GE stock, respectively. Table 3.4 gives Kendall's τ from four different methods and the corresponding computational time. All methods give the same Kendall's τ, but it takes remarkably less time to calculate $\tilde{\tau}_N$ and $\hat{\tau}_N$. In this case, the critical values at level $\alpha = 0.01$ for $N = 7,883$ and $11,639$ are 0.019 and 0.016, respectively. Therefore,

Table 3.4
Testing serial dependence: stock data.

		Ford		GE	
Statistics	K	Time	Estimate	Time	Estimate
τ_N	1	5.69	0.988	12.42	0.995
$\tilde{\tau}_N$	20	0.28	0.988	0.62	0.995
	60	0.10	0.988	0.21	0.995
	100	0.06	0.988	0.13	0.995
$\hat{\tau}_N$ $(B = 5)$	20	1.42	0.988	3.19	0.995
	60	0.48	0.988	1.05	0.995
	100	0.38	0.988	0.68	0.995
$\hat{\tau}_N$ $(B = 10)$	20	2.87	0.988	6.28	0.995
	60	0.98	0.988	2.10	0.995
	100	0.67	0.988	1.29	0.995

the tests are highly significant and we reject the null hypothesist that the observations in the data sets are independent.

In conclusion, tests based on τ_N, $\tilde{\tau}_N$ and $\hat{\tau}_N$ perform identically in their Type I error rate and power. But tests based on AU-statistics and AAU-statistics are computationally much more efficient than the one based on U-statistics.

4. An Application to Functional Regression Models

In this chapter, I will apply the aggregation method developed in Chapter 3 to expediate the computation of the functional regression models (FRM) [18]. FRM not only widens the class of existing regression models to accommodate new challenges in modelling real data, but also provides a general framework for unifying and generalizing popular non-parametric approaches for continuous as well as discrete data. The generalization occurs by replacing the single subject-based response y_i with a function $f(y_{i_1}, \cdots, y_{i_k})$ of several response $y_{i_1}, \cdots, y_{i_k}$ from multiple subjects $i_1, \cdots, i_k$. Then, a U-statistic based generalized estimating equation (UGEE) is constructed to estimate the parameters in the model. The details of the FRM will be given in Section 4.1.

However, the computational complexity of solving the estimating equation for the FRM is very high in general since the estimating equation is U-statistics based. In fact, the computation burden of the FRMs is even heavier than that of U-statistics, since numerical methods of solving equations involve many iterations and each iteration involves calculating a U-statistic. In this chapter, I will apply the aggregation scheme developed in Chapter 3 and propose an alternative AU-statistic based generalized estimating equation (AUGEE) which is computationally much more efficient than the original estimating equation. I will show that the estimator from the AUGEE

is asymptotically equivalent to the estimator to UGEE. Simulation studies also show that the estimator obtained from the AUGEE is nearly as efficient as the estimator obtained from the UGEE but computationally more efficient.

4.1 Functional Regression Models

The classic mean-based distribution-free generalized linear model (GLM) [36] only models the conditional mean of the response given the independent variables. Any inference about the variance is based on the model of the mean in the classic distribution-free GLM. When the model of the mean is not correctly specified, all the inference would be invalid or even misleading even if the model for the variance is correct. This limitation is exacerbated if our interest lies in modelling second and higher order moments. FRM provides a general framework for directly modelling of the second and higher order moments and it also unifies and generalizes existing nonparametric and semi-parametric models.

Suppose that $Z_1 = (Y_1, X_1), \cdots, Z_N = (Y_N, X_N)$ are independent observations. Let $\mathbf{f}$ and $\mathbf{g}$ be two known measurable q-dimensional vector-valued functions, which satisfies the following equation,

$$E[\mathbf{f}(Y_{i_1}, \cdots, Y_{i_m})|X_{i_1}, \cdots, X_{i_m}] = \mathbf{g}(X_{i_1}, \cdots, X_{i_m}; \boldsymbol{\theta}_0), \tag{4.1}$$

where $\boldsymbol{\theta}_0$ is a p-dimensional unknown parameter. Model (4.1) is called the functional regression model (FRM). Without loss of generality, we may assume the functions $\mathbf{f}$

and $\mathbf{g}$ are symmetric about their arguments. Otherwise, we can easily symmetrize them.

Suppose that $H(x_1, \cdots, x_m)$ is a measurable $p \times q$ dimensional matrix-valued function and is symmetric about its arguments and

$$\mathbf{h}(z_1, \cdots, z_m; \boldsymbol{\theta}) = H(x_1, \cdots, x_m)[\mathbf{f}(y_1, \cdots, y_m) - \mathbf{g}(x_1, \cdots, x_m; \boldsymbol{\theta})], \tag{4.2}$$

where $z_i = (y_i, x_i)$. The following UGEE is used to estimate $\boldsymbol{\theta}_0$ in the FRM,

$$\mathbf{U}_N(\boldsymbol{\theta}) = \binom{N}{m}^{-1} \sum_{1 \le i_1 < \cdots < i_m \le N} \mathbf{h}(Z_{i_1}, \cdots, Z_{i_m}; \boldsymbol{\theta}) = 0. \tag{4.3}$$

However, solving equation (4.3) is computationally expensive for $m \geq 2$ and relatively large N. In the next section, I will use the aggregation technique developed in Chapter 3 to reduce the computational complexity for solving (4.3).

4.2 AU-statistic Based Estimating Equations

In this section, I propose an alternative AUGEE for the FRM and show that the estimator obtained from the AUGEE is asymptotically equivalent to the estimator from the UGEE.

As before, we first partition the data set $\{Z_1, \cdots, Z_N\}$ into K subsets $\{Z_{k1}, \cdots, Z_{kn_k}\}$. Let $\mathbf{U}_k(\boldsymbol{\theta})$ be the U-statistic based function in (4.3) based on the kth subset. Then, we can solve the following alternative AUGEE to get an estimate of $\boldsymbol{\theta}_0$,

$$\tilde{\mathbf{U}}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{k=1}^{K} n_k \mathbf{U}_{kn_k}(\boldsymbol{\theta}) = \mathbf{0}. \tag{4.4}$$

Let $\tilde{\boldsymbol{\theta}}_{K,N}$ be the solution to the estimating equation (4.4). Note that $\tilde{\boldsymbol{\theta}}_{1,N}$ is just the solution to the UGEE (4.3). Because the estimating equation (4.4) uses less m-tuples than the estimating equation (4.3), the computational complexity of solving (4.4) would be much lower. If we use the Newton-Raphson algorithm and choose n_k to be the same, the computational complexity of solving the AUGEE (4.4) would be at the order of $O(N^m/K^{m-1})$ in each iteration, but the computational complexity of solving the UGEE (4.3) is at the order of $O(N^m)$ in each iteration. Therefore, the aggregation method tremendously reduces the computational burden of estimating FRMs when $m \geq 2$.

Let $\boldsymbol{\vartheta} = E[\mathbf{h}(X_1, \cdots, X_m)]$, $\mathbf{h}_k = E[\mathbf{h}(x_1, \ldots, x_k, X_{k+1}, \ldots, X_m)]$ and $\boldsymbol{\zeta}_k(\mathbf{h}) = \mathrm{Var}(\mathbf{h}_k(X_1, \cdots, X_k))$. Before presenting the asymptotic property of the estimator $\tilde{\boldsymbol{\theta}}_{K,N}$, I give the following conditions.

(C1) $E[\mathbf{h}(Z_1, \cdots, Z_m; \boldsymbol{\theta}_0)^T \mathbf{h}(Z_1, \cdots, Z_m; \boldsymbol{\theta}_0)] < \infty$ and $\boldsymbol{\zeta}_1(\mathbf{h}_{\boldsymbol{\theta}_0})$ is positive definite, where $\mathbf{h}_{\boldsymbol{\theta}_0}(z_1, \cdots, z_m) = \mathbf{h}(z_1, \cdots, z_m; \boldsymbol{\theta}_0)$.

(C2) $\mathbf{h}_(z_1, \cdots, z_m; \boldsymbol{\theta})$ is twice differentiable in a neighborhood of $\boldsymbol{\theta}_0$ and

$$B = E\big[\frac{\partial \mathbf{h}}{\partial \boldsymbol{\theta}}(Z_1, \cdots, Z_m; \boldsymbol{\theta}_0)\big]$$

is an invertible matrix.

(C3) Suppose that $h^s(z_1, \cdots, z_m; \boldsymbol{\theta})$ $(s = 1, \cdots, p)$ is the sth entry of the vector function $\mathbf{h}(z_1, \cdots, z_m; \boldsymbol{\theta})$ and $b(z_1, \cdots, z_m)$ is a measurable function which is symmetric about its argument and $E\big[b(Z_1, \cdots, Z_m)\big]^2 < \infty$. We have for all $s, i, j = 1, \cdots, p$

$$E[\frac{\partial h^s}{\partial \theta_j}(Z_1, \cdots, Z_m; \boldsymbol{\theta}_0)]^2 < \infty$$

and

$$\big|\frac{\partial^2 h^s}{\partial \theta_i \partial \theta_j}(z_1, \cdots, z_m, \boldsymbol{\theta})\big| \leq b(z_1, \cdots, z_m) \text{ in a neighborhood of } \boldsymbol{\theta}_0.$$

Theorem 8 *Let $\tilde{\boldsymbol{\theta}}_{K,N}$ be the solution to the AUGEE (4.4). If Conditions (C1), (C2) and (C3) are satisfied and $K = o(N)$, the estimator $\tilde{\boldsymbol{\theta}}_{K,N}$ is a consistent estimator of $\boldsymbol{\theta}_0$ and*

$$\sqrt{N}(\tilde{\boldsymbol{\theta}}_{K,N} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, m^2 G \boldsymbol{\zeta}_1(\mathbf{h}_{\boldsymbol{\theta}_0}) G^T), \tag{4.5}$$

where $G = B^{-1}$.

Note that Theorem 8 applies to the case of $K = 1$. A quick corollary of Theorem 8 is that the estimators $\tilde{\boldsymbol{\theta}}_{K,N}$ and $\tilde{\boldsymbol{\theta}}_{1,N}$ are asymptotically equivalent when $K =$

$o(N)$. Therefore, the aggregation method reduces the computational complexity while maintains the asymptotic efficiency of the estimator $\tilde{\boldsymbol{\theta}}_{1,N}$. The proof of Theorem 8 is given in Section 4.4.

4.3 Simulation Studies

In this section, I will show by simulation that the estimator obtained from the AUGEE is statistically equivalent to the estimator obtained from the UGEE, while the former is computationally more efficient.

Suppose that y_{1it}, y_{2it} are two measurement on subject i at time t ($i = 1, \cdots, n$ $t = 1, \cdots, T$). Assume that subjects are independent. Let σ_{kt}^2 be the variance of y_{kit} ($k = 1, 2$) and ρ_t be the correlation between the two measurements y_{1it} and y_{2it} at time t. By the independence assumption, it follows that

$$\begin{aligned} E[(y_{1it} - y_{1jt})^2/2] &= \sigma_{1t}^2 \\ E[(y_{2it} - y_{2jt})^2/2] &= \sigma_{2t}^2 \\ E[(y_{1it} - y_{1jt})(y_{2it} - y_{2jt})/2] &= \rho_t \sqrt{\sigma_{1t}^2}\sqrt{\sigma_{2t}^2} \end{aligned}$$

Let $\mathbf{y}_{it} = (y_{1it}, y_{2it})$ and $\mathbf{y}_i = (\mathbf{y}_{i1}, \cdots, \mathbf{y}_{iT})$. Denote $f_{kt}(\mathbf{y}_i, \mathbf{y}_j) = (y_{kit} - y_{kjt})^2/2$, $h_{kt} = \sigma_{kt}^2$ ($k = 1, 2$), $f_{3t}(\mathbf{y}_i, \mathbf{y}_j) = (y_{1it} - y_{1jt})(y_{2it} - y_{2jt})/2$ and $h_{3t} = \rho_t \sqrt{\sigma_{1t}^2}\sqrt{\sigma_{2t}^2}$. Then the FRM model is

$$E[f_{kt}(\mathbf{y}_i, \mathbf{y}_j)] = h_{kt} \quad k = 1, 2, 3 \;\; t = 1, \cdots, T.$$

Let $\mathbf{f}_t = (f_{1t}, f_{2t}, f_{3t})$, $\mathbf{h}_t = (h_{1t}, h_{2t}, h_{3t})$, $\mathbf{f} = (\mathbf{f}_1, \cdots, \mathbf{f}_T)$ and $\mathbf{h} = (\mathbf{h}_1, \cdots, \mathbf{h}_T)$. Then the FRM becomes $E[\mathbf{f}(\mathbf{y}_i, \mathbf{y}_j)] = \mathbf{h}$. Given observations $\mathbf{y}_1, \cdots, \mathbf{y}_N$, the following UGEE is used to estimate the parameters ρ_t, σ_{1t}^2 and σ_{2t}^2,

$$\mathbf{U}_N(\boldsymbol{\theta}) = \binom{N}{2}^{-1} \sum_{1 \le i < j \le N} \big(\mathbf{f}(\mathbf{y}_i, \mathbf{y}_j) - \mathbf{h}\big).$$

The FRM under consideration is based on a posttraumatic stress disorder (PTSD) study with a total of 95 women victims of sexual and non-sexual assault at the University of Pennsylvania Medical Center. The two measurements are PTSD Symptom Scale and Beck Depression Inventory at 5 time points. The goal is to longitudinally examine the correlations between the two measurements.

In the simulation, the number of time points is set as 3, i.e. $T = 3$. I generate 100 data sets and each data set has 100 observations. In every data set, I generate (y_{1it}, y_{2it}) are from a mean 0 bivariate normal distribution. The parameters are set as $\sigma_{1t}^2 = \sigma_{2t}^2 = 1$ and $\rho_t = 0.2$ for all $t = 1, 2, 3$. I compare the estimates from the UGEE and from the AUGEE with partition number $K = 5$, 10 and 20 using a program written in R. Figure 4.1 shows the box plots the 100 estimates of the correlation ρ_t from the UGEE and three AUGEEs and box plots from different estimating equations are similar. Table 4.1 compares the sample means and sample variances of the 100 estimates and average computation times using the four different estimating equations. All sample means are close for a given t. The variance increases as K increases, but the change is small. However, the AUGEE saves a considerable amount

of computation time compared with the UGEE. In all, the simulation clearly shows that AUGEEs provide estimators nearly as good as estimators obtained from UGEEs, while the computational burden of solving AUGEEs is much lower.

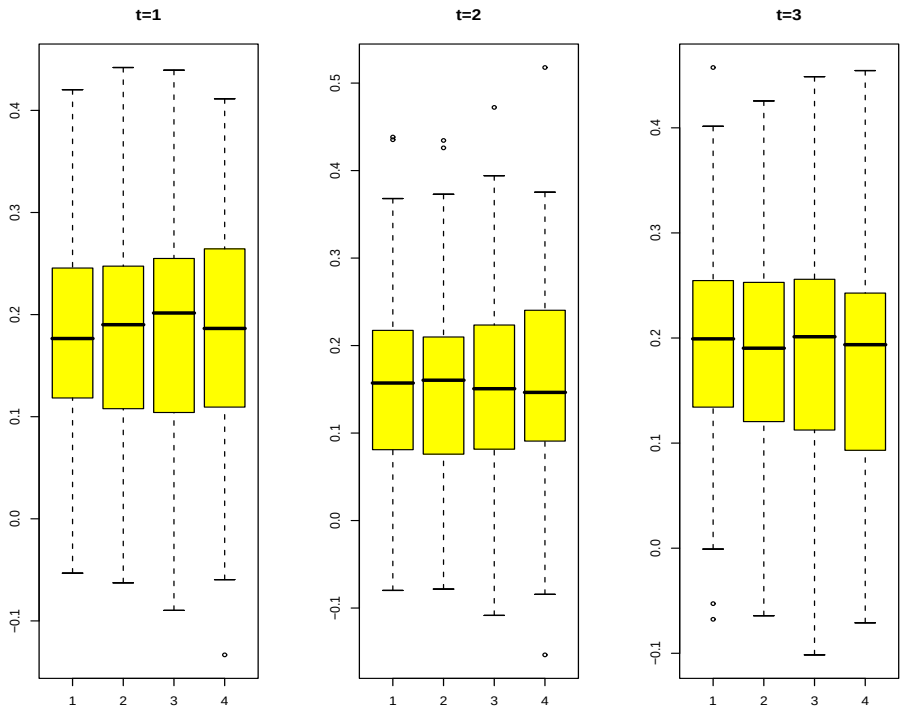

Figure 4.1. Box plots of correlation estimates from different types of estimating equations. 1: UGEE; 2: AUGEE ($K = 5$); 3: AUGEE ($K = 10$) and 4: AUGEE ($K = 20$).

Table 4.1
Comparison of estimates from different estimating equations

	K	Mean	Variance($\times 10^{-3}$)	Time (seconds)
	1	0.181	9.21	467.3
ρ_1 $(t=1)$	5	0.181	9.48	90.7
	10	0.182	10.2	44.5
	20	0.181	11.5	20.3
	1	0.154	11.0	467.3
ρ_2 $(t=2)$	5	0.155	10.7	90.7
	10	0.152	10.7	44.5
	20	0.159	12.0	20.3
	1	0.191	9.35	467.3
ρ_3 $(t=3)$	5	0.187	9.74	90.7
	10	0.183	10.6	44.5
	20	0.178	10.7	20.3

4.4 Proof of the Consistency and the Asymptotic Normality

In this section, I give the proof of Theorem 8. I first give a theorem about the asymptotic normality for the vector-valued AU-statistics, which itself is also of interest.

Theorem 9 *Suppose that* $\mathbf{h} = (h^1, \cdots, h^p)^T$ *is a* p*-dimensional vector-valued measurable functions which is symmetric about its arguments. Let* $\tilde{\mathbf{U}}_N$ *be the vector-valued AU-statistic with kernel* $\mathbf{h}$*. Suppose* $E[h^i(X_1, \cdots, X_m)]^2 < \infty$ *for all* $i = 1, \cdots, p$ *and* $\boldsymbol{\zeta}_1(\mathbf{h})$ *is positive definite. Then, if* $K = o(N)$*, one has*

$$\sqrt{N}[\tilde{\mathbf{U}}_N - \boldsymbol{\vartheta}] \xrightarrow{d} \mathcal{N}(0, m^2\boldsymbol{\zeta}_1(\mathbf{h})) \quad as \quad N \to \infty,$$

where $\boldsymbol{\vartheta} = E[\mathbf{h}(X_1, \cdots, X_m)]$.

Proof It is sufficient to prove that for any nonzero vector $\mathbf{c} = (c_1, \cdots, c_p)^T \in \mathbb{R}^p$, we have

$$\sqrt{N}[\mathbf{c}^T\tilde{\mathbf{U}}_N - \mathbf{c}^T\boldsymbol{\vartheta}] \xrightarrow{d} \mathcal{N}(0, m^2\mathbf{c}^T\boldsymbol{\zeta}_1(\mathbf{h})\mathbf{c}) \quad \text{as} \quad N \to \infty. \tag{4.6}$$

It is easy to see that $\mathbf{c}^T\tilde{\mathbf{U}}_N$ is an AU-statistic with kernel $g = \mathbf{c}^T\mathbf{h}$ and $E[g(X_1, \cdots, X_n)] = \mathbf{c}^T\boldsymbol{\vartheta}$. Since $E[h^i(X_1, \cdots, X_m)]^2 < \infty$ for all $i = 1, \cdots, p$, we have

$$E[g(X_1, \cdots, X_m)]^2 \;=\; \sum_{i,j=1}^{p} E[c_i c_j h^i(X_1, \cdots, X_m)h^j(X_1, \cdots, X_m)] < \infty.$$

At last, since $\zeta_1(g) = \mathbf{c}^T\boldsymbol{\zeta}_1(\mathbf{h})\mathbf{c} > 0$ and $K = o(N)$, we get the asymptotic normality (4.6) from Theorem 6 in Chapter 3. ■

Kantorovitch's theorem, whose proof can be found in [37], is needed in proving the consistency and asymptotic normality of the estimator $\tilde{\boldsymbol{\theta}}_{K,N}$. For ease of reference, I list Kantorovitch's theorem as the following lemma.

Lemma 6 (Kantorovitch's theorem) *Let* $\mathbf{a}_0$ *be a point in* $\mathbb{R}^p$, U *an open neighborhood of* $\mathbf{a}_0$ *and* $\mathbf{f} : U \mapsto \mathbb{R}^p$ *a differential mapping, with its derivative* $D\mathbf{f}(\mathbf{a}_0)$ *invertible. Define*

$$\mathbf{r}_0 = -D\mathbf{f}(\mathbf{a}_0)^{-1}\mathbf{f}(\mathbf{a}_0),\ \mathbf{a}_1 = \mathbf{a}_0 + \mathbf{r}_0,\ U_0 = \{\mathbf{x}|\ \|\mathbf{x} - \mathbf{a}_1\| \le \|\mathbf{r}_0\|\}$$

If the derivative $D\mathbf{f}(\mathbf{a}_0)$ *satisfies the Lipschitz condition*

$$\|D\mathbf{f}(\mathbf{x}_1) - D\mathbf{f}(\mathbf{x}_2)\| \leq M\|\mathbf{x}_1 - \mathbf{x}_2\| \quad \textit{for all points } \mathbf{x}_1, \mathbf{x}_2 \in U_0,$$

and if the inequality

$$\|\mathbf{f}(\mathbf{a}_0)\| \cdot \|D\mathbf{f}(\mathbf{a}_0)^{-1}\|^2 M \leq \frac{1}{2}$$

is satisfied, the equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ *has a unique solution in* U_0.

Proof [Proof of Theorem 8]

A. Consistency. Since $\mathbf{h}_{\boldsymbol{\theta}_0}$ satisfies Condition (C1), by Theorem 9 we have $\tilde{\mathbf{U}}_N(\boldsymbol{\theta}_0) = o_p(1)$. From Condition (C3), we get

$$\frac{\partial \tilde{\mathbf{U}}_N}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0) = E\big[\frac{\partial \mathbf{h}}{\partial \boldsymbol{\theta}}(Z_1, \cdots, Z_m; \boldsymbol{\theta}_0)\big] + o_p(1).$$

Then, $\frac{\partial \tilde{\mathbf{U}}_N}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0)$ is invertible in probability and $\mathbf{r}_N = -(\frac{\partial \tilde{\mathbf{U}}_N}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0))^{-1}\tilde{\mathbf{U}}_N(\boldsymbol{\theta}_0)$ tends to zero in probability. By Condition (C3), it is straightforward to show that there exists a neighborhood U of $\boldsymbol{\theta}_0$ and a constant M such that in probability

$$\left\| \frac{\partial \tilde{\mathbf{U}}_N}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_1) - \frac{\partial \tilde{\mathbf{U}}_N}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_2) \right\| \leq M\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$$

for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in U$. Again, since $\tilde{\mathbf{U}}_N(\boldsymbol{\theta}_0) = o_p(1)$ and $(\frac{\partial \tilde{\mathbf{U}}_N}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0))^{-1}$ is bounded in probability, we have $\|\tilde{\mathbf{U}}_N(\boldsymbol{\theta}_0)\|\|(\frac{\partial \tilde{\mathbf{U}}_N}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0))^{-1}\|^2 M \leq 1/2$ in probability. Then, by Kantorovitch's theorem, there exists a unique solution $\tilde{\boldsymbol{\theta}}_N$ in the neighborhood $U_N =$

$\{\boldsymbol{\theta}|\ \|\boldsymbol{\theta}-\boldsymbol{\theta}_N\| \le \mathbf{r}_N\}$ in probability, where $\boldsymbol{\theta}_N = \boldsymbol{\theta}_0 + \mathbf{r}_N$. Then, we have $\|\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\| \le 2\|\mathbf{r}_N\| = o_p(1)$ and the estimator $\tilde{\boldsymbol{\theta}}_{K,N}$ is a consistent estimator.

B. NORMALITY. Since $\tilde{\boldsymbol{\theta}}_N$ is the solution to Equation (4.4), we have $\tilde{\mathbf{U}}_N(\tilde{\boldsymbol{\theta}}_N) = \mathbf{0}$. Expand the vector-valued function $\tilde{\mathbf{U}}_N(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$ using Taylor's Theorem, we have

$$\mathbf{0} = \tilde{\mathbf{U}}_N(\tilde{\boldsymbol{\theta}}_N) = \tilde{\mathbf{U}}_N(\boldsymbol{\theta}_0) + \frac{\partial \tilde{\mathbf{U}}_N}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) + \mathbf{R}_2,$$

where $\mathbf{R}_2$ is the second order residual in the Taylor's expansion. Therefore, we have the following representation

$$\sqrt{N}(\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) = -\left(\frac{\partial \tilde{\mathbf{U}}_N}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right)^{-1} \sqrt{N}\tilde{\mathbf{U}}_N(\boldsymbol{\theta}_0) - \left(\frac{\partial \tilde{\mathbf{U}}_N}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right)^{-1} \sqrt{N}\mathbf{R}_2.$$

By Conditions (C2) and (C3), we have $\frac{\partial \tilde{\mathbf{U}}_N}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0) \to B$ in probability. Let V_k be the U-statistic with kernel $b(\cdot)$ based on the observations $\{Z_{k1}, \cdots, Z_{kn_k}\}$ and $\tilde{V}_N = \sum_{k=1}^{K} n_k V_k / N$ be the corresponding AU-statistic. Since $\tilde{\boldsymbol{\theta}}_N$ is a consistent estimator of $\boldsymbol{\theta}_0$, we have $\|\mathbf{R}_2\| \le C\tilde{V}_N \|\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\|^2$ in probability for some constant C. From the proof of Part A, we know that $\sqrt{N}\|\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\|^2 \le \sqrt{N}\|\mathbf{r}_N\|^2 = o_p(1)$. Furthermore, $\tilde{V}_N$ goes to $\mathrm{E}[b(Z_1, \cdots, Z_m)]$ in probability. Hence, $\|\sqrt{N}\mathbf{R}_2\| = o_p(1)$ and Theorem 8 is proved using the delta method. ■

5. Conclusion and Discussion

In this thesis, I introduced a general strategy, statistical aggregation, for performing statistical analysis on massive data sets. Statistical aggregation partitions the entire data set into small pieces and performs statistical analysis on each piece to obtain certain summary statistics. Then, it aggregates the resulted summary statistics together into the statistics of interest. Statistical aggregation provides a general scheme for OLAP of advanced statistical analyses in data cubes. Statistical aggregation distinguishes itself from previous technologies in data cubes by a looser but statistically satisfactory requirement, i.e. the resulted statistics from statistical aggregation should be asymptotically equivalent to the statistics of interest.

I applied the statistical aggregation strategy to two large families of statistics, EE estimators and U-statistics. The summary statistics for EE estimators are just the EE estimate plus an auxiliary matrix. The summary statistics for U-statistics are just U-statistics and sample sizes of subsets. The aggregation algorithms are simple weighted average for both EE estimators and U-statistics. I showed that the aggregated statistics are asymptotically equivalent to the original statistics. Simulation studies and real data examples also show the aggregation methods perform equally well and are computationally much faster. In the thesis, I also applied the aggregation method developed for U-statistics to FRMs. I proposed an alternative AUGEE

for the FRM and showed that the resulted estimator is asymptotically equivalent to the estimator obtained from the UGEE. Simulation studies validated this result and further revealed that a considerable amount of computation time could be saved by using the AUGEE.

The choice of summary statistics is not unique in general. For example, we can approximate the estimating equation based on the kth subsets by the second order Taylor series and define another set of summary statistics for the EE estimator. The aggregated estimating equation becomes a quadratic equation and the aggregation algorithm is to solve this quadratic equation. Preliminary simulation studies show that this version of aggregated EE estimators have smaller bias than the AEE estimators in Chapter 2 and hence enable us to partition the entire data set into more pieces. However, the asymptotic theory is much harder to develop.

There are potentially many more applications of the statistical aggregation strategy. For instance, we can apply the statistical aggregation to Bayes estimators such as posterior means. It is well-known that the posterior mean $\hat{\theta}_n$ asymptotically follows the normal distribution $N(0, I_{\theta_0}^{-1})$ under certain regularity conditions [17, 38], where I_{θ_0} is the Fisher information matrix. Suppose that the entire data set is partitioned into K subsets, each of which has n_k observations. Let $\hat{\theta}_{k,n_k}$ be the posterior mean and $N = \sum_{k=1}^{K} n_k$. Define the aggregated Bayes estimate $\tilde{\theta}_{K,N}$ as the weighted average of the posterior means $\hat{\theta}_{k,n_k}$, $\tilde{\theta}_{K,N} = \sum_{k=1}^{K} n_k \hat{\theta}_{k,n_k}/N$. One can prove that the aggregated Bayes estimate $\tilde{\theta}_{K,N}$ also asymptotically follows $N(0, I_{\theta_0}^{-1})$ based on Levy's continuity theorem (see [39] and [40] among many others). Simulation studies show

that the aggregated Bayes estimator performs equally well as the Bayes estimator. Though this compression and aggregation scheme cannot save computation time of the posterior mean in general, it is very useful for the OLAP of Bayesian estimation in data cubes.

Another possible application of the statistical aggregation is to the problems of ranking and ordering instances. In the ranking problem, one has to compare two or more different observations and provide a rank to each observation. For example, web search engines compare and sort hundreds of web pages by degree of relevance for a particular request, rather than simply classifying them as relevant or not. Recently, Clémençon et al. [41] proposed to learn the ranking rule by minimizing a ranking risk of the form of a U-statistic. Hence, the statistical aggregation method could be applied to reduce the computational costs of the minimization procedure.

One closely related work to this thesis is the binning technique [42, 43], which partitions the sample space into many bins, compresses the data into the averages of the bins and uses the compressed data to perform any statistical analysis. However, when the sample space is multidimensional, the bins can be defined in many different ways and it is not clear how to choose the best among them. The number of bins increases exponentially as the dimension of space increases and there will be a huge number of bins when the dimension of sample space is relatively large. Furthermore, the existing asymptotic results are mostly for the one-dimensional case, which makes the application of binning technique to multidimensional studies lack of theoretical foundation.

Another closely related work to statistical aggregation is the data squashing (DS) [44] technique. DS is a model-free data compression technique which could be useful when different users prefer different models for the same massive data set. The recipe of DS is to compress the raw data into pseudo-samples attached with certain weights. These weights are determined by matching the lower order moments. After DS, statistical analyses maybe performed only based on the pseudo-sample and their weights. However, the asymptotic behavior of the DS is largely unclear. Hence, the quality of the statistical analysis based on the squashed data is difficult to evaluate when some specific model is used. Likelihood-based DS (LDS) [45] is an extension of the DS and it compresses the raw data into pseudo-samples by approximating the likelihood of the raw data. Similar to the DS, the asymptotic theory of the LDS is also unclear and the quality of LDS is largely unguaranteed.

Bibliography

[1] G. Hulten, L. Spencer, and P. Domingos. Mining time-changing data streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 97–106, 2001.

[2] J. I. Munro and M. S. Paterson. Selection and sorting with limited storage. *Theoretical Computer Science*, 12:315–323, 1980.

[3] M. R. Henzinger, P. Raghavan, and S. Rajagopalan. Computing on data streams. Technical Report 1998-011, Digital Equipment Corporation, Systems Research Center, May 1998.

[4] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. *Data Mining and Knowledge Discovery*, 1:29–54, 1997.

[5] S. Agarwal, R. Agarwal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. In *Proceedings of the International Conference on Very Large Data Bases*, pages 506–521, 1996.

[6] Y. Zhao, P. M. Deshpande, and J. F. Naughton. An array-based algorithm for simultaneous multidimensional aggregates. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 159–170, 1997.

[7] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. *IEEE Transactions on Knowledge and Data Engineering*, 15:515–528, 2003.

[8] J. Beringer and E. Hüllermeier. Online clustering of parallel data streams. *Data and Knowledge Engineering*, 58(2):180–204, 2006.

[9] H. Wang, W. Fan, P. Yu, and J. Han. Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the 2003 ACM International Conference on Knowledge Discovery and Data Mining*, pages 226–235, 2003.

[10] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang. Multi-dimensional regression analysis of time-series data streams. In *Proceedings of the VLDB Conference*, pages 323–334, 2002.

[11] J. Han, Y. Chen, G. Dong, J. Pei, B. W. Wah, J. Wang, and Y. Cai. Steam cube: An architecture for multi-dimensional analysis of data streams. *Distributed and Parallel Databases*, 18(2):173–197, 2005.

[12] Y. Chen, G. Dong, J. Han, J. Pei, B. Wah, and J. Wang. Regression cubes with lossless compression and aggregation. *IEEE Transactions on Knowledge and Data Engineering*, 18:1585–1599, 2006.

[13] C. Liu, M. Zhang, M. Zheng, and Y. Chen. Step-by-step regression: A more efficient alternative for polynomian multiple linear regression in stream cube. In *Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 437–448, 2003.

[14] B. C. Chen, L. Chen, Y. Lin, and R. Ramakrishnan. Prediction cubes. In *Proceedings of the 31st VLDB Conference*, pages 982–993, 2005.

[15] S. Pang, S. Ozawa, and N. Kasabov. Incremental linear discriminant analysis for classification of data streams. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 35(2):905–914, 2005.

[16] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19:293–325, 1948.

[17] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, New Jersey, 2nd edition, 1998.

[18] J. Kowalski and X. M. Tu. *Modern Applied U-Statistics*. John Wiley & Sons, Hoboken, New Jersey, 2008.

[19] R. W. M. Wedderburn. Quasi-likelihood functions, generlized linear models, and the gaussian-newton method. *Biometrika*, 61:439–447, 1974.

[20] D. Relles. *Robust regression by modified least squares*. PhD thesis, Yale University, 1968.

[21] P. Huber. Robust regression. *The Annals of statistics*, 1:799–821, 1973.

[22] P. J. Huber. *Robust Statistics*. Wiley, New Jersey, 1981.

[23] K. Chen, I. Hu, and Z. Ying. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27:1155–1163, 1999.

[24] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In *Proceedings of the 29th International Conference on Very Large Data Bases*, pages 81–92, 2003.

[25] G. S. Wang, M. X. Wu, and Z. Z. Jia. *Matrix Inequalities*. Science Press, Beijing, 2006. in Chinese.

[26] X. He and Q. Shao. A general Bahadur representation of M-estimators and its applications to linear regression with non-stochastic designs. *The Annals of Statistics*, 24:2608–2630, 1996.

[27] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60, 1947.

[28] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.

[29] M. Kendall. A new measure of rank correlation. *Biomerika*, 30:81–89, 1938.

[30] R. H. Randles, M. A. Fligner, G. E. Policello, and D. A. Wolfe. An asymptotically distribution-free test for symmetry versus asymmetry. *Journal of the American Statistical Association*, 75:168–172, 1980.

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

[32] J. Shao. *Mathematical Statistics*. Springer, New York, 2003.

[33] A. J. Lee. *U-statistics*. Marcel Dekker Inc., New York, 1990.

[34] V. S. Koroljuk and Yu. V. Borovskich. *Theory of U-statistics*. Kluwer Academic Publishers, Norwell, MA, 1994.

[35] T. S. Ferguson, C. Genest, and M. Hallin. Kendall's tau for serial dependence. *The Canadian Journal of Statistics*, 28:587–604, 2000.

[36] P. McCullagh and J. A. Nelder. *Generlized linear models*. Chapman and Hall, London, 2nd edition, 1989.

[37] J. H. Hubbard and B. B. Hubbard. *Vector Calculus, Linear Algebra, and Differential Forms*. Prentice-Hall, New Jersey, 1999.

[38] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonprametrics*. Springer, New Jersey, 2002.

[39] K. L. Chung. *A Course in Probability Theory*. Elsevier, San Diego, California, 3rd edition, 2001.

[40] A. N. Shiryaev. *Probability*. Springer, New Jersey, 2nd edition, 1995.

[41] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36:844–874, 2008.

[42] P. Hall, B. U. Park, and B.A. Turlach. A note on design transformation and binning in nonparametric curve estimation. *Biometrika*, 85(2):469–476, 1998.

[43] T. Shi and B. Yu. Binning in gaussian kernel regularization. *Statistica Sinica*, 16:541–567, 2005.

[44] W. DuMouchel, C. Volinsky, T. Johnson, C. Cortes, and D. Pregibon. Squashing flat files flatter. In *Proceedings of the Fifth ACM Conference on Knowledge Discovery and Data Mining*, pages 6–15, 1999.

[45] D. Madigan, N. Raghavan, W. DuMouchel, M. Nason, C. Posse, and G. Ridgeway. Likelihood-based data squashing: A modeling approach to instance construction. *Data Mining and Knowledge Discovery*, 6:173–190, 2002.