Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Real-Time Regression Analysis of Streaming Clustered Data With Possible Abnormal Data Batches

Lan Luo, Ling Zhou & Peter X.-K. Song

To cite this article: Lan Luo, Ling Zhou & Peter X.-K. Song (2022): Real-Time Regression Analysis of Streaming Clustered Data With Possible Abnormal Data Batches, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2026778

To link to this article: https://doi.org/10.1080/01621459.2022.2026778

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–16: Theory and Methods
https://doi.org/10.1080/01621459.2022.2026778

Real-Time Regression Analysis of Streaming Clustered Data With Possible Abnormal Data Batches

Lan Luo[a], Ling Zhou[b], and Peter X.-K. Song[c]

[a]Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA; [b]Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, China; [c]Department of Biostatistics, University of Michigan, Ann Arbor, MI

ABSTRACT

This article develops an incremental learning algorithm based on quadratic inference function (QIF) to analyze streaming datasets with correlated outcomes such as longitudinal data and clustered data. We propose a renewable QIF (RenewQIF) method within a paradigm of renewable estimation and incremental inference, in which parameter estimates are recursively renewed with current data and summary statistics of historical data, but with no use of any historical subject-level raw data. We compare our renewable estimation method with both offline QIF and offline generalized estimating equations (GEE) approach that process the entire cumulative subject-level data all together, and show theoretically and numerically that our renewable procedure enjoys statistical and computational efficiency. We also propose an approach to diagnose the homogeneity assumption of regression coefficients via a sequential goodness-of-fit test as a screening procedure on occurrences of abnormal data batches. We implement the proposed methodology by expanding existing Spark's Lambda architecture for the operation of statistical inference and data quality diagnosis. We illustrate the proposed methodology by extensive simulation studies and an analysis of streaming car crash datasets from the National Automotive Sampling System-Crashworthiness Data System (NASS CDS). Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2019
Accepted January 2022

KEYWORDS
Abnormal data batch detection; Generalized estimating equation; Incremental statistical analysis; Online learning; Quadratic inference function; Spark computing platform

1. Introduction

When a car accident happens, driver and passengers in the same car would be all likely to get injured, and their degrees of injury are correlated within a car. Here a car is a sample unit which, in general, is referred to as a cluster. The National Automotive Sampling System-Crashworthiness Data System (NASS CDS) is a publicly accessible source of streaming datasets containing car accident information in the USA. Other examples of such streaming correlated data include cohorts of patients sequentially assembled at different clinical centers to periodically update national disease registry databases, where, for example, a family is the sample unit. In this article, we consider a problem where a series of independent clusters becomes available sequentially over data batches, and arrivals of data batches may be perpetual. Similar to the data of car accidents, each data batch consists of temporally correlated or cluster-correlated outcomes. The primary goal of processing such streaming data is to sequentially update some statistics of interest upon the arrival of a new data batch, in the hope to not only free up space for the storage of massive historical individual-level data, but also to provide real-time inference and decision making.

With the emergence of streaming data collection techniques, sequential data analytics have received much attention in the literature to address computational efficiency while preserving essential statistical properties. Arguably, the stochastic gradient descent (SGD) algorithm has been thus far the most well-known algorithm to analyze streaming data along the lines of stochastic approximations (Robbins and Monro 1951). Unfortunately, most of currently available online learning methods in the SGD paradigm and its variants, including online Newton SGD (Hazan, Agarwal, and Kale 2007) and quasi-Newton SGD (Schraudolph et al. 2007; Bordes, Bottou, and Gallinari 2009), have focused only on point estimation or prediction, and unfortunately precluded statistical inference. Toulis and Airold (2017) obtained some analytic expressions for the asymptotic variances in the implementation of an SGD algorithm to produce maximum likelihood estimation (MLE) with cross-sectional data, in which statistical inference was absent. Recently, Fang (2019) proposed a perturbation-based resampling method to construct confidence intervals in the framework of the averaged implicit SGD (AI-SGD) estimation proposed by Toulis and Airold (2017). Luo and Song (2020) demonstrated in simulation studies that Fang's resampling method for the AI-SGD may fail to provide desirable coverage probability, and thus, possibly leads to misleading inference, especially when the number of regression parameters is large. In the setting of linear models, since the least square estimation (LSE) of the regression parameters has a closed form expression for recursive calculations, it is possible to establish certain sequential updating schemes for both point estimation and

CONTACT Peter X.-K. Song *pxsong@umich.edu* Department of Biostatistics, University of Michigan, Ann Arbor, MI.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

related calculation of standard errors; see for example, Luo and Song (2020). In this case the resulting estimates can be exactly the same as those obtained by its offline counterpart (Stengel 1994). This property has been used to conduct real-time regression analysis with the linear model for streaming data by, for example, Nadungodage et al. (2011). However, according to Luo and Song (2020), this property of equality between online and offline estimates no longer holds beyond the linear model, for example, the logistic model for binary outcomes.

This article considers an important extension of renewable estimation and incremental inference in the class of generalized linear models (GLMs) developed by Luo and Song (2020) for cross-sectional data to the case of streaming data with repeatedly measured responses. This extension is substantial in both methodology and applications. In terms of methodological extensions, it relaxes not only the availability of likelihood functions in the renewable MLE to a general framework of estimating functions, but also the popular assumption of homogeneous marginal models. Model homogeneity refers to the situation where data streams are generated under a common set of parameters over the sequence of data batches. This homogeneity will be violated in the presence of "abnormal" data batches – those being generated under different sets of model parameters from the standard/default ones of primary interest. In real world applications, practitioners often encounter outlying data batches. In this case, continually updating results without noticing and removing abnormal data batches would lead to invalid statistical inference and misleading conclusions. To deal with this issue of practical importance, we develop a quality control (QC) type of monitoring scheme by the means of abnormal data batch detection and deletion.

We propose a new online regression methodology along the lines of the generalized estimating equation (GEE) approach proposed by Liang and Zeger (1986), one of the most widely used methods for the analysis of data with correlated outcomes. This quasi-likelihood approach is based only on the first two moments of the correlated data distribution with no need of specifying a parametric joint distribution. Such regression model is termed as marginal generalized linear model (MGLM) or population-average model in the literature of correlated data analysis (Song 2007, chap. 5). In this field, another quasi-likelihood inference method is quadratic inference function (QIF) (Qu, Lindsay, and Li 2000). QIF has several advantages in comparison to GEE: (i) QIF does not require more model assumptions than GEE; (ii) it provides a goodness-of-fit test for the first moment assumption, that is, the mean-model specification; (iii) QIF estimator is more efficient than the GEE estimator when the working correlation is misspecified; and (iv) it is more robust with a bounded influence function against large outliers (Qu and Song 2004).

Our key methodology contributions include: (i) we propose an online QIF method that allows to perform real-time regression analysis of correlated outcomes under fast recursive estimating equations with little reliance on data storage capacity; (ii) the proposed online estimation and inference, termed as RenewQIF, is asymptotically equivalent to the offline QIF estimator obtained from the full cumulative data, and thus, has no loss of statistical power in inference; (iii) our RenewQIF method can be implemented in the existing Spark's Lambda architecture to carry over incremental estimation and inference with desirable statistical and computational efficiency; and (iv) by adding a monitoring layer to the Lambda architecture, our method allows to detect and delete abnormal data batches in the real-time analysis of correlated data. A direct use of the offline QIF with cumulative data encounters fast-growing demands on hardward capacities over the course of perpetual data streams.

It is worth pointing out that the proposed RenewQIF is indeed different from the offline QIF that works for a single data batch. First, the RenewQIF is built upon a sequential updating paradigm, which is computationally much faster than the offline QIF. This gain in computational speed stems from to our new formulation of online search algorithm that is operated recursively under a different objective function from that of the original offline QIF. As shown in this article, the RenewQIF and offline QIF estimators are not the same but only stochastically equivalent. Second, unlike most of existing online methods, the proposed RenewQIF provides online statistical inference that, technically, depends on a fast recursive calculation of Godambe information matrix (or the sandwich covariance matrix). Such work has not been considered in the current literature. Third, theoretical justifications for large sample properties of the RenewQIF are different in the case where individual data batch sizes are fixed but the number of data batches tends to infinity. In contrast, the asymptotic results of the offline QIF do not hold if the sample size is fixed, which corresponds to the case that the number of data batches is only one rather than diverging to infinity. Thus, the technical treatments in both theoretical arguments and algorithm designs in the RenewQIF are more advanced than those given by Qu, Lindsay, and Li (2000). Technically, the offline QIF may be regarded as a special case of the RenewQIF.

We also propose an online screening method for a real-time diagnosis of abnormal data batches in the framework of RenewQIF. Most existing online monitoring procedures are based on certain metrics such as a test statistic (Page 1954; Shiryayev 1963; Roberts 1966; Goel and Wu 1971; Amin and Search 1991; Pollak and Siegmund 1991; Amin, Reynolds, and Bakir 1995). In a similar spirit to (Lai 2004), we establish a fast screening procedure based on Hansen's goodness-of-fit test statistic (Hansen 1982) to identify any abnormal data batch and exclude it from updating results of estimation and inference. To implement the RenewQIF in the presence of potential abnormal data batches, we expand the Rho architecture developed by (Luo and Song 2020) in the context of GLMs for cross-sectional data, by adding a new monitoring layer in the Spark's Lambda architecture. This monitoring layer houses a QIF-based testing procedure to check the compatibility of each arriving data batch with the normal data reference.

Essentially, we aim to develop a new online methodology with the following tasks: (i) to put forward RenewQIF estimation and incremental inference in MGLMs for correlated outcomes; and (ii) to study a QIF-based goodness-of-fit test statistic in the monitoring layer that enables to effectively detect abnormal data batches over data streams with no fixed ending point. This article is organized as follows to achieve these two aims. Section 2 provides both algorithms and theoretical guarantees for our RenewQIF method. Section 3 discusses an extended Lambda architecture with an addition of quality control layer

and pseudo code for numerical implementation, together with an analysis on algorithmic convergence and a monitoring procedure for abnormal data batches. Section 4 includes simulation results with comparisons of the proposed RenewQIF to the offline GEE, QIF and renewable GEE (RenewGEE) with or without abnormal data batches. Section 5 illustrates the proposed method by a real data analysis application. The proofs of the large-sample properties for the RenewQIF method are included in the Appendix. Derivation of RenewGEE as well as additional numerical results are included in the supplementary materials.

2. RenewQIF Methodology

2.1. Offline QIF

Consider independent streaming data batches consisting of cluster-correlated outcomes, sequentially generated from a common underlying population-averaged model (Zeger, Liang, and Albert 1988) or marginal generalized linear model (MGLM) (Song 2007, chap. 5) with an unknown regression parameter $\boldsymbol{\beta}_0 \in \Theta \subset \mathbb{R}^p$ where Θ is the parameter space for $\boldsymbol{\beta}$. For the ease of exposition, we assume an equal cluster size $m_i = m$. Our goal is to evaluate population-average effects of p covariates, denoted by $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^\top$ in an MGLM with the marginal mean and covariance given by

$$\begin{aligned} \boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y}|\boldsymbol{X}) &= \left[h(\boldsymbol{x}_1^\top \boldsymbol{\beta}_0), \ldots, h(\boldsymbol{x}_m^\top \boldsymbol{\beta}_0)\right]^\top, \\ \operatorname{cov}(\boldsymbol{y}|\boldsymbol{X}) &= \phi \boldsymbol{\Sigma}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}) = \phi \boldsymbol{A}^{1/2} \boldsymbol{R}(\boldsymbol{\alpha}) \boldsymbol{A}^{1/2}, \end{aligned} \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_m)^\top$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$ with $\mu_k = h(\boldsymbol{x}_k^\top \boldsymbol{\beta}_0)$ where $h(\cdot)$ is a known link function, and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)^\top$ with $\boldsymbol{x}_k = (x_{k1}, \ldots, x_{kp})^\top$, $k = 1, \ldots, m$. ϕ is a dispersion parameter, $\boldsymbol{A} = \operatorname{diag}\{v(\mu_1), \ldots, v(\mu_m)\}$ is a diagonal matrix with $v(\cdot)$ being a known variance function, and $\boldsymbol{R}(\boldsymbol{\alpha})$ is a working correlation matrix that is fully characterized by a correlation parameter vector $\boldsymbol{\alpha}$. Both dispersion ϕ and correlation $\boldsymbol{\alpha}$ are treated as nuisance parameters outside the space of parameters of interest, Θ.

In the context of streaming data, consider a time point $b \geq 2$ with a total of N_b clusters arriving sequentially in b data batches, $\{\mathcal{D}_1, \ldots, \mathcal{D}_b\}$, each containing $n_j = |\mathcal{D}_j|, j = 1, \ldots, b$, clusters. Let $\mathcal{D}_b^\star = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_b$ denote the cumulative collection of datasets up to data batch b where each sample unit corresponds to an m-element vector of cluster-correlated outcomes, and the cumulative sample size is $N_b = |\mathcal{D}_b^\star|$. For simplicity, $\mathcal{D}_b$ (a single data batch b) or $\mathcal{D}_b^\star$ (an aggregation of b data batches) is also used as respective sets of indices for clusters involved. For cluster i, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})^\top$ and $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{im})^\top$ be the correlated response vectors and associated covariates, $i = 1, \ldots, n_j, j = 1, \ldots, b$. According to Liang and Zeger (1986), an offline GEE estimator of $\boldsymbol{\beta}_0$ is a solution to the following generalized estimating equation for the cumulative data $\mathcal{D}_b^\star$ up to time point b:

$$\boldsymbol{\psi}_b^\star(\mathcal{D}_b^\star; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i \in \mathcal{D}_b^\star} \boldsymbol{D}_i^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}, \tag{2}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im})^\top, \boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}^\top$ is an $m \times p$ matrix and $\boldsymbol{\Sigma}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2}$ with $\boldsymbol{A}_i = \operatorname{diag}\{v(\mu_{i1}), \ldots, v(\mu_{im})\}$.

According to Qu, Lindsay, and Li (2000), the formulation of an offline QIF is based on an approximation to the inverse working correlation matrix by $\boldsymbol{R}^{-1}(\boldsymbol{\alpha}) \approx \sum_{s=1}^S \gamma_s \boldsymbol{M}_s$, where $\gamma_1, \ldots, \gamma_S$ are constants possibly dependent on $\boldsymbol{\alpha}$, and $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_S$ are known basis matrices with elements 0 and 1, which are determined by a given correlation matrix $\boldsymbol{R}(\boldsymbol{\alpha})$. In some cases, the above expansion can be exact. For example, as discussed in Qu, Lindsay, and Li (2000) and (Song 2007, chap. 5), the basis matrices for the compound symmetry working correlation matrix are $\boldsymbol{M}_1 = \boldsymbol{I}_m$, the identity matrix, and $\boldsymbol{M}_2^{cs}$, a matrix with all 0 on the diagonal and all 1 off the diagonal. Plugging such expansion into (2) leads to $\boldsymbol{\psi}_b^\star(\mathcal{D}_b^\star; \boldsymbol{\beta}, \alpha) = \sum_{i \in \mathcal{D}_b^\star} \sum_{s=1}^S \gamma_s \boldsymbol{D}_i^\top \boldsymbol{A}_i^{-1/2} \boldsymbol{M}_s \boldsymbol{A}_i^{-1/2} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}$, which may be regarded as a combination of the following extended score vector of pS dimension:

$$\begin{aligned} \boldsymbol{g}_b^\star(\boldsymbol{\beta}) &= \sum_{i \in \mathcal{D}_b^\star} \boldsymbol{g}(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{\beta}) \\ &= \sum_{i \in \mathcal{D}_b^\star} \begin{pmatrix} \boldsymbol{D}_i^\top \boldsymbol{A}_i^{-1/2} \boldsymbol{M}_1 \boldsymbol{A}_i^{-1/2} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ \boldsymbol{D}_i^\top \boldsymbol{A}_i^{-1/2} \boldsymbol{M}_S \boldsymbol{A}_i^{-1/2} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \end{pmatrix}. \end{aligned}$$

This is an over-identified estimating function, namely $\dim(\boldsymbol{g}_b^\star(\boldsymbol{\beta})) > \dim(\boldsymbol{\beta})$. To obtain an estimator of $\boldsymbol{\beta}_0$, following Hansen (1982)'s generalized method of moments (GMM), we take $\hat{\boldsymbol{\beta}}_b^\star = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} Q_b^\star(\boldsymbol{\beta})$ with

$$Q_b^\star(\boldsymbol{\beta}) = \boldsymbol{g}_b^\star(\boldsymbol{\beta})^\top \left\{\boldsymbol{C}_b^\star(\boldsymbol{\beta})\right\}^{-1} \boldsymbol{g}_b^\star(\boldsymbol{\beta}), \tag{3}$$

where $\boldsymbol{C}_b^\star(\boldsymbol{\beta}) = \sum_{i \in \mathcal{D}_b^\star} \boldsymbol{g}(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{\beta}) \boldsymbol{g}(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{\beta})^\top$ is the sample variance of $\boldsymbol{g}_b^\star(\boldsymbol{\beta})$. Note that the nuisance correlation parameter $\boldsymbol{\alpha}$ is not involved in (3) for the estimation of $\boldsymbol{\beta}_0$. Throughout the article, for a vector $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$, we denote the L_2-norm by $||\boldsymbol{a}|| := \sqrt{\sum_{i=1}^n a_i^2}$; for a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, we define $||\boldsymbol{A}|| := \sqrt{\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})}$ where $\lambda_{\max}$ denote the largest eigenvalue. Wherever applicable, denote the L_1-norm by $||\cdot||_1$.

2.2. Online QIF

Instead of processing the cumulative data $\mathcal{D}_b^\star$ once to obtain an offline QIF as shown above, we may conduct the online estimation and inference via a sequential and recursive updating scheme. In the proposed online estimation framework, let $\tilde{\boldsymbol{\beta}}_b$ be a renewable estimator, which is initialized by $\tilde{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} Q_1(\boldsymbol{\beta})$, namely the QIF estimate obtained with the first data batch. When data batch $\mathcal{D}_b$ arrives, a previous estimator $\tilde{\boldsymbol{\beta}}_{b-1}$ is renewed or updated to $\tilde{\boldsymbol{\beta}}_b$ using historical summary statistics of previous data batches $\mathcal{D}_{b-1}^\star$ and full observations of current data batch $\mathcal{D}_b$. This sequentially renewed estimator is termed as the renewable QIF or RenewQIF. After the completion of this updating, individual-level data of $\mathcal{D}_b$ is no longer accessible for the sake of data storage, but only the updated estimate $\tilde{\boldsymbol{\beta}}_b$ and summary statistics are carried forward in future calculations. For the empirical version, we use $\boldsymbol{g}_b(\boldsymbol{\beta}) = \sum_{i \in \mathcal{D}_b} \boldsymbol{g}(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{\beta})$ to denote the extended score vector of data batch $\mathcal{D}_b$, clearly $\boldsymbol{g}_b^\star(\boldsymbol{\beta}) = \sum_{j=1}^b \boldsymbol{g}_j(\boldsymbol{\beta})$.

The negative gradient and sample variance matrix of $\boldsymbol{g}_b(\boldsymbol{\beta})$ are denoted by $\boldsymbol{G}_b(\boldsymbol{\beta}) = -\sum_{i\in\mathcal{D}_b} \partial \boldsymbol{g}(\boldsymbol{y}_i;\boldsymbol{X}_i,\boldsymbol{\beta})/\partial\boldsymbol{\beta}^\top$ and $\boldsymbol{C}_b(\boldsymbol{\beta}) = \sum_{i\in\mathcal{D}_b} \boldsymbol{g}(\boldsymbol{y}_i;\boldsymbol{X}_i,\boldsymbol{\beta})\boldsymbol{g}(\boldsymbol{y}_i;\boldsymbol{X}_i,\boldsymbol{\beta})^\top$, respectively. In the theoretical framework, let the population variability matrix and sensitivity matrix be $\mathbb{C}(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\beta}}\left\{\boldsymbol{g}(\boldsymbol{y};\boldsymbol{X},\boldsymbol{\beta})\boldsymbol{g}^\top(\boldsymbol{y};\boldsymbol{X},\boldsymbol{\beta})\right\}$ and $\mathbb{G}(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\beta}}\left\{-\partial \boldsymbol{g}(\boldsymbol{y};\boldsymbol{X},\boldsymbol{\beta})/\partial\boldsymbol{\beta}^\top\right\}$ (Godambe and Kale 1991). In the process of RenewQIF, the same basis matrices are used in all data batches.

We begin the derivation of RenewQIF with two batches, the second one $\mathcal{D}_2$ arriving after the first $\mathcal{D}_1$. This simple scenario can be easily generalized to the case of an arbitrary number of data batches with little effort. According to Qu, Lindsay, and Li (2000), a QIF estimator, $\hat{\boldsymbol{\beta}}_1 = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, Q_1(\boldsymbol{\beta})$ with $Q_1(\boldsymbol{\beta}) = \boldsymbol{g}_1^\top(\boldsymbol{\beta})\{\boldsymbol{C}_1(\boldsymbol{\beta})\}^{-1}\boldsymbol{g}_1(\boldsymbol{\beta})$, may be obtained from the estimating equation: $\boldsymbol{G}_1^\top(\hat{\boldsymbol{\beta}}_1)\left\{\boldsymbol{C}_1(\hat{\boldsymbol{\beta}}_1)\right\}^{-1}\boldsymbol{g}_1(\hat{\boldsymbol{\beta}}_1) = \boldsymbol{0}$ with a higher-order term of $O_p(n_1^{-1})$ being dropped from the gradient. When $\mathcal{D}_2$ arrives, we aim to obtain the offline QIF estimator, $\hat{\boldsymbol{\beta}}_2^\star$, based on the accumulated data $\mathcal{D}_2^\star$, satisfying the estimating equation $\boldsymbol{G}_2^\star(\hat{\boldsymbol{\beta}}_2^\star)^\top\boldsymbol{C}_2^\star(\hat{\boldsymbol{\beta}}_2^\star)^{-1}\boldsymbol{g}_2^\star(\hat{\boldsymbol{\beta}}_2^\star) = \boldsymbol{0}$, or equivalently

$$\left\{\boldsymbol{G}_1(\hat{\boldsymbol{\beta}}_2^\star) + \boldsymbol{G}_2(\hat{\boldsymbol{\beta}}_2^\star)\right\}^\top \left\{\boldsymbol{C}_1(\hat{\boldsymbol{\beta}}_2^\star) + \boldsymbol{C}_2(\hat{\boldsymbol{\beta}}_2^\star)\right\}^{-1} \left\{\boldsymbol{g}_1(\hat{\boldsymbol{\beta}}_2^\star) + \boldsymbol{g}_2(\hat{\boldsymbol{\beta}}_2^\star)\right\} = \boldsymbol{0}. \tag{4}$$

Clearly, solving (4) for $\hat{\boldsymbol{\beta}}_2^\star$ involves subject-level data from both batches $\mathcal{D}_1$ and $\mathcal{D}_2$ where $\mathcal{D}_1$ may no longer be accessible. Our RenewQIF estimation is able to handle this issue. To proceed, heuristically, we may take the first-order Taylor expansions of the terms $\boldsymbol{g}_1(\hat{\boldsymbol{\beta}}_2^\star)$, $\boldsymbol{G}_1(\hat{\boldsymbol{\beta}}_2^\star)$ and $\boldsymbol{C}_1(\hat{\boldsymbol{\beta}}_2^\star)$ element-wise in (4) around $\hat{\boldsymbol{\beta}}_1$. Given that $\boldsymbol{g}_1$ is continuously differentiable and $\boldsymbol{G}_1$ is Lipschitz continuous in Θ, it follows that $\frac{n_1}{N_2}\boldsymbol{g}_1(\hat{\boldsymbol{\beta}}_2^\star) = \frac{n_1}{N_2}\boldsymbol{g}_1(\hat{\boldsymbol{\beta}}_1) + \frac{n_1}{N_2}\boldsymbol{G}_1(\hat{\boldsymbol{\beta}}_1)(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2^\star) + O_p\left(\frac{n_1}{N_2}||\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2^\star||^2\right)$, $\frac{n_1}{N_2}\boldsymbol{G}_1(\hat{\boldsymbol{\beta}}_2^\star) = \frac{n_1}{N_2}\boldsymbol{G}_1(\hat{\boldsymbol{\beta}}_1) + O_p\left(\frac{n_1}{N_2}||\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2^\star||\right)$, and $\frac{n_1}{N_2}\boldsymbol{C}_1(\hat{\boldsymbol{\beta}}_2^\star) = \frac{n_1}{N_2}\boldsymbol{C}_1(\hat{\boldsymbol{\beta}}_1) + O_p\left(\frac{n_1}{N_2}||\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2^\star||\right)$. The error terms $O_p\left(\frac{n_1}{N_2}||\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2^\star||\right)$ and $O_p\left(\frac{n_1}{N_2}||\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2^\star||^2\right)$ may be asymptotically ignorable if N_2 is large enough. Dropping such error terms, we propose a new QIF estimator $\tilde{\boldsymbol{\beta}}_2$ as a solution to the estimating equation: $\tilde{\boldsymbol{G}}_2(\tilde{\boldsymbol{\beta}}_2)^\top\tilde{\boldsymbol{C}}_2(\tilde{\boldsymbol{\beta}}_2)^{-1}\tilde{\boldsymbol{g}}_2(\tilde{\boldsymbol{\beta}}_2) = \boldsymbol{0}$, or equivalently,

$$\left\{\boldsymbol{G}_1(\hat{\boldsymbol{\beta}}_1) + \boldsymbol{G}_2(\tilde{\boldsymbol{\beta}}_2)\right\}^\top \left\{\boldsymbol{C}_1(\hat{\boldsymbol{\beta}}_1) + \boldsymbol{C}_2(\tilde{\boldsymbol{\beta}}_2)\right\}^{-1} \left\{\boldsymbol{g}_1(\hat{\boldsymbol{\beta}}_1) + \boldsymbol{G}_1(\hat{\boldsymbol{\beta}}_1)(\hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_2) + \boldsymbol{g}_2(\tilde{\boldsymbol{\beta}}_2)\right\} = \boldsymbol{0}, \tag{5}$$

where $\tilde{\boldsymbol{g}}_2$, $\tilde{\boldsymbol{G}}_2$ and $\tilde{\boldsymbol{C}}_2$ are, respectively, the resulting adjusted extended score vector, the online aggregated negative gradient, and the online sample variance matrix, none of which is calculated with the subject-level raw data $\mathcal{D}_1$. Thus, Equation (5) updates the initial $\hat{\boldsymbol{\beta}}_1$ to a new estimate with no use of the raw data in $\mathcal{D}_1$. Thus, $\tilde{\boldsymbol{\beta}}_2$ is called *a RenewQIF estimator* of $\boldsymbol{\beta}_0$, and Equation (5) is termed as *an incremental QIF estimating equation*. Furthermore, it is straightforward to find the RenewQIF estimator $\tilde{\boldsymbol{\beta}}_2$ numerically via the Newton–Raphson algorithm. That is, at the $(r+1)$th iteration,

$$\tilde{\boldsymbol{\beta}}_2^{(r+1)} = \tilde{\boldsymbol{\beta}}_2^{(r)} + \left\{\tilde{\boldsymbol{G}}_2(\tilde{\boldsymbol{\beta}}_2^{(r)})^\top\tilde{\boldsymbol{C}}_2(\tilde{\boldsymbol{\beta}}_2^{(r)})^{-1}\tilde{\boldsymbol{G}}_2(\tilde{\boldsymbol{\beta}}_2^{(r)})\right\}^{-1} \tilde{\boldsymbol{G}}_2(\tilde{\boldsymbol{\beta}}_2^{(r)})^\top\tilde{\boldsymbol{C}}_2(\tilde{\boldsymbol{\beta}}_2^{(r)})^{-1}\tilde{\boldsymbol{g}}_2(\tilde{\boldsymbol{\beta}}_2^{(r)}),$$

where $\tilde{\boldsymbol{G}}_2(\tilde{\boldsymbol{\beta}}_2^{(r)}) = \boldsymbol{G}_1(\hat{\boldsymbol{\beta}}_1) + \boldsymbol{G}_2(\tilde{\boldsymbol{\beta}}_2^{(r)})$ and $\tilde{\boldsymbol{C}}_2(\tilde{\boldsymbol{\beta}}_2^{(r)}) = \boldsymbol{C}_1(\hat{\boldsymbol{\beta}}_1) + \boldsymbol{C}_2(\tilde{\boldsymbol{\beta}}_2^{(r)})$. Here, the gradient is also an approximation that is traditionally used in the offline QIF with higher-order terms being dropped (Qu, Lindsay, and Li 2000). Once again, it is worth pointing out that the above iterations do not require the subject-level data of $\mathcal{D}_1$, but only the historical summary statistics, including estimate $\hat{\boldsymbol{\beta}}_1$, its negative gradient $\boldsymbol{G}_1(\hat{\boldsymbol{\beta}}_1)$ and sample variance matrix $\boldsymbol{C}_1(\hat{\boldsymbol{\beta}}_1)$. In the QIF estimation above, the nuisance correlation parameter $\boldsymbol{\alpha}$ is not involved in the iterations, either.

Extending the above renewable procedure to a general setting of streaming datasets, we now define a renewable QIF estimation of $\boldsymbol{\beta}_0$ as follows. Let $\hat{\boldsymbol{\beta}}_b^\star$ be the offline QIF estimator of $\boldsymbol{\beta}_0$ with the cumulative data $\mathcal{D}_b^\star = \cup_{j=1}^b \mathcal{D}_j$ obtained from the offline QIF estimating equation $\boldsymbol{G}_b^\star(\hat{\boldsymbol{\beta}}_b^\star)^\top\boldsymbol{C}_b^\star(\hat{\boldsymbol{\beta}}_b^\star)^{-1}\boldsymbol{g}_b^\star(\hat{\boldsymbol{\beta}}_b^\star) = \boldsymbol{0}$. A renewable estimator $\tilde{\boldsymbol{\beta}}_b$ of $\boldsymbol{\beta}_0$ is defined as a solution to the incremental QIF estimating equation: $\tilde{\boldsymbol{G}}_b(\tilde{\boldsymbol{\beta}}_b)^\top\tilde{\boldsymbol{C}}_b(\tilde{\boldsymbol{\beta}}_b)^{-1}\tilde{\boldsymbol{g}}_b(\tilde{\boldsymbol{\beta}}_b) = \boldsymbol{0}$, which is equivalent to

$$f_b(\tilde{\boldsymbol{\beta}}_b) = \left\{\sum_{j=1}^{b-1}\boldsymbol{G}_j(\tilde{\boldsymbol{\beta}}_j) + \boldsymbol{G}_b(\tilde{\boldsymbol{\beta}}_b)\right\}^\top \left\{\sum_{j=1}^{b-1}\boldsymbol{C}_j(\tilde{\boldsymbol{\beta}}_j) + \boldsymbol{C}_b(\tilde{\boldsymbol{\beta}}_b)\right\}^{-1} \times \left\{\tilde{\boldsymbol{g}}_{b-1}(\tilde{\boldsymbol{\beta}}_{b-1}) + \sum_{j=1}^{b-1}\boldsymbol{G}_j(\tilde{\boldsymbol{\beta}}_j)(\tilde{\boldsymbol{\beta}}_{b-1} - \tilde{\boldsymbol{\beta}}_b) + \boldsymbol{g}_b(\tilde{\boldsymbol{\beta}}_b)\right\} = \boldsymbol{0}, \tag{6}$$

where $\tilde{\boldsymbol{G}}_b = \sum_{j=1}^b \boldsymbol{G}_j(\tilde{\boldsymbol{\beta}}_j)$ is the sequentially aggregated negative gradient matrix, $\tilde{\boldsymbol{C}}_b = \sum_{j=1}^b \boldsymbol{C}_j(\tilde{\boldsymbol{\beta}}_j)$ is the sequentially aggregated sample variance matrix, and $\tilde{\boldsymbol{g}}_b(\tilde{\boldsymbol{\beta}}_b) = \tilde{\boldsymbol{g}}_{b-1}(\tilde{\boldsymbol{\beta}}_{b-1}) + \sum_{j=1}^{b-1}\boldsymbol{G}_j(\tilde{\boldsymbol{\beta}}_j)(\tilde{\boldsymbol{\beta}}_{b-1} - \tilde{\boldsymbol{\beta}}_b) + \boldsymbol{g}_b(\tilde{\boldsymbol{\beta}}_b)$ is the adjusted extended score vector. In effect, Equation (6) may be rewritten as a recursive formula:

$$\tilde{\boldsymbol{\beta}}_b = \tilde{\boldsymbol{\beta}}_{b-1} + N_b^{-1}\boldsymbol{H}_b^{-1}\tilde{\boldsymbol{U}}_b(\tilde{\boldsymbol{\beta}}_b), \text{ with } \tilde{\boldsymbol{U}}_b(\tilde{\boldsymbol{\beta}}_b) = \tilde{\boldsymbol{G}}_b^\top\tilde{\boldsymbol{C}}_b^{-1}\left(\tilde{\boldsymbol{g}}_{b-1} + \boldsymbol{g}_b(\tilde{\boldsymbol{\beta}}_b)\right), \tag{7}$$

where $\boldsymbol{H}_b = N_b^{-1}\tilde{\boldsymbol{G}}_b^\top\tilde{\boldsymbol{C}}_b^{-1}\tilde{\boldsymbol{G}}_{b-1}$. Even though the first expression in (7) takes a similar form to that of the traditional SGD (Robbins and Monro 1951; Sakrison 1965; Toulis and Airold 2017), it is not an SGD as $\tilde{\boldsymbol{\beta}}_b$ is involved in both $\tilde{\boldsymbol{U}}_b(\cdot)$ and $\boldsymbol{H}_b$. It may be regarded as a second-order method involving Hessian inversion calculation useful for statistical inference. In comparison to SGD, (7) not only achieves efficient estimation but also provides relevant inferential quantities; the latter is of great importance in statistical analyses of cluster-correlated data streams. This RenewQIF estimate $\tilde{\boldsymbol{\beta}}_b$ given in (7) is clearly different from the offline QIF that requires to access the entire cumulative data $\mathcal{D}_b^\star$ to obtain the parameter estimate $\hat{\boldsymbol{\beta}}_b^\star$, and the latter may not be computationally feasible when the number of data batches $b \to \infty$.

Solving (6) can be easily carried out via the following Newton–Raphson iterations:

$$\tilde{\boldsymbol{\beta}}_b^{(r+1)} = \tilde{\boldsymbol{\beta}}_b^{(r)} + \left\{\tilde{\boldsymbol{G}}_b(\tilde{\boldsymbol{\beta}}_b^{(r)})^\top \tilde{\boldsymbol{C}}_b(\tilde{\boldsymbol{\beta}}_b^{(r)})^{-1}\tilde{\boldsymbol{G}}_b(\tilde{\boldsymbol{\beta}}_b^{(r)})\right\}^{-1} \tilde{\boldsymbol{G}}_b(\tilde{\boldsymbol{\beta}}_b^{(r)})^\top \tilde{\boldsymbol{C}}_b(\tilde{\boldsymbol{\beta}}_b^{(r)})^{-1}\tilde{\boldsymbol{g}}_b(\tilde{\boldsymbol{\beta}}_b^{(r)}). \quad (8)$$

In algorithm (8), clearly we only use the subject-level data of current data batch $\mathcal{D}_b$ and summary statistics $\{\tilde{\boldsymbol{\beta}}_{b-1}, \tilde{\boldsymbol{g}}_{b-1}, \tilde{\boldsymbol{G}}_{b-1}, \tilde{\boldsymbol{C}}_{b-1}\}$ from historical data batches up to $b-1$ rather than subject-level raw data of $\mathcal{D}^\star_{b-1}$. Thus, our proposed RenewQIF is indeed an online estimation procedure. In addition, a consistent estimator of parameter ϕ is given by $\tilde{\phi}_b = \frac{mN_{b-1}-p}{mN_b-p}\tilde{\phi}_{b-1} + \frac{mn_b-p}{mN_b-p}\widehat{\phi}_b$, where $\widehat{\phi}_b = \frac{1}{mn_b-p}\sum_{i\in\mathcal{D}_b}\sum_{k=1}^{m}\frac{(y_{ik}-\widehat{\mu}_{ik})^2}{v(\widehat{\mu}_{ik})}$ with $\widehat{\mu}_{ik} = h(\boldsymbol{x}_{ik}^\top\tilde{\boldsymbol{\beta}}_b)$.

2.3. Large Sample Properties

In our discussion of large sample properties, the cumulative sample size $N_b \to \infty$ may arise from one of the following three scenarios: **(S1)** n_j is finite for $j = 1, \ldots, b$ but the number of data batches $b \to \infty$; **(S2)** the initial data batch size $n_1 \to \infty$ (aggregate the first few batches to create a large initial data batch prior to the renewable updating), and subsequent batch sizes n_j's may be either finite or infinite for $j = 2, \ldots, b$, but the number of data batches b is finite; and **(S3)** the initial data batch size $n_1 \to \infty$ and other n_j's are either finite or infinite for $j = 2, \ldots, b$, and the number of data batches $b \to \infty$.

Following the regularity conditions given in Hansen (1982)'s theory of GMM, we postulate the regularity conditions to establish some key large sample properties for RenewQIF.

(C1) The true parameter $\boldsymbol{\beta}_0$ lies in the interior of parameter space $\Theta \subset \mathbb{R}^p$, and the space Θ is compact.
(C2) The incremental estimating function $\boldsymbol{f}_b(\boldsymbol{\beta})$ is unbiased such that $\mathbb{E}_{\boldsymbol{\beta}}\{\boldsymbol{f}_b(\boldsymbol{y};\boldsymbol{X},\boldsymbol{\beta})\} = \boldsymbol{0}$ if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.
(C3) The extended score vector $\boldsymbol{g}(\boldsymbol{y};\boldsymbol{X},\boldsymbol{\beta})$ is twice continuous differentiable with respect to parameter $\boldsymbol{\beta}$, and the sensitivity matrix $\mathbb{G}(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\beta}}\boldsymbol{G}(\boldsymbol{y};\boldsymbol{X},\boldsymbol{\beta})$ is of full column rank for $\boldsymbol{\beta} \in \Theta$.
(C4) $\mathbb{E}_{\boldsymbol{\beta}}\left[||\boldsymbol{g}(\boldsymbol{y};\boldsymbol{X},\boldsymbol{\beta})||^2\right] < \infty$ for all $\boldsymbol{\beta} \in \Theta$.
(C5) The variability matrix $\mathbb{C}(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\beta}}\boldsymbol{g}(\boldsymbol{y};\boldsymbol{X},\boldsymbol{\beta})\boldsymbol{g}(\boldsymbol{y};\boldsymbol{X},\boldsymbol{\beta})^\top$ is positive-definite for $\boldsymbol{\beta} \in \Theta$.

Remark 1. (C1)–(C5) are mild conditions required to establish asymptotic consistency for scenarios S2 and S3. Condition (C4) indicates that the sample variance matrix $N_b^{-1}\tilde{\boldsymbol{C}}_b$ is finite for all $\boldsymbol{\beta} \in \Theta$. (C5) of positive-definite $\mathbb{C}(\boldsymbol{\beta})$ on the whole parameter space Θ is required specifically in scenario S1, while in scenarios S2 and S3, (C5) may be relaxed to hold in a neighborhood with radius $o(\sqrt{n_1})$.

Note that we do not include the convergence condition on a sequence of weighting matrices as in Hansen (1982) because the online QIF uses the sample variance matrix where $N_b^{-1}\tilde{\boldsymbol{C}}_b \xrightarrow{\text{a.s.}} \mathbb{C}$ holds under N_b iid samples and conditions (C1)–(C4).

We first establish estimation consistency for the RenewQIF in scenarios (S2)–(S3).

Theorem 1. In scenarios (S2)–(S3), under regularity conditions (C1)–(C5), the renewable estimator $\tilde{\boldsymbol{\beta}}_b$ given in (6) is consistent, namely $\tilde{\boldsymbol{\beta}}_b \xrightarrow{p} \boldsymbol{\beta}_0$, as $N_b = \sum_{j=1}^b n_j \to \infty$.

Under the same regularity conditions, we can further establish the asymptotic normality for the RenewQIF in the two scenarios (S2)–(S3).

Theorem 2. In scenarios (S2)–(S3), under regularity conditions (C1)–(C5), the renewable estimator $\tilde{\boldsymbol{\beta}}_b$ in (6) is asymptotically normally distributed, namely $\sqrt{N_b}(\tilde{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \mathbb{J}^{-1}(\boldsymbol{\beta}_0)\right)$, as $N_b \to \infty$, where Godambe information $\mathbb{J}(\boldsymbol{\beta}_0) = \mathbb{G}^\top(\boldsymbol{\beta}_0)\mathbb{C}^{-1}(\boldsymbol{\beta}_0)\mathbb{G}(\boldsymbol{\beta}_0)$.

The proof of Theorems 1 and 2 for scenarios (S2)–(S3) are given in Section 3 in the supplementary materials. To establish consistency and asymptotic normality for scenario (S1), we need the following additional conditions.

(C6) Define $\boldsymbol{f}_1(\boldsymbol{\beta}) := \boldsymbol{G}_1(\boldsymbol{\beta})^\top\boldsymbol{C}_1(\boldsymbol{\beta})^{-1}\boldsymbol{g}_1(\boldsymbol{\beta})$. Both matrices $\boldsymbol{G}_1(\boldsymbol{\beta})^\top\boldsymbol{C}_1(\boldsymbol{\beta})^{-1}\boldsymbol{G}_1(\boldsymbol{\beta})$ and $-\partial\boldsymbol{f}_1(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^\top$ are positive-definite for $\boldsymbol{\beta} \in \Theta$.
(C7) $\boldsymbol{H}_b = N_b^{-1}\tilde{\boldsymbol{G}}_b^\top\tilde{\boldsymbol{C}}_b^{-1}\tilde{\boldsymbol{G}}_{b-1}$ is positive-definite.

Remark 2. The condition (C6) requiring positive-definiteness of $\boldsymbol{G}_1(\boldsymbol{\beta})^\top\boldsymbol{C}_1(\boldsymbol{\beta})^{-1}\boldsymbol{G}_1(\boldsymbol{\beta})$ and $-\partial\boldsymbol{f}_1(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^\top$ appears to be mild as long as the first data batch size n_1 is moderately large. Since matrix $\boldsymbol{H}_b$ approximates the sample covariance matrix of $\tilde{\boldsymbol{\beta}}_b$ as $b \to \infty$, the condition (C7) of $\boldsymbol{H}_b$ being positive-definite is also modest.

Theorem 3. In scenario (S1), under regularity conditions (C1)–(C7), the renewable estimator $\tilde{\boldsymbol{\beta}}_b$ given in (6) is consistent, namely $\tilde{\boldsymbol{\beta}}_b \xrightarrow{p} \boldsymbol{\beta}_0$, as $N_b = \sum_{j=1}^b n_j \to \infty$.

Theorem 4. In scenario (S1), under regularity conditions (C1)–(C7), the renewable estimator $\tilde{\boldsymbol{\beta}}_b$ given in (6) is asymptotically normally distributed, namely, $\sqrt{N_b}(\tilde{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \mathbb{J}^{-1}(\boldsymbol{\beta}_0)\right)$, as $N_b \to \infty$, where Godambe information $\mathbb{J}(\boldsymbol{\beta}_0) = \mathbb{G}^\top(\boldsymbol{\beta}_0)\mathbb{C}^{-1}(\boldsymbol{\beta}_0)\mathbb{G}(\boldsymbol{\beta}_0)$.

The proof of Theorems 3 and 4 are given in Appendix A.1 and A.2, respectively.

It is interesting to notice that the asymptotic covariance matrix of the renewable estimator $\tilde{\boldsymbol{\beta}}_b$ given in both Theorems 2 and 4 is exactly the same as that of the offline estimator $\hat{\boldsymbol{\beta}}^\star_b$. This implies that the proposed RenewQIF estimator achieves the same asymptotic distribution as the offline QIF estimator. With no access to any historical subject-level data in the computation, using only the online aggregated matrices $\tilde{\boldsymbol{G}}_b = \sum_{j=1}^b \boldsymbol{G}_j(\mathcal{D}_j;\tilde{\boldsymbol{\beta}}_j)$ and $\tilde{\boldsymbol{C}}_b = \sum_{j=1}^b \boldsymbol{C}_j(\mathcal{D}_j;\tilde{\boldsymbol{\beta}}_j)$, we can calculate the estimated asymptotic covariance matrix as $\widetilde{\boldsymbol{\Sigma}}_b(\boldsymbol{\beta}_0) = \left\{N_b^{-1}\tilde{\boldsymbol{J}}_b\right\}^{-1} = N_b\left\{\tilde{\boldsymbol{G}}_b^\top\tilde{\boldsymbol{C}}_b^{-1}\tilde{\boldsymbol{G}}_b\right\}^{-1}$. It follows that the online estimated asymptotic variance matrix for the RenewQIF $\tilde{\boldsymbol{\beta}}_b$ is

$$\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{\beta}}_b) := \widetilde{\text{var}}(\tilde{\boldsymbol{\beta}}_b) = \frac{1}{N_b}\widetilde{\boldsymbol{\Sigma}}_b(\boldsymbol{\beta}_0) = \left\{\tilde{\boldsymbol{G}}_b^\top\tilde{\boldsymbol{C}}_b^{-1}\tilde{\boldsymbol{G}}_b\right\}^{-1}. \quad (9)$$

Positive-definiteness on $\mathbb{C}(\boldsymbol{\beta})$ is required by (C5) for all $\boldsymbol{\beta} \in \Theta$, but the sample version $N_b^{-1}\tilde{\boldsymbol{C}}_b$ may be noninvertible. In the implementation, we invoke a generalized inverse (Moore–Penrose inverse). Other possible remedy approaches include (i) a linear shrinkage estimator to replace the sample covariance matrix (Han and Song 2011); (ii) a selection of moment conditions (Cho and Qu 2015); and (iii) a principle component dimension reduction (Pearson 1901) to ensure that the components in $\boldsymbol{g}$ are not (nearly) linearly dependent.

In addition, it is noteworthy that in (S2) or (S3), our method can be used as an alternative to parallelized computing methods. However, with our method, the convergence rate is $O_p(N_b^{-1/2})$ which is based on the cumulative sample size N_b. This indicates a faster convergence rate than parallelized distributed estimation where the convergence rate is based on the sample size of the smallest single dataset (Zhou and Song 2017) $\min_j \sqrt{n_j}$.

3. Computing and Monitoring

3.1. Computing Implementation of RenewQIF

We expand the existing Spark's Lambda architecture to reduce computing burden in the proposed framework of RenewQIF methodology. The iterative calculation in (8) can be implemented in the speed and inference layers in an extended Lambda architecture shown in Figure 1. Here, relevant inferential statistics include the aggregated extended score vector $\tilde{\boldsymbol{g}}$ and two inferential matrices $\tilde{\boldsymbol{G}}$ (aggregated negative gradient) and $\tilde{\boldsymbol{C}}$ (aggregated sample variance matrix). If data batch $\mathcal{D}_b$ passes the scrutiny, we update $\tilde{\boldsymbol{\beta}}_{b-1}$ to $\tilde{\boldsymbol{\beta}}_b$ at the speed layer and update $\tilde{\boldsymbol{g}}_{b-1}$, $\tilde{\boldsymbol{G}}_{b-1}$, $\tilde{\boldsymbol{C}}_{b-1}$ to $\tilde{\boldsymbol{g}}_b$, $\tilde{\boldsymbol{G}}_b$ and $\tilde{\boldsymbol{C}}_b$ at the inference layer. Otherwise, skip all updating steps and proceed to next data batch $\mathcal{D}_{b+1}$. Algorithm 1 lists the pseudo code for the implementation of the RenewQIF via the paradigm of the extended Lambda architecture shown in Figure 1. Some explanations are given below.

1. Line 1: the population-averaged model or MGLM has been specified in Section 2.1. It is worth noting that the estimating functions used in the construction of the QIF objective function involves neither the correlation parameter $\boldsymbol{\alpha}$ nor the dispersion parameter ϕ. The updating of the dispersion parameter ϕ takes place outside of the QIF loop.
2. Line 2: the outputs include RenewQIF estimates of the regression coefficients and the corresponding estimated asymptotic covariance matrix at each time point b, and the latter is needed for statistical inference.
3. Line 3: set certain initial values for the regression coefficients, for example, get the initial estimate $\tilde{\boldsymbol{\beta}}_0$ by fitting $\mathcal{D}_1$ to R function `glm()`.
4. Line 4: run through the sequential updating procedure along data streams.
5. Line 6: before updating $\tilde{\boldsymbol{\beta}}_{b-1}$ with current $\mathcal{D}_b$, first check its compatibility with the normal reference data batch $\mathcal{D}_1$. QIF estimator $\check{\boldsymbol{\beta}}_b$ is obtained by minimizing the quadratic inference function based only on these two data batches, $\mathcal{D}_1 \cup \mathcal{D}_b$. The goodness-of-fit test statistic Λ_b will be discussed in detail in Section 3.3.
6. Line 7: if the current data batch $\mathcal{D}_b$ does not pass the scrutiny test, we don't use it in the updating and jump to Line 16 directly.
7. Lines 8–11: if the test concludes the current data batch $\mathcal{D}_b$ passes the scrutiny, at the inference layer, use the prior estimate $\tilde{\boldsymbol{\beta}}_{b-1}$ and current batch $\mathcal{D}_b$ to calculate $\tilde{\boldsymbol{g}}_b$, $\tilde{\boldsymbol{G}}_b$ and $\tilde{\boldsymbol{C}}_b$ through a communication with the speed layer.
8. Line 12: run the Newton–Raphson algorithm to renew $\tilde{\boldsymbol{\beta}}_{b-1}$ to $\tilde{\boldsymbol{\beta}}_b$.
9. Line 13: the convergence criterion and the distance between theoretical estimator $\tilde{\boldsymbol{\beta}}_b$ and experimental estimator $\tilde{\boldsymbol{\beta}}_b^{(r)}$ will be further discussed in Section 3.2.
10. Line 14: the inference layer performs statistical inference with $\tilde{\boldsymbol{G}}_b$ and $\tilde{\boldsymbol{C}}_b$.
11. Line 15: the speed layer updates the estimate of the dispersion parameter ϕ outside the QIF loop. The proposed estimate

Algorithm 1: RenewQIF for streaming cluster-correlated data in the extended Lambda architecture.

1 **Inputs:** Sequentially arrived datasets $\mathcal{D}_1,\ldots,\mathcal{D}_b,\ldots$from an MGLM with mean $\mathbb{E}(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{\mu}$ and covariance $\text{cov}(\boldsymbol{y}|\boldsymbol{X}) = \phi\boldsymbol{\Sigma}$ specified in Equation (1);
2 **Outputs:** $\tilde{\boldsymbol{\beta}}_b$, $\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{\beta}}_b)$ and $\tilde{\phi}_b$, for $b = 1, 2, \ldots$;
3 **Initialize:** Initial values $\tilde{\boldsymbol{\beta}}_0$, $\tilde{\phi}_0$, $\tilde{\boldsymbol{g}}_0 = \boldsymbol{0}_{pS}$, $\tilde{\boldsymbol{G}}_0 = \boldsymbol{0}_{pS\times p}$ and $\tilde{\boldsymbol{C}}_0 = \boldsymbol{0}_{pS\times pS}$;
4 **for** $b = 1, 2, \ldots$ **do**
5 Read in dataset $\mathcal{D}_b$;
6 At the monitoring layer, if $b \geq 2$, calculate $\Lambda_b = Q_1(\check{\boldsymbol{\beta}}_b) + Q_b(\check{\boldsymbol{\beta}}_b)$;
7 if $\Lambda_b \geq \chi^2_{df,\alpha}$, say $\alpha = 0.05$, set $\tilde{\boldsymbol{\beta}}_b = \tilde{\boldsymbol{\beta}}_{b-1}$, $\tilde{\phi}_b = \tilde{\phi}_{b-1}$, $\tilde{\boldsymbol{g}}_b = \tilde{\boldsymbol{g}}_{b-1}$, $\tilde{\boldsymbol{G}}_b = \tilde{\boldsymbol{G}}_{b-1}$, $\tilde{\boldsymbol{C}}_b = \tilde{\boldsymbol{C}}_{b-1}$ and jump to Line 16;
8 otherwise, start iterations with $(\tilde{\boldsymbol{\beta}}_b, \tilde{\phi}_b)$ initialized by $(\tilde{\boldsymbol{\beta}}_{b-1}, \tilde{\phi}_{b-1})$;
9 **repeat**
10 At the inference layer, calculate $\tilde{\boldsymbol{g}}_b^{(r)} = \tilde{\boldsymbol{g}}_{b-1} + \tilde{\boldsymbol{G}}_{b-1}(\tilde{\boldsymbol{\beta}}_{b-1} - \tilde{\boldsymbol{\beta}}_b^{(r)}) + \boldsymbol{g}_b(\mathcal{D}_b; \tilde{\boldsymbol{\beta}}_b^{(r)})$,
11 $\tilde{\boldsymbol{G}}_b^{(r)} = \tilde{\boldsymbol{G}}_{b-1} + \boldsymbol{G}_b(\mathcal{D}_b; \tilde{\boldsymbol{\beta}}_b^{(r)})$ and $\tilde{\boldsymbol{C}}_b^{(r)} = \tilde{\boldsymbol{C}}_{b-1} + \boldsymbol{C}_b(\mathcal{D}_b; \tilde{\boldsymbol{\beta}}_b^{(r)})$;
12 At the speed layer, $\tilde{\boldsymbol{\beta}}_b^{(r+1)} = \tilde{\boldsymbol{\beta}}_b^{(r)} + \left\{\tilde{\boldsymbol{G}}_b^{(r)\top}\tilde{\boldsymbol{C}}_b^{(r)^{-1}}\tilde{\boldsymbol{G}}_b^{(r)}\right\}^{-1}\tilde{\boldsymbol{G}}_b^{(r)\top}\tilde{\boldsymbol{C}}_b^{(r)^{-1}}\tilde{\boldsymbol{g}}_b^{(r)}$;
13 **until** *convergence*;
14 At the inference layer, calculate $\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{\beta}}_b) = \left\{\tilde{\boldsymbol{G}}_b^\top\tilde{\boldsymbol{C}}_b^{-1}\tilde{\boldsymbol{G}}_b\right\}^{-1}$;
15 At the speed layer, calculate $\tilde{\phi}_b = \frac{mN_{b-1}-p}{mN_b-p}\tilde{\phi}_{b-1} + \frac{mn_b-p}{mN_b-p}\widehat{\phi}_b$, where $\widehat{\phi}_b = \frac{1}{mn_b-p}\sum_{i\in\mathcal{D}_b}\sum_{k=1}^{m}\frac{(y_{ik}-\widehat{\mu}_{ik})^2}{v(\widehat{\mu}_{ik})}$;
16 Save $\tilde{\boldsymbol{\beta}}_b$, $\tilde{\phi}_b$, $\tilde{\boldsymbol{g}}_b$, $\tilde{\boldsymbol{G}}_b$ and $\tilde{\boldsymbol{C}}_b$ at the speed and inference layers, respectively;
17 Release dataset $\mathcal{D}_b$ from the memory.
18 **end**
19 **Return** $\tilde{\boldsymbol{\beta}}_b$, $\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{\beta}}_b)$ and $\tilde{\phi}_b$, for $b = 1, 2, \ldots$.

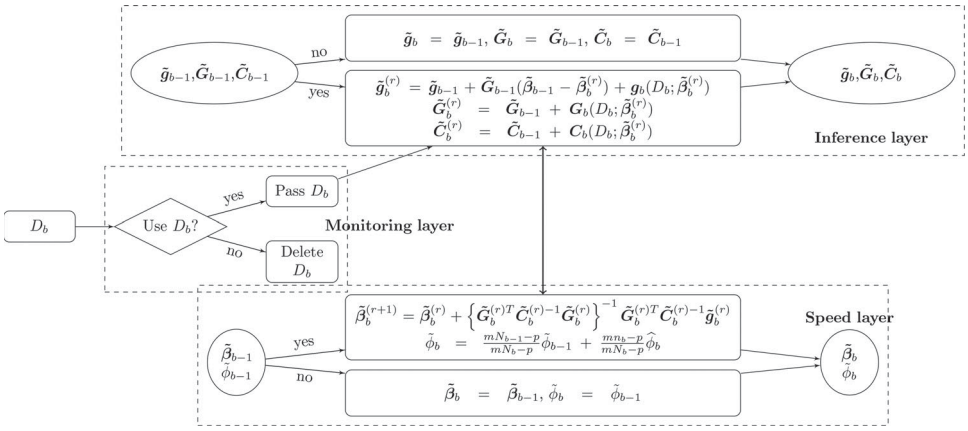

Figure 1. Diagram of an extended Lambda architecture with the addition of both monitoring and inference layers to the standard speed layer.

is guaranteed to be positive. In the cases of logistic and Poisson models, since the dispersion parameters are known, this updating step is omitted.

3.2. Monitoring of Algorithmic Convergence

Note that the large sample properties developed in Section 2.3 centers at the theoretical root, $\tilde{\boldsymbol{\beta}}_b$, our proposed incremental estimating equation in (6), that is, $\boldsymbol{f}_b(\tilde{\boldsymbol{\beta}}_b) = \mathbf{0}$. In the implementation, $\tilde{\boldsymbol{\beta}}_b^{(r)}$ is the actual estimate harvested numerically at the r-th iteration of the Newton–Raphson algorithm (8). Following the conventional practice, we set up a stringent convergence criterion, so that the two quantities $\tilde{\boldsymbol{\beta}}_b$ and $\tilde{\boldsymbol{\beta}}_b^{(r)}$ are numerically close enough that their difference is ignored. According to Nocedal and Wright (2006, chap. 3), the convergence rate of the algorithm (8) is at least quadratic if the following conditions are satisfied: (a) the starting point $\tilde{\boldsymbol{\beta}}_b^{(0)}$ is sufficiently close to the root $\tilde{\boldsymbol{\beta}}_b$; (b) the negative gradient $\tilde{\boldsymbol{J}}_b := -\partial \boldsymbol{f}_b(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^\top = \tilde{\boldsymbol{G}}_b^\top \tilde{\boldsymbol{C}}_b^{-1} \tilde{\boldsymbol{G}}_b$ is Lipschitz continuous in a neighborhood of the root $\tilde{\boldsymbol{\beta}}_b$; and (c) $\boldsymbol{f}_b(\boldsymbol{\beta})$ is differentiable and $\tilde{\boldsymbol{J}}_b$ is positive-definite in a neighborhood of the root $\tilde{\boldsymbol{\beta}}_b$. These three conditions are all satisfied by our algorithm (8). This is because condition (a) holds for large b in renewable updating due to the initialization of $\tilde{\boldsymbol{\beta}}_b^{(0)} = \tilde{\boldsymbol{\beta}}_{b-1}$. Being an example, the right panel of Figure 2 shows the closeness to the root $\tilde{\boldsymbol{\beta}}_b^{(r)}$ improves over the course of increased b. Conditions (b) and (c) are also satisfied by the regularity Conditions (C3) and (C5) in Section 2.3.

Line 13 in Algorithm 1 stems from the convergence criteria in the Newton's decrement given by $\Delta_r = \boldsymbol{f}_b^\top(\tilde{\boldsymbol{\beta}}_b^{(r)}) \left(\tilde{\boldsymbol{J}}_b(\tilde{\boldsymbol{\beta}}_b^{(r)})\right)^{-1} \boldsymbol{f}_b(\tilde{\boldsymbol{\beta}}_b^{(r)}) < 10^{-6}$, which measures the proximity of $\tilde{\boldsymbol{\beta}}_b^{(r)}$ to $\tilde{\boldsymbol{\beta}}_b$. Clearly, $\Delta_r = 0$ when $\tilde{\boldsymbol{\beta}}^{(r)} = \tilde{\boldsymbol{\beta}}_b$ should be monitored over iterations of algorithm (8). The algorithm will stop once Δ_r satisfies the stopping rule $\Delta_r < 10^{-6}$. To control the computation time, we also terminate the algorithm when the number of iterations reaches a prefixed threshold. Based on our extensive empirical experience, in our current implementation, we set the maximum number of iterations at 50. If this threshold is reached with failure of $\Delta_r < 10^{-6}$, a warning message "algorithm reached 'maxit' but did not reach the convergence criteria" will be given as an output. All simulation studies in Section 4 have shown this criterion is satisfactory with zero warning message. As shown in the left panel in Figure 2, the number of iterations required to reach the criteria $\Delta_r < 10^{-6}$ is all less than 10, and with the sequential addition of data batches, the number of iterations run by algorithm (8) decreases monotonically as b increases.

3.3. Monitoring of Abnormal Data Batches

For the case of high throughput data streams in practice, it is highly likely to encounter abnormal data batches. To address this issue, we relax the RenewQIF method to a situation where abnormal data batches may occur over the sequence of data streams, $\mathcal{D}_2, \ldots, \mathcal{D}_b$. A data batch $\mathcal{D}_\tau$, $\tau \in \{2, \ldots, b\}$, is regarded as being abnormal if $\mathcal{D}_\tau$ is generated from a model whose regression parameters, say $\boldsymbol{\beta}_\tau$'s, are different from those of the underlying main model of interest, $\boldsymbol{\beta}_0$ of the true model, that is, $\boldsymbol{\beta}_\tau \neq \boldsymbol{\beta}_0$. In other words, $\mathcal{D}_\tau$ is an outlying data batch, which is incompatible with the data batches generated from the true model. Let $\Gamma_q = \{\tau_1, \ldots, \tau_q\}$ denote the set of indices for q abnormal data batches. In reality, we do not know set Γ_q in advance but want to find them out during the collection of data streams. For convenience, we assume that the first data batch $\mathcal{D}_1$ is the normal reference, which is generated from a model with "the normal" parameter $\boldsymbol{\beta}_0$. At each subsequent time point b ($b \geq 2$), we propose a diagnostic procedure via hypothesis testing of mean-zero assumption for the pair of extended scores $H_0 : \mathbb{E}_{\boldsymbol{\beta}}(\boldsymbol{g}_1) = \mathbb{E}_{\boldsymbol{\beta}}(\boldsymbol{g}_b) = \mathbf{0}$, where $\boldsymbol{g}_1$ and $\boldsymbol{g}_b$ are the extended score vectors for data batches $\mathcal{D}_1$ and $\mathcal{D}_b$, respectively, similar to the one given in (4). This goodness-of-fit test essentially enables to check the compatibility between data batch $\mathcal{D}_b$ under investigation and the normal reference $\mathcal{D}_1$. If H_0 is rejected, we would not use data batch $\mathcal{D}_b$ to renew $\tilde{\boldsymbol{\beta}}_{b-1}$, and set $\tilde{\boldsymbol{\beta}}_b$ equal to $\tilde{\boldsymbol{\beta}}_{b-1}$; otherwise, execute an update from $\tilde{\boldsymbol{\beta}}_{b-1}$ to $\tilde{\boldsymbol{\beta}}_b$ using RenewQIF. Next we proceed to test H_0 with next data batch $\mathcal{D}_{b+1}$.

We construct a test statistic along the line of Hansen's (1982) seminal goodness-of-fit test. The quadratic inference function has useful chi-squared properties for hypothesis testing (Lindsay and Qu 2003). For checking for data compatibility, we con-

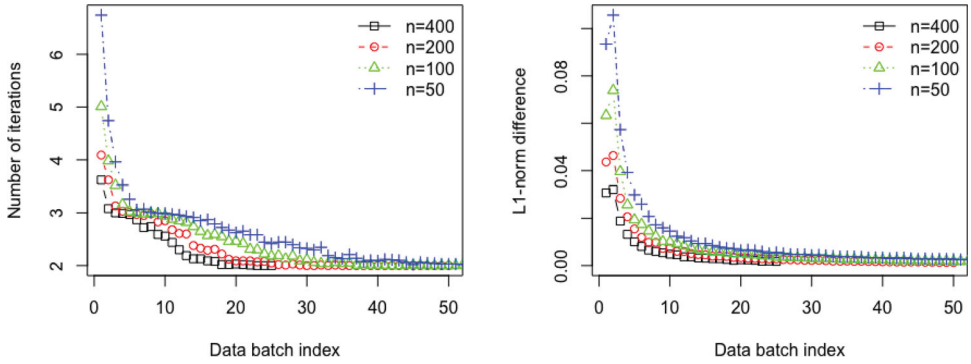

Figure 2. The left panel shows the number of iterations to reach $\Delta_r < 10^{-6}$ with different data batch size n_b; the right panel indicates the L_1-norm difference $||\tilde{\boldsymbol{\beta}}_b - \tilde{\boldsymbol{\beta}}_{b-1}||_1$ decreases fast as b increases. Both plots are generated under a marginal logistic model specified in Section 4.1 with a fixed $N_B = 10^4$ but varying batch size n_b.

sider a quadratic inference function of the following form:

$$\Lambda_b(\boldsymbol{\beta}) = \begin{pmatrix} \boldsymbol{g}_1(\boldsymbol{\beta}) \\ \boldsymbol{g}_b(\boldsymbol{\beta}) \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{C}_1(\boldsymbol{\beta}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}_b(\boldsymbol{\beta}) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{g}_1(\boldsymbol{\beta}) \\ \boldsymbol{g}_b(\boldsymbol{\beta}) \end{pmatrix}, \tag{10}$$

where $\boldsymbol{C}_1$ and $\boldsymbol{C}_b$ are the estimated sample covariances of extended scores $\boldsymbol{g}_1$ and $\boldsymbol{g}_b$, respectively. Note that the form of block-diagonal covariance in Λ_b is due to the independence between $\mathcal{D}_1$ and $\mathcal{D}_b$. Let $\check{\boldsymbol{\beta}}_b = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Lambda_b(\boldsymbol{\beta})$, under H_0, that is, $b \notin \Gamma_q$, test statistic $\Lambda_b(\check{\boldsymbol{\beta}}_b) \xrightarrow{d} \chi^2_{df}$ with $df = \text{rank}(\boldsymbol{C}_1) + \text{rank}(\boldsymbol{C}_b) - p$; under H_1, for an index $\tau \in \Gamma_q$, and given local alternatives in the form of $\boldsymbol{\beta}_\tau = \boldsymbol{\beta}_0 + (n_1 + n_\tau)^{-1/2}\boldsymbol{d}$, $\boldsymbol{d} \in \mathbb{R}^p$, test statistic $\Lambda_\tau(\check{\boldsymbol{\beta}}_\tau) \xrightarrow{d} \chi^2_{df}(\lambda)$, with $df = \text{rank}(\boldsymbol{C}_1) + \text{rank}(\boldsymbol{C}_\tau) - p$ and the noncentrality parameter $\lambda = \boldsymbol{d}^\top \mathbb{J}(\boldsymbol{\beta}_0)\boldsymbol{d}$ where $\mathbb{J}$ is the Godambe information matrix given in Theorems 4 and 2. Moreover, it is easy to show that Power $= P_{H_1}\left(\Lambda_\tau(\check{\boldsymbol{\beta}}_\tau) > \chi^2_{df,\alpha}\right) \to 1$, as $(n_1 + n_\tau) \to \infty$, which implies that the proposed test Λ_τ is consistent. Under a finite sample size, with fixed $\boldsymbol{d}$, the power of Λ_τ depends on both statistical significance level α and abnormal data batch size n_τ. Larger α leads to higher power and smaller Type II error, but also a higher chance to produce false alarms. Obviously, increasing data batch size n_b will help increase power. The above monitoring procedure is based on the asymptotic properties under scenarios S2 and S3 with $n_j \to \infty$ for some j. However, if the reference data batch size is small as assumed in scenario S1, one may carry out the proposed diagnostic test against an augmented reference data that combines several normal data batches to reach a desirable sample size.

In practice the use of the first data batch $\mathcal{D}_1$, which is the normal reference sampled from the true model, may be replaced by any data batch that is deemed appropriate. This choice is obviously somewhat subjective and mostly made by practitioners base on their prior experience and existing knowledge on data quality. We do not recommend frequently changing the reference batch, but combining several normal data batches to form a larger reference one is useful to reach more stable performance of the diagnostic test Λ_b. For example, adaptively using the adjacent data batch as the reference, we show in the left panel of Figure 3 that the monitoring diagnostic test suffers from an inflated Type I error once an abnormal data batch is mistakenly set as the reference. Therefore, fixing a single or an augmented reference data batch in the monitoring leads to reliable diagnoses for abnormal data batches, as shown in the right panel of Figure 3 with the normal reference fixed at the first data batch.

4. Simulation Experiments

4.1. Setup

We conduct simulation experiments to assess the performances of the proposed RenewQIF estimation and inference, as well as of the diagnostic procedure for abnormal data batches, in the setting of marginal generalized linear models (MGLMs) for cluster-correlated data streams. We compare the proposed RenewQIF method with (i) the offline GEE estimator obtained by processing the entire cumulative data once, (ii) the offline QIF estimator obtained by processing the entire data once, and (iii) renewable GEE estimation method (RenewGEE) that is similar to RenewQIF (see the relevant derivation in Section 2 in the supplementary materials).

In the first part of comparisons to be presented below, we consider the following criterion related to both parameter estimation and inference: (a) averaged absolute bias (A.bias), (b) averaged asymptotic standard error (ASE), (c) empirical standard error (ESE), and (d) coverage probability (CP). Both offline GEE and offline QIF estimates are yielded from the R packages `gee` and `qif`. Computational efficiency is assessed by (e) computation time (C.Time) and (f) running time (R.Time). R.Time accounts only algorithm execution time, while C.Time includes time spent on both data loading and algorithm execution. In the second part of comparisons, we will first evaluate the Type I error and power of the proposed goodness-of-fit test for data compatibility with different significance level α and data batch size n_b. In addition, the criteria for parameter estimation and inference will be reported thoroughly on the competing methods with and without quality control.

In the simulation studies, we set a terminal point B, and generate a full dataset $\mathcal{D}_B^\star$ with N_B independent cluster-correlated observations of an m-dimensional MGLM, consisting of marginal mean $\mathbb{E}(\boldsymbol{y}_i|\boldsymbol{X}_i) = \left[h(\boldsymbol{x}_{i1}^\top\boldsymbol{\beta}_0), \ldots, h(\boldsymbol{x}_{im}^\top\boldsymbol{\beta}_0)\right]^\top$ with $\boldsymbol{\beta}_0 = (0.2, -0.2, 0.2, -0.2, 0.2)^\top$, and covariance matrix $\text{cov}(\boldsymbol{y}_i|\boldsymbol{X}_i) = \phi\boldsymbol{\Sigma}_i = \phi\boldsymbol{A}_i^{1/2}\boldsymbol{R}(\alpha_y)\boldsymbol{A}_i^{1/2}$, $i = 1, \ldots, N_B$, where four covariates $\boldsymbol{x}_{ij[2:5]} \overset{\text{iid}}{\sim} \mathcal{N}_4(\boldsymbol{0}, \boldsymbol{R}(\alpha_x))$ and intercept $\boldsymbol{x}_{ij[1]} = 1$,

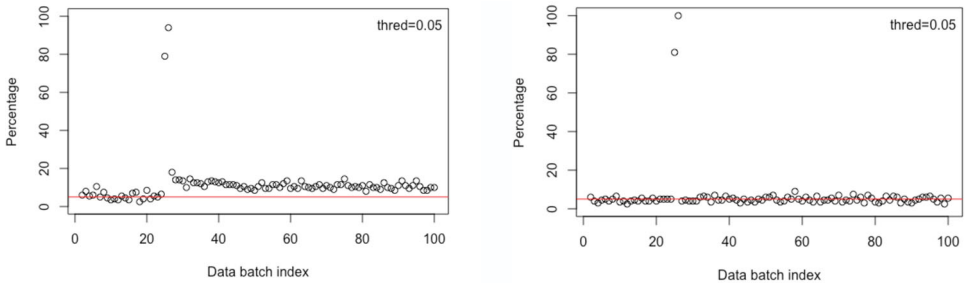

Figure 3. These two plots are generated under the marginal logistic model specified in Section 4.1 with a fixed total sample size $N_B = 10^4$ and data batch size $n_b = 100$. Two abnormal data batches are $\mathcal{D}_{25}$ and $\mathcal{D}_{26}$. The *y*-axis is the empirical proportion of rejections over 500 replications, and *x*-axis is the index of test statistic Λ_b (also the data batch index).

Table 1. Simulation results under the linear and logistic MGLMs are summarized over 500 replications, with fixed $N_B = 10^5$ and $p = 5$ with increasing B.

	Linear MGLM											
	offline GEE			RenewGEE			offline QIF			RenewQIF		
B	100	500	2000	100	500	2000	100	500	2000	100	500	2000
A.bias$\times 10^{-3}$	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10
ASE$\times 10^{-3}$	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42	1.42
ESE$\times 10^{-3}$	1.40	1.40	1.40	1.40	1.40	1.40	1.40	1.40	1.40	1.40	1.40	1.40
CP	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
C.Time(s)	9.27	12.89	20.56	2.70	4.08	8.45	2.60	5.36	13.91	1.39	2.77	6.63
R.Time(s)	8.53	9.49	8.53	2.31	2.64	3.62	1.86	1.96	1.88	1.06	1.65	2.95
	Logistic MGLM											
	offline GEE			RenewGEE			offline QIF			RenewQIF		
B	100	500	2000	100	500	2000	100	500	2000	100	500	2000
A.bias$\times 10^{-3}$	2.70	2.70	2.70	2.70	2.70	2.70	2.70	2.70	2.70	2.70	2.70	2.70
ASE$\times 10^{-3}$	3.31	3.31	3.31	3.31	3.31	3.31	3.31	3.31	3.31	3.31	3.31	3.31
ESE$\times 10^{-3}$	3.37	3.37	3.37	3.37	3.37	3.37	3.37	3.37	3.37	3.37	3.37	3.37
CP	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
C.Time(s)	10.73	14.04	27.77	2.05	2.45	3.32	3.22	6.51	20.25	1.14	1.47	2.37
R.Time(s)	9.85	9.86	9.81	1.86	2.07	2.35	2.34	2.33	2.30	0.99	1.23	1.86

NOTE: "A.bias," "ASE," "ESE" and "CP" stand for the mean absolute bias, the mean asymptotic standard error, the empirical standard error, and the coverage probability, respectively. "A.bias$\times 10^{-3}$" indicates the scale of number in the cell, for example, $1.10 \times 10^{-3} = 0.0011$. "C.Time" and "R.Time" denote computation time and running time, and the unit of both is second.

$j = 1, \ldots, m$. Here both correlation matrices $\boldsymbol{R}(\alpha_x)$ and $\boldsymbol{R}(\alpha_y)$ are set as compound symmetry with $\alpha_x = 0.5$ and $\alpha_y = 0.7$, respectively. The dispersion parameter $\phi = 1$ and the cluster size is $m = 5$. We consider both marginal linear model for continuous y_{ij} with $h(\mu_{ij}) = \mu_{ij}$ and marginal logistic model for binary y_{ij} with $h(\mu_{ij}) = \log(\mu_{ij}/(1-\mu_{ij}))$. For all four methods in comparison, the working correlation matrix is specified to be compound symmetry which is also the true correlation structure.

4.2. Evaluation of Parameter Estimation

Scenario 1: fixed N_B but varying batch size n_b

We begin with the comparison of the four methods for the effect of data batch size n_b on their point estimation and computational efficiency. A collection of B data batches specified in Section 4.1 are generated, each with data batch size n_b and a total of $N_B = |D_B^\star| = 10^5$ independent clusters, from an m-variate Gaussian linear model and an m-dimensional logistic model (using R package `SimCorMultRes`). Tables 1 reports the results of both linear and logistic MGLMs, over 500 rounds of simulations.

Bias and coverage probability. In linear and logistic MGLMs, Table 1 shows that both RenewGEE and RenewQIF have similar bias and coverage probability in comparison to the two offline methods. This confirms the theoretical results given in Theorem 2; the RenewQIF as well as RenewGEE are stochastically equivalent to the offline QIF and offline GEE, respectively. It is easy to see that both bias and coverage probability in both the linear and logistic models are not affected by individual data batch size n_b. In other words, their performances seem to depend only on N_B.

Computation time. Two metrics are used to evaluate computation efficiency: "C.Time" in Table 1 refers to the total amount of time required by data loading and algorithm execution. With an increased B, both RenewGEE and RenewQIF show clearly advantageous for much lower computation time over the offline GEE or offline QIF, due to the fact that the two offline methods are much more time consuming to load in full datasets.

Scenario 2: fixed batch size n_b but varying B

Now we turn to a streaming setting where B data batches arrive sequentially. For convenience, we fix single batch size $n_b = 100$, but let N_B increase from 10^3 to 10^6 (or B from 10 to 10^4). Tables 2 and 3 list the simulation results under the linear and logistic MGLMs.

Bias and coverage probability. As the number of batches B rises from 10 to 10^4, both RenewGEE and RenewQIF confirm

Table 2. Compare renewable estimators and offline ones in the linear MGLM model with fixed single batch size $n_b = 100$ and $p = 5$, where B increases from 10 to 10^4.

	$B = 10, N_B = 10^3$				$B = 100, N_B = 10^4$			
	GEE		QIF		GEE		QIF	
Criterion	offline	Renew	offline	Renew	offline	Renew	offline	Renew
A.bias$\times 10^{-3}$	11.06	11.06	11.08	11.08	3.64	3.64	3.64	3.64
ASE$\times 10^{-3}$	14.19	14.16	14.15	14.13	4.49	4.49	4.49	4.49
ESE$\times 10^{-3}$	13.82	13.83	13.85	13.85	4.51	4.51	4.51	4.51
CP	0.956	0.955	0.952	0.952	0.949	0.947	0.946	0.946
C.Time(s)	0.033	0.028	0.019	0.023	0.69	0.25	0.28	0.18
R.Time(s)	0.030	0.024	0.015	0.019	0.58	0.25	0.16	0.14
	$B = 10^3, N_B = 10^5$				$B = 10^4, N_B = 10^6$			
	GEE		QIF		GEE		QIF	
Criterion	offline	Renew	offline	Renew	offline	Renew	offline	Renew
A.bias$\times 10^{-3}$	1.11	1.11	1.11	1.11	0.35	0.35	0.35	0.35
ASE$\times 10^{-3}$	1.42	1.42	1.42	1.42	0.45	0.45	0.45	0.45
ESE$\times 10^{-3}$	1.40	1.40	1.40	1.40	0.44	0.44	0.44	0.44
CP	0.952	0.954	0.952	0.952	0.955	0.955	0.955	0.955
C.Time(s)	15.38	5.70	8.57	4.26	781.46	62.30	704.11	51.38
R.Time(s)	8.72	2.96	1.90	2.18	99.21	32.12	21.85	25.21

Table 3. Compare renewable estimators and offline ones in the logistic MGLM model with fixed single batch size $n_b = 100$ and $p = 5$, where B increases from 10 to 10^4.

	$B = 10, N_B = 10^3$				$B = 100, N_B = 10^4$			
	GEE		QIF		GEE		QIF	
Criterion	offline	Renew	offline	Renew	offline	Renew	offline	Renew
A.bias$\times 10^{-3}$	25.92	25.82	26.07	26.01	8.17	8.16	8.17	8.16
ASE$\times 10^{-3}$	33.08	33.06	33.03	33.07	10.45	10.45	10.45	10.45
ESE$\times 10^{-3}$	32.48	32.36	32.67	32.60	10.31	10.30	10.32	10.31
CP	0.953	0.952	0.950	0.952	0.950	0.952	0.951	0.952
C.Time(s)	0.048	0.029	0.024	0.023	1.09	0.23	0.32	0.17
R.Time(s)	0.045	0.026	0.021	0.020	0.99	0.20	0.22	0.14
	$B = 10^3, N_B = 10^5$				$B = 10^4, N_B = 10^6$			
	GEE		QIF		GEE		QIF	
Criterion	offline	Renew	offline	Renew	offline	Renew	offline	Renew
A.bias$\times 10^{-3}$	2.71	2.71	2.71	2.71	–	0.82	0.82	0.82
ASE$\times 10^{-3}$	3.31	3.31	3.31	3.31	–	1.05	1.05	1.05
ESE$\times 10^{-3}$	3.39	3.39	3.39	3.39	–	1.04	1.04	1.04
CP	0.948	0.948	0.948	0.948	–	0.946	0.946	0.948
C.Time(s)	22.41	5.84	9.99	4.49	–	57.47	856.70	47.44
R.Time(s)	15.59	3.02	3.18	2.35	–	31.08	45.83	21.31

NOTE: The dashed line in the column for "offline GEE" when $N_B = 10^6$ indicates the standard `gee` package in R does not produce output due to the excessive computational burden.

the asymptotic properties in Theorem 4: their average absolute bias decreases rapidly as the cumulative sample size accumulates, and the coverage probability stays robustly around the nominal level 95%.

Computation time. Both online RenewGEE and RenewQIF methods show more and more advantageous as N_B increases: the combined amount of time for data loading and algorithm execution only takes less than 5 sec, whereas the offline GEE and offline QIF, when processing a dataset of 10^5 clusters once, requires more than 20 sec. This gain of 5-fold faster computation by the proposed RenewGEE and RenewQIF methods sacrifice little price of estimation precision or inferential power. One thing worth mentioning for Table 3 is that when $N_B = 10^6$, to run the logistic MGLM, the offline GEE is computationally too intensive to produce convergent results within 12 hr using the existing R package `gee`.

4.3. Evaluation of Monitoring Procedure

We also evaluate the performance of the proposed diagnostic procedure using the goodness-of-fit test Λ_b in Equation (10) to detect abnormal data batches. First, we check the properties of this test statistic with respect to both Type I error and power of detection abnormal data batches. Then, we compare the estimation and inference performance of the RenewQIF methods with and without the use of monitoring procedure in terms of the following four criterion: (a) A.bias, (b) ASE, (c) ESE, and (d) CP, as define above. The abnormal data batches are created by altering the true parameters via a local departure on β_{02}, that is $\boldsymbol{\beta}_\tau = (0.2, -(0.2 + d), 0.2, -0.2, 0.2)^\top$, $\tau \in \Gamma_q$. We set Γ_2, containing two positions ($q = 2$) at which two abnormal data batches occurs, respectively, at $\tau_1 = 0.25B$ and $\tau_2 = 0.75B$. Simulation studies have showed that Type I errors are very close

Table 4. Performances with and without the monitoring procedure.

	Without monitoring procedure			
n_b	50	100	200	400
A.bias$\times 10^{-3}$	12.36	16.74	28.68	56.58
ASE$\times 10^{-3}$	14.61	14.62	14.69	14.79
ESE$\times 10^{-3}$	14.25	14.02	15.26	15.03
CP	0.932	0.838	0.534	0.030

	With monitoring procedure									
	$n_b = 50$					$n_b = 100$				
$\alpha \times 10^{-3}$	100	50	10	1	0.005	100	50	10	1	0.005
A.bias$\times 10^{-3}$	**11.29**	8.923	7.970	8.312	8.802	9.773	8.433	7.962	8.147	9.564
ASE$\times 10^{-3}$	10.18	9.558	9.259	9.220	9.219	10.21	9.604	9.311	9.263	9.256
ESE$\times 10^{-3}$	20.90	12.08	9.860	9.782	9.568	16.67	11.37	9.834	9.340	9.736
CP	0.852	0.896	0.920	0.906	0.890	0.926	0.942	0.942	0.940	0.882
N_0/N_B	0.876	0.939	0.988	0.997	0.999	0.883	0.933	0.976	0.987	0.992
	$n_b = 200$					$n_b = 400$				
$\alpha \times 10^{-3}$	100	50	10	1	0.005	100	50	10	1	0.005
A.bias$\times 10^{-3}$	8.836	8.122	7.485	7.757	9.641	9.317	8.371	7.723	7.656	7.771
ASE$\times 10^{-3}$	10.16	9.709	9.426	9.369	9.347	10.39	9.958	9.638	9.584	9.573
ESE$\times 10^{-3}$	11.73	10.41	9.305	9.551	10.40	12.41	10.79	9.690	9.612	9.795
CP	0.938	0.948	0.956	0.946	0.886	0.934	0.948	0.958	0.956	0.952
N_0/N_B	0.863	0.913	0.952	0.962	0.972	0.823	0.872	0.911	0.918	0.920

NOTE: Fixed total number of samples $N_B = 10^4$ with varying data batch size n_b. $\tau_1 = 0.25B$ and $\tau_2 = 0.75B$. In the table "With monitoring procedure," N_0/N_B denotes the proportion of used samples in the renewable estimation and inference.

to the nominal level α, and that the power of detecting abnormal data batches drops as α becomes smaller, but increases as n_b increases. See Table 1 in the supplementary materials.

Without monitoring procedure. With fixed $N_B = 10^4$ and $\Gamma_2 = \{0.25B, 0.75B\}$, the upper panel in Table 4 shows that larger data batch size n_b leads to a larger bias due to the increased number of contaminated clusters generated from the incompatible data model. A.bias increases almost linearly with n_b. At similar levels of ASE and ESE, the coverage probability deviates more from the nominal level 95% as n_b increases; it drops from 93.2% to 3.0% as n_b rises from 50 to 200 due to more severe data contamination.

With monitoring procedure. For the purpose of quality control, larger α increases the sensitivity of rejection, so many small departures may be detected, which would be consequently ignored in the online updating. These are clearly shown in the lower panel of Table 4 due to the reduced proportion of used samples, defined by N_0/N_B. See also the last subplot in Figure 1 in the supplementary materials where N_0 is the number of clusters that passed the diagnostic test. The price to pay in this case is that the resulting bias and standard error would be larger than they would be if the false positives may be avoided. In contrast, choosing a small α may elevate Type II error (false negative) and thus, can lose power in detecting abnormal data batches. In this case, the price to pay is not only increased bias but also decreased coverage probability. The latter is indeed a more serious problem as far as inference concerns. This phenomenon is evident when n_b is small as shown in Table 4. As an extreme case of $\alpha = 5 \times 10^{-6}$, the detection power is greatly lost with $n_b = 200$, 100, or 50, and the coverage probability reduces to lower than 90%. In practice, with high throughput data streams, where cumulative sample sizes increase rapidly, using a larger α, say $\alpha = 0.05$, is much safer and recommended in practice for the purpose of monitoring, resulting in a more protective scenario by effectively and cautiously avoiding abnormal data batches.

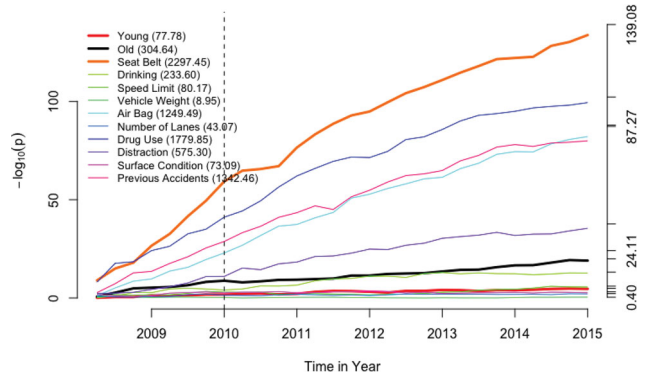

Figure 4. Trace plots of $-\log_{10}(p)$ over quarterly data batches from January, 2009 to December, 2015, each for one regression coefficient. Dashed vertical line indicates the location of detected abnormal data batch.

5. Analysis of NASS CDS Data

In regard to injuries involved in car accidents, we are interested in not only the extent of injuries in drivers but also in passengers. Apparently, injury levels of driver and passengers within the same vehicle are correlated, and such within-cluster correlation needs to be taken into account in the analysis. In this real data application, we focus on the analysis of a series of car crash datasets from the National Automotive Sampling System-Crashworthiness Data System (NASS CDS) from January, 2009 to December, 2015. Our primary interest was to evaluate the effectiveness of graduated driver licensing (GDL) on overall driving safety with respect to injury levels in both driver and

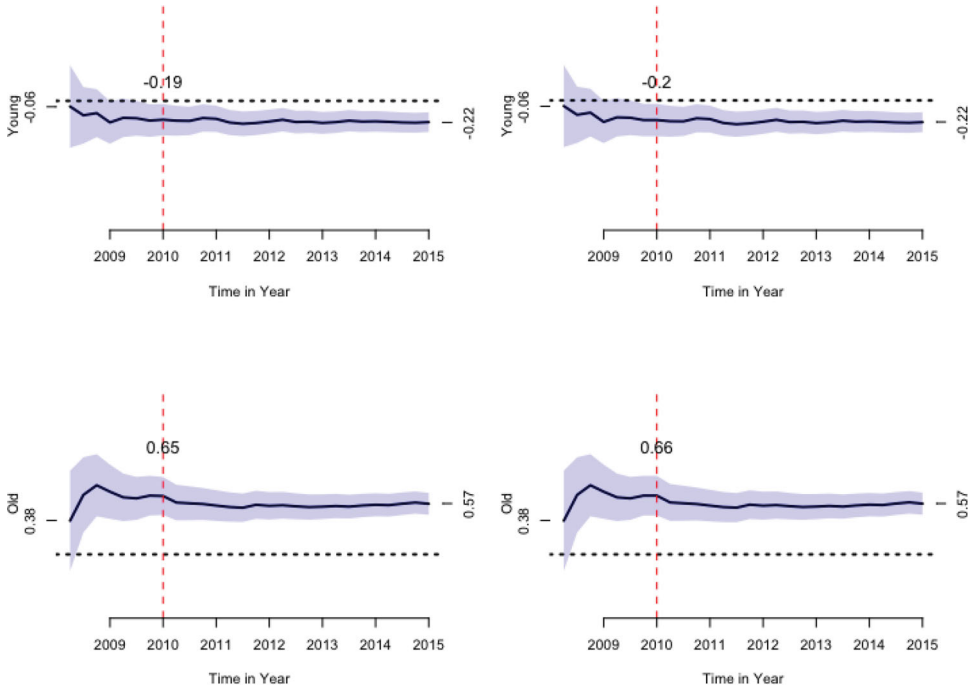

Figure 5. Trace plots for the estimates of coefficients and 95% point-wise confidence bands of "Young" and "Old." Numerical numbers on two sides denote the estimated regression coefficients after the arrivals of first and last batches, while the ones above the traces denote the estimates at the eighth data batch.

passengers. GDL is a nationwide legislature on novice drivers of age 21 or younger with various conditions of vehicle operation. In contrast, under the current law, there are no restrictions on vehicle operation for older drivers with age, *say*, older than 65. Thus, we want to compare drivers' age groups with respect to the extent of injury when a car accident happens. We first categorized the "Age" variable into three age groups: "Age<21" represents the young group under a restricted GDL, and "Age≥65" indicates the old group with a regular full driver's license, while those of age in between is treated as the reference group. Extent of injury in a crash is a binary variable with 1 for a moderate or severe injury, and 0 for minor or no injury. This outcome variable was created from the variable of Maximum Known Occupant Ais (MAIS), which indicates the single most severe injury level reported for each occupant. Other potential risk factors are also considered in the model, including seat belt use (Seat Belt, 1 for used and 0 for no), drinking (Drinking, 1 for yes and 0 for no), speed limit (Speed Limit), vehicle weight (Vehicle Weight, 0 for $\leq$ 3000, 1 for 3000∼4000, 2 for $\geq$4000), air bag system deployed (Air Bag, 1 for yes and 0 for no), number of lanes (Number of Lanes, 0 for $\leq$ 2 and 1 for else), drug involvement in this accident (Drug Use, 1 for yes and 0 for no), driver's distraction/inattention to driving (Distraction, 1 for attentive and 0 for else), roadway surface condition (Surface Condition, 1 for dry and 0 for else), and vehicle has been in previous accidents (Previous Accidents, 0 for no and 1 for else).

Streaming data are formed by quarterly accident data from the period of 7 years from January, 2009 to December, 2015, with $B = 28$ data batches and a total of $N_B = 18{,}832$ crashed vehicles that contain 26,330 occupants with complete records. Each vehicle is treated as a cluster, and the cluster size varies from 1 to 10 with an average of two occupants. We invoke RenewQIF to fit a marginal logistic regression model with the compound symmetry correlation to account for the within vehicle correlation. In the analysis, we are interested in a 7-year average risk assessment and thus, assuming constant associations between extent of injuries and risk factors over time. Since this sequence of data streams arrive at low speed and large data batch size, we would expect to have high power to detect the abnormal data batch even if we choose a small α, say $\alpha = 0.01$. Additionally, since samples from NASS CDS have gone through extensive data cleaning and preprocessing steps, a stringent α is a reasonable choice to make a full use of samples. At $\alpha = 0.01$, our proposed monitoring procedure identifies data batch $\mathcal{D}_8$ to be incompatible, corresponding to the data collected during the fourth quarter in year 2010. Estimated coefficients, standard errors and p-values are reported in Table 2 in the supplementary materials.

Figure 4 shows the trajectories of $-\log_{10}(p)$ values of the online Wald test developed for RenewQIF, each for one regression coefficient over 28 quarters. Even though the total sample size N_B increases over time, not all of them show steep monotonic increasing trends in evidence against the null $H_0 : \beta_j = 0$. "Seat Belt" turns out to have the strongest association to the odds of injury in a crash among all covariates included in the model. This is an overwhelming confirmation to the enforcement of policy "buckle up" when sitting in a moving vehicle. For the convenience of comparison, we report a summary statistic as of the area under the p-value curve for each covariate. "Seat Belt" (2297.45), "Drug Use" (1779.85), "Air Bag" (1249.49) and "Previous Accidents" (1342.46) appear well separated from the other risk factors. Their ranking is well aligned with the ranking of p-values obtained at the end time of streaming data availability, namely December, 2015.

The trajectories of both young and old age groups are evaluated in Figure 5 with or without data batch $\mathcal{D}_8$. The trace of

the estimators for the young age group (Age<21) stays below 0 over the 28-quarter period, indicating that it has lower adjusted odds of moderate/severe injury than the reference age group. This finding confirms the effectiveness of GDL in protecting young novice drivers. Unfortunately, in contrast, the old age group (Age≥65) turns out to suffer from significantly higher adjusted odds of moderate/severe injury outcomes comparing to the middle age group. This suggests a need of policy-making to protect older drivers from injuries when an accident happens. Furthermore, the abnormal data batch seems to affect marginally the estimates for age groups if we compare the plots with (right) and without (left) monitoring procedure (see the red vertical dashed line).

Applying the proposed RenewQIF to the above CDS data analysis enables us to visualize time-course patterns of data evidence accrual as well as stability and reproducibility of inference. As shown in Figure 5, at the early stage of data streams, due to limited sample sizes and possibly sampling bias, both parameter estimates and test power may be unstable and even possibly misleading. These potential shortcomings can be overcome when estimates and inferential quantities are continuously updated along with data streams, which eventually reach stability and reliable conclusions.

6. Concluding Remarks

Due to the advantage of technologies in data storage and data collection, streaming data arise from many practical areas such as healthcare (Peek, Holmes, and Sun 2014). Healthcare data are typically measurements in forms of clusters or time series, such as patients from the same clinic, or data from personal wearable devices (Sahoo et al. 2016). The traditional methods for clustered/longitudinal data analysis such as generalized estimating equations (GEE) and quadratic inference functions (QIF) that process the entire dataset once may be greatly challenged due to the following reasons: (i) they become computationally prohibitive as the total sample size accumulates too fast and too large, so to exceed the available computational power; see Table 3 where GEE fails to produce numerical outputs; and (ii) historical subject-level data may no longer be accessible due to storage, time, or privacy issues. This type of problem has been extensively tackled in the framework of online updating where stochastic gradient descent algorithms are the primary methods of choice to provide fast updating with no use of historical data. However, most online learning algorithms have not considered statistical inference or the diagnosis of contaminated data batches.

Such gaps have been filled in this article by the RenewQIF methodology. The proposed RenewQIF method provides a new paradigm of renewable estimation and incremental inference, in which parameter estimates are recursively updated with current data and inferential statistics of historical data, but does not require the accessibility to any historical subject-level data. To achieve efficient communications between current data and historical summary statistics, we design an extended Spark's Lambda architecture to execute both data storage and analysis updates. Both proposed statistical methodology and computational algorithms have been investigated for their theoretical guarantees and examined numerically via extensive simulation studies. The proposed RenewQIF has been shown to be much more computationally efficient with no loss of inferential power in comparison to the offline GEE or offline QIF. Additionally, we propose a diagnostic procedure to detect abnormal data batches in data streams for the proposed RenewQIF. The utilization of a goodness-of-fit test in the QIF framework enabled us to check the compatibility of current data batch with a normal reference effectively and efficiently. The proposed monitoring procedure has been integrated into an extended Lambda architecture with an additional monitoring layer.

The formulation of RenewQIF is under the assumption that clusters arrive independently over data streams, and when cluster size $m = 1$, it reduces to generalized linear model as a special case. A direction of interest is to consider the case of intercorrelated batches, such as serially dependent data streams generated by individual wearable devices. Such types of data streams are pervasive in healthcare where thousands of physiological measurements are recorded per second, such as body temperature, heart rate, respiratory rate and blood pressure (Priyanka and Kulennavar 2014). Therefore, developing the analytic tools for the analysis of serially dependent data streams is an important future research as part of new analytics for handling massive data volumes and making behavioral interventions.

Appendix

A.1. Consistency

From Equation (7), we have

$$||\tilde{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_0||^2 = ||\tilde{\boldsymbol{\beta}}_{b-1} - \boldsymbol{\beta}_0||^2 + 2N_b^{-1}(\tilde{\boldsymbol{\beta}}_{b-1} - \boldsymbol{\beta}_0)^\top \boldsymbol{H}_b^{-1}\tilde{\boldsymbol{U}}_b(\tilde{\boldsymbol{\beta}}_b) + N_b^{-2}||\boldsymbol{H}_b^{-1}\tilde{\boldsymbol{U}}_b(\tilde{\boldsymbol{\beta}}_b)||^2 := I_1 + I_2 + I_3.$$

For I_2, we have

$$\begin{aligned}\mathbb{E}[I_2] = {} & 2\mathbb{E}\left\{N_b^{-1}(\tilde{\boldsymbol{\beta}}_{b-1} - \boldsymbol{\beta}_0)^\top \boldsymbol{H}_b^{-1}\boldsymbol{f}_b(\tilde{\boldsymbol{\beta}}_{b-1})\right\} \\ & + 2\mathbb{E}\left\{N_b^{-1}(\tilde{\boldsymbol{\beta}}_{b-1} - \boldsymbol{\beta}_0)^\top \boldsymbol{H}_b^{-1} \right. \\ & \times \left[\tilde{\boldsymbol{G}}_b^\top \tilde{\boldsymbol{C}}_b^{-1} - \left(\tilde{\boldsymbol{G}}_{b-1} + \boldsymbol{G}_b(\tilde{\boldsymbol{\beta}}_{b-1})\right)^\top \right. \\ & \left.\left.\left(\tilde{\boldsymbol{C}}_{b-1} + \boldsymbol{C}_b(\tilde{\boldsymbol{\beta}}_{b-1})\right)^{-1}\right]\left(\tilde{\boldsymbol{g}}_{b-1}(\tilde{\boldsymbol{\beta}}_{b-1}) + \boldsymbol{g}_b(\tilde{\boldsymbol{\beta}}_{b-1})\right)\right\} \\ & + 2\mathbb{E}\left\{N_b^{-1}(\tilde{\boldsymbol{\beta}}_{b-1} - \boldsymbol{\beta}_0)^\top \boldsymbol{H}_b^{-1}\tilde{\boldsymbol{G}}_b^\top \tilde{\boldsymbol{C}}_b^{-1} \right. \\ & \left.\left(\boldsymbol{g}_b(\tilde{\boldsymbol{\beta}}_b) - \boldsymbol{g}_b(\tilde{\boldsymbol{\beta}}_{b-1})\right)\right\} := I_{21} + I_{22} + I_{23}.\end{aligned}$$

Denote $z_{b-1} := \boldsymbol{H}_b^{-1/2}\tilde{\boldsymbol{\beta}}_{b-1}$. Assume there exists a function $\tilde{\ell}_b(z_{b-1})$ such that $\nabla\tilde{\ell}(z_{b-1}) = \partial\tilde{\ell}_b(z_{b-1})/\partial z_{b-1}^\top := -N_b^{-1}\boldsymbol{H}_b^{-1/2}\boldsymbol{f}_b(\tilde{\boldsymbol{\beta}}_{b-1})$. Then the second order derivative of $\tilde{\ell}_b(z_{b-1})$ is

$$\begin{aligned}\nabla^2\tilde{\ell}_b(z_{b-1}) := \frac{\partial^2\tilde{\ell}_b(z_{b-1})}{\partial z_{b-1}\partial z_{b-1}^\top} &= N_b^{-1}\boldsymbol{H}_b^{-1/2}\frac{-\partial \boldsymbol{f}_b(\boldsymbol{H}_b^{1/2}z_{b-1})}{\partial z_{b-1}^\top} \\ &= N_b^{-1}\boldsymbol{H}_b^{-1/2}\frac{-\partial \boldsymbol{f}_b(\boldsymbol{H}_b^{1/2}z_{b-1})}{\partial (\boldsymbol{H}_b^{1/2}z_{b-1})^\top}\frac{\partial (\boldsymbol{H}_b^{1/2}z_{b-1})^\top}{\partial z_{b-1}^\top} \\ &= N_b^{-1}\boldsymbol{H}_b^{-1/2}\left[\left(\tilde{\boldsymbol{G}}_{b-1} + \boldsymbol{G}_b(\tilde{\boldsymbol{\beta}}_{b-1})\right)^\top\right.\end{aligned}$$

$$\left(\tilde{\boldsymbol{C}}_{b-1}+\boldsymbol{C}_b(\tilde{\boldsymbol{\beta}}_{b-1})\right)^{-1}\left(\tilde{\boldsymbol{G}}_{b-1}+\boldsymbol{G}_b(\tilde{\boldsymbol{\beta}}_{b-1})\right)\Big]\boldsymbol{H}_b^{1/2}+O_p\left(N_b^{-1}||\boldsymbol{f}_b(\tilde{\boldsymbol{\beta}}_{b-1})||\right).$$

By the continuity of $\boldsymbol{f}_b(\cdot)$, $\boldsymbol{f}_b(\tilde{\boldsymbol{\beta}}_b)=\mathbf{0}$ and $||\tilde{\boldsymbol{\beta}}_b-\tilde{\boldsymbol{\beta}}_{b-1}||=O_p(N_{b-1}^{-1/2})$ in Lemma 4, we have $||\boldsymbol{f}_b(\tilde{\boldsymbol{\beta}}_{b-1})||=o_p(1)$. It implies that $\nabla^2\tilde{\ell}_b(z_{b-1})$ is positive-definite and $\tilde{\ell}_b(z_{b-1})$ is a strong convex function. By Condition 2.3, we have

$$\begin{aligned}\mathbb{E}[\tilde{\ell}_b(z_0)]&=\mathbb{E}\Big[\tilde{\ell}_b(z_{b-1})+(z_0-z_{b-1})^\top\nabla\tilde{\ell}_b(z_{b-1})\\&\quad+\frac{1}{2}(z_{b-1}-z_0)^\top\nabla^2\tilde{\ell}_b(z_m)(z_{b-1}-z_0)\Big]\\&\le\mathbb{E}[\tilde{\ell}_b(z_{b-1})],\end{aligned}$$

where $z_m:=\boldsymbol{H}_b^{-1/2}\boldsymbol{\beta}_m$ with $\boldsymbol{\beta}_m=t\tilde{\boldsymbol{\beta}}_{b-1}+(1-t)\boldsymbol{\beta}_0$ for some $t\in(0,1)$, and $z_0:=\boldsymbol{H}_b^{-1/2}\boldsymbol{\beta}_0$. It follows that

$$\begin{aligned}I_{21}&=-2\mathbb{E}\left\{(z_{b-1}-z_0)^\top\nabla\tilde{\ell}_b(z_{b-1})\right\}\\&\le-\mathbb{E}\left\{(z_{b-1}-z_0)^\top\nabla^2\ell_b(z_m)(z_{b-1}-z_0)\right\}\\&\le-\lambda_{\min}\left(\nabla^2\ell_b(z_m)\right)\mathbb{E}(||z_{b-1}-z_0||^2)\\&\le-\frac{\lambda_{\min}(\nabla^2\ell_b(z_m))}{2\lambda_{\max}(\boldsymbol{H}_b)}\mathbb{E}(||\tilde{\boldsymbol{\beta}}_{b-1}-\boldsymbol{\beta}_0||^2),\end{aligned}\tag{A.1}$$

where $\lambda_{\min}$ and λ_{max} deonte the smallest and largest eigenvalues. Furthermore, since $\boldsymbol{g}_b$ is twice continuous differentiable w.r.t. $\boldsymbol{\beta}\in\Theta$, we have $\boldsymbol{G}_b(\tilde{\boldsymbol{\beta}}_b)=\boldsymbol{G}_b(\boldsymbol{\beta}_m)+O_p(n_b)$ and $\boldsymbol{C}_b(\tilde{\boldsymbol{\beta}}_b)=\boldsymbol{C}_b(\boldsymbol{\beta}_m)+O_p(n_b)$. It follows that

$$\begin{aligned}\nabla^2\tilde{\ell}_b(z_m)&=N_b^{-1}\boldsymbol{H}_b^{-1/2}\Big[\left(\tilde{\boldsymbol{G}}_{b-1}+\boldsymbol{G}_b(\boldsymbol{\beta}_m)\right)^\top\\&\quad\left(\tilde{\boldsymbol{C}}_{b-1}+\boldsymbol{C}_b(\boldsymbol{\beta}_m)\right)^{-1}\left(\tilde{\boldsymbol{G}}_{b-1}+\boldsymbol{G}_b(\boldsymbol{\beta}_m)\right)\Big]\boldsymbol{H}_b^{1/2}\\&\quad+o_p(1/N_b)\\&=\boldsymbol{H}_b^{-1/2}\Big[N_b^{-1}\left(\tilde{\boldsymbol{G}}_{b-1}+\boldsymbol{G}_b(\tilde{\boldsymbol{\beta}}_b)\right)^\top\\&\quad\left(\tilde{\boldsymbol{C}}_{b-1}+\boldsymbol{C}_b(\tilde{\boldsymbol{\beta}}_b)\right)^{-1}\tilde{\boldsymbol{G}}_{b-1}+O_p(n_b/N_b)\Big]\boldsymbol{H}_b^{1/2}\\&\quad+o_p(1/N_b)\\&=\boldsymbol{H}_b^{-1/2}\boldsymbol{H}_b\boldsymbol{H}_b^{1/2}+O_p(n_b/N_b)=\boldsymbol{H}_b+O_p(N_b^{-1}).\end{aligned}$$

Therefore, Equation (A.1) becomes

$$I_{21}\le-\frac{\lambda_{\min}(\boldsymbol{H}_b)}{2\lambda_{\max}(\boldsymbol{H}_b)}\mathbb{E}(||\tilde{\boldsymbol{\beta}}_{b-1}-\boldsymbol{\beta}_0||^2)+O_p(N_b^{-1}),\tag{A.2}$$

For I_{22} and I_{23}, we apply $||\tilde{\boldsymbol{\beta}}_b-\tilde{\boldsymbol{\beta}}_{b-1}||=O_p(N_{b-1}^{-1/2})$ in Lemma 4 which leads to

$$\begin{aligned}I_{22}&\le c_1'\mathbb{E}\Big\{||\tilde{\boldsymbol{\beta}}_{b-1}-\boldsymbol{\beta}_0||||\boldsymbol{H}_b^{-1}||||\tilde{\boldsymbol{\beta}}_b-\tilde{\boldsymbol{\beta}}_{b-1}||\\&\quad\Big\|N_b^{-1}\left(\tilde{\boldsymbol{g}}_{b-1}(\tilde{\boldsymbol{\beta}}_{b-1})+\boldsymbol{g}_b(\tilde{\boldsymbol{\beta}}_{b-1})\right)\Big\|\Big\}\\&\le c_1N_b^{-1/2}\mathbb{E}(||\tilde{\boldsymbol{\beta}}_{b-1}-\boldsymbol{\beta}_0||),\\I_{23}&\le c_3'N_b^{-1}\mathbb{E}\left(||\tilde{\boldsymbol{\beta}}_{b-1}-\boldsymbol{\beta}_0||||\boldsymbol{H}_b^{-1}||||\tilde{\boldsymbol{\beta}}_b-\tilde{\boldsymbol{\beta}}_{b-1}||\right)\\&\le c_3N_b^{-3/2}\mathbb{E}(||\tilde{\boldsymbol{\beta}}_{b-1}-\boldsymbol{\beta}_0||).\end{aligned}\tag{A.3}$$

where c_1, c_1', c_3 and c_3' are constants that do not depend on b.

Combining inequalities (A.2) and (A.3), we have

$$\mathbb{E}[I_2]\le-\frac{\lambda_{\min}(\boldsymbol{H}_b)}{2\lambda_{\max}(\boldsymbol{H}_b)}\mathbb{E}(||\tilde{\boldsymbol{\beta}}_{b-1}-\boldsymbol{\beta}_0||^2)+c_1N_b^{-1/2}\mathbb{E}(||\tilde{\boldsymbol{\beta}}_{b-1}-\boldsymbol{\beta}_0||^2).$$

Similarly, for I_3 we have $\mathbb{E}[I_3]=N_b^{-2}\mathbb{E}\left\|\boldsymbol{f}_b(\tilde{\boldsymbol{\beta}}_b)-\tilde{\boldsymbol{G}}_b^\top\tilde{\boldsymbol{C}}_b^{-1}\tilde{\boldsymbol{G}}_{b-1}(\tilde{\boldsymbol{\beta}}_{b-1}-\tilde{\boldsymbol{\beta}}_b)\right\|^2\le c_2N_b^{-1}$, where c_2 is a constant. Putting together, the error bound is

$$\begin{aligned}\mathbb{E}\left(||\tilde{\boldsymbol{\beta}}_b-\boldsymbol{\beta}_0||^2\right)&\le\left(1-\frac{\lambda_{\min}(\boldsymbol{H}_b)}{2\lambda_{\max}(\boldsymbol{H}_b)}+c_1N_b^{-1/2}\right)\\&\quad\mathbb{E}\left(||\tilde{\boldsymbol{\beta}}_{b-1}-\boldsymbol{\beta}_0||^2\right)+c_2N_b^{-1}.\end{aligned}$$

Therefore, by $0<\lambda_{\min}(\boldsymbol{H}_b)\le\lambda_{\max}(\boldsymbol{H}_b)$, we have $\mathbb{E}\left(||\tilde{\boldsymbol{\beta}}_b-\boldsymbol{\beta}_0||^2\right)\to0$.

A.2. Asymptotic Normality

For the sake of simplicity, we only show a sketch of the proof here. For more technical details on finding the recursion, please refer to Sections 1 and 2 in the supplementary materials. From Equation (5) in the supplementary materials, we first obtain the recursive relationship between $(\tilde{\boldsymbol{\beta}}_b-\boldsymbol{\beta}_0)$ and $(\tilde{\boldsymbol{\beta}}_{b-1}-\boldsymbol{\beta}_0)$:

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_b-\boldsymbol{\beta}_0&=\mathcal{A}_{b,b}^{-1}\mathcal{A}_{b,b-1}(\tilde{\boldsymbol{\beta}}_{b-1}-\boldsymbol{\beta}_0)+\sum_{j=1}^{b-2}\mathcal{A}_{b,b}^{-1}\mathcal{A}_{b,j}(\tilde{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_0)\\&\quad+\mathcal{A}_{b,b}^{-1}N_b^{-1}\boldsymbol{H}_b^{-1}\mathcal{M}_b\left\{\sum_{j=1}^{b}\boldsymbol{g}_j(\boldsymbol{\beta}_0)\right\}.\end{aligned}$$

Let $\tilde{\mathcal{A}}_{b,j}^{(1)}:=\mathcal{A}_{b,b-1}\mathcal{A}_{b-1,b-1}^{-1}\mathcal{A}_{b-1,j}+\mathcal{A}_{b,j}$ for $1\le j\le b-2$, and a one-step back to $(\tilde{\boldsymbol{\beta}}_{b-2}-\boldsymbol{\beta}_0)$ is

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_b-\boldsymbol{\beta}_0&=\mathcal{A}_{b,b}^{-1}\tilde{\mathcal{A}}_{b,b-2}^{(1)}(\tilde{\boldsymbol{\beta}}_{b-2}-\boldsymbol{\beta}_0)+\sum_{j=1}^{b-3}\mathcal{A}_{b,b}^{-1}\tilde{\mathcal{A}}_{b,j}^{(1)}(\tilde{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_0)\\&\quad+\mathcal{A}_{b,b}^{-1}N_b^{-1}\boldsymbol{H}_b^{-1}\mathcal{M}_b\boldsymbol{g}_b(\boldsymbol{\beta}_0)\\&\quad+\mathcal{A}_{b,b}^{-1}\Big(\mathcal{A}_{b,b-1}\mathcal{A}_{b-1,b-1}^{-1}N_{b-1}^{-1}\boldsymbol{H}_{b-1}^{-1}\mathcal{M}_{b-1}\\&\quad+N_b^{-1}\boldsymbol{H}_b^{-1}\mathcal{M}_b\Big)\left\{\sum_{j=1}^{b-1}\boldsymbol{g}_j(\boldsymbol{\beta}_0)\right\}.\end{aligned}$$

Similarly, by recursively defining $\tilde{\mathcal{A}}_{b,j'}^{(b-j-1)}:=\tilde{\mathcal{A}}_{b,j+1}^{(b-j-2)}\mathcal{A}_{j+1,j+1}^{-1}\mathcal{A}_{j+1,j'}+\tilde{\mathcal{A}}_{b,j'}^{(b-j-2)}$ for $1\le j'\le j\le b-2$, $\tilde{\mathcal{A}}_{b,b-1}^{(0)}:=\mathcal{A}_{b,b-1}$ and $\tilde{\mathcal{A}}_{b,b}^{(-1)}:=\mathcal{A}_{b,b}$, we finally obtain that

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_b-\boldsymbol{\beta}_0=&\mathcal{A}_{b,b}^{-1}\tilde{\mathcal{A}}_{b,1}^{(b-2)}(\tilde{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_0)+\mathcal{A}_{b,b}^{-1}\\&\left\{\left(\sum_{j'=2}^{b}\tilde{\mathcal{A}}_{b,j'}^{(b-j'-1)}\mathcal{A}_{j',j'}^{-1}N_{j'}^{-1}\boldsymbol{H}_{j'}^{-1}\mathcal{M}_{j'}\right)\boldsymbol{g}_1(\boldsymbol{\beta}_0)\right\}\\&+\mathcal{A}_{b,b}^{-1}\sum_{j=2}^{b}\left\{\left(\sum_{j'=j}^{b}\tilde{\mathcal{A}}_{b,j'}^{(b-j'-1)}\mathcal{A}_{j',j'}^{-1}N_{j'}^{-1}\boldsymbol{H}_{j'}^{-1}\mathcal{M}_{j'}\right)\boldsymbol{g}_j(\boldsymbol{\beta}_0)\right\}.\end{aligned}\tag{A.4}$$

Let $a_j = \sum_{j'=j}^{b} \tilde{\mathcal{A}}_{b,j'}^{(b-j'-1)} \mathcal{A}_{j',j'}^{-1} N_{j'}^{-1} \boldsymbol{H}_{j'}^{-1} \mathcal{M}_{j'}$, then we have

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_0 &= \mathcal{A}_{b,b}^{-1}\tilde{\mathcal{A}}_{b,1}^{(b-2)}(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_0) + \mathcal{A}_{b,b}^{-1}a_2\boldsymbol{g}_1(\boldsymbol{\beta}_0) \\ &\quad + \mathcal{A}_{b,b}^{-1}\sum_{j=2}^{b} a_j\boldsymbol{g}_j(\boldsymbol{\beta}_0) \\ &= \mathcal{A}_{b,b}^{-1}\Bigg[\tilde{\mathcal{A}}_{b,1}^{(b-2)}\left\{\int_0^1 -\nabla_{\boldsymbol{\beta}}\boldsymbol{f}_1(\boldsymbol{\beta}_{m,1})d\delta_1\right\}^{-1} \\ &\quad \boldsymbol{G}_1(\boldsymbol{\beta}_0)^\top \boldsymbol{C}_1(\boldsymbol{\beta}_0)^{-1} + a_2\Bigg]\boldsymbol{g}_1(\boldsymbol{\beta}_0) \\ &\quad + \mathcal{A}_{b,b}^{-1}\sum_{j=2}^{b} a_j\boldsymbol{g}_j(\boldsymbol{\beta}_0). \end{aligned} \tag{A.5}$$

Since $\boldsymbol{H}_b = O_p\left(\frac{N_{b-1}}{N_b}\right)$, $\mathcal{A}_{b,b} = O_p\left(\frac{N_b}{N_{b-1}}\right)$, and $\mathcal{A}_{b,j} = O_p\left(\frac{n_j}{N_{b-1}}||\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_0||\right)$ for $j \le b-1$. After some calculations, we have

$$a_j = O_p\left\{\frac{1}{N_{b-1}}\prod_{k=1}^{b-j}\left(1 + \frac{n_{b-k}||\tilde{\boldsymbol{\beta}}_{b-k} - \boldsymbol{\beta}_0||}{N_{b-k}}\right)\right\},\ 2 \le j \le b.$$

Then the order of the terms in Equation (A.5) becomes

$$||\tilde{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_0|| = O_p\left\{\sqrt{n_1}a_2 + \frac{N_{b-1}}{N_b}\left(\sum_{j=2}^{b} n_j a_j^2\right)^{1/2}\right\}, \tag{A.6}$$

where the second term in Equation (A.6) is an upper bound on $||\sum_{j=2}^{b} a_j\boldsymbol{g}_j(\boldsymbol{\beta}_0)||$ by Cauchy–Schwarz inequality.

Since we have shown the consistency of $\tilde{\boldsymbol{\beta}}_b$ as $b \to \infty$, we further assume $||\tilde{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_0|| = O_p(N_b^{-\tau})$ for some $\tau > 0$. By Lemma 3, the product $\prod_{k=1}^{b-1}\left\{1 + n_{b-k}N_{b-k}^{-(1+\tau)}\right\}$ converges if and only if $\sum_{k=1}^{b-1} n_{b-k}N_{b-k}^{-(1+\tau)}$ converges. For $\tau > 0$, and uniformly bounded n_{b-k}'s in Scenario (S1), the over-harmonic series $\sum_{k=1}^{b-1} n_{b-k}N_{b-k}^{-(1+\tau)}$ converges, and it follows that $\prod_{k=1}^{b-1}\left\{1 + n_{b-k}N_{b-k}^{-(1+\tau)}\right\}$ converges. Note that for any $x \ge 0$, $\log(1+x) \le x$. Thus, we have

$$\sum_{k=1}^{b-1}\log\left(1 + n_{b-k}N_{b-k}^{-(1+\tau)}\right) \le \sum_{k=1}^{b-1} n_{b-k}N_{b-k}^{-(1+\tau)}.$$

Then, it follows that

$$\begin{aligned}\sum_{j=2}^{b} n_j a_j^2 &\le O_p\left[\frac{1}{N_{b-1}^2}\sum_{j=2}^{b} n_j \exp\left\{\left(\sum_{k=1}^{b-j} n_{b-k}N_{b-k}^{-(1+\tau)}\right)^2\right\}\right] \\ &\le O_p\left[\frac{1}{N_{b-1}^2}\sum_{j=2}^{b} n_j \exp\left\{\left(\sum_{k=1}^{b-2} n_{b-k}N_{b-k}^{-(1+\tau)}\right)^2\right\}\right] \\ &= O_p\left(\frac{N_b - n_1}{N_{b-1}^2}\right) = O_p(N_b^{-1}).\end{aligned}$$

Furthermore, since $a_2 = O_p(N_{b-2}^{-1})$, and according to (A.6) we have $||\tilde{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_0|| = O_p(N_b^{-1/2})$. Finally, using the expression (5) in supplementary materials and inequality (1) in Lemma 2, we have

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_0 &= \mathcal{A}_{b,b}^{-1}N_b^{-1}\boldsymbol{H}_b^{-1}\mathcal{M}_b\left\{\sum_{j=1}^{b}\boldsymbol{g}_j(\boldsymbol{\beta}_0)\right\} \\ &\quad + O_p\left(\left\|\sum_{j=2}^{b-2}\mathcal{A}_{b,b}^{-1}\mathcal{A}_{b,j}(\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_0)\right\|\right) \\ &= \mathcal{A}_{b,b}^{-1}N_b^{-1}\boldsymbol{H}_b^{-1}\mathcal{M}_b\left\{\sum_{j=1}^{b}\boldsymbol{g}_j(\boldsymbol{\beta}_0)\right\} \\ &\quad + O_p\left(\frac{1}{N_b}\log\frac{N_{b-2}}{n_1}\right).\end{aligned}$$

By the weak law of large numbers, the asymptotic covariance is

$$\begin{aligned}&N_b^{-1}\boldsymbol{H}_b^{-1}\mathcal{M}_b\left\{\sum_{j=1}^{b}\boldsymbol{C}_j(\boldsymbol{\beta}_0)\right\}\mathcal{M}_b^\top(\boldsymbol{H}_b^{-1})^\top \\ &= (\mathbb{G}^\top\mathbb{C}^{-1}\mathbb{G})^{-1}\mathbb{G}^\top\mathbb{C}^{-1}\mathbb{C}\mathbb{C}^{-1}\mathbb{G}(\mathbb{G}^\top\mathbb{C}^{-1}\mathbb{G})^{-1} + o_p(1) \\ &= (\mathbb{G}^\top\mathbb{C}^{-1}\mathbb{G})^{-1} + o_p(1).\end{aligned}$$

It then follows from the Slutsky's Theorem and the Central Limit Theorem,

$$\begin{aligned}\sqrt{N_b}(\tilde{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_0) &= \left\{\frac{1}{N_b}\sum_{j=1}^{b}\boldsymbol{G}_j(\boldsymbol{\beta}_0)\right\}^{-1}\left\{\frac{1}{\sqrt{N_b}}\sum_{j=1}^{b}\boldsymbol{g}_j(\boldsymbol{\beta}_0)\right\} \\ &\quad + O_p\left(\frac{\log b}{\sqrt{N_b}}\right) \xrightarrow{d} \mathcal{N}_p(\boldsymbol{0}, \mathbb{J}^{-1}(\boldsymbol{\beta}_0)),\end{aligned}$$

where $\mathbb{J}(\boldsymbol{\beta}_0) = \mathbb{G}^\top(\boldsymbol{\beta}_0)\mathbb{C}^{-1}(\boldsymbol{\beta}_0)\mathbb{G}(\boldsymbol{\beta}_0)$ is the Godambe information matrix.

Supplementary Materials

This file includes the proof of some useful lemmas in Section 1, additional details in the proof under scenario (S1) in Section 2, proof of large sample properties in scenarios (S2) and (S3) in Section 3, and the analysis of cumulative error bound in Section 4. In Section 5, we include one table and one figure from simulation studies and an additional table from real data analysis. Additionally in Section 6, we includes "Renewable GEE" with derivation of renewable estimation and incremental inference method in the generalized estimating equations. (PDF)

Acknowledgments

The authors are grateful to editor, associate editor and three anonymous reviewers for their insightful comments and constructive suggestions that help greatly improve the manuscript.

Funding

This research was partially supported by the National Science Foundation grants DMS1811734 and DMS2113564 to Song, and the National Natural Science (Nos. 11901470, 11931014 and 11829101) and the Fundamental Research Funds for the Central Universities (Nos. JBK190904 and JBK1806002) to Zhou.

References

Amin, R., and Search, A. J. (1991), "A Nonparametric Exponentially Weighted Moving Average Control Scheme," *Communications in Statistics – Simulation and Computation*, 20, 1049–1072. [2]

Amin, R. W., Reynolds, M. R., and Bakir, S. T. (1995), "Nonparametric Quality Control Charts based on the Sign Statistic," *Communications in Statistics – Theory and Methods*, 24, 1597–1623. [2]

Bordes, A., Bottou, L., and Gallinari, P. (2009), "SGD-QN: Careful Quasi-Newton Stochastic Gradient Descent," *Journal of Machine Learning Research*, 10, 1737–1754. [1]

Cho, H., and Qu, A. (2015), "Efficient Estimation for Longitudinal Data by Combining Large-Dimensional Moment Conditions," *Electronic Journal of Statistics*, 9, 1315–1334. [6]

Fang, Y. (2019), "Scalable Statistical Inference for Averaged Implicit Stochastic Gradient Descent," *Scandinavian Journal of Statistics*, 46, 987–1002. [1]

Godambe, V. P., and Kale, B. K. (1991), *Estimating Functions: An Overview*, Oxford: Oxford University Press. [4]

Goel, A. L., and Wu, S. M. (1971), "Determination of A.R.L. and a Contour Nomogram for Cusum Charts to Control Normal Mean," *Technometrics*, 13, 221–230. [2]

Han, P., and Song, P. X.-K. (2011), "A Note on Improving Quadratic Inference Functions Using a Linear Shrinkage Approach," *Statistics and Probability Letters*, 81, 438–445. [6]

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054. [2,3,5,7]

Hazan, E., Agarwal, A., and Kale, S. (2007), "Logarithmic Regret Algorithms for Online Convex Optimization," *Journal of Machine Learning Research*, 69, 169–192. [1]

Lai, T. L. (2004), "Likelihood Ratio Identities and Their Applications to Sequential Analysis," *Sequential Analysis*, 23, 467–497. [2]

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22. [2,3]

Lindsay, B. G., and Qu, A. (2003), "Inference Functions and Quadratic Score Tests," *Statistical Science*, 18, 394–410. [7]

Luo, L., and Song, P. X.-K. (2020), "Renewable Estimation and Incremental Inference in Generalized Linear Models with Streaming Datasets," *Journal of the Royal Statistical Society*, Series B, 82, 69–97. [1,2]

Nadungodage, C. H., Xia, Y., Li, F., Lee, J. J., and Ge, J. (2011), "StreamFitter: A Real Time Linear Regression Analysis System for Continuous Data Streams," *Database Systems for Advanced Applications*, 6588, 458–461. [2]

Nocedal, J., and Wright, S. J. (2006), *Numerical Optimization* (2nd ed.), New York: Springer. [7]

Page, E. S. (1954), "Continuous Inspection Schemes," *Biometrika*, 41, 100–115. [2]

Pearson, K. (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, 2, 559–572. [6]

Peek, N., Holmes, J. H., and Sun, J. (2014), "Technical Challenges for Big Data in Biomedicine and Health: Data Sources, Infrastructure, and Analytics," *IMIA Yearbook of Medical Informatics*, 9, 42–47. [13]

Pollak, M., and Siegmund, D. (1991), "Sequential Detection of a Change in a Normal Mean When the Initial Value is Unknown," *The Annals of Statistics*, 19, 394–416. [2]

Priyanka, K., and Kulennavar, N. (2014), "A Survey On Big Data Analytics in Health Care," *International Journal of Computer Science and Information Technologies*, 5, 5865–5868. [13]

Qu, A., Lindsay, B.G., and Li, B. (2000), "Improving Generalised Estimating Equations Using Quadratic Inference Functions," *Biometrika*, 87, 823–876. [2,3,4]

Qu, A., and Song, P. X.-K. (2004), "Assessing Robustness of Generalised Estimating Equations an Quadratic Inference Functions," *Biometrika*, 91, 447–459. [2]

Robbins, H., and Monro, S. (1951), "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, 22, 400–407. [1,4]

Roberts, S. W. (1966), "A Comparison of Some Control Chart Procedures," *Technometrics*, 8, 411–430. [2]

Sahoo, P. K., Mohapatra, S. K., and Wu, S.-L. (2016), "Analyzing Healthcare Big Data With Prediction for Future Health Condition," *IEEE Access*, 4, 9786–9799. [13]

Sakrison, D. J. (1965), "Efficient Recursive Estimation: Application to Estimating the Parameter of a Covariance Function," *International Journal of Engineering Science*, 3, 461–483. [4]

Schraudolph, N. N., Yu, J., and Günter, S. (2007), "A Stochastic Quasi-Newton Method for Online Convex Optimization," in *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AIstats)*, Workshop and Conference Proceedings (Vol. 2), eds. M. Meila and X. Shen, pp. 436–443. San Juan, Puerto Rico: Journal of Machine Learning Research. [1]

Shiryayev, A. N. (1963), "On Optimal Method in Earliest Detection Problems," *Theory of Probability and Its Applications*, 8, 26–51. [2]

Song, P. X.-K. (2007), *Correlated Data Analysis*, Springer Series in Statistics, New York: Springer. [2,3]

Stengel, R. F. (1994), *Optimal Control and Estimation*, New York: Dover Publications Inc. [2]

Toulis, P., and Airold, E. M. (2017), "Asymptotic and Finite-Sample Properties of Estimators Based on Stochastic Gradients," *The Annals of Statistics*, 45, 1694–1727. [1,4]

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics*, 44, 1049–1060. [3]

Zhou, L., and Song, P. X.-K. (2017), "Scalable and Efficient Statistical Inference with Estimating Functions in the MapReduce Paradigm for Big Data," arXiv:1709.04389. [6]