Statistics and Its Interface Volume 4 (2011) 73–83

Aggregated estimating equation estimation

Nan Lin* and Ruibin Xi†

Motivated by the recent active research on online analytical processing (OLAP), we develop a computation and storage efficient algorithm for estimating equation (EE) estimation in massive data sets using a "divide-and-conquer" strategy. In each partition of the data set, we compress the raw data into some low dimensional statistics and then discard the raw data. Then, we obtain an approximation to the EE estimator, the aggregated EE (AEE) estimator, by solving an equation aggregated from the saved low dimensional statistics in all partitions. Such low dimensional statistics are taken as the EE estimates and first-order derivatives of the estimating equations in each partition.

We show that, under proper partitioning and some regularity conditions, the AEE estimator is strongly consistent and asymptotically equivalent to the EE estimator. A major application of the AEE technique is to support fast OLAP of EE estimations for data warehousing technologies such as data cubes and data streams. It can also be used to reduce the computation time and conquer the memory constraint problem posed by massive data sets. Simulation studies show that the AEE estimator provides efficient storage and remarkable deduction in computational time, especially in its applications to data cubes and data streams.

Keywords and phrases: Massive data sets, Estimating equation, Data compression, Aggregation, Consistency, Asymptotic normality, Data cube.

1. INTRODUCTION

Two major challenges in analyzing massive data sets are storage and computational efficiency. In recent years, there have been active researches on developing compression and aggregation schemes to support fast online analytical processing (OLAP) of various statistical analyses, such as linear regression [7, 14], general multiple linear regression [6, 19], logistic regression analysis [26], predictive filters [6], naive Bayesian classifiers [4] and linear discriminant analysis [22]. The OLAP analysis is usually associated with data warehousing technologies such as data cubes [1, 12, 27] and data streams [16, 21], where fast responses to queries are often needed. The response time of any OLAP tool should be in the order of seconds, at most minutes, even if complex statistical analyses are involved.

Most of the current OLAP tools can only support simple analyses that are essentially linear operators [7, 6, 14, 19]. However, many advanced statistical analyses are nonlinear and thus most of the current OLAP tools cannot be used to support these advanced analyses. In this paper, we developed a compression and aggregation strategy to support fast OLAP analysis for estimating equation (EE) estimators. The EE estimators are a very large family of estimators and many statistical estimation techniques can be unified into the framework of EE estimators, including the ordinary least square (OLS) estimator, the quasi-likelihood estimator (QLE) [25] and the robust M-estimator [17]. The scheme developed in this paper can not only support fast OLAP of EE estimation, but also can be used to reduce the computation time of the EE estimates and solve the memory constraint problem imposed by massive data sets.

The compression and aggregation technique developed in this paper is based on the "divide-and-conquer" strategy. We first partition the massive data sets into K 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 reconstructing 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. We show in theory that the proposed aggregated EE (AEE) estimator is asymptotically equivalent to the EE estimator if the number of partitions K does not go to infinity too fast. Simulation studies validate the theory and show that the AEE estimator is computationally very efficient. Our results also show that the AEE estimator provides more accurate estimates than estimates from a subsample of the entire data set, which is commonly done for static massive data sets.

The remainder of the paper is organized as follows. We first review regression cube [6] in Section 2 and then present the AEE estimator in Section 3 with its asymptotic properties given in Section 4. In Section 5, we study the application of the AEE estimator to QLE and provide asymptotic properties for the resulted aggregated QLE. Sections 6 and 7 study the performance of the AEE estimator and its applications to data cubes and data streams through simulation studies. And at last, Section 8 concludes the paper and provides some discussion. All proofs are given in the Appendix.

*Corresponding author.

†This work was done when Ruibin Xi was a PhD student in the Department of Mathematics, Washington University in St. Louis.

2. AGGREGATION FOR LINEAR REGRESSION

In this section, we review the regression cube technique [6] to illustrate the idea of aggregation for linear regression analysis.

Suppose that we have N independent observations $(y_1, \mathbf{x}_1), \ldots, (y_N, \mathbf{x}_N)$, 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 $\hat{\boldsymbol{\beta}}_k = (\mathbf{X}_k^T\mathbf{X}_k)^{-1}\mathbf{X}_k^T\mathbf{y}_k$ 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

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

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 thanks to 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, $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{k=1}^K \mathbf{X}_k^T(\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}) = 0$.

3. THE AEE ESTIMATOR

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

Given independent observations $\{\mathbf{z}_i, i = 1, \ldots, N\}$, 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 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 analysis, 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, the estimating equation is linear in $\boldsymbol{\beta}$ and the resulting estimator is the OLS estimator. 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 (1). Therefore, we adjust our aim to 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 (1) and hence the proper low-dimensional summary statistics.

Again, consider partitioning the entire data set into K subsets. To simplify our notation, we assume that all subsets are of equal size n. This condition is not necessary for our theory, though. Denote the observations in the kth subset by $\mathbf{z}_{k1}, \ldots, \mathbf{z}_{kn}$. The EE estimate $\hat{\boldsymbol{\beta}}_{nk}$ based on observations in the kth subset is then the solution to the following estimating equation,

$$(2) \qquad \mathbf{M}_k(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{\psi}(\mathbf{z}_{ki}, \boldsymbol{\beta}) = 0.$$

Let

$$(3) \qquad \mathbf{A}_k = -\sum_{i=1}^n \frac{\partial \boldsymbol{\psi}(\mathbf{z}_{ki}, \hat{\boldsymbol{\beta}}_{nk})}{\partial \boldsymbol{\beta}}.$$

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

$$(4) \qquad \tilde{\boldsymbol{\beta}}_{NK} = \left(\sum_{k=1}^K \mathbf{A}_k\right)^{-1} \sum_{k=1}^K \mathbf{A}_k\hat{\boldsymbol{\beta}}_{nk}.$$

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 $k = 1, \ldots, K$.
3. **Aggregation**. Calculate the AEE estimator $\tilde{\boldsymbol{\beta}}_{NK}$ using (4).

This implementation makes it feasible to process massive data sets on regular computers as long as each partition is manageable to the computer. It also provides a very efficient storage solution because only $K(p^2 + p)$ numbers need to be stored after compressing the data.

4. ASYMPTOTIC PROPERTIES

In this section, we give the consistency of the AEE estimator. Theorem 4.1 gives the strong consistency the AEE estimator for finite K. Theorem 4.2 further shows that when K goes to infinity not too fast, the AEE estimator is a consistent estimator under some regularity conditions. Theorem 4.2 is very useful to prove the asymptotic equivalence of the AEE estimator and the EE estimator. In the next section, we apply Theorem 4.2 to the aggregated quasi-likelihood estimators (QLE) and show its asymptotic equivalence to its original QLE. Let the score function be $\boldsymbol{\psi}(\mathbf{z}_i, \boldsymbol{\beta}) = (\psi_1(\mathbf{z}_i, \boldsymbol{\beta}), \ldots, \psi_p(\mathbf{z}_i, \boldsymbol{\beta}))^T$. We 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 $k = 1, \ldots, K$, 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 (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. $\|\frac{\partial^2 \boldsymbol{\psi}_j(\mathbf{z}_i,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\| \leq C_2$ for all i, j, where C_2 is a constant.

(C6) There exists 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.

Under Condition (C2), the matrices $\mathbf{A}_k$ is positive definite in probability and therefore the AEE estimator $\tilde{\boldsymbol{\beta}}_{NK}$ is well-defined in probability. 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) is useful in showing 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 observations goes to infinity. In Section 5, we will show that Condition (C6) is satisfied for the quasi-likelihood estimators considered in [5] under some regularity conditions.

Theorem 4.1. *Let* $k_0 = \text{argmax}_{1\leq k\leq K}\{\|\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0\|\}$. *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* $\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N\| \leq C_1(\|\hat{\boldsymbol{\beta}}_{nk_0} - \boldsymbol{\beta}_0\|^2 + \|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\|^2)$ *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 4.1 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 4.2. *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 4.2 tells us 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 4.2 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.

5. THE AGGREGATED QLE

In this section, we demonstrate the applicability of the AEE technique to quasi-likelihood estimation and call the resulted estimator the aggregated quasi-likelihood estimator (AQLE). We consider a simplified version of the QLE discussed in [5]. Suppose that we have N independent observations $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, N$, where y is a scalar response and $\mathbf{x}$ is a p-dimensional vector of explanatory variables. Let μ be a continuously differentiable function such that $\dot{\mu}(t) = d\mu/dt > 0$ for all t. Suppose that we have

$$E(y_i) = \mu(\boldsymbol{\beta}_0^T\mathbf{x}_i) \qquad i = 1, \ldots, N. \tag{5}$$

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

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^N [y_i - \mu(\boldsymbol{\beta}^T\mathbf{x}_i)]\mathbf{x}_i = 0, \tag{6}$$

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

Theorem 5.1. *Consider a generalized linear model specified by* (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 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 [5] are satisfied under Conditions (B1)–(B2). Hence, by Theorem 1 in [5] the QLEs $\hat{\boldsymbol{\beta}}_{nk}$ are strongly consistent. Theorem 5.1 implies that the QLEs $\hat{\boldsymbol{\beta}}_{nk}$ satisfy Condition (C6) under Conditions (B1)–(B3). Therefore, the conclusions in Theorem 4.1 and Theorem 4.2 hold for the AQLE under Conditions (B1)–(B3). Furthermore, the AQLE $\tilde{\boldsymbol{\beta}}_{NK}$ has the following asymptotic normality.

Theorem 5.2. *Let* $\mathbf{\Sigma}_N = \sum_{i=1}^{N}\sigma_i^2\mathbf{x}_i\mathbf{x}_i^T$ *and* $\mathbf{D}_N(\boldsymbol{\beta}) = -\sum_{i=1}^{N}\dot{\mu}(\mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i^T\mathbf{x}_i$. *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* $\mathbf{\Sigma}_N^{-1/2}\mathbf{D}_N(\boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}}_{NK}-\boldsymbol{\beta}_0)\xrightarrow{d}\mathcal{N}(\mathbf{0}, I_p)$ *and* $\tilde{\boldsymbol{\beta}}_{NK}$ *is asymptotically equivalent to the QLE* $\hat{\boldsymbol{\beta}}_N$.

6. SIMULATION STUDIES AND REAL DATA ANALYSIS

6.1 Simulation

In this section, we illustrate the computational advantages of the AEE estimator by simulation studies. We consider computing the maximum likelihood estimator (MLE) of the regression coefficients in logistic regression with five predictors $x_1, \ldots, x_5$. Let y_i be the binary response and $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i5})^T$. In a logistic regression model, we have

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

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

We then compute $\tilde{\boldsymbol{\beta}}_{NK}$, the AEE estimate of $\boldsymbol{\beta}$, with different partition numbers for $K = 1,000, 950, \ldots, 100, 90, \ldots, 10$. In compressing the subsets, we use the Newton-Raphson method to calculate the MLE $\hat{\boldsymbol{\beta}}_{nk}$ in every subset k, $k = 1, \ldots, K$. For comparison, we 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 1.6GHz Pentium processor and 512MB memory.

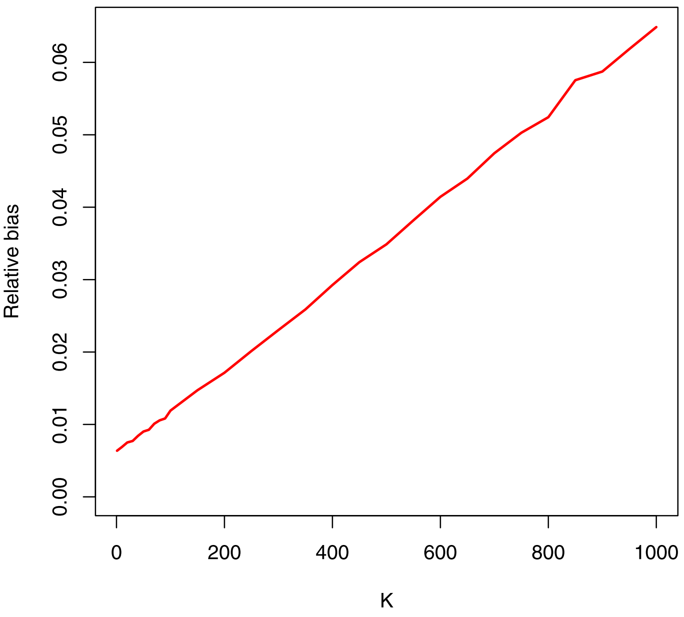

Figure 1. Relative bias against number of partitions.

Figure 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 our theory. In Section 4.1, we 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 [5], 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, $\|\hat{\boldsymbol{\beta}}_N-\boldsymbol{\beta}_0\|$ is fixed and so $\|\tilde{\boldsymbol{\beta}}_{NK}-\boldsymbol{\beta}_0\|$ will roughly increase linearly with K.

Figure 2 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 computational time 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, we 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 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 a relative bias larger than that of the $\tilde{\beta}_{NK}$, which clearly shows that the AEE estimator is more accurate than estimators based on sub-sampling.

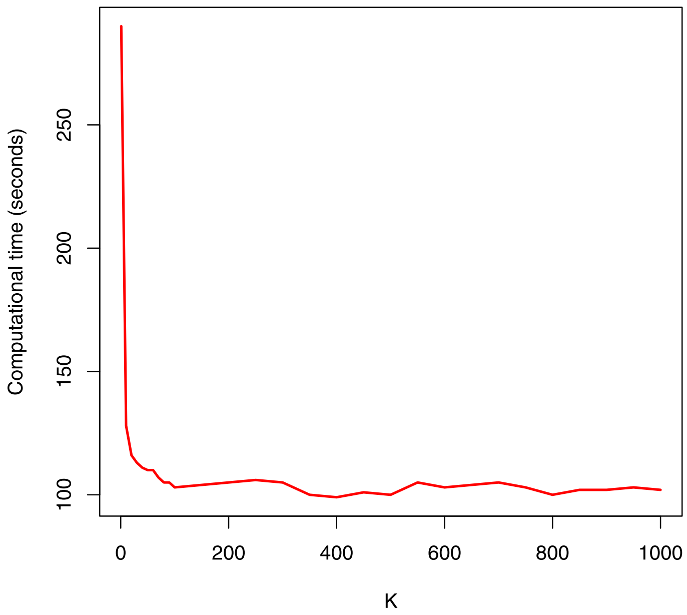

Figure 2. Computation time against number of partition K.

Table 1. Performance of $\hat{\boldsymbol{\beta}}_{nk}$

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

6.2 Real data analysis

In this section, we apply our aggregation on a real data set. In [8], Chiang et al. used next-generation sequencing (NGS) to detect copy number variation in the sample genome. It is known that the current NGS platforms have various biases [9], for example, GC-bias can lead to uneven distribution of short reads on the genome. Another important factor that can influence the read distribution is the mappability [23] of the genomic positions. Specifically, due to the existence of the segmental duplication and repeat sequences, a short sequence (e.g. 35 bp short sequence) starting from a genomic position may have many copies in the reference genome, making this genomic position not uniquely mappable. Hence, variation of the mappability across the reference genome will also lead to uneven distribution of uniquely mapped reads. Here, we are interested in how the number of reads in a certain genomic window is related with factors like GC-content and mappability.

We use the sequencing data of the matched normal genome of the cell line H2347 in [8] to study how the number of reads relate with other factors. We first binned the uniquely mapped reads into 1,000 bp bins and counted the number of reads in each bin. Then, for each bin, we counted how many nucleotides are G, C and A. Since the bin size is known, once we know nucleotide counts for G, C and A, we basically know how many nucleotides are T. For each bin,

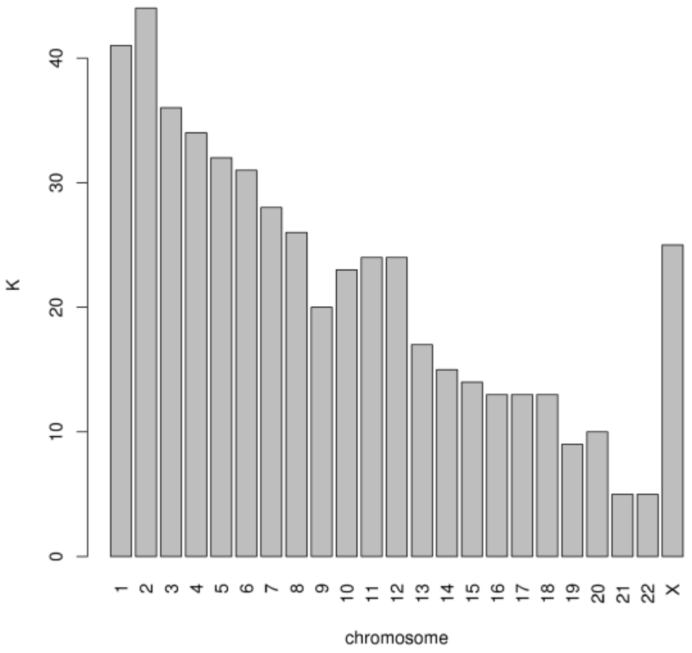

Figure 3. The number of subsets K used for each chromosome.

we also counted how many genomic positions are uniquely mappable (35 bp short sequence). Assume that the number of reads in the ith bin follow a Poisson distribution with parameter λ_i. We consider the following model

$$\log(\lambda_i) = \beta_0 + \beta_1 \log(G) + \beta_2 \log(C) + \beta_3 \log(A) + \beta_4 \log(M),$$

where G, C, A are the G, C, and A count in the ith bin, and M is the proportion of the uniquely mappable positions. To avoid taking logarithm of zero, we added a small number on G, C and A count (0.1) and the mappability (0.0001). Then, for each chromosome (chromosome $1, \ldots, 22$ and X), we compared the MLE $\hat{\boldsymbol{\beta}}$ with its corresponding AEE estimate $\tilde{\boldsymbol{\beta}}$ of the Poisson regression model. To calculate the AEE estimate, for each chromosome, we partitioned the data set into K subsets such that each subset had 5,000 data points (maybe except one subset). Figure 3 shows the number of subsets K used for each chromosome. Then, for each chromosome, we calculated the relative bias $\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\|/\|\hat{\boldsymbol{\beta}}\|$ (Figure 4). From Figure 4, we see that the MLE and its corresponding AEE estimates are very close, showing that our aggregation performs well in this data set.

7. APPLICATIONS: DATA CUBES AND DATA STREAMS

In this section, we discuss applications of the AEE estimator in two massive data environments: data cubes and data streams. Analysis in both environments require performing the same analysis for different subsets while the raw data often can not be saved permanently. Efficient compression

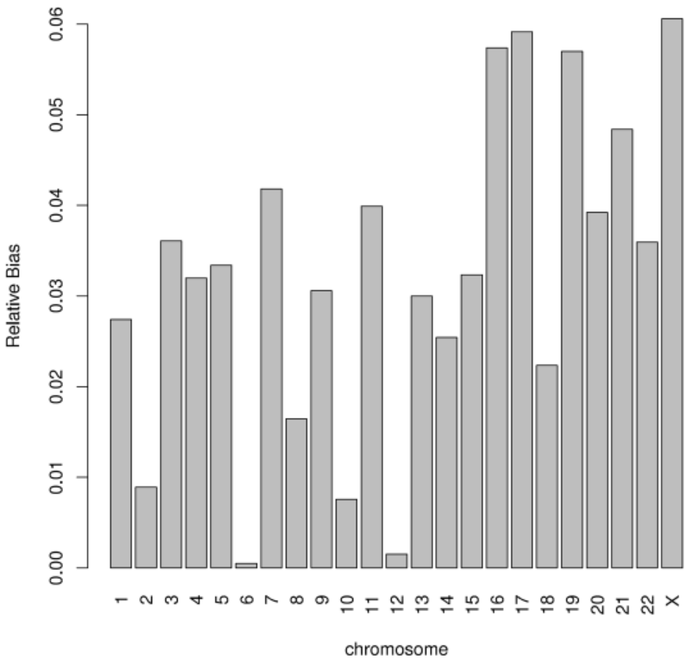

Figure 4. Relative Bias of the AEE estimate for the sequencing data.

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.

7.1 Application to data cubes

Data cubes [12] are multidimensional extensions of two-dimensional tables, which are common in massive business transaction data sets. As a data warehouse tool, it has its advantage over relational databases when fast OLAP is desired. Very often, some dimensions can take values at multiple levels that form a hierarchical structure. For example, a business's transaction location can be defined at any of the three levels: country, state or city, and its time at any of the four levels: year, month, week or day. Hence, a subset of the entire data set is determined when values of these dimensions are given at a certain level, and we call it a cube. The lowest level cubes in the hierarchical structure are called base cells. For example, data records in a particular city at a particular day in the aforementioned example form a base cell. Then any cube can be obtained from aggregating base cells. Business analysts are often interested in performing the same analysis in different cubes. Even if these dimension attributes have only a moderate number of distinct values at their lowest level, the total number of all cubes is enormous. If the analysis needs to access the raw data in a cube everytime, a huge amount of time is needed as accessing subsets in a massive data set is time-consuming, and oftentimes the raw data are too massive to store permanently.

Using the AEE method, we first compress the raw data in each base cell into the EE estimate $\hat{\boldsymbol{\beta}}_{nk}$ and A_k in (3). This only requires scanning the raw data once and then we can discard the raw data. And the EE estimate in any cube can be approximated by computing the AEE estimate using the aggregation in (4). This aggregation is very fast since only simple operations are needed. Consequently, fast computation and efficient storage are both achieved when EE estimation is needed for many different cubes.

7.2 Application to data streams

Data streams are data records coming rapidly along time. Examples include phone records in large call centers, web search activities, and network traffic. Formally, a data stream is a sequence of data items $z_1, \ldots, z_t, \ldots, z_N$ such that the items are read once in increasing order of the indices t [16]. In reality, enormous amounts of data accumulate quickly, which makes permanent storage of the raw data impossible. 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.

While analyses such as clustering [13, 3] and classification [24] for data streams have been extensively studied, parametric estimation such as EE estimation is still an untouched area. The AEE method provides a natural solution to EE estimation for data streams. We 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. (2003) [2] 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 $k = 1, \ldots, K$. At each time point m_k, we calculate and store the EE estimate $\hat{\boldsymbol{\beta}}_{nk}$ and A_k based on data items $z_{m_{k-1}}, \ldots, z_{m_k}$ in the time interval $[m_{k-1}, m_k]$. According to the property of the pyramidal time frame in [2], we can obtain a good approximation to the EE estimate in any time interval by computing the AEE estimator using (4).

7.3 Simulation studies

We again consider 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, we 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 $1,000 = 50 \times 20$ base cells in total. In reality, this data set can be business transaction records in 50 months for 20 cities. We suppose that there are 500 records for each city in each month. We consider the situation where a business analyst is interested in computing the MLE in 100 different

Table 2. Comparison of computational time

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

cubes. We simulate each of these 100 cubes by first randomly selecting D from $\{1, \ldots, 1,000\}$ as the number of base cells contained in a cube, and then randomly choosing D base cells from the 1,000 base cells.

We compare the computation time of the AEE estimates with that of computing the EE estimates directly from the raw data. Table 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.

7.4 Change detection

The purpose to perform the same analysis for different cubes or for a data stream in different time intervals is often to detect whether any change occurs [18, 10]. When changes are detected between two cubes or two time intervals, it becomes inappropriate to aggregate further as aggregating inhomogeneous groups may lead to misleading conclusions such as Simpson's Paradox. The AEE method also provides a way to test the non-homogeneity. Consider AEE estimation in the data cube context. Suppose that each base cell is compressed into the EE estimate $\hat{\boldsymbol{\beta}}_{nk}$ and the weight matrix A_k. Let $\chi = (\hat{\boldsymbol{\beta}}_{k_1} - \hat{\boldsymbol{\beta}}_{k_2})^T (A_{k_1}^{-1} + A_{k_2}^{-1})^{-1} (\hat{\boldsymbol{\beta}}_{k_1} - \hat{\boldsymbol{\beta}}_{k_2})$. Then from Theorem 5.2, if the data in the k_1th cell and the k_2th cell are homogenous, the statistic χ is asymptotically χ_p^2 distributed, where p is the dimension of $\boldsymbol{\beta}$. Hence we can use χ as a test statistic to test the homogeneity between the k_1th cell and the k_2th cell.

8. CONCLUSIONS AND DISCUSSIONS

We develop the AEE estimator to overcome the memory constraint or storage limit for EE estimation in massive data sets based on first-order approximation to the estimating equation. It accurately approximates the original EE estimator with significant time reduction, especially in its applications to data cubes and data streams. The AEE method compresses the raw data nearly losslessly as our asymptotic theory shows the asymptotic equivalence between the original EE estimator and the AEE estimator under mild conditions. This efficient compression avoids accessing the raw data everytime when the EE estimate needs to be computed for different subsets, and provides remarkable value for the AEE estimator to be used in data cubes and data stream. Our results also show that the AEE estimator outperforms the common practice of computing the EE estimator based on a random sample of the entire data set.

Closely related work to our AEE method includes data squashing (DS) [11] and its extension, likelihood-based DS (LDS) [20]. Both methods compress the raw data lossly into a much smaller set of "squashed" values with proper weights attached to achieve efficient storage. The weights are taken so that the weighted moments or the weighted likelihood of the squashed data equate (or approximate) those of the raw data. Despite the success of DS and LDS shown by the examples in [11, 20], no general theory guarantees EE estimators based on the squashed data always accurately approximating the original EE estimators. In addition, while the set of squashed values is very important to the performance of DS and LDS, choosing squashed values properly is difficult as the optimal set depends on the shape of the likelihood function, which is usually unavailable when an EE estimator is needed.

ACKNOWLEDGMENTS

The work was partly supported by grant NSF-DMS0906023.

APPENDIX A. PROOFS

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

By the definition of the above matrix norm, it is easy to prove the following two facts.

Fact A.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* $C \leq \lambda$*.*

Fact A.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* $\|\mathbf{A}^{-1}\| \leq c^{-1}$*.*

In the following, we will give the proofs for theorems in Sections 4 and 5.

Proof of Theorem 1. From Conditions (C2) and (C5), we know that matrix $\mathbf{A}_k$ is positive definite for each $k = 1, \ldots, K$ when n is sufficiently large. Hence, $\sum_{k=1}^K \mathbf{A}_k$ is a positive definite matrix. In particular, $(\sum_{k=1}^K \mathbf{A}_k)^{-1}$ exists and Equation (4) is valid. Subtracting $\boldsymbol{\beta}_0$ from both sides of (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,

$$(7)\qquad \|\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0\| \le \sum_{k=1}^{K} \left\| \left(\sum_{k=1}^{K} \mathbf{A}_k \right)^{-1} \mathbf{A}_k (\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0) \right\| \le \sum_{k=1}^{K} \|\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0\|.$$

The second inequality comes from the fact $\|(\sum_{k=1}^K \mathbf{A}_k)^{-1}\mathbf{A}_k\| \le 1$. Hence the first part of Theorem 4.1 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 $\mathbf{v}^T \frac{1}{n}\mathbf{A}_k \mathbf{v} \ge \mathbf{v}^T \boldsymbol{\Lambda}_1 \mathbf{v} \ge \lambda_1 \|\mathbf{v}\|^2$. Hence, $\mathbf{v}^T \frac{1}{nK}\sum_{k=1}^K \mathbf{A}_k \mathbf{v} \ge \lambda_1\|\mathbf{v}\|^2$. Then from Facts A.1 and A.2, we have $\|(\frac{1}{nK}\sum_{i=1}^K \mathbf{A}_k)^{-1}\| \le \lambda_1{}^{-1}$. Then since $\|n^{-1}\mathbf{A}_k\| \le \|\boldsymbol{\Lambda}_2\| \le \lambda_2$, it follows that

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

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

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

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

$$(8)\qquad \mathbf{M}_k(\hat{\boldsymbol{\beta}}_N) = \mathbf{M}_k(\hat{\boldsymbol{\beta}}_{nk}) + \mathbf{A}_k(\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_{nk}) + \mathbf{R}_{nk},$$

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 $\|\mathbf{R}_{nk}\| \le Cn\|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_{nk}\|^2 \le 2Cn(\|\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0\|^2 + \|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\|^2)$ 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 (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 $\hat{\boldsymbol{\beta}}_N - \tilde{\boldsymbol{\beta}}_{NK} = (\sum_{k=1}^K \mathbf{A}_k)^{-1}\sum_{k=1}^K \mathbf{R}_{nk}$. Then similar to the first part of the proof, we get $\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N\| \le C_1(\|\boldsymbol{\beta}_{nk_0} - \boldsymbol{\beta}_0\|^2 + \|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\|^2)$ for some constant C_1. □

Proof of Theorem 4.2. Suppose that all the random variables are defined on a probability space $(\Omega, \mathcal{F}, P)$. Let $\Omega_{n,k,\eta} = \{\omega \in \Omega :\ n^\alpha\|\hat{\boldsymbol{\beta}}_{nk} - \boldsymbol{\beta}_0\| \le \eta\}$, $\Omega_{N,\eta} = \{\omega \in \Omega :\ N^\alpha\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\| \le \eta\}$ and $\Gamma_{N,K,\eta} = \cap_{k=1}^K \Omega_{n,k,\eta} \cap \Omega_{N,\eta}$. From Condition (C6), for any $\eta > 0$, we have

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

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 4.1. For all $\omega \in \Gamma_{N,K,\eta}$, we have $\boldsymbol{\beta}_k^* \in B_\eta(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} \in \mathbb{R}^p :\ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \eta\}$ 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 $\|\mathbf{R}_{nk}\| \le C_2 pn\|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_{nk}\|^2$ 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,

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

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 $C\eta_\delta^2 < \delta$. 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 $\sqrt{N}\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N\| \le \sqrt{N}n^{-2\alpha}\delta = O(n^{(1+\gamma-4\alpha)/2})\delta$. Therefore, when n is large enough, we have $\Gamma_{N,K,\eta_\delta} \subset \{\omega \in \Omega :\ \sqrt{N}\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N\| \le \delta\}$ and hence, $P(\sqrt{N}\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N\| > \delta) \le P(\Gamma_{N,K,\eta_\delta}^c) \to 0$ as $n \to \infty$. □

To prove Theorem 5.1, we need the following two lemmas. The proof of Lemma A.2 could be found in [5].

Lemma A.1. *Suppose that A, B are two $p \times p$ positive definite matrices. Then*

(1) *$A \ge B$ if and only if $A^{-1} \le B^{-1}$*
(2) *If we have $AB = BA$, then $A \ge B$ implies $A^2 \ge B^2$.*

Lemma A.2. *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 of Theorem 5.1. 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,

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

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 A.1, $(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T)^{-2} \ge M^{-2}N^{-2}I_p$. 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

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

i.e. $\|G_N(\boldsymbol{\beta})\| \ge m_\eta C\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|/M$ for $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \eta$. In particular, $\|G_N(\boldsymbol{\beta})\| \ge m_\eta C\eta/M$ 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 A.2, if $\|\mathbf{a}_N\| \le m_\eta C\eta/M$, there exists an $\hat{\boldsymbol{\beta}}_N \in \mathbb{R}^p$, $\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\| \le \eta$, such that $G_N(\hat{\boldsymbol{\beta}}_N) = \mathbf{a}_N$.

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

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

Furthermore,

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

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

$$\begin{aligned}&tr\left[\left(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\left(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T\sigma_i^2\right)\left(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\right] \\ &\quad \le tr\left[M\left(\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\right] \le pM(CN)^{-1}.\end{aligned}$$

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

For $\omega \in W_{N,\eta}$, $\|\mathbf{a}_N\| \le m_\eta C\eta/M$. By Lemma A.2, there exists an $\hat{\boldsymbol{\beta}}_N \in \mathbb{R}^p$, $\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\| \le \eta$, such that $G_N(\hat{\boldsymbol{\beta}}_N) = \mathbf{a}_N$. Furthermore, for $\omega \in W_{N,\eta}$ we have $N^\alpha\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\| \le N^\alpha(\frac{m_\eta C}{M})^{-1}\|\mathbf{a}_N\| \le \eta$. Hence,

$$W_{N,\eta} \subseteq \Omega_{N,\eta} = \{\omega \in \Omega : N^\alpha\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\| \le \eta\}.$$

At last we get

$$\begin{aligned}P(N^\alpha\|\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0\| > \eta) &= P(\Omega_{N,\eta}^c) \\ &\le P(W_{N,\eta}^c) \le pM^3C^{-3}(m_\eta\eta)^{-2}N^{2\alpha-1}.\end{aligned}$$

□

Proof of Theorem 5.2. We first prove

$$\text{(9)} \qquad \boldsymbol{\Sigma}_N^{-1/2}\mathbf{M}(\boldsymbol{\beta}_0) = \boldsymbol{\Sigma}_N^{-1/2}\sum_{i=1}^N \mathbf{x}_i[y_i - \mu(\mathbf{x}_i^T\boldsymbol{\beta}_0)] \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_p).$$

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 $\mathrm{E}(\xi_{ni}) = 0$, $i = 1, \dots, N$, and $\mathrm{Var}(\xi_N) = 1$. From the Cramér-Wold theorem and the Linderberg central limit theorem, to prove (9), we only need to prove that, for any $\epsilon > 0$, $g_N(\epsilon) := \sum_{i=1}^N \mathrm{E}(|\xi_{Ni}|^2 I(|\xi_{Ni}| > \epsilon)) \to 0$ as $N \to \infty$. 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 $\boldsymbol{\Sigma}_N > c_1^2\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T$, i.e. $\boldsymbol{\Sigma}_N - c_1^2\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T$ is a positive definite matrix, and hence,

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

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

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

Therefore, we have proved (9). It is easy to check that all the conditions in Corollary 2.2 in [15] are satisfied here, the QLE $\hat{\boldsymbol{\beta}}_N$ has the following Badahur representation $\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0 = -\mathbf{D}_N^{-1}(\boldsymbol{\beta}_0)\sum_{i=1}^N \mathbf{x}_i[y_i - \mu(\mathbf{x}_i^T\boldsymbol{\beta}_0)] + O(N^{-3/4}(\log N)^3)$ 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 $\boldsymbol{\Sigma}_N^{-1/2} = O(N^{-1/2})$ and $\mathbf{D}_N(\boldsymbol{\beta}_0) = O(N)$, we get

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

For the AQLE, we have

$$\begin{aligned}
&-\boldsymbol{\Sigma}_N^{-1/2}\mathbf{D}_N(\boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0) \\
&\quad= -\boldsymbol{\Sigma}_N^{-1/2}\mathbf{D}_N(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0 + \tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N).
\end{aligned}$$

Since $\|-\boldsymbol{\Sigma}_N^{-1/2}\mathbf{D}_N(\boldsymbol{\beta}_0)\| = O(N^{-1/2})$, Theorem 4.2 and Theorem 5.1 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 $\gamma < \min\{1 - 2\alpha, 4\alpha - 1\}$. □

Received 2 March 2010

REFERENCES

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

[2] Aggarwal, C. C., Han, J., Wang, J., and Yu, P. S. (2003). A framework for clustering evolving data streams. In *Proceedings of the 29th international conference on very large data bases.* 81–92.

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

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

[5] Chen, K., Hu, I., and Ying, Z. (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics 27*, 1155–1163. MR1740117

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

[7] Chen, Y., Dong, G., Han, J., Wah, B. W., and Wang, J. (2002). Multi-dimensional regression analysis of time-series data streams. *Proceedings of the International Conference on Very Large Data Bases*, 323–334.

[8] Chiang, D., Getz, G., Jaffe, D. Kelly, M., Zhao, X., Carter, S., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E. (2002). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods*, *6*, 99–103

[9] Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. In *Nucleic Acids Research 6*:e105

[10] Dong, G., Han, J., Lam, J., Pei, J., and Wang, K. (2001). Mining multi-dimensional constrained gradient in data cubes. In *Proceedings of the International Conference on Very Large Data Bases.* 321–330.

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

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

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

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

[15] He, X. and Shao, Q. (1996). A general bahadur representation of m-estimators and it's applications to linear regression with non-stochastic designs. *The Annals of Statistics 24*, 2608–2630. MR1425971

[16] Henzinger, M. R., Raghavan, P., and Rajagopalan, S. (1998). Computing on data streams. Tech. Rep. 1998-011, Digital Equipment Corporation, Systems Research Center. May. MR1730706

[17] Huber, P. J. (1981). *Robust Statistics.* Wiley, New Jersey. MR0606374

[18] Imielinski, T., Khachiyan, L., and Abdulghani, A. (2002). Cubergrades: generalizing association rules. *Data Mining and Knowledge Discovery 6*, 219–257. MR1917925

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

[20] Madigan, D., Raghavan, N., DuMouchel, W., Nason, M., Posse, C., and Ridgeway, G. (2002). Likelihood-based data squashing: A modeling approach to instance construction. *Data Mining and Knowledge Discovery 6*, 173–190. MR1917936

[21] Munro, J. I. and Paterson, M. S. (1980). Selection and sorting with limited storage. *Theoretical Computer Science 12*, 315–323. MR0589312

[22] Pang, S., Ozawa, S., and Kasabov, N. (2005). Incremental linear discriminant analysis for classification of data streams. In *IEEE Transactions on Systems, Man, and Cybernetics, Part B 35(5)*, 905–14.

[23] Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. and Gerstein, M. B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. In *Nature Biotechnology 27*, 66–75.

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

[25] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generlized linear models, and the Gaussian-Newton method. *Biometrika 61*, 439–447. MR0375592

[26] Xi, R., Lin, N., and Chen, Y. (2009). Compression and aggregation for logistic regression analysis in data cubes. *IEEE Transactions on Knowledge and Data Engineering 21(4)*, 479–492.

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

Nan Lin
Department of Mathematics
Washington University in St. Louis
USA
E-mail address: nlin@math.wustl.edu

Ruibin Xi
Center for Biomedical Informatics
Harvard Medical School
USA
E-mail address: Ruibin_Xi@hms.harvard.edu