Received: 15 June 2021 | Accepted: 19 August 2022

DOI: 10.1111/rssc.12597

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Series C Applied Statistics

Sequential one-step estimator by sub-sampling for customer churn analysis with massive data sets

Feifei Wang[1,2] | **Danyang Huang**[1,2] | **Tianchen Gao**[3] | **Shuyuan Wu**[4] | **Hansheng Wang**[4]

[1]Center for Applied Statistics, Renmin University of China, Beijing, China

[2]School of Statistics, Renmin University of China, Beijing, China

[3]School of Economics, Xiamen University, Xiamen, Fujian, China

[4]Guanghua School of Management, Peking University, Beijing, China

Correspondence
Danyang Huang, Center for Applied Statistics, School of Statistics, Renmin University of China, No.59 Zhongguancun Street, 100872 Beijing, China.
dyhuang@ruc.edu.cn

Shuyuan Wu, Guanghua School of Management, Peking University, No.5 Yiheyuan Road, 100871 Beijing, China.
shuyuan.w@pku.edu.cn

Funding information
Chinese National Statistical Science Research Project, Grant/Award Number: 2022LD06; Foundation from Ministry of Education of China, Grant/Award Number: 20JZD023; National Natural Science Foundation of China, Grant/Award Numbers: 11701560, 11831008, 11971504, 12071477, 72001205, 72171229; Renmin University of China

Abstract

Customer churn is one of the most important concerns for large companies. Currently, massive data are often encountered in customer churn analysis, which bring new challenges for model computation. To cope with these concerns, sub-sampling methods are often used to accomplish data analysis tasks of large scale. To cover more informative samples in one sampling round, classic sub-sampling methods need to compute *non-uniform* sampling probabilities for all data points. However, this method creates a huge computational burden for data sets of large scale and therefore, is not applicable in practice. In this study, we propose a sequential one-step (SOS) estimation method based on repeated sub-sampling data sets. In the SOS method, data points need to be sampled only with *uniform* probabilities, and the sampling step is conducted repeatedly. In each sampling step, a new estimate is computed via one-step updating based on the newly sampled data points. This leads to a sequence of estimates, of which the final SOS estimate is their average. We theoretically show that both the bias and the standard error of the SOS estimator can decrease with increasing sub-sampling sizes or sub-sampling times. The finite

Danyang Huang and Shuyuan Wu have contributed equally to this study.

sample SOS performances are assessed through simulations. Finally, we apply this SOS method to analyse a real large-scale customer churn data set in a securities company. The results show that the SOS method has good interpretability and prediction power in this real application.

KEYWORDS

customer churn, massive data, one-step estimator, sequential estimator, sub-sampling

1 INTRODUCTION

Customer churn is one of the most important concerns for large companies (Ascarza et al., 2018; Kayaalp, 2017). With increasingly fierce competition, it is important for companies to retain existing customers and prevent potential customer churn. Customer churn often refers to a situation in which customers no longer buy or use a company's products or services. For each customer, to churn or not to churn is a typical binary indicator. Therefore, customer churn prediction is often considered as a classification task, and several classifier models have been applied to address this issue (Ahmad et al., 2019). Among the previous models, logistic regression is widely used owing to its good interpretability (Ahn et al., 2019; Maldonado et al., 2021). Therefore, in this work, we employ logistic regression models to handle the task of customer churn analysis.

With the arrival of the 'Big Data' era, massive data are often encountered in customer churn analysis. For example, the real data analysed in this work comprise 12 million transaction records from 230,000 customers, which take up 300 GB in total. Such large amounts of data present great challenges for customer churn analysis. The first concern is memory constraint. The data can be too large to be stored in a computer's memory, and hence, have to remain stored on a hard drive. The second challenge is computational cost. For massive data, even a simple analysis (e.g., mean calculation) can take a long time. These challenges create large barriers for customer churn analysis with massive data.

To cope with these challenges, modern statistical analysis for massive data is developing fast. Basically, there are two streams of approaches for massive data. The first stream considers storing data in a distributed system and then applying the 'divide-and-conquer' strategy to accomplish data analysis tasks of huge scales. See McDonald et al. (2009), Lee et al. (2017), Battey et al. (2018), Jordan et al. (2019), Zhu et al. (2021), Wang et al. (2021), and the references therein. Another stream uses sub-sampling techniques that conduct calculations on sub-samples drawn from the whole data set to reduce both memory cost and computational cost. Important works include, but are not limited to, Dhillon et al. (2013), Ma et al. (2015), Wang et al. (2018), Quiroz et al. (2019), Ma et al. (2020), and Yu et al. (2020).

It is remarkable that there are big differences in the two streams of approaches. The distributed methods usually require typical equipment support. For example, a distributed system is often required, which consists of one central computer served as *Master* and all other computers served as *Workers*. Then, the goal of distributed methods is to obtain an estimator based on the whole sample. However, the sub-sampling methods can be conducted using one single computer. They focus on the approximation of the whole sample estimator based on sub-samples with limited resources. In this customer churn application, we do not have equipment support for distributed analysis. Therefore, in this work, we focus on sub-sampling methods to handle the customer churn analysis with massive data.

Classic sub-sampling methods require only one round of data sampling. To cover more informative samples in the single sampling round, *non-uniform* sampling probabilities are often specified for each data point, so that more informative data points can be sampled with higher probabilities. Typical approaches include leverage score-based sub-sampling (Drineas et al., 2011; Ma et al., 2015; Ma & Sun, 2015), information-based optimal sub-data selection (Wang et al., 2019), the optimal Poisson sampling method, and its distributed implementation (Wang et al., 2021), among others. These sub-sampling estimators have been proved to be consistent and asymptotically normal under appropriate regularity conditions; see Ma et al. (2020) for an important example. However, for these non-uniform sub-sampling methods, the step of evaluating non-uniform sampling probabilities for the whole data set would create a huge computational burden. For example, Wang et al. (2018) proposed optimal sub-sampling methods motivated by the A-optimality criterion (OSMAC) for large sample logistic regression. To find the optimal sub-sampling probabilities in the OSMAC method, the computational complexity is $O(Nd)$ for a data set with N observations and d-dimensional covariates. Consequently, this optimal sub-sampling algorithm could be computationally very expensive when N is very large. In the customer churn application, the whole data size is about 12 million. It would not be computationally feasible for the previous sub-sampling methods to handle this task.

An obvious way to address this problem would be to sample sub-data with *uniform* probabilities while operating the sub-sampling step repeatedly. By sampling with uniform probabilities, we are free from computing probabilities for the whole data set in advance, which can largely alleviate the computational burden. By sampling repeatedly, the sampled data would be close to the whole data set. In an ideal situation, if we were to conduct sub-sampling without replacement, then sampling N/n times would cover the whole data set, where N and n are the sizes of the whole data and sub-data, respectively. It is as if the whole data were stored in a distributed system.

The sub-sampling cost cannot be negligible, especially for repeated sub-sampling methods. It is remarkable that the time needed to sample one data point from the hard drive mainly consists of two parts. The first is the *addressing cost*, which is the time taken to identify the target data point on the hard drive. The second is the *I/O cost*, which is the time needed to read the target data point into the computer memory. Both are hard drive sampling costs, which cannot be ignored when we apply sub-sampling methods to massive data sets. Therefore, inspired by Pan et al. (2020), we adopt the sequential addressing sampling method to reduce the hard drive sampling cost. When data are randomly distributed on a hard drive, for each sub-data, only one starting point is selected and the other data points can be obtained sequentially from the starting point. In this way, the addressing cost can be reduced substantially for sub-sampling methods.

Based on the repeated sub-sampling data sets from the whole customer churn data, it is worthwhile to consider how to obtain an efficient customer churn estimation for both statistical property and computational cost. To this end, we propose a sequential one-step (SOS) method, whose estimation bias and variance can both be reduced by increasing the total sub-sampling times K. Specifically, in the first sub-sampling step, we can obtain an estimate $\hat{\beta}_1$ based on the first sub-data using, for example, the traditional Newton–Raphson algorithm. Then, in the second sub-sampling step, we regard $\hat{\beta}_1$ as the initial value, and conduct only one-step updating based on the second sub-data. This leads to $\hat{\beta}_2$. In the next step, the average of the first two estimates, that is, $\overline{\beta}_2 = (\hat{\beta}_1 + \hat{\beta}_2)/2$, is regarded as the initial value, and one-step updating based on the third sub-data is conducted again to obtain $\hat{\beta}_3$. In summary, in the $(k+1)$th step, the average of all previous estimates $\left(\text{i.e., } \overline{\beta}_k = \sum_{l=1}^{k} \hat{\beta}_l\right)$ serves as the initial value, and one-step updating based on the newly sampled sub-data is conducted to obtain the estimate $\hat{\beta}_{k+1}$. The final estimate of K sub-sampling steps is the average of all estimates, that is, $\hat{\beta}^{\text{SOS}} = \sum_{k=1}^{K} \hat{\beta}_k / K$.

It is noteworthy that, except for the first sub-sampling step, SOS requires only one round of updating in the subsequent sub-sampling steps. Therefore, it is computationally efficient. We also establish theoretical properties of the SOS estimator. We prove that both the bias and variance of the SOS estimator can be reduced as the sampling times K increase. We conduct extensive numerical studies based on simulated data sets to verify our theoretical findings. Finally, the SOS method is applied to the customer churn data set in a securities company to demonstrate its application.

The rest of this article is organised as follows. Section 2 introduces the SOS method. Section 3 presents simulation analysis to demonstrate the finite sample performance of the SOS estimators. Section 4 presents a real application for customer churn analysis using the SOS method. Section 5 concludes with a brief discussion.

2 | SOS ESTIMATOR BY SUB-SAMPLING

2.1 | Basic notations

Assume we have all the sample observations, $\mathcal{S} = \{1, 2, \cdots, N\}$, where N is defined as the whole sample size. Define (Y_i, X_i) to be the observation collected from the ith $(1 \le i \le N)$ observation, where $Y_i \in \mathbb{R}^1$ is the response and $X_i \in \mathbb{R}^p$ is the associated predictor. Conditional on X_i, assume that Y_i is independently and identically distributed with density function $f(Z_i; \beta)$, where $Z_i = (Y_i, X_i)$, $\beta \in \Theta$ is the unknown parameter, and Θ is an open subset in $\mathbb{R}^d$. We assume $p = d$ for convenience. To estimate the unknown parameter β, the log-likelihood function can be spelled out as

$$\ell(\beta) = \sum_{i \in \mathcal{S}} \ell(Z_i; \beta),$$

where $\ell(\cdot; \beta) = \log f(\cdot; \beta)$ is the log-likelihood function. For convenience, we use $\ell(\beta)$ to denote $\ell(Z_i; \beta)$ hereafter.

Note that when N is quite large, the whole data set is often kept on a hard drive, and cannot be read into the memory as a whole. Then it would be time consuming to select a sample randomly from the hard drive into the computer memory. To save time, we apply the sequential addressing sub-sampling (SAS) method (Pan et al., 2020) to conduct the sub-sampling directly on hard drive, not memory. To apply the SAS method, the whole data should be randomly distributed on the hard drive. Otherwise, a shuffle operator would be needed to make data randomly distributed. The detailed implementation process of conducting shuffle operation can be found in Appendix A in Data S1. Then, the sub-sampling steps can be performed iteratively based on the randomly distributed data to obtain sub-samples. Next, we describe the SAS method in detail.

First, one data point should be randomly selected on the hard drive as a starting point, that is, m $(1 \le m \le N - n + 1)$, where n is the desired sub-data size. This yields marginal addressing cost, as only one data point is chosen. Second, with a fixed starting point, the sub-sample with size n is selected sequentially with index $\{m, m + 1, \cdots, m + n - 1\} \in \mathcal{S}$. These selected sub-samples are collected as $\mathcal{T}_m = \{(X_m, Y_m), \cdots \ (X_{m+n-1}, Y_{m+n-1})\}$. Except for the first starting point indexed by m, the remaining data points are sampled sequentially. Therefore, no additional addressing cost is required, and the total sampling cost can be reduced substantially.

It is notable that, although the SAS method serves as a preparing step for the SOS method, there are fundamental differences between SAS and SOS. The SAS method is actually a sub-sampling technique, which can sample data directly from the hard drive and save much addressing cost. Based on the SAS sub-samples, Pan et al. (2020) also study the theoretical properties of some basic statistics estimators (e.g., sample mean). However in SOS, we focus on the estimation problem of logistic regression and discuss updating strategy to exploit the sub-samples obtained by SAS. It is remarkable that, the SAS method only serves as a tool for fast sampling, and the theoretical properties of the SOS estimator could still be guaranteed without this sampling step.

2.2 | SOS estimator

Assume the whole data set is randomly distributed on the hard drive. Then, the SAS method can be applied for fast sub-sampling. Recall that the sub-sample size is n. By the SAS method, a total of $M = N - n + 1$ different sequential sub-samples can be generated. Assuming that the sub-sampling is repeated K times, in the kth $(1 \le k \le K)$ sub-sampling, the sub-sample $\mathcal{T}_{(k)} \in \{\mathcal{T}_1, \cdots, \mathcal{T}_M\}$ can be selected with replacement. We further denote $\mathcal{S}_k$ as the indexes of data points in $\mathcal{T}_{(k)}$. Based on $\mathcal{S}_k$ with $1 \le k \le K$, we propose the SOS method for efficient estimation of β.

To obtain the SOS estimator, an initial estimator $\overline{\beta}_1$ is first calculated based on one of the SAS sub-samples with the index set denoted as $\mathcal{S}_1$. For example, it may be a maximum likelihood estimator (MLE). Then, in the $(k$+1)th sub-sampling step with $1 \le k \le K - 1$, a new SAS sub-sample can be obtained as $\mathcal{T}_{(k+1)}$. Based on the $(k$+1)th sub-sample, we conduct the following two steps iteratively to obtain the SOS estimator. The details of the SOS method are shown in Algorithm 1. Recall we have assumed that, the whole data are already shuffled or randomly distributed before we begin the SOS procedure.

Step 1: One-step update. Assume that the initial estimator in this step is $\overline{\beta}_k$. Then, we conduct one-step updating based on $\overline{\beta}_k$ to obtain the one-step updated estimator $\hat{\beta}_{k+1}$ in this step, that is,

$$\hat{\beta}_{k+1} = \overline{\beta}_k - \left\{ \ddot{\ell}_{\mathcal{S}_{k+1}}(\overline{\beta}_k) \right\}^{-1} \dot{\ell}_{\mathcal{S}_{k+1}}(\overline{\beta}_k), \tag{1}$$

where $\dot{\ell}_{\mathcal{S}_{k+1}}(\overline{\beta}_k)$ and $\ddot{\ell}_{\mathcal{S}_{k+1}}(\overline{\beta}_k)$ are the first- and second-order derivatives of the likelihood function based on the kth sub-sample, respectively.

Step 2: Average. The SOS sub-sampling estimator for the $(k$+1)th step can be calculated as

$$\overline{\beta}_{k+1} = \frac{1}{k+1} \left\{ k\overline{\beta}_k + \hat{\beta}_{k+1} \right\} = \frac{1}{k+1} \sum_{l=1}^{k+1} \hat{\beta}_l.$$

Conduct Steps 1 and 2 iteratively. The estimator obtained in the Kth step is the final SOS estimator, that is, $\hat{\beta}^{\text{SOS}} = \overline{\beta}_K$. See Figure 1 for an illustration of calculating the SOS estimator based on the SAS sub-sampling method.

Algorithm 1. SOS estimation algorithm

```
STEP 1: Compute the initial estimator
    1. Randomly select $m_1$ from $\{1, \cdots, N\}$ and obtain the SAS sub-sample with index set
       $\mathcal{S}_1$, where the predictors and responses are denoted by $\mathcal{T}_{(1)}$.
    2. Conduct MLE based on the sample with index $\mathcal{S}_1$ to obtain $\hat{\beta}_1$. Further define $\bar{\beta}_1 = \hat{\beta}_1$.
STEP 2: Compute the SOS estimate
    1. for $k = 1, ..., K-1$ do:
        (1) Randomly select $m_{k+1}$ from $\{1, \cdots, N\}$, and obtain the SAS sub-sample with
            index set $\mathcal{S}_{k+1}$, where the predictors and responses are denoted by $\mathcal{T}_{(k+1)}$.
        (2) Compute $\hat{\beta}_{k+1}$ with the one-step update as
                $$\hat{\beta}_{k+1} = \bar{\beta}_k - \left\{\ddot{\ell}_{\mathcal{S}_{k+1}}(\bar{\beta}_k)\right\}^{-1} \dot{\ell}_{\mathcal{S}_{k+1}}(\bar{\beta}_k),$$
        (3) Calculate the sequential estimator in the $(k+1)$th step as
                $$\bar{\beta}_{k+1} = \frac{1}{k+1}\left\{k\bar{\beta}_k + \hat{\beta}_{k+1}\right\}.$$
    2. The final SOS estimate is $\hat{\beta}^{\mathrm{SOS}} = \bar{\beta}_K$.
```

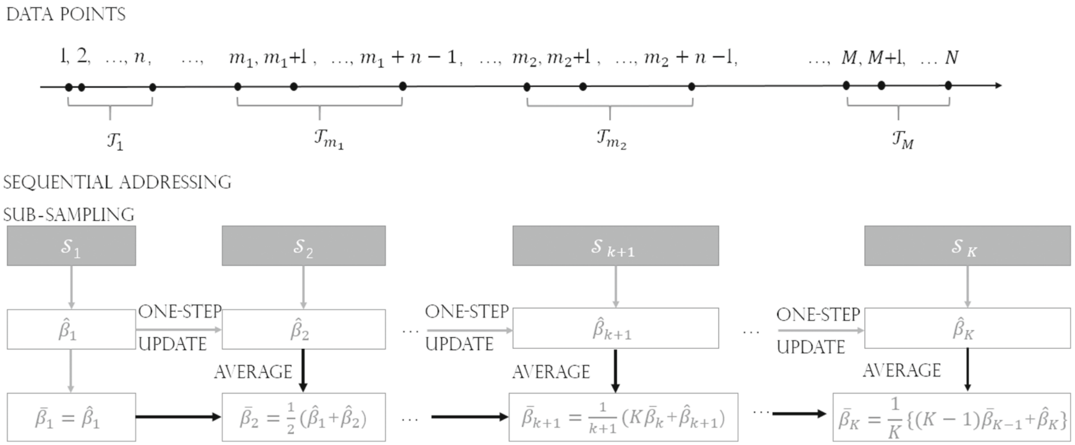
DATA POINTS
1, 2, ..., n, ..., m_1, m_1+1 , ..., $m_1 + n - 1$, ..., m_2, m_2+1 , ..., $m_2 + n - 1$, ..., M, M+1, ... N
$\mathcal{T}_1$
$\mathcal{T}_{m_1}$
$\mathcal{T}_{m_2}$
$\mathcal{T}_M$
SEQUENTIAL ADDRESSING
SUB-SAMPLING
$\mathcal{S}_1$
$\mathcal{S}_2$
$\mathcal{S}_{k+1}$
$\mathcal{S}_K$
$\hat{\beta}_1$
ONE-STEP
UPDATE
$\hat{\beta}_2$
...
ONE-STEP
UPDATE
$\hat{\beta}_{k+1}$
...
ONE-STEP
UPDATE
$\hat{\beta}_K$
AVERAGE
AVERAGE
AVERAGE
$\bar{\beta}_1 = \hat{\beta}_1$
$\bar{\beta}_2 = \frac{1}{2}(\hat{\beta}_1+\hat{\beta}_2)$
...
$\bar{\beta}_{k+1} = \frac{1}{k+1}(K\bar{\beta}_k+\hat{\beta}_{k+1})$
...
$\bar{\beta}_K = \frac{1}{K}\{(K-1)\bar{\beta}_{K-1}+\hat{\beta}_K\}$

FIGURE 1 Illustration of the sequential one-step estimator based on random addressing sub-sampling

Remark: We offer two remarks here about the SOS estimator. First, for each sub-sample, only the one-step update is conducted by (1). This may yield marginal computational cost because (1) the sample size n is relatively small; and (2) no Newton–Raphson-type iterations are involved. It is notable that, there is no need to achieve fully Newton–Raphson convergence for each sub-sample. This is because the Newton–Raphson algorithm is not affected by the initial value when it goes to convergence. Therefore, if the sub-sample estimator $\hat{\beta}_{k+1}$ is not one-step updating, but obtained with fully Newton–Raphson convergence, then the initial value $\overline{\beta}_k$ would not affect the resulting estimator in the $(k+1)$th sub-sample. Consequently, the SOS estimator degenerates to the one-shot (OS) estimator, which we theoretically compare in the next sub-section. Second, each $\overline{\beta}_k$ can be viewed as the average of the one-step estimator in form (1) for the first k steps. This leads to some nice properties: (1) a total of K sub-samples are used in the estimation; and (2) the standard error of the final estimator can be reduced by averaging. More weighting schemes could be considered in the SOS updating strategy; see the weighting scheme in the aggregated estimating equation (AEE) method (Lin & Xi, 2011) for an example.

It is also notable that, our proposed SOS method is an extension of the classical one-step estimator (Shao, 2003; Zou & Li, 2008) in the field of sub-sampling. For example, Zhu et al. (2021) has developed a distributed least squares approximation (DLSA) method to solve regression problems on distributed systems. In the DLSA method, once the weighted least squares estimator (WLSE) is obtained by the Master, it would be broadcast to all Workers. Then each Worker would perform a one-step iteration using the WLSE as the initial value to obtain a new estimator. Another work is Wang et al. (2021), who propose a one-step upgraded pilot method for non-uniformly and non-randomly distributed data. However, both the two methods are divide-and-conquer (DC) type methods, which are quite different from the sub-sampling methods. For example, one typical difference is that, in DC methods, the estimators from different Workers can be regarded as independent. However, in our SOS method, the estimators $\overline{\beta}_1$ to $\overline{\beta}_K$ are sequentially obtained, which makes them dependent with each other. This leads to challenges in investigation of the asymptotic theory for the SOS estimator. More differences between the divide-and-conquer type methods and sub-sampling methods can be found in the Appendix B in Data S1.

2.3 | Theoretical properties of the SOS estimator

We further investigate the properties of the SOS estimator in this subsection. To establish the theoretical properties of the SOS estimator, assume that Θ is an open subset in the Euclidean space, and we have the following conditions.

(C1) The sub-sample size n, the whole sample size N, and the sub-sampling steps K satisfy that as $n \to \infty$, $n/N \to 0$, $K \to \infty$, and $\log K = o(n)$.

(C2) Assume that the first- and second-order derivatives of log-likelihood $\ell(\beta)$ satisfy equations $E\left\{\partial\ell(\beta)/(\partial\beta_j)\right\} = 0$, and $-E\left\{\partial^2\ell(\beta)/(\partial\beta_{j_1}\partial\beta_{j_2})\right\} = E\left[\left\{\partial\ell(\beta)/\partial\beta_{j_1}\right\} \quad \left\{\partial\ell(\beta)/\partial\beta_{j_2}\right\}\right]$, for $1 \le j, j_1, j_2 \le p$.

(C3) Assume that $E\left[\partial\ell_i(\beta)/\partial\beta\{\partial\ell_i(\beta)/\partial\beta\}^\top\right] = \Sigma^{-1}$ is finite and positive definite at $\beta = \beta_0$, where β_0 is the true parameter.

(C4) There is an open subset ω of Θ that contains the true parameter β_0, such that for all Z_is, $\partial^3 f(Z_i, \beta)/(\partial\beta_{j_1}\partial\beta_{j_2}\partial\beta_{j_3})$ exists for all $\beta \in \omega$, and $1 \le j, j_1, j_2, j_3 \le p$. Moreover, assume function $M(\cdot)$ exists, such that for any $\beta \in \Theta$, we have $EM(Y_i) < C$, where C is a constant. For $\beta \in \omega$ and $1 \le j, j_1, j_2, j_3 \le p$, we have $\left|\partial\ell(Y_i, \beta)/(\partial\beta_{j_1}\partial\beta_{j_2}\partial\beta_{j_3})\right| \le M(Y_i)$.

(C5) Assume the covariates X_{ij}s independently follow Gaussian distributions.

Condition (C1) restricts the relationships of (n, K) and (n, N). By the condition, we know that the sub-sampling times K should not grow too fast in the sense that $\log K = o(n)$, and the sub-sampling size n should not increase too fast in the sense that $n/N \to 0$. Conditions (C2)–(C4) are standard regularity conditions. They are commonly adopted to guarantee asymptotic normality of the ordinary maximum likelihood estimates; see, for example, Lehmann and

Casella (1998) and Fan and Li (2001). Condition (C5) is a classical assumption on covariates (Wang, 2009).

With the conditions satisfied, we can establish the properties of $\hat{\beta}^{\mathrm{SOS}}$, which equals $\overline{\beta}_K$. Define $U_{(k)} = \left\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\right\}^{-1}\left\{n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0)\right\}$, and $\bar{U}_K = K^{-1}\sum_{k=1}^{K} U_{(k)}$. Then, the following theorem holds.

Theorem 1. *Assume conditions (C1)–(C5) hold. Then, we have (1)* $\overline{\beta}_K - \beta_0 = -\bar{U}_K + \Delta$, *with* $E\bar{U}_K = 0, \mathrm{var}(\bar{U}_K) = \{1/(nK) + 1/N\}\,\Sigma\{1 + o(1)\}$, *and* $\Delta = O_p\left[(\log K/n)\{1/(nK) + 1/N\}\right]^{1/2}$. (2) $\{1/(nK) + 1/N\}^{-1/2}(\overline{\beta}_K - \beta_0) \rightarrow_d N(0, \Sigma)$.

The proof of Theorem 1 is in Appendix B.1. As shown in Theorem 1, we separate the difference between the SOS estimator $\overline{\beta}_K$ and the true parameter β_0 into two parts, the bias term and variance term. One could note that the bias term Δ and variance term $\mathrm{var}(\overline{\beta}_K)$ are both determined by two main parts. The first part is related to the whole sample size N. This part cannot disappear by using the SOS procedure. The second term is $(nK)^{-1}$, which is affected by the SOS procedure and can decrease with larger K or n. In addition, the SOS estimator satisfies asymptotic normality with asymptotic variance $\{1/(nK) + 1/N\}^{-1}\Sigma$. In particular, when nK is much larger than N, it could achieve the same statistical efficiency as the global estimator. Note that the practical demand for estimation precision is usually limited. On the contrary, the budget for sampling cost is very valuable. Then, it may be more appealing to sacrifice the statistical efficiency to some extent for lower sampling cost. Therefore, in practice, we often expect the SOS method to be implemented with reasonably large n and K (i.e., $nK \ll N$) as long as the desired statistical precision is achieved.

For theoretical comparison, we introduce a simple alternative method. For each sub-sample $\mathcal{S}_k$, we separately compute the MLE $\hat{\beta}_{k,\mathrm{mle}}$. Then, all sub-sample estimators are simply averaged to obtain the OS estimator. Let $\overline{\beta}_K^{OS} = K^{-1}\sum_{k=1}^{K}\hat{\beta}_{k,\mathrm{mle}}$ denote the OS estimator. We obtain the following conclusion.

Proposition 1. *For the OS estimator, under the same conditions in Theorem 1, we have* $\overline{\beta}_K^{OS} - \beta_0 = -\bar{U}_K + \Delta_{os}$, *where* $\Delta_{os} = O_p(1/n)$. *Further assume* $n^2/N \rightarrow \infty$; *then, we have* $\{1/(nK) + 1/N\}^{-1/2}(\overline{\beta}_K^{OS} - \beta_0) \rightarrow_d N(0, \Sigma)$.

The proof of Proposition 1 is in Appendix B.2. Comparing Theorem 1 and Proposition 1, we find that the leading terms for the variance of both the SOS and the OS estimators are identical. However, the bias term of the OS estimator is of order $O_p(1/n)$, which cannot be improved as K increases. By contrast, the bias of the SOS estimator is $O_p\left[(\log K/n)\{1/(nK) + 1/N\}\right]^{1/2}$, which can be significantly reduced as K increases. Therefore, compared with the SOS estimator $\overline{\beta}_K$, $\overline{\beta}_K^{OS}$ requires a more stringent condition $n^2/N \rightarrow \infty$, such that it could achieve the same asymptotic normality as the global estimator (Huang & Huo, 2019; Jordan et al., 2019). However, this condition is not necessary for the SOS estimator.

Next, to make an automatic inference, we discuss the estimation of the standard error of the SOS estimator. Specifically, based on the SOS procedure, we construct the following statistic as

$$\widehat{\mathrm{SE}}^2(\overline{\beta}_K) = \frac{n}{K-1}\left(\frac{1}{nK} + \frac{1}{N}\right)\sum_{k=1}^{K}\left(U_{(k)} - \bar{U}_K\right)\left(U_{(k)} - \bar{U}_K\right)^{\top}. \tag{2}$$

The properties of $\widehat{\mathrm{SE}}^2(\overline{\beta}_K)$ are presented in the following theorem.

Theorem 2. *Under the same conditions in Theorem 1, we have*

$$E\left\{\widehat{\mathrm{SE}}^2(\overline{\beta}_K)\right\} = \Sigma\left(\frac{1}{nK}+\frac{1}{N}\right)\{1+o(1)\}$$
$$\mathrm{var}(\overline{\beta}_K) - E\left\{\widehat{\mathrm{SE}}^2(\overline{\beta}_K)\right\} = O\left(\frac{n}{N^2}\right)$$

The proof of Theorem 2 is in Appendix B.3. We conclude that the leading orders of $\mathrm{var}(\overline{\beta}_K)$ and $\widehat{\mathrm{SE}}^2(\overline{\beta}_K)$ are the same. However, as an estimator of $\mathrm{var}(\overline{\beta}_K)$, $\widehat{\mathrm{SE}}^2(\overline{\beta}_K)$ is biassed, but the order of the bias is $O(nN^{-2})$, which decreases as N increases or n decreases. Practically, the unknown parameter β_0 in $U_{(k)}$ can be replaced by $\overline{\beta}_k$ to obtain $\hat{U}_{(k)}$ and $\hat{\overline{U}}_K = K^{-1}\sum_{k=1}^{K}\hat{U}_{(k)}$. Then, $\widehat{\mathrm{SE}}_*^2(\overline{\beta}_K)$ can be calculated based on $\hat{U}_{(k)}$ and $\hat{\overline{U}}_K$ in the form of (2). Note that by Theorem 1, the leading term for the variance of $\overline{\beta}_K$ is $\{1/(nK)+1/N\}\,\Sigma$. Such term can be consistently estimated by the proposed SE estimator $\widehat{\mathrm{SE}}_*(\overline{\beta}_K)$. Its asymptotic property is presented in the following theorem.

Theorem 3. *Under the same conditions in Theorem 1, we have*

$$\left(\frac{1}{nK}+\frac{1}{N}\right)^{-1}\widehat{\mathrm{SE}}_*^2(\overline{\beta}_K) \to_p \Sigma. \tag{3}$$

The proof of Theorem 3 is in Appendix B.4. Combining the results of Theorems 1 and 3, we immediately obtain that $\left\{\widehat{\mathrm{SE}}_*(\overline{\beta}_K)\right\}^{-1}(\overline{\beta}_K-\beta_0) \to_d N(0, I_p)$. As a result, both the estimator and statistical inference could be easily and efficiently derived by our SOS procedure. We illustrate the performance of the SOS estimator and $\widehat{\mathrm{SE}}_*^2(\overline{\beta}_K)$ in the next section.

3 | SIMULATION STUDIES

3.1 | Simulation design

To demonstrate the finite sample performance of the SOS estimator, we present a variety of simulation studies. Assume that the whole data set contains $N = 150{,}000$ observations. For $i = 1, \ldots, N$, we generate each observation (X_i, Y_i) under the logistic regression model. We choose the logistic regression model because it is a specific model used for customer churn analysis. Given that the SOS method can be extended easily to other generalised regression models, we also choose the Poisson regression model as an example to test the effectiveness of the SOS method. The specific settings for the two model examples are as follows.

Example 1 (Logistic regression). Logistic regression is used to model binary responses. In this example, we consider $p = 5$ exogenous covariates $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5})^\top$, where each covariate is generated from a standard normal distribution $N(0, 1)$. We set the coefficients for X_i as $\beta = (0, -0.2, -0.1, 0.1, 0.2)^\top$. Then, the response $Y_i (1 \le i \le N)$ is generated from a Bernoulli distribution with the probability given as

$$P(Y_i = 1 | X_i, \beta) = \frac{\exp\left(X_i^\top \beta\right)}{1+\exp\left(X_i^\top \beta\right)}.$$

Example 2 (Poisson regression). Poisson regression is used to deal with count responses. We also consider $p = 5$ exogenous covariates, which are all generated from standard normal distribution. The corresponding coefficients are set as $\beta = (-3, -2, -1, 1, 2)^\top$. Then, the response Y_i $(1 \le i \le N)$ is generated from a Poisson distribution given as

$$P(Y_i|X_i, \beta) = \frac{\lambda_i^{Y_i}}{Y_i!} \exp(\lambda_i), \text{where } \ \lambda_i = \exp\left(X_i^\top \beta\right).$$

After obtaining N observations, we consider combinations of different sub-sampling size and different sub-sampling times. In both logistic and Poisson regression examples, we set $n = (100, 200, 400)$. Then in the logistic regression example, we consider cases of small K, and set $K = (10, 20, 30, 40, 50, 100)$. In the Poisson regression example, we consider cases of big K, and set $K = (100, 200, 300, 400, 500, 1000)$. In each combination of n and K, we repeat the experiment $B = 1000$ times.

3.2 | Comparison with repeated sub-sampling methods

In this sub-section, we compare the proposed SOS estimator with the OS estimator, which is representative of the repeated sub-sampling methods. Specifically, in each simulated data set, we obtain the SOS estimator using Algorithm 1. For the OS estimator, we first randomly obtain K sub-data, and then independently apply the Newton–Raphson method to each sub-data. Specifically, in the kth sub-data, we set the initial value as $\beta_{\text{ini}} = (0, 0, 0, 0, 0)^\top$, and then fully conduct the Newton–Raphson method to obtain estimate $\hat{\beta}_{k,\text{mle}}$. The final OS estimator is calculated as $\overline{\beta}_K^{\text{OS}} = \sum_{k=1}^K \hat{\beta}_{k,\text{mle}}/K$. For one particular method (i.e., SOS and OS), we define $\hat{\beta}^{(b)} = \left(\hat{\beta}_j^{(b)}\right)_{j=1}^p$ as the estimator in the bth $(1 \le b \le B)$ replication. Then, to evaluate the estimation efficiency of each estimator, we calculate the bias as $\flat = |\beta - \overline{\beta}|$, where $\overline{\beta} = B^{-1} \sum_b \hat{\beta}^{(b)}$. The standard error $\left(\widehat{\text{SE}}^{(b)}\right)$ can be estimated based on Theorem 2 for the SOS method. We report the average $\widehat{\text{SE}} = B^{-1} \sum_b \widehat{\text{SE}}^{(b)}$. Then, we compare $\widehat{\text{SE}}$ with the Monte Carlo SD of $\hat{\beta}^{(b)}$, which is calculated by $\text{SE} = \left\{B^{-1} \sum_b (\hat{\beta}^{(b)} - \overline{\beta})^2\right\}^{1/2}$. Next, we construct a 95% confidence interval for β as $\text{CI}^{(b)} = \left(\hat{\beta}^{(b)} - z_{0.975}\widehat{\text{SE}}^{(b)}, \hat{\beta}^{(b)} + z_{0.975}\widehat{\text{SE}}^{(b)}\right)$, where z_α is the αth lower quantile of a standard normal distribution. Then, the coverage probability is computed as $\text{ECP} = B^{-1} \sum_b I\left(\beta \in \text{CI}^{(b)}\right)$, where $I(\cdot)$ is the indicator function. Last, we compare the computational efficiency of the two methods. It is notable that, for a fixed sample size n, the computational time consumed by each Newton–Raphson iteration is the same for the SOS and OS methods. Therefore, we use the average round of Newton–Raphson iterations to compare the computational efficiency of the SOS and OS methods.

Tables 1 and 2 present the simulation results for estimation performance under the logistic regression and Poisson regression, respectively. In general, the simulation results under the two examples are similar. We draw the following conclusions. First, as the sub-sampling times K increases, the bias of the SOS estimator becomes much smaller than that of the OS estimator. This is because the bias of the SOS estimator decreases with increasing K, while the bias of the OS estimator is always $O(1/n)$. Second, the SE and $\widehat{\text{SE}}$ of both estimators decrease with increasing n or K, implying that the two estimators are consistent. Third, as the bias and SE of the SOS estimator can decrease with n and K, the bias always behaves negligibly compared with SE. However, the bias of the OS estimator is comparable to or even larger than its SE; see $n = 100, K = 300$ in Table 2 for an example. Next, the empirical coverage probabilities of the SOS

TABLE 1 Simulation results for estimation performance under logistic regression

	Bias × 100		**SE × 100**		**$\widehat{\text{SE}}$ × 100**		**ECP**		**ROUND**	
K	**SOS**	**OS**	**SOS**	**OS**	**SOS**	**OS**	**SOS**	**OS**	**SOS**	**OS**
$n = 100$										
10	0.532	0.903	6.598	7.237	6.545	6.596	0.952	0.912	1.875	7.145
20	0.336	0.963	4.694	5.109	4.687	4.717	0.947	0.879	1.923	7.492
30	0.204	0.943	3.815	4.142	3.857	3.88	0.950	0.858	1.884	7.146
40	0.170	0.976	3.278	3.559	3.359	3.379	0.950	0.858	1.821	7.240
50	0.150	0.915	2.997	3.247	3.015	3.032	0.954	0.823	2.122	7.654
100	0.058	0.921	2.124	2.293	2.162	2.173	0.955	0.755	1.942	7.149
$n = 200$										
10	0.291	0.545	4.709	4.923	4.539	4.554	0.946	0.927	1.873	7.364
20	0.160	0.479	3.233	3.374	3.283	3.293	0.952	0.919	1.917	7.381
30	0.098	0.420	2.682	2.793	2.707	2.714	0.950	0.912	1.921	7.399
40	0.065	0.453	2.328	2.422	2.363	2.370	0.954	0.906	1.893	7.416
50	0.043	0.438	2.080	2.161	2.125	2.130	0.951	0.904	1.834	7.433
100	0.041	0.448	1.550	1.609	1.553	1.556	0.948	0.882	1.985	7.450
$n = 400$										
10	0.119	0.245	3.198	3.269	3.200	3.205	0.950	0.938	1.881	7.467
20	0.055	0.239	2.285	2.331	2.322	2.325	0.952	0.938	1.954	7.484
30	0.046	0.224	1.891	1.926	1.926	1.928	0.951	0.936	2.016	7.501
40	0.035	0.218	1.681	1.714	1.689	1.692	0.949	0.934	1.982	7.519
50	0.031	0.226	1.529	1.558	1.532	1.534	0.947	0.929	2.104	7.536
100	0.020	0.237	1.148	1.170	1.150	1.151	0.947	0.912	2.163	7.553

Notes: The bias, SE, $\widehat{\text{SE}}$, and ECP are reported for the sequential one-step (SOS) and one-shot (OS) estimators, respectively. The average Newton–Raphson rounds (representing the computational time) of the two estimators are also reported.

estimator are all around the nominal level of 95%, which suggests that the true SE can be well approximated by its estimators derived in Theorem 2. However, the empirical coverage probabilities of the OS estimator show a deceasing trend when enlarging K. This is because the bias of the OS estimator cannot be negligible when K is large. Last, regarding the computational time, we compare the average rounds of Newton–Raphson iterations consumed by the two methods. Except for the initial sub-sample estimator, the SOS method uses only one Newton–Raphson update for each subsequent sub-sample. However, on average, the OS estimator requires seven or eight rounds of Newton–Raphson updates to obtain convergence. Therefore, the SOS estimator is computationally more efficient than the OS estimator.

3.3 | Comparison with non-uniform sub-sampling methods

In this sub-section, we compare the proposed SOS method with the non-uniform sub-sampling methods. To this end, we choose the OSMAC method (Wang et al., 2018) for the comparison,

TABLE 2 Simulation results for estimation performance under Poisson regression

	Bias ×100		SE×100		$\widehat{\text{SE}} \times 100$		ECP		ROUND	
K	SOS	OS	SOS	OS	SOS	OS	SOS	OS	SOS	OS
$n = 100$										
100	0.063	3.656	3.237	3.437	3.237	3.505	0.951	0.820	1.892	8.669
200	0.042	3.649	2.329	2.478	2.365	2.555	0.951	0.711	1.934	8.673
300	0.035	3.594	1.947	2.085	1.989	2.148	0.955	0.634	1.910	8.671
400	0.033	3.599	1.720	1.839	1.770	1.910	0.959	0.571	1.907	8.671
500	0.031	3.597	1.564	1.667	1.625	1.754	0.959	0.524	1.907	8.671
1000	0.033	3.593	1.222	1.294	1.286	1.387	0.960	0.426	1.893	8.670
$n = 200$										
100	0.039	1.275	2.096	2.138	2.103	2.167	0.950	0.912	1.884	8.326
200	0.034	1.241	1.555	1.584	1.575	1.622	0.957	0.879	1.892	8.323
300	0.030	1.234	1.315	1.337	1.353	1.393	0.958	0.856	1.927	8.322
400	0.026	1.241	1.194	1.214	1.226	1.262	0.961	0.834	1.962	8.323
500	0.026	1.241	1.115	1.135	1.143	1.176	0.960	0.816	2.043	8.322
1000	0.023	1.239	0.923	0.936	0.956	0.984	0.958	0.765	1.895	8.321
$n = 400$										
100	0.038	0.485	1.466	1.476	1.470	1.489	0.947	0.932	1.884	8.153
200	0.035	0.485	1.131	1.137	1.145	1.160	0.956	0.932	1.892	8.153
300	0.032	0.498	1.006	1.012	1.013	1.026	0.953	0.920	1.903	8.153
400	0.034	0.485	0.935	0.939	0.940	0.952	0.951	0.915	1.906	8.153
500	0.032	0.488	0.882	0.886	0.894	0.905	0.953	0.917	1.891	8.152
1000	0.026	0.486	0.769	0.772	0.792	0.802	0.954	0.910	1.902	8.152

Notes: The bias, SE, $\widehat{\text{SE}}$, and ECP are reported for the sequential one-step (SOS) and one-shot (OS) estimators, respectively. The average Newton–Raphson rounds (representing the computational time) of the two estimators are also reported.

which is particularly designed for large sample logistic regression. The OSMAC method applies a two-step algorithm for the model estimation. In the first step, a pilot sample of size r_0 is randomly chosen to obtain a pilot estimate. Then, the pilot estimate is used to compute the optimal sub-sampling probabilities for the whole data. In the second step, a new sub-sample of size r is chosen based on the optimal sub-sampling probabilities. Then, the final OSMAC estimate is obtained using the total $r_0 + r$ samples. We compare the two methods under the logistic regression example. To mimic a large data set, we consider the whole sample size $N = (1, 2, 5, 10) \times 10^5$. For fixed N, the whole data set is generated under the logistic regression model following the procedures in Section 3.1.

Below, we compare the SOS method and OSMAC method under one specific situation. That is, the compute memory is limited, which could only support building a logistic regression model for a sample with size $n = 400$. In this situation, the sub-sample size used in SOS is fixed as $n = 400$, and we vary the sub-sampling times as $K = (1, 5, 10, 20, 40)$. As for the OSMAC method, we set

TABLE 3 The mean squared error (MSE) and time costs (in seconds) are obtained by the sequential one-step (SOS) and optimal sub-sampling methods motivated by the A-optimality criterion (OSMAC) methods for different sample sizes N under the logistic regression model.

		MSE	**Time**	**MSE**	**Time**	**MSE**	**Time**	**MSE**	**Time**
Method		$\boldsymbol{N = 100,000}$		$\boldsymbol{N = 200,000}$		$\boldsymbol{N = 400,000}$		$\boldsymbol{N = 800,000}$	
OSMAC		0.0460	0.0213	0.0460	0.0410	0.0452	0.0875	0.0476	0.1679
SOS	$K = 1$	0.2346	0.0058	0.2320	0.0070	0.2300	0.0064	0.2274	0.0086
	$K = 5$	0.0415	0.0101	0.0419	0.0090	0.0414	0.0124	0.0418	0.0150
	$K = 10$	0.0410	0.0171	0.0405	0.0168	0.0419	0.0175	0.0406	0.0256
	$K = 20$	0.0404	0.0318	0.0409	0.0322	0.0405	0.0378	0.0403	0.0437
	$K = 40$	0.0401	0.0574	0.0398	0.0618	0.0400	0.0740	0.0400	0.0804

$r_0 = 200$ and $r = 400$. The OSMAC method is implemented using the corresponding R package provided by Wang et al. (2018).

For a reliable comparison, the experiment is randomly replicated for $B = 200$ times for each experimental setup. We use the mean squared error (MSE) to evaluate the statistical efficiencies of the two methods. Specifically, for one particular method (i.e., SOS and OSMAC), we define $\hat{\beta}^{(b)} = \big(\hat{\beta}_j^{(b)}\big)_{j=1}^{p}$ as the estimator in the bth ($1 \leq b \leq B$) replication. Then, the MSE of each estimator is computed as $B^{-1}\sum_{b=1}^{B}\sum_{j=1}^{p}\big(\hat{\beta}_j^{(b)} - \beta_j\big)^2$. We also compare the computational time of the two methods. All experiments are conducted on a server with 12× Xeon Gold 6271 CPU and 64 GB RAM. The total time costs consumed by different methods under each experimental setup are averaged over $B = 200$ random replications. The detailed results are displayed in Table 3.

Based on the results presented in Table 3, we draw the following conclusions. First, the OSMAC method achieves better estimation performance than the SOS method with $K = 1$. This is because the OSMAC method can find an optimal sub-sample, while the SOS method only randomly selects the sub-sample. Second, as the sub-sampling times K increases, the estimation performance of the SOS method improves by achieving smaller MSE values. This finding is consistent with our theoretical results in Theorem 1. It is also notable that a relatively small K (e.g., $K = 5$) makes the SOS method achieve better estimation performance than the OSMAC method. Third, the computational time consumed by the OSMAC method increases largely with the whole sample size N. This is because the OSMAC method computes the sub-sampling probabilities for all N samples, which results in a large computational cost. Meanwhile, the computational cost of the SOS method mainly results from the repeated sub-sampling strategy. Then, with an increase of K, the computational time consumed by SOS is enlarged. However, with appropriately chosen K, the SOS method can achieve both better estimation performance and smaller computational time than the OSMAC method.

3.4 | Comparison with all-sample methods

Finally, to complete our empirical comparison, we compare the SOS method with methods using the whole sample. We first compare the SOS method with DC methods, which are also commonly used to accomplish data analysis tasks of huge scale. The key idea of DC methods is to divide a large-scale data set into multiple sub-data sets, each of which is then estimated separately to obtain a local estimate. Then, all local estimates are reassembled together to obtain the final estimate. Different from sub-sampling methods, DC methods in fact exploit the whole data information. Therefore, they often have good statistical efficiency but high computational cost. In this regard, we take the AEE method (Lin & Xi, 2011) as a representative example. Another method to consider is the mini-batch gradient descent (MGD) estimation method (Duchi et al., 2011). The MGD method splits the whole data set into several mini-batches. Each mini-batch is then read into the memory and estimated sequentially. Different from the SOS method, MGD is not a Newton–Raphson-type method. Instead, it applies the stochastic gradient descent strategy for parameter estimation.

To undertake a comprehensive evaluation, we consider the whole sample size as $N = (10, 12, 14, 16, 18, 20) \times 10^4$. For fixed N, we generate the data set under the logistic regression model following the procedures described in Section 3.1. For the AEE method, we assume there are a total of $J = 100$ workers. For the MGD method, we assume the total number of mini-batches is also $J = 100$. Then, the whole data set is randomly and evenly divided into $J = 100$ sub-data sets, and each sub-data set has $n = N/J$ observations. For comparison, the sub-sample size in the SOS method is fixed as $n = N/J$. For all experiments, we assume sub-sampling times of $K = 50$. Theoretically, the information exploited by the SOS method is smaller than the DC methods. We repeat the experiment $B = 200$ times under each experimental setup. The statistical efficiencies of the three methods are evaluated by MSE. We also compare the computational efficiency of the three methods. The detailed results are displayed in Figure 2.

As shown by Figure 2a, compared with AEE and MGD, the SOS method performs statistically less efficiently by achieving the largest MSE values. This finding is obvious, because the AEE and MGD methods exploit the full data information. Therefore, in theory, the two estimators are both $\sqrt{N}$-consistent. However as suggested by Theorem 2, the SE of the SOS estimator is $O\{1/(nK) + 1/N\}$. Then, the SOS estimator should be statistically less efficient when nK is smaller than N. Although the SOS estimator has the worst statistical efficiency, its MSE has already achieved 10^{-4}, indicating satisfactory estimation precision in practice.

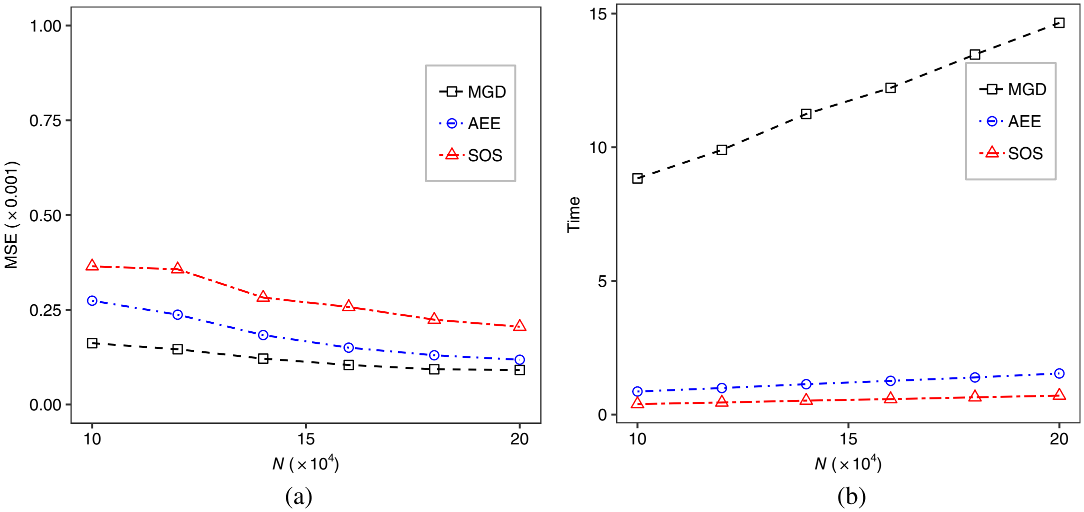

FIGURE 2 The mean squared error and time costs (in logarithm) are obtained by the sequential one-step, mini-batch gradient descent, and aggregated estimating equation methods for different sample sizes N under the logistic regression model. (a) Estimation performance; (b) Computation performance

We then compare the computational efficiency of the three methods. We fix the learning rate as 0.2 in the MGD method. All experiments are conducted on a server with 12× Xeon Gold 6271 CPU and 64 GB RAM. The total time costs (in s) consumed by the different methods under different sample sizes are averaged over $B = 200$ random replications. Then, the averaged time costs are plotted in Figure 2b in log-scale. As shown, the MGD method takes the most computational time. The AEE and SOS methods are much more computationally efficient than the MGD method. Furthermore, the SOS method takes less computational time than the AEE method. In general, the time costs consumed by the SOS method are only half those of the AEE method. These empirical findings confirm that the SOS method is computationally more efficient than the AEE and MGD methods.

4 | APPLICATION TO CUSTOMER CHURN ANALYSIS

4.1 | Data description and pre-processing

We apply the SOS method to a large-scale customer churn data set, which is provided by a well-known securities company in China. The original data set contains 12 million transaction records from 230,000 customers from September to December 2020. This data set is originated from 10 files, which are directly exported from the company's database system. The 10 files record different aspects of users. Specifically, the 10 files include: user basic information, behaviour information on the company's APP, daily asset information, daily market information, inflow-outflow information, debt information, trading information, fare information and information of holding financial products or service products. In total, the 10 files contain 398 variables describing the asset and non-asset information of a specific customer on a specific trading day. The asset information contains 325 variables related to customer transactions, such as assets, stock value, trading volume, profit and debit. The non-asset information contains 73 variables about detailed customer information, such as customer ID, gender, age, education and login behaviour. The whole data set takes up about 300 GB on a hard drive.

The research goal of this study is to provide an early warning of customer churn status, which may help the securities company to retain customers. According to the common practice of the securities company, a customer is defined as lost when the following two criteria are met: (1) the customer has less than 20,000 floating assets in 20 trading days; and (2) the customer logs in less than three times in 20 trading days. Based on this definition, a new binary variable *Churn* is used to indicate whether the customer is lost ($Churn = 1$) or not ($Churn = 0$). Given the response is a binary variable (churn or not churn), a logistic regression model can be built to help customer churn prediction. To predict the customer churn status, we compute both asset-related and non-asset-related covariates for each customer using transaction information ahead of 20 trading days. In other words, all used covariates can help forecast the churn status of customers 20 trading days in advance.

Before model building, we conduct several steps to preprocess the original data set. First, we check the missing value proportions for all variables in the data set, and then discard variables whose missing value proportions were larger than 80%. Second, basic summary statistics (e.g., mean and SD) are computed for each variable to help detect potential outliers. Third, we check the stability for each variable to detect potential changepoints. Fortunately, we find the daily basic statistics for most variables are stable from September to December. Therefore, all daily observations are pooled together in the subsequent analysis.

TABLE 4 The detailed information about responses and covariates

Variables		Source File	Description
Response	*Churn*	Asset	Whether the customers churn or not. Yes: 13.7%; No: 86.3%.
Asset	*MaxMVS*	Market	The maximum market value of stock.
	StdTF	Fare	The SD of total fare.
	MaxFA	Asset	The maximum floating assets.
	StdTD	Debt	The SD of total debt.
	WheTVAM	Trading	Whether the trading volume of A-shares is missing or not. Yes: 69.1%; No: 30.9%.
	WheIFM	Inflow	Whether the inflow of funds is missing or not. Yes: 81.3%; No: 18.7%.
Non- asset	*Age*	Basic	The age of customers (4 levels). <40: 20.6%; 40–50: 25.8%; 50–60: 29.3%; >60: 24.3%.
	WheHFP	FProduct	Whether the customers hold financial products or not. Yes: 71.5%; No: 28.5%.
	WheHSP	SProduct	Whether the customers hold service products or not. Yes: 86.0%; No: 14%.
	WheL	Behaviour	Whether the customers Login or not. Yes: 16.8%; No:83.2%.

Note: “FProduct” and “SProduct” represent source files of holding financial products or service products.

Preliminary analysis shows that strong correlations exist among most variables in the original data set. Therefore, we design a practical procedure for variable selection, borrowing ideas from the independence screening method (Fan & Song, 2010). Specifically, we first classify all covariates into 10 groups based on their source files. Then a logistic regression model with each single covariate is conducted for variable selection. It is notable that, the prediction performance of the customer churn model should be evaluated on a test data set. Therefore, we first order all observations by time and then split the whole data set into the training data set (the first 70% observations) and the test data set (the last 30% observations). Then we build a logistic regression model with each single covariate on the training data set, and the resulting AUC value is recorded to measure the prediction ability of the specific covariate. Next in each variable group, the covariate with the largest AUC value is chosen. This leads to the final predictor set consisting of 10 covariates. Table 4 shows the detailed information about the responses and the 10 selected covariates.

We then explore the relationship between the responses and the covariates. For illustration, we take *MaxFA* and *WheL* as examples of continuous and categorical covariates, respectively. Among the whole data, the percentage of churn customers is 13.7%. Because the continuous variable *MaxFA* has a highly right-skewed distribution, logarithmic transformation is applied. Figure 3a presents the boxplot of *MaxFA* (in log-scale) under different levels of *Churn*. As shown, customers with fewer maximum floating assets are more likely to churn. As for the categorical variable *WheL*, Figure 3b presents the spinogram of this variable under

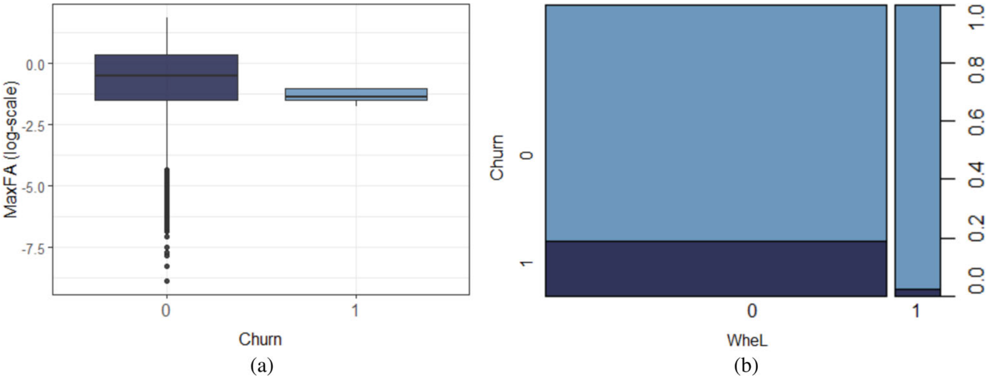

FIGURE 3 The boxplot of *Maximum Floating Assets* (in log-scale) (a) and the spinogram of *Whether to Login* (b) under the response *Churn* = 0 (non-churn status) and *Churn* = 1 (churn status)

different levels of *Churn*. As shown, customers who do not log into the system are more likely to churn.

4.2 | Customer churn prediction using SOS

We build a logistic regression model to investigate influential factors in the churn status of customers. The whole data set is quite large and cannot be analysed in the computer memory directly. To handle this huge data set, we do not consider DC methods for the lack of distributed systems in hand. We also do not consider the non-informative sub-sampling methods, because they are usually statistically less efficient than the SOS method. In addition, they require computing the optimal sampling probabilities for the entire data set, which would incur high computational cost. Based on these considerations, the SOS method is applied to estimate the logistic regression model. To evaluate the predictive ability of the SOS method, we ordered all observations by time and then split the whole data set into the training data (the first 70% observations) and the test data (the last 30% observations). Below, we would build a logistic regression model on the training data set, and then evaluate the prediction performance on the test data set.

Before applying the SOS method on the training data set, we need to set the sub-sampling size n and the sub-sampling times K. To balance between K and n, we first select the sub-sampling size n, and then determine the total sub-sampling times K based on n. Specifically, the sub-sampling size n is mainly decided by the computation resources. In this real application, we use a server with 12*Xeon Gold 6271 CPU and 64 GB RAM for computation. In addition, the securities company requires fast computation speed for updating the customer churn model conveniently day by day in the future. Based on the limited computation resources and the fast computation requirement, we fix $n = 10,000$.

For the sub-sampling times K, we apply an iterative strategy to select an appropriative value. We define $\overline{\beta}_K$ as the SOS estimate obtained with K times sub-sampling. Then, with increasing K, we compare $\overline{\beta}_{K-1}$ with $\overline{\beta}_K$. An appropriate K is chosen when the l_2-norm $\|\overline{\beta}_K - \overline{\beta}_{K-1}\|_2$ is smaller than a pre-defined threshold. Other selection methods can also be considered, such as the five-fold cross-validation method. Specifically, we can monitor the out-sample prediction accuracy under each K, and then select an optimal value that can balance the prediction accuracy and the computational time. In this work, we vary K from 10 to 100, with a step of 10. Then, we use

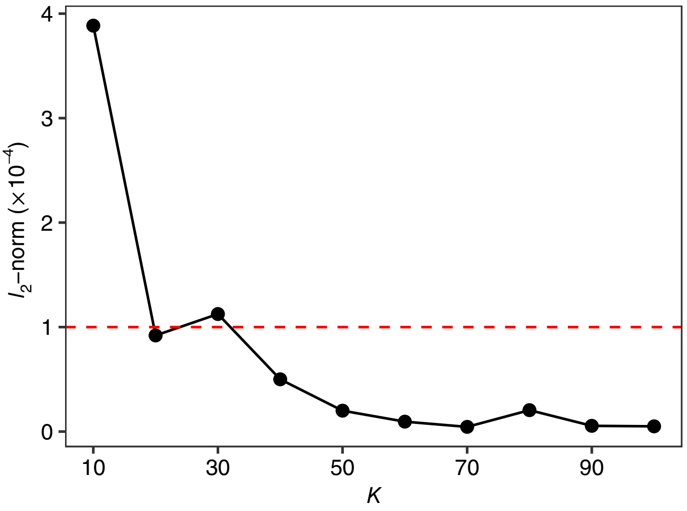

FIGURE 4 The value of l_2-norm $\|\overline{\beta}_K - \overline{\beta}_{K-1}\|_2$ under different K

the iterative strategy to select K. Under each K, we compute the l_2-norm $\|\overline{\beta}_K - \overline{\beta}_{K-1}\|_2$ and the corresponding values are plotted in Figure 4. Based on some preliminary analysis, we find the coefficients of variables are not very small. Therefore, we set a threshold 10^{-4} to find stable estimated coefficients. By this threshold, we choose $K = 20$. In addition, as shown by Figure 4, $K = 20$ is also the point with the fastest decline speed of $\|\overline{\beta}_K - \overline{\beta}_{K-1}\|_2$. Based on the above considerations, we finally choose $K = 20$.

Table 5 presents the detailed regression results for the SOS methods on the training data set. For comparison purpose, we also report the regression results on the full data set in Table 5. In general, the regression results under the training data and full data are similar. Specifically, the variable *MaxFA* plays a significantly negative role in churn status, which implies that the fewer the maximum floating assets, the more likely the customers would be to churn. This is in accordance with the descriptive results shown in Figure 3a. The variable *WheTVAM* plays a significantly positive role in the churn status, which implies that customers with no volume are more likely to churn. Similarly, customers with no inflow of funds are more likely to churn. As for non-asset-related variables, the variable *WheHSP* plays a significantly negative role in the churn status. This result indicates customers who do not hold service products are more likely to churn. In addition, the variable *Age* has significant influence on the churn status for some age groups. Specifically, customers in the age groups 50–60, and >60 are more likely to churn than those in the age group <40. Finally, the variable *WheL* has a significantly negative effect, indicating that customers who do not log in to the system in the past 20 trading days are more likely to churn. For the purpose of robustness check, we also conduct the same experiment on four daily data sets, each of which is randomly selected from September to December, respectively. The detailed results are shown in Appendix D in Data S1, which suggest stable coefficient estimates across time.

Finally, we evaluate the predictive ability of the model. In the above, we have obtained the logistic regression model on the training data set. Then, the estimated model is used to predict the churn status of customers in the test data set. We use the receiver operating characteristic (ROC) curve combined with the area-under-curve (AUC) value to measure the predictive

TABLE 5 The estimation results for logistic regression using the sequential one-step method on the training data set and full data set, respectively

		Training data				Full data			
Variable		**Est.**	**SE**	***p*-Value**	**Sig.**	**Est.**	**SE**	***p*-Value**	**Sig.**
Intercept		−23.753	0.351	<0.001	***	−23.482	0.331	<0.001	***
MaxMVS		−11.736	0.405	<0.001	***	−11.646	0.336	<0.001	***
StdTF		2.759	0.313	<0.001	***	2.829	0.261	<0.001	***
StdTD		−3.568	0.316	<0.001	***	−3.614	0.263	<0.001	***
MaxFA		−27.699	0.708	<0.001	***	−27.604	0.588	<0.001	***
WheTVAM: Yes		1.293	0.306	<0.001	***	1.333	0.255	<0.001	***
WheIFM: Yes		0.529	0.253	0.037	*	0.544	0.255	0.033	*
Age	40–50	0.426	0.306	0.164		0.450	0.255	0.077	
	50–60	0.732	0.282	0.009	**	0.693	0.255	0.007	**
	>60	1.131	0.308	<0.001	***	1.155	0.257	<0.001	***
WheL: Yes		−2.700	0.312	<0.001	***	−2.622	0.259	<0.001	***
WheHFP: Yes		0.476	0.274	0.083		0.451	0.254	0.076	
WheHSP: Yes		−0.468	0.226	0.039	*	−0.523	0.255	0.041	*

Notes: We report the estimated coefficient $\overline{\beta}_K$, standard error ($\widehat{\text{SE}}_*(\overline{\beta}_K)$), and *p*-values for all variables. The symbols *, **, *** represent a significant influence under the significance level 5%, 1% and 0.1%, respectively.

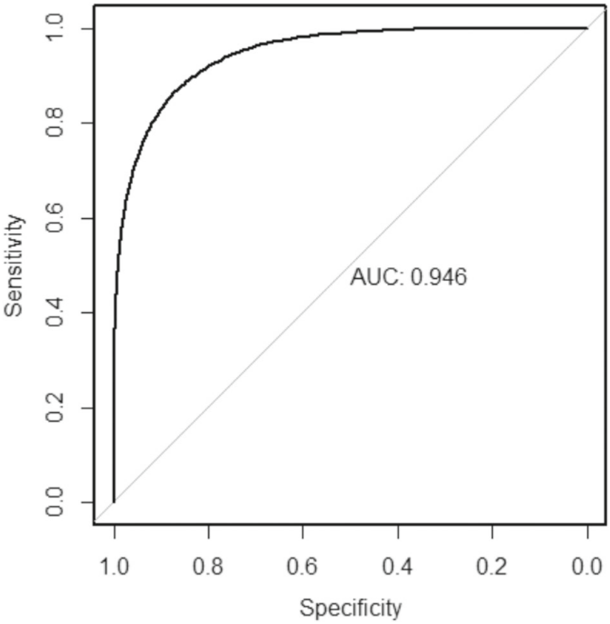

FIGURE 5 The receiver operating characteristic curve of the logistic regression using the sequential one-step method on test data

accuracy, which is shown in Figure 5. As shown, the corresponding AUC value in this data split is 0.946, which suggests very good predictive ability of the proposed model in classifying customers as churn or non-churn. For comparison purpose, we also apply the OS method with $K = 20$ and $n = 10,000$ on the training data set. The AUC value of the OS method on the test data is 0.938, which is smaller than the SOS method. We also compare the computational time for the two methods. On a server with 12× Xeon Gold 6271 CPU and 64 GB RAM, the computational time for the SOS and OS methods are 3.02 and 10.01 s, respectively. It is obvious that the SOS method behaves more computationally efficiently than the OS method does.

Finally, we present a practical customer recovery strategy using the established model in Table 5. First, we sort all customers in the test data set by their predicted churn probabilities using our model. Then, we divide all customers into 10 groups of equal size. Specifically, *group 1* consists of customers with the highest predicted churn probabilities, which we refer to as the high-risk group; and *group 10* contains customers with the lowest predicted churn probabilities, which is regarded as the low risk group. In each of the 10 groups, we calculate the ratio of truly churned customers. As shown in Figure 6, the churn ratio of all customers is only 13.7%, but the churn ratio of *group 1* is as high as 81.5%. This result verifies the predictive power of the established model. In other words, customers with high predicted churn probabilities tend to churn in reality. This finding suggests that we need to pay more attention to this group of customers and employ active strategies to retain them, such as face-to-face visits, reducing commissions, and providing exclusive services. It is also notable that *group 2* shows higher churn ratios (i.e., 31.0%) than that of all customers (13.7%). Therefore, *group 2* requires close attention and continuous monitoring.

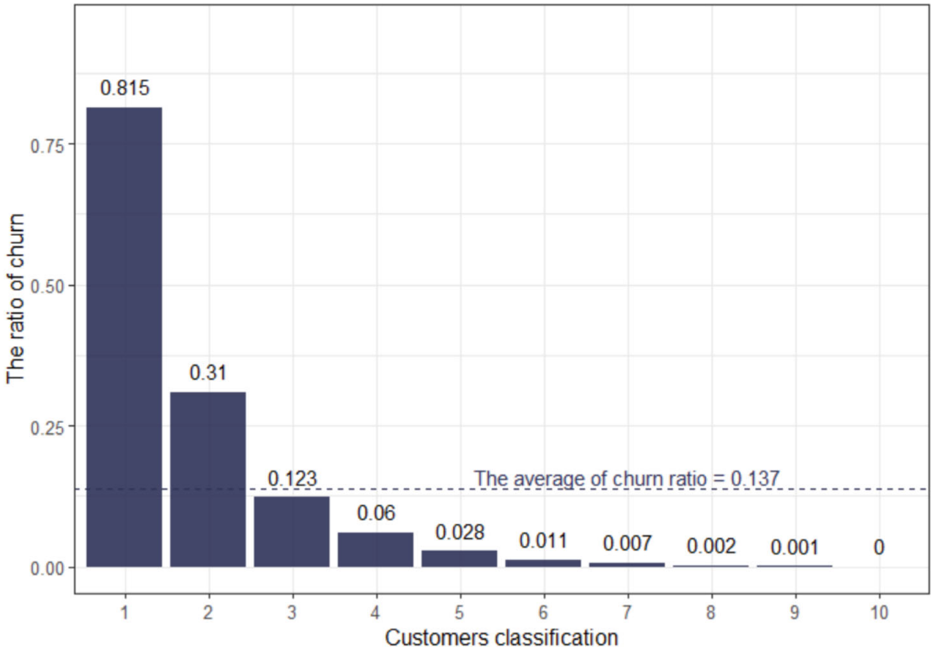

FIGURE 6 The churn ratios of 10 groups divided by their predicted churn probabilities under the sequential one-step method

5 | CONCLUDING REMARKS

In this work, we propose a sampling-based method for customer churn analysis with massive data sets. Classic sub-sampling methods require only one round of sub-sampling, but it is necessary to calculate *non-uniform* sampling probabilities of all data points. This often makes classic sub-sampling methods computationally inefficient. To address this issue, we propose the SOS method, which considers sampling data points with *uniform* probabilities but operates the sub-sampling step repeatedly. In this way, the sub-sampling cost can be largely reduced. Based on the SOS method, a sequence of estimators is computed, each of which is calculated using one-step updating based on the newly sampled sub-data. The final SOS estimate is the average of all estimators. We establish the theoretical properties of the SOS estimator. Both its bias and SE can be reduced by increasing the sub-sampling times or the sub-sample size. The performance of the SOS estimator is elaborated by simulation studies. Finally, we apply the SOS method to a real customer churn data set of a securities company. By using the SOS method, we can handle the large-scale data set, obtain useful factors that influence costumers' churn status, and achieve a high prediction accuracy for latent churn customers. It is remarkable that, although the proposed SOS method is designed for estimation of logistic regression, it can be easily extended to other generalised regression models.

We consider directions for future study. First, the SOS estimator still depends on multiple rounds of sub-sampling. New sub-sampling methods could be designed to reduce the number of sub-sampling times, which could help reduce the computational cost further. Second, the weights of previous estimators in the final SOS step are the same. However, in reality, one could consider larger weights for estimators in later steps because they have better performance. Finally, a good topic for future study when dynamic massive data are available is how to extend the SOS method for data streams.

FUNDING INFORMATION

This work was supported by the National Natural Science Foundation of China (grant numbers 72001205, 72171229, 11971504, 12071477, 11701560, 11831008); fund for building world-class universities (disciplines) of Renmin University of China, Chinese National Statistical Science Research Project (2022LD06), Foundation from Ministry of Education of China (20JZD023). This research was supported by Public Computing Cloud, Renmin University of China.

DATA AVAILABILITY STATEMENT

The data could not be made public. Because the cooperation with the company is based on the confidentiality of the original data.

ORCID

Danyang Huang https://orcid.org/0000-0002-4053-2680

REFERENCES

Ahmad, A.K., Jafar, A. & Aljoumaa, K. (2019) Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6, 1–24.

Ahn, Y., Kim, D. & Lee, D.-J. (2019) Customer attrition analysis in the securities industry: a large-scale field study in Korea. *International Journal of Bank Marketing*, 38(3), 561–577.

Ascarza, E., Neslin, S.A., Netzer, O., Anderson, Z., Fader, P.S., Gupta, S. et al. (2018) In pursuit of enhanced customer retention management: review, key issues, and future directions. *Customer Needs and Solutions*, 5, 65–81.

Battey, H., Fan, J., Liu, H., Lu, J. & Zhu, Z. (2018) Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46, 1352–1382.

Dhillon, P., Lu, Y., Foster, D.P. & Ungar, L. (2013) New subsampling algorithms for fast least squares regression. *Proceedings of the International Conference on Neural Information Processing Systems*.

Drineas, P., Magdon-Ismail, M., Mahoney, M.W. & Woodruff, D.P. (2011) Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13, 3475–3506.

Duchi, J., Hazan, E. & Singer, Y. (2011) Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12, 257–269.

Fan, J. & Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. & Song, R. (2010) Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics*, 38, 3567–3604.

Huang, C. & Huo, X. (2019) A distributed one-step estimator. *Mathematical Programming*, 174, 41–76.

Jordan, M.I., Lee, J.D. & Yang, Y. (2019) Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114, 668–681.

Kayaalp, F. (2017) Review of customer churn analysis studies in telecommunications industry. *Karaelmas Science and Engineering Journal*, 7, 696–705.

Lee, J.D., Liu, Q., Sun, Y. & Taylor, J.E. (2017) Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18, 115–144.

Lehmann, E. & Casella, G. (1998) *Theory of point estimation*, 2nd edition. New York: Springer-Verlag.

Lin, N. & Xi, R. (2011) Aggregated estimating equation estimation. *Statistics and Its Interface*, 1, 73–83.

Ma, P., Mahoney, M.W. & Yu, B. (2015) A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16, 861–919.

Ma, P. & Sun, X. (2015) Leveraging for big data regression. *Wiley Interdisciplinary Reviews Computational Statistics*, 7, 70–76.

Ma, P., Zhang, X., Xing, X., Ma, J. & Mahoney, M.W. (2020) Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. *AISTATS*, 108, 1026–1035.

Maldonado, S., Domínguez, G., Olaya, D. & Verbeke, W. (2021) Profit-driven churn prediction for the mutual fund industry: a multisegment approach. *Omega*, 100, 102380.

McDonald, R., Mohri, M., Silberman, N., Walker, D. & Mann, G.S. (2009) Efficient large-scale distributed training of conditional maximum entropy models. *Advances in Neural Information Processing Systems*, 22, 1231–1239.

Pan, R., Zhu, Y., Guo, B., Zhu, X. & Wang, H. (2020) *A sequential addressing subsampling method for massive data analysis under memory constraint*. https://doi.org/10.48550/arXiv.2110.00936

Quiroz, M., Kohn, R., Villani, M. & Tran, M.-N. (2019) Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, 114, 831–843.

Saulis, L. & Statulevičius, V. (2012) *Limit theorems for large deviations*. New York: Springer Science & Business Media.

Shao, J. (2003) *Mathematical statistics*. *Springer texts in statistics*. New York: Springer.

Wang, F., Zhu, Y., Huang, D., Qi, H. & Wang, H. (2021) Distributed one-step upgraded estimation for non-uniformly and non-randomly distributed data. *Computational Statistics & Data Analysis*, 162, 107265.

Wang, H. (2009) Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488), 1512–1524.

Wang, H., Zhu, R. & Ma, P. (2018) Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113, 829–844.

Wang, H.Y., Yang, M. & Stufken, J. (2019) Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114, 393–405.

Yu, J., Wang, H., Ai, M. & Zhang, H. (2020) Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, 117(537), 265–276.

Zhu, X., Li, F. & Wang, H. (2021) Least squares approximation for a distributed system. *Journal of Computational and Graphical Statistics*, 30(4), 1004–1018.

Zou, H. & Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4), 1509–1533.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Wang, F., Huang, D., Gao, T., Wu, S. & Wang, H. (2022) Sequential one-step estimator by sub-sampling for customer churn analysis with massive data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1–34. Available from: https://doi.org/10.1111/rssc.12597

APPENDIX A. USEFUL LEMMAS

In this section, we prove some useful lemmas.

Lemma 1. *Considering the convergence rate of $\overline{\beta}_1$, we have $\overline{\beta}_1 - \beta_0 = O_p(n^{-1/2})$.*

For generalised linear models, under conditions (C2) and (C3), the objective function $\ell_{S_1}(\beta)$ is a strictly concave function in β_0. As a result, to verify that $\overline{\beta}_1 = \hat{\beta}_1$ is $\sqrt{n}$-consistent, it suffices to follow the technique of Fan and Li (2001) to show that, for any $\epsilon > 0$, there exists a finite constant $C > 0$ such that,

$$\lim_n \sup P\left\{ \sup_{|u|=C} \ell_{S_1}(\beta_0 + n^{-1/2}u) < \ell_{S_1}(\beta_0) \right\} \geq 1 - \epsilon. \tag{A1}$$

To this end, we define $\beta_u = \beta_0 + Cu/\sqrt{n}$ where $C > 0$ is a fixed constant and $u \in \mathbb{R}^p$ is a p-dimensional vector with unit length (i.e., $\|u\| = 1$). Then, we apply the Taylor expansion and obtain

$$\begin{aligned} \sup_{|u|=1} \left\{\ell_{S_1}(\beta_0 + Cn^{-1/2}u) - \ell_{S_1}(\beta_0)\right\} &= n^{-1/2}Cu^\top \dot{\ell}_{S_1}(\beta_0) + (2n)^{-1}C^2u^\top \ddot{\ell}_{S_1}(\beta_0)u + o_p(1) \\ &= Cu^\top \mathcal{R}_1 - C^2u^\top \mathcal{R}_2 u/2 + o_p(1), \end{aligned} \tag{A2}$$

where $\mathcal{R}_1 = \dot{\ell}_{S_1}(\beta_0)/\sqrt{n}$ and $\mathcal{R}_2 = -\ddot{\ell}_{S_1}(\beta_0)/n$.

Next, we compute $E(\mathcal{R}_1)$ and $\text{var}(\mathcal{R}_1)$ as follows. First, we consider $E(\mathcal{R}_1)$.

$$\begin{aligned} E(\mathcal{R}_1) &= E\left\{E(\mathcal{R}_1|\mathcal{T})\right\} \\ &= \sqrt{n}E\left[E\left\{\frac{1}{n}\dot{\ell}_{S_1}(\beta_0)|\mathcal{T}\right\}\right] \\ &= \sqrt{n}E\left\{N^{-1}\sum_{i=1}^N S_i(\beta_0)\right\} = o(1), \end{aligned}$$

where $\mathcal{T} = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)\}$, and $S_i(\beta_0)$ is the score function of the ith observation for $1 \le i \le N$.

Second, we compute $\text{var}(\mathcal{R}_1)$. It can be calculated that

$$\begin{aligned}\text{var}(\mathcal{R}_1) &= E\left\{\text{var}(\mathcal{R}_1|\mathcal{T})\right\} + \text{var}\left\{E(\mathcal{R}_1|\mathcal{T})\right\}\\ &= E\left\{\frac{1}{n}\text{var}\left(\sum_{i\in\mathcal{S}_1} S_i(\beta_0)|\mathcal{T}\right)\right\} + \text{var}\left\{\frac{\sqrt{n}}{N}\sum_{i=1}^{N} S_i(\beta_0)\right\}\\ &= E\big(\widehat{\Sigma}^{-1}\big) + \frac{n}{N}\Sigma^{-1}, \end{aligned} \tag{A3}$$

where $\widehat{\Sigma}^{-1} = N^{-1}\sum_{i=1}^{N} S_i(\beta_0)S_i^{\top}(\beta_0)$ and $\Sigma^{-1} = E\big(\widehat{\Sigma}^{-1}\big) = E\left\{S_i(\beta_0)S_i^{\top}(\beta_0)\right\}$. Note that under condition (C1), $n/N \to 0$, then (A3) converges to Σ^{-1} as $n \to \infty$. This suggests that $\mathcal{R}_1$ is an $O_p(1)$. By a similar technique, it can be verified that $\mathcal{R}_2 \to_p \Sigma^{-1}$. Consequently, as long as C is sufficiently large, the quadratic term in (A2) dominates its linear term. Therefore, $\ell_{\mathcal{S}_1}(\beta_0 + Cn^{-1/2}u) - \ell_{\mathcal{S}_1}(\beta_0) < 0$ with probability tending to 1 as $n \to \infty$. This suggests that with probability tending to 1, a local optimiser (i.e., $\hat{\beta}_1 = \overline{\beta}_1$) exists, such that $\overline{\beta}_1 - \beta_0 = O_p(n^{-1/2})$. The optimiser satisfies $\dot{\ell}_{\mathcal{S}_1}\left(\overline{\beta}_1\right) = 0$. The conclusion is thus proved.

Lemma 2. *Denote the mth $(1 \le m \le M)$ sequential sub-sample as $\mathcal{T}_m = \{(X_m, Y_m), \cdots (X_{m+n-1}, Y_{m+n-1})\}$. Thus, given β_0, define $U_m = \left\{n^{-1}\ddot{\ell}_{\mathcal{T}_m}(\beta_0)\right\}^{-1}\left\{n^{-1}\dot{\ell}_{\mathcal{T}_m}(\beta_0)\right\}$, and $\bar{U} = M^{-1}\sum_{m=1}^{M} U_m$. The expectation and variance of $\bar{U}$ are $E(\bar{U}) = 0$ and*

$$\text{var}(\bar{U}) = \frac{1}{N}\Sigma\left\{1 + \frac{2n}{3N} + o\left(\frac{n}{N}\right)\right\}. \tag{A4}$$

Furthermore, recall that $U_{(k)} = \left\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\right\}^{-1}\left\{n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0)\right\}$, and $\bar{U}_{k^*} = k^{*-1}\sum_{k=1}^{k^*} U_{(k)}$. For any $2 \le k^* \le K$, we have

$$E(\bar{U}_{k^*}) = 0, \ \text{ and } \ \text{var}(\bar{U}_{k^*}) = \frac{1}{nk^*}\Sigma + \left(1 - \frac{1}{k^*}\right)\frac{1}{N}\Sigma\{1 + o(1)\}. \tag{A5}$$

Proof. The lemma is proved in the following two steps. In the first step, we verify $E(\bar{U}) = 0$ and Equation (A4). In the second step, we prove Equation (A5).

Step 1. For the expectation, we have $E(U_m) = E\left\{E(U_m|\mathbb{X}_m)\right\}$=0, where $\mathbb{X}_m = \{X_m, \cdots, X_{m+n-1}\}$. and

$$\begin{aligned}E(U_mU_m^{\top}) &= E\left(E\left[\left\{n^{-1}\ddot{\ell}_{\mathcal{T}_m}(\beta_0)\right\}^{-1}\left\{n^{-1}\dot{\ell}_{\mathcal{T}_m}(\beta_0)\right\}\left\{n^{-1}\dot{\ell}_{\mathcal{T}_m}(\beta_0)\right\}^{\top}\left\{n^{-1}\ddot{\ell}_{\mathcal{T}_m}(\beta_0)\right\}^{-1}\right]|\mathbb{X}_m\right)\\ &= \frac{1}{n}E\left\{n^{-1}\ddot{\ell}_{\mathcal{T}_m}(\beta_0)\right\}^{-1} = \frac{1}{n}\Sigma. \end{aligned} \tag{A6}$$

Next, we consider the variance of $\bar{U}$. We have $\text{var}(\bar{U}) = E(\bar{U}\bar{U}^{\top}) - E(\bar{U})E(\bar{U})^{\top} = E(\bar{U}\bar{U}^{\top})$, as $E\bar{U} = 0$. Furthermore,

$$E(\bar{U}\bar{U}^{\top}) = \frac{1}{M^2}E(U_1 + \cdots + U_M)^2 = \frac{1}{M^2}\left\{\sum_{m=1}^{M} E\big(U_mU_m^{\top}\big) + \sum_{m_1\neq m_2} E\big(U_{m_1}U_{m_2}^{\top}\big)\right\}.$$

It can be verified that $\sum_{m=1}^{M} E\big(U_m U_m^\top\big) = Mn^{-1}\Sigma$. Next, we focus on the calculation of $M^{-2}\sum_{m_1 \neq m_2} E\big(U_{m_1} U_{m_2}^\top\big)$. We assume that $n/N \to 0$, and $M = N - n + 1$ should be much larger than the sub-sample size n. We now discuss the two cases.

First, when $|m_1 - m_2| \geq n$, we have $E(U_{m_1} U_{m_2}^\top) = 0$. There are $(M - n)(M - n + 1)$ pairs of m_1 and m_2 in this case. Second, let $|m_1 - m_2| = m'$; when $0 < m' < n$, we have $E\big(U_{m_1} U_{m_2}^\top\big) = n^{-1}(n - m')\Sigma$, and there are $2(M - m')$ pairs of m_1 and m_2 in this case. As a result, $\sum_{0<m'<n} E\big(U_{m_1} U_{m_2}^\top\big) = 2n^{-1}c_1\Sigma$, where $c_1 = n(n-1)(3M - n - 1)/6$. We can derive that $E(\bar{U}\bar{U}^\top) = (nM)^{-2}(nM + 2c_1)\Sigma$. Then, we have

$$\begin{aligned} E(\bar{U}\bar{U}^\top) &= \frac{(N-n+1) + (n-1)(3N - 4n + 2)/3}{n(N-n+1)^2}\Sigma \\ &= \frac{1}{N}\left\{1 + \frac{2n}{3M} + o\left(\frac{n}{N}\right)\right\}\Sigma. \end{aligned}$$

This finishes the first step.

Step 2. By a similar proof technique to that for Step 1, we immediately have $E(\bar{U}_{k^*}) = 0$. We then focus on the computations of $\mathrm{var}(\bar{U}_{k^*})$. To study the variance of $\bar{U}_{k^*}$, define $E^*(\cdot)$ and $\mathrm{var}^*(\cdot)$ as the conditional expectation and conditional variance, respectively, given $\mathcal{M} = \{U_1, U_2, \cdots, U_M\}$. We know that $\mathrm{var}(\bar{U}_{k^*}) = E\{\mathrm{var}^*(\bar{U}_{k^*})\} + \mathrm{var}\{E^*(\bar{U}_{k^*})\}$. Then, we study $E\{\mathrm{var}^*(\bar{U}_{k^*})\}$ and $\mathrm{var}\{E^*(\bar{U}_{k^*})\}$ separately.

First, we compute $E\{\mathrm{var}^*(\bar{U}_{k^*})\}$, which is

$$\begin{aligned} E\{\mathrm{var}^*(\bar{U}_{k^*})\} &= \frac{1}{k^*}\sum_{k=1}^{k^*} E\left\{\mathrm{var}^*(U_{(k)})\right\} = \frac{1}{k^*}E\left\{E^*(U_{(k^*)} - \bar{U})(U_{(k^*)} - \bar{U})^\top\right\} \\ &= \frac{1}{k^*M}E\left\{\sum_{m=1}^{M}(U_m - \bar{U})(U_m - \bar{U})^\top\right\} \\ &= \frac{1}{k^*}\left(EU_mU_m^\top - E\bar{U}\bar{U}^\top\right). \end{aligned}$$

Second, we consider $\mathrm{var}\{E^*(\bar{U}_{k^*})\}$. We have

$$E^*(\bar{U}_{k^*}) = \frac{1}{k^*}\sum_{k=1}^{k^*} E^*\{U_{(k)}\} = \frac{1}{m}\sum_{m=1}^{M} U_m = \bar{U}.$$

Then, $\mathrm{var}\left\{E^*(\bar{U}_{k^*})\right\} = \mathrm{var}(\bar{U}) = E(\bar{U}\bar{U}^\top)$. Thus,

$$\begin{aligned} \mathrm{var}(\bar{U}_{k^*}) &= E\left\{\mathrm{var}^*(\bar{U}_{k^*})\right\} + \mathrm{var}\left\{E^*(\bar{U}_{k^*})\right\} \\ &= \frac{1}{k^*}EU_mU_m^\top - \frac{1}{k^*}E\bar{U}\bar{U}^\top + E\bar{U}\bar{U}^\top \\ &= \frac{1}{nk^*}\Sigma + \left(1 - \frac{1}{k^*}\right)\mathrm{var}(\bar{U}). \end{aligned} \tag{A7}$$

This completes the proof.

Lemma 3. *Define R_{1k} and R_{2k} as follows,*

$$R_{1k} = \left\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\overline{\beta}_{k-1})\right\}^{-1}\left[(\overline{\beta}_{k-1}-\beta_0)^\top\left\{n^{-1}\Delta_{k-1,j}\right\}(\overline{\beta}_{k-1}-\beta_0)\right], k \geq 2$$

$$R_{2k} = \left[\left\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\right\}^{-1} - \left\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\overline{\beta}_{k-1})\right\}^{-1}\right] n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0), k \geq 2.$$

Here $\Delta_{k,j} = (\Delta_{k,j,i_1i_2}) \in \mathbb{R}^{p\times p}$ for $1 \leq j \leq p$, and $\Delta_{k,j,i_1i_2} = \partial\dot{\ell}_{j,\mathcal{S}_{k+1}}(\beta)/\partial\beta_{i_1}\partial\beta_{i_2}|_{\beta=\tilde{\beta}_k}$, $\dot{\ell}_{j,\mathcal{S}_{k+1}}(\beta)$ is the j th element of $\dot{\ell}_{\mathcal{S}_{k+1}}(\beta)$ and $\tilde{\beta}_k = \eta\hat{\beta}_k + (1-\eta)\beta_0$ for some $0 \leq \eta \leq 1$. In particular, $R_{11} = \left\{n^{-1}\ddot{\ell}_{\mathcal{S}_1}(\beta_0)\right\}^{-1}\left[(\overline{\beta}_1-\beta_0)^\top\left\{n^{-1}\Delta_{0,j}\right\}(\overline{\beta}_1-\beta_0)\right]$, with $\Delta_{0,j,i_1,i_2} = \partial\dot{\ell}_{j,\mathcal{S}_1}(\beta)/\partial\beta_{i_1}\partial\beta_{i_2}|_{\beta=\tilde{\beta}_1}$ and $R_{21} = 0$. Then, if $\|\overline{\beta}_k - \beta_0\| \leq \kappa_1\left\{\sqrt{1/(nk)+1/N}\right\}\{1+o_p(1)\}$, then for any $k^* \geq 2$, we have $\Delta_{k^*} = {k^*}^{-1}\sum_{k=1}^{k^*}(R_{1k}+R_{2k}) \leq \kappa_2[(\log k^*/n)^{1/2}\{1/\ (nk^*)+1/N\}^{1/2}]\{1+o_p(1)\}$. Here, κ_1, κ_2 are some positive constants.

Proof. By condition (C4), it can be calculated that

$$\begin{aligned}\|\Delta_{k^*}\| \leq{}& C\max_{1\leq k\leq k^*}\lambda_{\max}(\hat{\Sigma}_k)\left[\frac{1}{k^*}\sum_{k=1}^{k^*}(\overline{\beta}_{k-1}-\beta_0)^\top(\overline{\beta}_{k-1}-\beta_0)\right] \\ &+ C\frac{1}{k^*}\sum_{k=1}^{k^*}\|\overline{\beta}_{k-1}-\beta_0\|\|n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0)\|. \end{aligned} \tag{A8}$$

Here, $\lambda_{\max}(M)$ denotes the largest absolute eigenvalue of M, $\hat{\Sigma}_k = \left\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\overline{\beta}_{k-1})\right\}^{-1}$, $\Sigma_k = \left\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\right\}^{-1}$, and we define $\overline{\beta}_0 = \overline{\beta}_1$. To analyse Equation (A8), we then study the two terms on the right of the equation separately in the following two steps.

Step 1. First, we are going to show that $\max_{1\leq k\leq k^*}\lambda_{\max}(\hat{\Sigma}_k)$ is an $O_p(1)$. It is sufficient to show that with probability tending to 1, we have

$$\tau_{\min} \leq \min_{1\leq k\leq k^*}\lambda_{\min}(\hat{\Sigma}_k) \leq \max_{1\leq k\leq k^*}\lambda_{\max}(\hat{\Sigma}_k) \leq \tau_{\max}, \tag{A9}$$

for some positive constants $\tau_{\min} < \tau_{\max} < \infty$. By condition (C3), we can find two positive constants, such that $2\tau_{\min} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 2^{-1}\tau_{\max}$, for two positive constants $\tau_{\min} < \tau_{\max} < \infty$. Thus, we know immediately that $2\tau_{\min} \leq \inf_{\|r\|=1} r^\top\Sigma r \leq \sup_{\|r\|=1} r^\top\Sigma r \leq 2^{-1}\tau_{\max}$ for a p-dimensional vector r. Thus, the desired conclusion (A9) is implied by

$$P\left(\max_{1\leq k\leq k^*}\sup_{\|r\|=1}\left|r^\top(\hat{\Sigma}_k-\Sigma)r\right| > \epsilon\right) \to 0, \tag{A10}$$

where $\epsilon > 0$ is an arbitrary positive number. Then, the left-hand side of (A10) is bounded by $\sum_{1\leq k\leq k^*} P\left(\sup_{\|r\|=1}|r^\top(\hat{\Sigma}_k-\Sigma)r| > \epsilon\right)$. Note that for any k, we have

$$|r^\top(\hat{\Sigma}_k-\Sigma)r| \leq \sum_{j_1j_2}|r_{j_1}|\times|r_{j_2}|\times|\hat{\sigma}_{j_1j_2,k}-\sigma_{j_1j_2}|$$

$$
\begin{aligned}
&\leq \max_{1\leq j_1,j_2\leq p}|\hat{\sigma}_{j_1j_2,k}-\sigma_{j_1j_2}|\sum_{1\leq j_1,j_2\leq p}|r_{j_1}|\times|r_{j_2}| \\
&\leq \max_{1\leq j_1,j_2\leq p}|\hat{\sigma}_{j_1j_2,k}-\sigma_{j_1j_2}|\left(\sum_j |r_j|\right)^2 \\
&\leq p\max_{1\leq j_1,j_2\leq p}|\hat{\sigma}_{j_1j_2,k}-\sigma_{j_1j_2}|.
\end{aligned}
$$

Consequently, the left-hand side of (A10) can be further bounded by

$$
\begin{aligned}
&\leq \sum_{1\leq k\leq k^*} P\left(\max_{1\leq j_1,j_2\leq p}|\hat{\sigma}_{j_1j_2,k}-\sigma_{j_1j_2}|>\frac{\epsilon}{p}\right) \\
&\leq \sum_{1\leq k\leq k^*}\sum_{1\leq j_1,j_2\leq p} P\left(|\hat{\sigma}_{j_1j_2,k}-\sigma_{j_1j_2}|>\frac{\epsilon}{p}\right). \qquad \text{(A11)}
\end{aligned}
$$

Next, under condition (C5), it can be proved that $P(|\hat{\sigma}_{j_1j_2,k}-\sigma_{j_1j_2}|>\epsilon)\leq C_1\exp(-C_2n\epsilon^2)$ for two positive constants C_1 and C_2 by theorem 3.2 on p. 45 of Saulis and Statulevičius (2012) and the proof technique of Wang (2009). Thus, (A11) can be bounded by

$$
\begin{aligned}
&\leq k^*p^2C_1\exp\left(-C_2n\frac{\epsilon^2}{p^2}\right) \\
&= C_1\exp\left\{\log(k^*)+2\log(p)-C_2n\frac{\epsilon^2}{p^2}\right\} \\
&= C_1\exp\left\{\log(k^*)\left(1-C_2\frac{n\epsilon^2}{\log(k^*)p^2}\right)+2\log(p).\right\}. \qquad \text{(A12)}
\end{aligned}
$$

As under condition (C1), $\log(K)=o(n)$, the right-hand side of (A14) converges to 0 as $n\to\infty$. This implies that $\max_{1\leq k\leq k^*}\lambda_{\max}(\hat{\Sigma}_k)$ is an $O_p(1)$.

Step 2. In the second step, we investigate $\max_{1\leq k\leq k^*}\|n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0)\|$. Similar to Step 1, it can be proved that $P\left(\max_{1\leq k\leq k^*}\|n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0)\|>\epsilon\right)$ can be bounded by

$$
\begin{aligned}
&\leq \sum_{1\leq k\leq k^*}\sum_{1\leq j\leq p} P\left(|n^{-1}\ell_{j,\mathcal{S}_k}(\beta_0)|>\frac{\epsilon}{p}\right). \qquad \text{(A13)}\\
&\leq k^*pC_1\exp\left(-C_2n\frac{\epsilon^2}{p^2}\right) \\
&= C_1\exp\left\{\log(k^*)+\log(p)-C_2n\frac{\epsilon^2}{p^2}\right\} \\
&= C_1\exp\left\{\log(k^*)\left(1-C_2\frac{n\epsilon^2}{\log(k^*)p^2}\right)+\log(p)\right\}. \qquad \text{(A14)}
\end{aligned}
$$

If we replace ϵ by $\gamma_n\epsilon'$, then we can verify that if $\gamma_n=\sqrt{\log(k^*)/n}$, there exists ϵ' such that the right-hand side of (A14) could be arbitrarily small. This leads to $\max_{1\leq k\leq k^*}\|n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0)\|=O_p(\sqrt{\log(k^*)/n})$.

As a result, the right-hand side of (A8) is further bounded by

$$\begin{aligned}
&\le C\left(\sum_{1\le k\le k^*}\|\overline{\beta}_{k-1}-\beta_0\|^2/k^*\right)+\left(\sum_{2\le k\le k^*}\|\overline{\beta}_{k-1}-\beta_0\|/k^*\right)O_p(\sqrt{\log(k^*)/n})\\
&=O_p\left(\frac{\log k^*}{nk^*}\right)+O_p\left\{\frac{\log k^*}{n}\left(\frac{1}{nk^*}+\frac{1}{N}\right)\right\}^{1/2}=O_p\left\{\frac{\log k^*}{n}\left(\frac{1}{nk^*}+\frac{1}{N}\right)\right\}^{1/2}.
\end{aligned}$$

This completes the proof.

APPENDIX B. PROOF OF THE THEOREMS

In this section, we provide the detailed proof of the theorems and proposition to establish the theoretical properties of the proposed estimator.

B.1 Proof of Theorem 1

This theorem is to be proved in two parts. In the first part, we prove Theorem 1 (1). In the second part, we verify the asymptotic normality of the SOS estimator.

Part 1. We prove Theorem 1 (1) in the following two steps. In the first step, we show that for any $k^*\ge 2$, $\overline{\beta}_k^* = \beta_0 + {k^*}^{-1}\sum_{k=1}^{k^*}(R_{1k}+R_{2k}) - {k^*}^{-1}\sum_{k=1}^{k^*}U_{(k)}$, where R_{1k} and R_{2k} are defined in Lemma 3. In the second step, denote $\Delta = K^{-1}\sum_{k=1}^{K}(R_{1k}+R_{2k})$, and recall that $U_{(k)}=\left\{n^{-1}\ddot{\ell}_{S_k}(\beta_0)\right\}^{-1}\left\{n^{-1}\dot{\ell}_{S_k}(\beta_0)\right\}$, $\bar{U}_K = K^{-1}\sum_{k=1}^{K}U_{(k)}$, we then verify that $E\bar{U}_K=0$, $\mathrm{var}(\overline{\beta}_k)=\mathrm{var}(\bar{U}_K)\{1+o(1)\}=\{1/(nK)+1/N\}\,\Sigma\{1+o(1)\}$ and $\Delta = O_p\left[\sqrt{(\log k^*/n)\{1/(nk^*)+1/N\}}\right]$.

Step 1. By the Taylor expansion, we have

$$\begin{aligned}
\dot{\ell}_{S_{k+1}}(\beta_0) &= \dot{\ell}_{S_{k+1}}(\overline{\beta}_k)+\ddot{\ell}_{S_{k+1}}(\overline{\beta}_k)(\beta_0-\overline{\beta}_k)\\
&\quad+\begin{pmatrix}(\overline{\beta}_k-\beta_0)^\top\Delta_{k,1}(\overline{\beta}_k-\beta_0)\\ \dots\\ (\overline{\beta}_k-\beta_0)^\top\Delta_{k,p}(\overline{\beta}_k-\beta_0)\end{pmatrix},
\end{aligned}\tag{B1}$$

Thus, based on (1), we have

$$\begin{aligned}
\hat{\beta}_{k+1} &= \overline{\beta}_k-\left\{\ddot{\ell}_{S_{k+1}}(\overline{\beta}_k)\right\}^{-1}\left\{\dot{\ell}_{S_{k+1}}(\overline{\beta}_k)-\dot{\ell}_{S_{k+1}}(\beta_0)\right\}-\left\{\ddot{\ell}_{S_{k+1}}(\overline{\beta}_k)\right\}^{-1}\dot{\ell}_{S_{k+1}}(\beta_0) \qquad \text{(B2)}\\
&=\overline{\beta}_k-\left\{\ddot{\ell}_{S_{k+1}}(\overline{\beta}_k)\right\}^{-1}\left\{\ddot{\ell}_{S_{k+1}}(\overline{\beta}_k)(\overline{\beta}_k-\beta_0)-\begin{pmatrix}(\overline{\beta}_k-\beta_0)^\top\Delta_{k,1}(\overline{\beta}_k-\beta_0)\\ \dots\\ (\overline{\beta}_k-\beta_0)^\top\Delta_{k,p}(\overline{\beta}_k-\beta_0)\end{pmatrix}\right\}\\
&\quad-\left\{\ddot{\ell}_{S_{k+1}}(\overline{\beta}_k)\right\}^{-1}\dot{\ell}_{S_{k+1}}(\beta_0)\\
&=\beta_0+\left\{n^{-1}\ddot{\ell}_{S_{k+1}}(\overline{\beta}_k)\right\}^{-1}\left[(\overline{\beta}_k-\beta_0)^\top\left\{n^{-1}\Delta_{k,j}\right\}(\overline{\beta}_k-\beta_0)\right]\\
&\quad+\left[\left\{n^{-1}\ddot{\ell}_{S_{k+1}}(\beta_0)\right\}^{-1}-\left\{n^{-1}\ddot{\ell}_{S_{k+1}}(\overline{\beta}_k)\right\}^{-1}\right]n^{-1}\dot{\ell}_{S_{k+1}}(\beta_0)\\
&\quad-\left\{n^{-1}\ddot{\ell}_{S_{k+1}}(\beta_0)\right\}^{-1}n^{-1}\dot{\ell}_{S_{k+1}}(\beta_0). \qquad \text{(B2)}
\end{aligned}$$

By (B3), we can rewrite $\overline{\beta}_{k^*}$ as

$$\overline{\beta}_k^* = \beta_0 + \frac{1}{k^*}\sum_{k=1}^{k^*}(R_{1k} + R_{2k}) - \frac{1}{k^*}\sum_{k=1}^{k^*} U_{(k)}, \quad \text{for any } k^* \geq 2. \tag{B4}$$

Step 2. This step is decomposed into two sub-steps. In Step 2.1, we verify that $\|\beta_{k^*} - \beta_0\| \leq \kappa_1(\sqrt{1/(nk^*) + 1/N})\{1 + o_p(1)\}$ for some constant κ_1 in a deductive way. In Step 2.2, we prove the remaining results.

Step 2.1. First, we consider $k^* = 2$. By (B4), it can be verified that $\overline{\beta}_2 = \beta_0 + 2^{-1}$ $\left(\left\{n^{-1}\ddot{\ell}_{S_2}(\overline{\beta}_1)\right\}^{-1}[(\overline{\beta}_1 - \beta_0)^\top\{n^{-1}\Delta_{1,j}\}(\overline{\beta}_1 - \beta_0)] + \left\{n^{-1}\ddot{\ell}_{S_1}(\beta_0)\right\}^{-1}\left[(\overline{\beta}_1 - \beta_0)^\top\{n^{-1}\Delta_{0,j}\}\$\overline{\beta}_1 - \beta_0\right)\right]$ $+ \left[\left\{n^{-1}\ddot{\ell}_{S_2}(\beta_0)\right\}^{-1} - \left\{n^{-1}\ddot{\ell}_{S_2}(\overline{\beta}_1)\right\}^{-1}\right] n^{-1}\dot{\ell}_{S_2}(\beta_0)\big) - \bar{U}_2$. From Lemma 2, we have $E(\bar{U}_2) = 0$ and $\operatorname{var}(\bar{U}_2) = (2n)^{-1}\Sigma + (1 - 1/2)N^{-1}\Sigma\{1 + o(1)\}$. Consequently, $\|\bar{U}_2\|^2 \leq 2\left\{(2n)^{-1}\Sigma + (1 - 1/2)N^{-1}\Sigma\right\}\{1 + o_p(1)\}$. Furthermore, by Lemma 1, we have $2^{-1}\{R_{11} + R_{12}\} = O_p(1/n)$ and $2^{-1}\{R_{21} + R_{22}\} = O_p(1/n)$. As a result, $2^{-1}\sum_{k=1}^{2}(R_{1k} + R_{2k}) = o_p(\bar{U}_2)$.

Next, we assume that for any $2 \leq k \leq k^* - 1$, we have $\|\bar{U}_k\|^2 \leq 2\left\{(kn)^{-1}\Sigma + (1 - 1/k)N^{-1}\Sigma\right\}\{1 + o_p(1)\}$ and $k^{-1}\sum_{\tilde{k}=1}^{k}(R_{1\tilde{k}} + R_{2\tilde{k}}) = o_p(\bar{U}_k)$. This suggests that $\|\overline{\beta}_k - \beta_0\| \leq \kappa_1(\sqrt{1/nk + 1/N})\{1 + o_p(1)\}$ for some $\kappa_1 > 0$. By Lemma 2, we know that $E(\bar{U}_{k^*}) = 0$ and $\operatorname{var}(\bar{U}_{k^*}) = \frac{1}{k^*n}\Sigma + (1 - 1/k^*)\frac{1}{N}\Sigma\{1 + o(1)\}$. Furthermore, by Lemma 3, we have $\frac{1}{k^*}\sum_{k=1}^{k^*}(R_{1k} + R_{2k}) = O_p\left[\sqrt{(\log k^*/n)\{1/(nk^*) + 1/N\}}\right] = o_p(\sqrt{1/(nk^*) + 1/N}) = o_p(\bar{U}_{k^*})$. The penultimate equality holds, as $\log k^* = o(n)$. As a result, we have proved $\|\overline{\beta}_{k^*} - \beta_0\| \leq \kappa_1(\sqrt{1/(nk^*) + 1/N})\{1 + o_p(1)\}$.

Step 2.2. Finally, by (B4), we know

$$\overline{\beta}_K = \beta_0 + \frac{1}{K}\sum_{k=1}^{K}(R_{1k} + R_{2k}) - \bar{U}_K,$$

by the results of *Step 2.1* and Lemma 3, we have $K^{-1}\sum_{k=1}^{K}(R_{1k} + R_{2k}) = O_p[(\log K/n)\{1/(nK) + 1/N\}]^{1/2}$. In addition, by Lemma 2, we have $E\bar{U}_K = 0$, and $\operatorname{var}(\bar{U}_K) = \Sigma\left\{1/(nK) + N^{-1}(1 - 1/K)\right\}\{1 + o(1)\}$. This accomplishes the proof of *Step 2.2*. Combining the results of Steps 1 and 2, we finish the first part.

Part 2. In the second part, we verify the asymptotic normality of the SOS estimator. As from Part 1, we have verified that $K^{-1}\sum_{k=1}^{K}(R_{1k} + R_{2k}) = o_p(\bar{U}_K)$, it suffices to study $\bar{U}_K$. To this end, we decompose $\bar{U}_K$ into $\bar{U}_K = \bar{U}_K^{(1)} + \bar{U}_K^{(2)}$ with $\bar{U}_K^{(1)} = K^{-1}\sum_{K=1}^{K}\Sigma\left\{n^{-1}\dot{\ell}_{S_k}(\beta_0)\right\}$ and $\bar{U}_K^{(2)} = K^{-1}\sum_{K=1}^{K}\left[\left\{n^{-1}\ddot{\ell}_{S_k}(\beta_0)\right\}^{-1} - \Sigma\right]\left\{n^{-1}\dot{\ell}_{S_k}(\beta_0)\right\}$. We then calculate the two terms separately.

We first compute $\bar{U}_K^{(2)}$, by the same analysis of Lemma 2, it can be proved that $E(\bar{U}_K^{(2)}) = 0$, and $\operatorname{var}(\bar{U}_K^{(2)}) = \left\{(nK)^{-1} + N^{-1}(1 - 1/K)\right\}o(1) = o(1/N)$ when $nK/N \to \infty$ and $K \to \infty$. This suggests that $\bar{U}_K^{(2)} = o_p(1/\sqrt{N})$. Next, we prove the normality of $\sqrt{N}\bar{U}_K^{(1)}$. To this end, it suffices to show that its characteristic function $f(t) = E\left[\exp(it^\top K^{-1}\sum_{K=1}^{K}\left\{n^{-1}\dot{\ell}_{S_k}(\beta_0)\right\}\right] \to \exp\left\{-t^\top\Sigma^{-1}t/2\right\}$. Denote $\overline{L} = M^{-1}\sum_{m=1}^{M}\left\{n^{-1}\dot{\ell}_{\mathcal{T}_m}(\beta_0)\right\}$. Then, we have

$$f(t) = E\left[\exp\left\{\frac{it^\top}{\tau K}\sum_{k=1}^{K}\left[\{n^{-1}\dot{\ell}_{S_k}(\beta_0)\} - \overline{L}\right]\right\}\exp\left(\frac{it^\top}{\tau}\overline{L}\right)\right]$$

$$= E\left[\exp\left\{\frac{it^\top}{\tau nK}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_k}\left(\dot{\ell}_i(\beta_0) - \overline{L}\right)\right\}\exp\left(\frac{it^\top}{\tau}\overline{L}\right)\right]$$
$$= E\left[\exp\left\{\frac{it^\top}{\tau\sqrt{nK}}Z_1\right\}\exp\left(\frac{it^\top}{\tau\sqrt{N}}Z_2\right)\right],$$

where $Z_1 = (nK)^{-1/2}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_k}\left\{\dot{\ell}_i(\beta_0) - \overline{L}\right\}$ and $Z_2 = \sqrt{N}\overline{L}$. Subsequently, we consider the following three cases to prove the convergence of $f(t)$.

Case 1. We first consider the case of $N/(nK) \to 0$. Then, we have $\tau^2 nK = (1 + nK/N) \to \infty$. Note that by a similar analysis technique to that in Lemma 2, we have $E(Z_1) = 0$ and $\text{var}(Z_1) = O(1)$. Consequently, we have $Z_1 = O_p(1)$. This leads to $\exp\left\{it^\top(\tau\sqrt{nK})^{-1}Z_1\right\} \to_p 1$. Accordingly, $f(t)$ shares the same asymptotic limit with $E\left[\exp\left\{it^\top(\tau\sqrt{N})^{-1}Z_2\right\}\right]$. Note that $E\left[\exp\left\{it^\top(\tau\sqrt{N})^{-1}Z_2\right\}\right] \to \exp\left(-t^2/2\right)$ due to the following two reasons: (1) $\tau^2 N = (N/nK + 1) \to 1$; and (2) $Z_2 = \sqrt{N}(Mn)^{-1}\sum_{i=1}^{M} \quad n\dot{\ell}_i(\beta_0)\{1 + o_p(1)\} = N^{-1/2}\sum_{i=1}^{N}\dot{\ell}_i(\beta_0)\{1 + o_p(1)\} \to_d N(0, \Sigma^{-1})$ by the central limit theorem, as $n/N \to 0$. This finishes the proof of Case 1.

Case 2. We next consider the case of $N/(nK) \to \infty$. Because $\tau^2 N = (N/nK + 1) \to 0$, and $Z_2 \to_d N(0, \Sigma)$, we should have $(\tau\sqrt{N})^{-1}Z_2 \to_p 0$ and it leads to $E\left[\exp\{it^\top(\tau\sqrt{N})^{-1}Z_2\}\right] \to 1$. Then, by the dominated convergence theorem, $f(t)$ has the same asymptotic limit as $E\left[\exp\{it^\top(\tau\sqrt{nK})^{-1}Z_1\}\right]$. This limit term could be verified to be $\exp(-t^2/2)$, due to the following two reasons: (1) $\tau^2 nK = (1 + nK/N) \to 1$; (2) $Z_1 \to_d N(0, 1)$ by the Central Limit Theorem. This finishes the proof of CASE 2.

Case 3. We finally consider the case that $N/(nK) \to \kappa$ for some constant $\kappa > 0$. We then decompose $f(t)$ into $f_1(t) + f(t) - f_1(t)$ with $f_1(t) = E\left[\tilde{\Delta}_2 \exp\left\{it^\top(\tau\sqrt{N})^{-1} \; Z_2\right\}\right]$ and $f(t) - f_1(t) = E\left[(\tilde{\Delta}_1 - \tilde{\Delta}_2)\exp\left\{it^\top(\tau\sqrt{N})^{-1}Z_2\right\}\right]$, where $\tilde{\Delta}_1 = E\left[\exp\left\{it^\top(\tau\ \sqrt{nK}\,)^{-1}Z_1|\mathcal{T}\right\}\right]$ and $\tilde{\Delta}_2 = \exp\left[-t^\top\Sigma^{-1}t/(2nK\tau^2)\right]$. Since $Z_1 \to_d N(0, 1)$, we then have $\tilde{\Delta}_1 - \tilde{\Delta}_2 \to_p 0$ conditional on $\mathcal{T}$. Thus by the dominated convergence theorem, we have $f(t) - f_1(t) \to 0$. Consequently, $f(t)$ shares the same asymptotical limit with $f_1(t)$. This implies that it suffices to verify that $f_1(t) \to \exp(-t^\top\Sigma^{-1}t/2)$. Note that $\tau^2 nK = 1 + 1/\kappa$. Then, we have $\tilde{\Delta}_2 \to \exp[-t^\top\Sigma^{-1}t/\{2(1 + 1/C)\}]$. Meanwhile, as $\tau^2 N \to 1 + \kappa$ and $Z_2 \to N(0, 1)$, we then should have $E\left[\exp\left\{it^\top(\tau\sqrt{N})^{-1}Z_2\right\}\right] \to \exp\{-t^\top\Sigma^{-1}t/2(1 + \kappa)\}$. It can be verified that $f_1(t) \to \exp\left(-t^\top\Sigma^{-1}t/\{2(1 + \kappa)\} - t^\top\Sigma^{-1}t/[2\{1 + 1/\kappa\}]\right) = \exp(-t^\top\Sigma^{-1}t/2)$. This completes the proof of Case 3 and Part 2. Combining the results of Part 1 and Part 2, we accomplish the whole theorem proof.

B.2 Proof of Proposition 1

To prove Proposition 1, we first verify $\overline{\beta}_K^{\text{OS}} - \beta_0 = \Delta_{\text{os}} - \bar{U}_K$ with $\Delta_{\text{os}} = O_p(1/n)$. Subsequently, by Theorem 1, it immediately leads to the asymptotic normality of $\overline{\beta}_K^{\text{OS}}$.

Recall that $\overline{\beta}_K^{\text{OS}} = K^{-1}\sum_{k=1}^{K}\hat{\beta}_{k,\text{mle}}$. By Taylor's expansion, we know

$$\hat{\beta}_{k,\text{mle}} - \beta_0 = -\left\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\right\}^{-1}\left\{n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0)\right\} + \Delta_{os}^{(k)}\{1 + o_p(1)\}.$$

Here $\Delta_{os}^{(k)} = \left\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\right\}^{-1}\left\{(\hat{\beta}_{k,\mathrm{mle}} - \beta_0)^\top \dots \ell_{\mathcal{S}_k}(\theta_0)(\hat{\beta}_{k,\mathrm{mle}} - \beta_0)\right\}$, and $\left\{(\hat{\beta}_{k,\mathrm{mle}} - \beta_0)^\top \ \dots \ell_{\mathcal{S}_k} (\theta_0)(\hat{\beta}_{k,\mathrm{mle}} - \beta_0)\right\}$ is defined similarly to that in Equation (B1). Note that by Lemma 1, we have $\hat{\beta}_{k,\mathrm{mle}} - \beta_0 = O_p(1/\sqrt{n})$, and by Condition (C4), we know that $\Delta_{os}^{(k)} = O_p(1/n)$ is the bias term. Then, it holds that

$$\overline{\beta}_K^{\mathrm{OS}} - \beta_0 = K^{-1}\sum_{k=1}^{K} \Delta_{os}^{(k)} - \bar{U}_K = O_p(1/n) - \bar{U}_K,$$

where $\Delta_{\mathrm{os}} = K^{-1}\sum_{k=1}^{K}\Delta_{os}^{(k)}$. Furthermore, if one requires $n^2/N \to \infty$, then we have $\Delta_{os} = o_p(1/\sqrt{N}) = o_p\left\{\sqrt{1/(nK) + 1/N}\right\}$. Because we have verified that $\{1/(nK) + 1/N\}^{-1/2}\bar{U}_K \to_d N(0, \Sigma)$ in Appendix B.1, this accomplishes the whole proof.

B.3 Proof of Theorem 2

First, we consider the expectation of $\widehat{\mathrm{SE}}^2(\overline{\beta}_K)$. We have

$$\begin{aligned}
\widehat{\mathrm{SE}}^2(\overline{\beta}_K) &= \frac{c_0}{K-1}\sum_{k=1}^{K}\left(U_{(k)} - \bar{U} + \bar{U} - \bar{U}_K\right)\left(U_{(k)} - \bar{U} + \bar{U} - \bar{U}_K\right)^\top \\
&= \frac{c_0}{K-1}\left\{\sum_{k=1}^{K}(U_{(k)} - \bar{U})(U_{(k)} - \bar{U})^\top - K(\bar{U} - \bar{U}_K)(\bar{U} - \bar{U}_K)^\top\right\} \\
&= \frac{c_0}{K-1}(A_1 - A_2),
\end{aligned}$$

where $A_1 = \left\{\sum_{k=1}^{K}(U_{(k)} - \bar{U})(U_{(k)} - \bar{U})^\top\right\}$ and $A_2 = K\left\{(\bar{U} - \bar{U}_K)(\bar{U} - \bar{U}_K)^\top\right\}$, and $c_0 = n\left\{(nK)^{-1} + N^{-1}\right\}$. We next consider $E(A_1)$ and $E(A_2)$ separately. Given $\mathcal{M}$, all $U_{(k)}$s can be seen as independent. Then, we derive the following:

$$\begin{aligned}
E(A_1) &= E\left[E^*\left\{\sum_{k=1}^{K}(U_{(k)} - \bar{U})(U_{(k)} - \bar{U})^\top\right\}\right] \\
&= KE\{E^*(U_{(k)} - \bar{U})(U_{(k)} - \bar{U})^\top\} = \frac{K}{M}E\left\{\sum_{m=1}^{M}(U_m - \bar{U})(U_m - \bar{U})^\top\right\},
\end{aligned} \tag{B5}$$

and we also have

$$\begin{aligned}
E(A_2) &= KE\left[E^*\left\{(\bar{U}_K - \bar{U})(\bar{U}_K - \bar{U})^\top\right\}\right] \\
&= E\left\{E^*(U_{(k)} - \bar{U})(U_{(k)} - \bar{U})^\top\right\} = \frac{1}{M}E\left\{\sum_{m=1}^{M}(U_m - \bar{U})(U_m - \bar{U})^\top\right\}.
\end{aligned} \tag{B6}$$

Combining the above, we have

$$E\left\{\widehat{\mathrm{SE}}^2(\overline{\beta}_K)\right\} = \frac{c}{M}E\left\{\sum_{m=1}^{M}(U_m - \bar{U})(U_m - \bar{U})^\top\right\}$$

$$
\begin{aligned}
&= n\left(\frac{1}{nK}+\frac{1}{N}\right)\left\{n^{-1}\Sigma-\mathrm{var}(\bar{U})\right\} &\text{(B7)}\\
&= \left(\frac{1}{nK}+\frac{1}{N}\right)\left\{1+O(\frac{n}{N})\right\}\Sigma.
\end{aligned}
$$

Second, we consider the bias of $\widehat{\mathrm{SE}}^2(\bar{\beta}_K)$. Together with (A7) and (B7), we have

$$
\mathrm{var}(\bar{U}_K)-E\left\{\widehat{\mathrm{SE}}^2(\bar{\beta}_K)\right\}=(1+\frac{n}{N})\mathrm{var}(\bar{U})-\frac{1}{N}\Sigma=O(\frac{n}{N^2})\Sigma.
$$

By a similar proof technique to that for Lemma 3, we can conclude that $\|R_1+R_2\|$ is sufficiently small compared with $\bar{U}_K$. Thus, the desired result can be obtained. This completes the proof.

B.4 Proof of Theorem 3

The purpose of this proof is to verify (3). Recall that $\widehat{\mathrm{SE}}^2_*(\bar{\beta}_K)=n(K-1)^{-1}\left(1/(nK)+1/N\right)\sum_{k=1}^K\left(\hat{U}_{(k)}-\hat{\bar{U}}_k\right)\left(\hat{U}_{(k)}-\hat{\bar{U}}_k\right)^\top$. Then we have

$$
\begin{aligned}
\left(\frac{1}{nK}+\frac{1}{N}\right)^{-1}\widehat{\mathrm{SE}}^2_*(\bar{\beta}_K)=\frac{n}{K-1}\Bigg\{&\sum_{k=1}^K\left(\hat{U}_{(k)}-\bar{U}\right)\left(\hat{U}_{(k)}-\bar{U}\right)^\top\\
&-K\left(\hat{\bar{U}}_k-\bar{U}\right)\left(\hat{\bar{U}}_k-\bar{U}\right)^\top\Bigg\}=\frac{nK}{K-1}(B_1-B_2),
\end{aligned}
$$

where $B_1=K^{-1}\sum_{k=1}^K(\hat{U}_{(k)}-\bar{U})(\hat{U}_{(k)}-\bar{U})^\top$, and $B_2=(\hat{\bar{U}}_k-\bar{U})^\top\left(\hat{\bar{U}}_k-\bar{U}\right)^\top$. To verify Equation (3), it suffices to prove that $nB_1\to_p\Sigma$, and $nB_2\to_p 0$ since $K/(K-1)\to 1$. Then we consider analysing B_1 and B_2 separately.

It then could be verified that

$$
\begin{aligned}
B_1&=K^{-1}\sum_{k=1}^K\left(\hat{U}_{(k)}-U_{(k)}\right)\left(\hat{U}_{(k)}-U_{(k)}\right)^\top+K^{-1}\sum_{k=1}^K\left(U_{(k)}-\bar{U}\right)\left(U_{(k)}-\bar{U}\right)^\top+\mathcal{O}\\
&=B_{11}+B_{12}+\mathcal{O},
\end{aligned}
$$

where $B_{11}=K^{-1}\sum_{k=1}^K\left(\hat{U}_{(k)}-U_{(k)}\right)\left(\hat{U}_{(k)}-U_{(k)}\right)^\top$, $B_{12}=K^{-1}\sum_{k=1}^K\left(U_{(k)}-\bar{U}\right)\left(U_{(k)}-\bar{U}\right)^\top$, and $\mathcal{O}=2K^{-1}\sum_{k=1}^K\left(\hat{U}_{(k)}-U_{(k)}\right)\left(U_{(k)}-\bar{U}\right)$ is the cross term. Next, we are going to investigate the three terms in the following two steps. First, we verify that B_{12} is the leading term with $nB_{12}\to_p\Sigma$. In addition, we prove that B_{11} and $\mathcal{O}$ are ignorable terms, more precisely, they are both of the order $o_p(1/n)$.

Step 1. We first show that $nB_{12}\to_p\Sigma$. Then the consistency can be verified in (1) $E(nB_{12})\to\Sigma$ and (2) $\mathrm{var}(nB_{12})\to 0$. We next calculate the expectation and variance separately.

(1) Proof of $E(nB_{12})\to\Sigma$: Note that, $B_{12}=A_1/K$, where A_1 is defined in Appendix B.3. Then by (B5), we have $E(B_{12})=n^{-1}\Sigma+O(1/N)\Sigma=n^{-1}\Sigma\{1+o(1)\}$.

(2) Proof of $\mathrm{var}(nB_{12})\to 0$: Because $\mathrm{var}(B_{12})=\mathrm{var}\{E^*(B_{12})\}+E\{\mathrm{var}^*(B_{12})\}$, we then investigate the two terms respectively. We compute $\mathrm{var}\{E^*(B_{12})\}$ first. It could be verified that

$$
E^*(B_{12})=E^*\left\{(U_{(k)}-\bar{U})(U_{(k)}-\bar{U})^\top\right\}=\frac{1}{M}\sum_{m=1}^M(U_m-\bar{U})(U_m-\bar{U})^\top. \qquad \text{(B8)}
$$

Take variance on both sides of (B8), by the same technique used in **Step (1)** of Lemma 2, we have

$$\begin{aligned}\mathrm{var}\{E^*(B_{12})\} &= M^{-1}\mathrm{var}\left\{(U_m-\bar{U})(U_m-\bar{U})^\top\right\} + \sum_{m_1\neq m_2}\mathrm{cov}\left\{(U_{m_1}-\bar{U})(U_{m_2}-\bar{U})^\top\right\}\\ &\leq M^{-1}\mathrm{var}\left\{(U_m-\bar{U})(U_m-\bar{U})^\top\right\} \qquad \text{(B9)}\\ &\quad+\frac{n\{2M-(n-1)\}}{M^2}\mathrm{var}\left\{(U_m-\bar{U})(U_m-\bar{U})^\top\right\}\\ &\leq \frac{M+n\{2M-(n-1)\}}{M^2}E\big\{(U_m-\bar{U})(U_m-\bar{U})^\top\big\}^2. \qquad \text{(B9)}\end{aligned}$$

Here, the first inequality holds because there are $2\sum_{m'=1}^{n-1}(M-m')$ pairs of m_1 and m_2 when the covariance is not equal to zero. Furthermore, it could be verified that $E\big\{(U_m-\bar{U})(U_m-\bar{U})^\top\big\}^2 \leq 8E\big\{U_mU_m^\top\big\}^2 = O(1/n^2)$. By the similar analysis with (A6), (B9) could be rewritten as $\mathrm{var}\{E^*(B_{12})\} = O(n/M)O(1/n^2) = o(1/n^2)$.

We next calculate $E\{\mathrm{var}^*(B_{12})\}$. Because $\mathrm{var}^*(B_{12}) = K^{-1}\mathrm{var}^*\left\{(U_{(k)}-\bar{U})(U_{(k)}-\bar{U})^\top\right\}$, it could be proved that

$$\begin{aligned}E\{\mathrm{var}^*(B_{12})\} &= \frac{1}{K}E\left\{\mathrm{var}^*\left\{(U_{(k)}-\bar{U})(U_{(k)}-\bar{U})^\top\right\}\right\}\\ &\leq \frac{1}{K}E\big\{(U_m-\bar{U})(U_m-\bar{U})^\top\big\}^2 = O\left\{(Kn^2)^{-1}\right\} = o(1/n^2).\end{aligned}$$

Combining the above results, we have verified that $\mathrm{var}(nB_{12}) \to 0$.

Step 2. We next show that $nB_{11} \to_p 0$.

By definition of $\hat{U}_{(k)}$, we have $\hat{U}_{(k)} = \left[\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\overline{\beta}_k)\}^{-1} - \{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\}^{-1}\right]\left\{n^{-1}\dot{\ell}_{\mathcal{S}_k}(\overline{\beta}_k)\right\} + \{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\}^{-1}\left\{n^{-1}\dot{\ell}_{\mathcal{S}_k}(\overline{\beta}_k)\right\}$. Then it could be shown that

$$\begin{aligned}\hat{U}_{(k)} - U_{(k)} &= \left[\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\overline{\beta}_k)\}^{-1} - \{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\}^{-1}\right]\left\{n^{-1}\dot{\ell}_{\mathcal{S}_k}(\overline{\beta}_k)\right\}\\ &\quad+\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\}^{-1}\left\{n^{-1}\dot{\ell}_{\mathcal{S}_k}(\overline{\beta}_k) - n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0)\right\}\\ &= U_{(k1)} + U_{(k2)},\end{aligned}$$

where $U_{(k1)} = \left[\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\overline{\beta}_k)\}^{-1} - \{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\}^{-1}\right]\left\{n^{-1}\dot{\ell}_{\mathcal{S}_k}(\overline{\beta}_k)\right\}$, $U_{(k2)} = \{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\}^{-1}\left\{n^{-1}\dot{\ell}_{\mathcal{S}_k}(\overline{\beta}_k) - n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0)\right\}$. Consequently, to verify $nB_{11} \to_p 0$, it suffices to prove that $nK^{-1}\sum_{k=1}^K U_{(k1)}U_{(k1)}^\top \to_p 0$ and $nK^{-1}\sum_{k=1}^K U_{(k2)}U_{(k2)}^\top \to_p 0$. Then we analysis the two terms separately.

(1) Proof of $nK^{-1}\sum_{k=1}^K U_{(k1)}U_{(k1)}^\top \to_p 0$: Note that, $U_{(k1)}$ could be further re-written as $U_{(k1)} = \left[\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\overline{\beta}_k)\}^{-1} - \{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\}^{-1}\right]\left\{n^{-1}\dot{\ell}_{\mathcal{S}_k}(\overline{\beta}_k) - n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0)\right\} + \left[\{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\overline{\beta}_k)\}^{-1} - \{n^{-1}\ddot{\ell}_{\mathcal{S}_k}(\beta_0)\}^{-1}\right]n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0)$. By condition (C4), we have $\|U_{(k1)}\| \leq C\max_k \lambda_{\max}(\widehat{\Sigma}_k)\|\overline{\beta}_k - \beta_0\|^2 + C\|n^{-1}\dot{\ell}_{\mathcal{S}_k}(\beta_0)\|\|\overline{\beta}_k - \beta_0\|$. Furthermore, by the same analytical skills used in Lemma 3, it could be proved that $\|K^{-1}\sum_{k=1}^K U_{(k1)}U_{(k1)}^\top\| = O_p\left(n^{-1}\left\{(nK)^{-1} + N^{-1}\right\}\log K\right) = o_p(1/n)$. This infers $nK^{-1}\sum_{k=1}^K U_{(k1)}U_{(k1)}^\top \to_p 0$.

(2) Proof of $nK^{-1}\sum_{k=1}^{K} U_{(k2)}U_{(k2)}^{\top} \to_p 0$: Similarly, because $\|U_{k2}\| \le \left\{ \max_k \lambda_{\max}(\widehat{\Sigma}_k)\right\}^2 \|\overline{\beta}_k - \beta_0\|$, we could verify that $\|\sum_{k=1}^{K} U_{(k2)}U_{(k2)}^{\top}\| = O_p\left(n^{-1}\left\{(nK)^{-1} + N^{-1}\right\}\log K\right) = o_p(1/n)$.

Combining the two above proofs, we then finish **Step 2**. Subsequently, by the Cauchy–Schwarz inequality, we have $n\mathcal{O} \to_p 0$. This completes the proof of $(K-1)^{-1}(nK)B_1 \to_p \Sigma$.

Finally, we calculate B_2. By a similar proof technique used in analysing B_1, we can conclude that $B_2 = (\bar{U}_K - \bar{U})(\bar{U}_K - \bar{U})^{\top} + o_p(1/n)$. Then by B6, we have $E(\bar{U}_K - \bar{U})(\bar{U}_K - \bar{U})^{\top} = (nK)^{-1}\Sigma\{1 + o(1)\}$. As a consequence, it could be shown that $E\left\{n(\bar{U}_K - \bar{U})(\bar{U}_K - \bar{U})^{\top}\right\} \to 0$, which leads to $nA_2 \to_p 0$. The desired results can be obtained. This completes the whole theorem proof.