Computational Statistics and Data Analysis 171 (2022) 107459

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Asymptotic covariance estimation by Gaussian random perturbation

Jing Zhou [a], Wei Lan [b,*], Hansheng Wang [c]

[a] Center for Applied Statistics, School of Statistics, Renmin University of China, China
[b] School of Statistics and Center of Statistical Research, Southwestern University of Finance and Economics, China
[c] Guanghua School of Management, Peking University, China

ARTICLE INFO

Article history:
Received 13 May 2021
Received in revised form 20 February 2022
Accepted 26 February 2022
Available online 9 March 2022

Keywords:
Covariance matrix estimation
Gaussian random perturbation
M-estimators

ABSTRACT

In most cases, the asymptotic covariance matrix of an *M*-estimator is in a sandwich form. This sandwich form involves calculations of the first and second order derivatives of the loss function, which is intractable if the loss function is complex. To alleviate this problem, we propose in this article a novel method called Gaussian random perturbation. This method can be used to estimate the asymptotic covariance matrix of a general *M*-estimator without derivative calculations. The idea can be summarized as follows. We first generate a small random perturbation around the *M*-estimator. Then, we re-evaluate the loss function at the randomly perturbed *M*-estimator and obtain the estimators of the first and second order derivatives of the loss function via Taylor series expansion. This leads to a novel estimator for the asymptotic covariance matrix. We then rigorously show that the resulting covariance estimator is statistically consistent with two elegant characteristics. First, it involves no computation of derivatives. This makes it easier to estimate the covariance matrix of an *M*-estimator with a complex loss function. Second, it is convenient for parallel computing and thus attractive for massive data analysis. The consistency of the proposed asymptotic covariance estimator is demonstrated under appropriate regularity conditions. The practical usefulness of the method is further demonstrated with both simulation studies and real data analysis.

1. Introduction

M-estimators refer to a large class of estimators that are obtained by minimizing (or maximizing) an appropriately defined loss function (Lehmann and Casella, 1983). There are many estimates that fall into this category. Consider for example, a loss function of a traditional linear regression model is the summation of the squares of the residuals (Casella et al., 2015). The resulting ordinary least squares estimator is an *M*-estimator. In addition to that, the generalized least squares estimator is also an *M*-estimator. In fact, the *M*-estimator is also known as the generalization of a maximum likelihood estimator. Thereafter, all the maximum likelihood estimators are *M*-estimators. Finally, if the parameters are identified by a set of moment conditions, the popularly used generalized method of moment can be equivalently formulated as a minimization problem. Consequently, it is also an *M*-estimator (Wooldridge, 2001).

* Corresponding author.
E-mail address: facelw@gmail.com (W. Lan).

https://doi.org/10.1016/j.csda.2022.107459
0167-9473/

In most cases, carefully defined M-estimators are consistent and asymptotically normal under appropriate conditions. The asymptotic covariance matrix of an M-estimator is usually in the form of a sandwich (Shao, 2003). In some cases (e.g., the maximum likelihood estimator), the sandwich form can be further simplified into a non-sandwich form. For an asymptotically valid statistical inference (e.g., hypothesis testing and confidence interval), the sandwich-type asymptotic covariance matrix needs to be estimated. One way to solve this problem is to obtain its analytical formula, and then replace the unknown parameters with appropriate estimates. This method is effective if the formula of the asymptotic covariance matrix is analytically simple. For example, in most cases, there are simple analytical solutions for various least squares estimates and maximum likelihood estimates.

However, the asymptotic covariance matrix involves the calculation of the first and second order derivatives of the loss function, which is tedious if the loss function is complex. A typical example is a regression model with missing data. In this case, the full likelihood function involves some unknown nuisance parameters. Particularly, these nuisance parameters are usually related to the missing mechanism and need to be integrated out (Shao and Wang, 2002; Wang and Dai, 2008; Lin et al., 2021; Zhou et al., 2022). Consequently, the resulting asymptotic covariance matrix involves the first and second order derivatives of an objective function with complicated integration and is thus difficult to calculate and estimate (Chen et al., 2015; Zhao and Shao, 2015). Estimating the asymptotic covariance matrix without knowing its analytical formula then becomes a problem of great importance. To alleviate this problem, re-sampling type methods such as bootstrapping and jackknifing have been proposed and are popularly used (Efron and Stein, 1981; Efron and Gong, 1983; Efron and Tibshirani, 1986; Efron, 1994; Jiao and Han, 2020). They estimate the asymptotic covariance matrix consistently without knowing its analytical formula. This avoids evaluating the derivatives of some complex integral functions. However, such re-sampling methods also suffer from computational complexity. This might not be a problem for traditional data analysis, when the sample size is not very large and the data dimensions are relatively low. However, this could be a serious burden for massive datasets. In the latter case, computing the M-estimator itself is already computationally expensive, and any further replication is practically infeasible.

To solve the problem detailed above, we propose a novel method called Gaussian random perturbation. The key idea is summarized as follows. First, for a given loss function and its M-estimator, we generate a small random perturbation around the M-estimator. The random perturbation is generated from a multivariate normal distribution with tiny variability. Accordingly, the randomly perturbed M-estimator still stays very close to the original M-estimator locally. Second, we re-evaluate the loss function on those locally and randomly perturbed M-estimators. Through rigorous mathematic derivation with Taylor series expansion, we find that the first and second order derivatives of the loss function evaluated at the M-estimator can be further approximated by two components. They are, respectively, the loss function evaluated at the M-estimator and the re-evaluated loss function at the randomly perturbed M-estimator. This suggests that the elements in the sandwich-type asymptotic covariance matrix can be approximately estimated by using loss functions, instead of computing the derivatives. This leads to a novel estimator for the asymptotic covariance matrix. We then rigorously demonstrate that the resulting covariance estimator is statistically consistent.

It is remarkable that the above proposed covariance estimator enjoys two important features. First, it involves no computation of the first and second order derivatives for the loss function at the M-estimator. Thus, the asymptotic covariance matrix can be consistently estimated automatically without knowing the analytical formula of an M-estimator. Second, the proposed covariance estimator can be expressed in a vector form. A vector can be naturally decomposed into different elements. Those elements can then be separately processed by different computers simultaneously. That makes vector forms more convenient for parallel computing. By using a parallel strategy, we can break a large-scale computation problem into many small pieces and then solve them in a parallel way (Battey et al., 2018; Jordan et al., 2018; Fan et al., 2019; Li et al., 2020). This makes the method particularly attractive for massive data analysis when the computation complexity and privacy protection are of great importance.

The remainder of this article is organized as follows. Section 2 introduces the proposed Gaussian random perturbation method. We then illustrate the idea using two types of M-estimators: a traditional M-estimator and an M-estimator with nuisance parameters. We then rigorously show that both estimators are statistically consistent for the two cases. Simulation studies and an empirical example are presented in Section 3, and Section 4 concludes the article with a discussion. All theoretical proofs are relegated to the Appendices.

2. Methodology

To illustrate the usefulness of the proposed Gaussian random perturbation method, we consider the asymptotic covariance for two types of M-estimators. The first is the traditional M-estimator, described in subsection 2.1. The second is an M-estimator with unknown nuisance parameters, which we will address in subsection 2.2.

2.1. Asymptotic covariance for M-estimator

Let $X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p$ be an independent and identically distributed observation collected from the i-th ($1 \le i \le N$) subject. Our main focus is to make an inference for θ based on the observed data X_i for $1 \le i \le N$, where $\theta \in \Theta \subset \mathbb{R}^p$ is the unknown parameter with $p < \infty$. Here, Θ is the parameter space, and it is an open set in $\mathbb{R}^p$. The corresponding loss function is defined as $\ell(X_i, \theta)$. Then, an M-estimator is proposed as $\hat{\theta} = \text{argmin}_{\theta \in \Theta} \mathcal{L}(\theta)$, where $\mathcal{L}(\theta) = N^{-1} \sum_{i=1}^{N} \ell(X_i, \theta)$.

Accordingly, we should have $\dot{\mathcal{L}}(\hat{\theta}) = 0$, where $\dot{\mathcal{L}}(\theta)$ stands for the first order derivative of $\mathcal{L}(\theta)$ with respect to θ. Let θ_0 be the true value, $\dot{\ell}(X_i, \theta_0)$ and $\ddot{\ell}(X_i, \theta_0)$ be the first and second order derivatives of $\ell(X_i, \theta_0)$, respectively. Under appropriate regularity conditions, we should have

$$\hat{\theta} - \theta_0 = \Big\{N^{-1}\sum_{i=1}^{N}\ddot{\ell}(X_i, \theta_0)\Big\}^{-1}\Big\{N^{-1}\sum_{i=1}^{N}\dot{\ell}(X_i, \theta_0)\Big\}\{1 + o_p(1)\},$$

and $\sqrt{N}(\hat{\theta} - \theta_0) \to_d N(0, \Sigma_2^{-1}\Sigma_1\Sigma_2^{-1})$. Here, the two unknown matrices are given by $\Sigma_1 = E\Big[\dot{\ell}(X_i, \theta_0)\dot{\ell}^{\top}(X_i, \theta_0)\Big]$ and $\Sigma_2 = E\Big[\ddot{\ell}(X_i, \theta_0)\Big]$. It should be noted that when $\hat{\theta}$ is the maximum likelihood estimator obtained under a correctly specified likelihood function, we should have $\Sigma_1 = \Sigma_2 = \Sigma$ for some positive definite matrix Σ. Accordingly, the asymptotic covariance of $\sqrt{N}(\hat{\theta} - \theta_0)$ becomes Σ^{-1}. As one can see, the key problem here is estimating Σ_1 and Σ_2.

We next consider how to estimate both Σ_1 and Σ_2 without knowing their analytical formulas. Specifically, for a given loss function $\ell(X_i, \theta)$ evaluated at the associated M-estimator $\hat{\theta}$, we generate a small random perturbation $\delta^{(k)}$ around $\hat{\theta}$. The random perturbation $\delta^{(k)}$ is generated from a multivariate normal distribution with mean 0 and covariance $\sigma^2 I_p \in \mathbb{R}^{p\times p}$ for $k = 1, \cdots, K$, where K is a pre-specified replication number. Here I_p stands for a $p \times p$ identity matrix. The variance σ^2 is a carefully selected small positive number. The value of σ should be selected as small as possible so that the randomly perturbed loss function $\ell(X_i, \hat{\theta} + \delta^{(k)})$ stays closely to the original one, that is, $\ell(X_i, \hat{\theta})$.

We start with Σ_1 first. A natural estimator for Σ_1 is $N^{-1}\sum_{i=1}^{N}\{\dot{\ell}(X_i, \hat{\theta})\dot{\ell}^{\top}(X_i, \hat{\theta})\}$. Through Taylor series expansion, one can verify that

$$\ell(X_i, \hat{\theta} + \delta^{(k)}) \approx \ell(X_i, \hat{\theta}) + \delta^{(k)\top}\dot{\ell}(X_i, \hat{\theta}). \tag{2.1}$$

The approximation holds because $\delta^{(k)}$ is selected to be sufficiently small. Multiplying $\delta^{(k)}$ on both sides of (2.1), we obtain $\delta^{(k)}\{\ell(X_i, \hat{\theta} + \delta^{(k)}) - \ell(X_i, \hat{\theta})\} \approx \delta^{(k)}\delta^{(k)\top}\dot{\ell}(X_i, \hat{\theta})$. Using the fact that $E(\delta^{(k)}\delta^{(k)\top}) = \sigma^2 I_p$, we replace $\delta^{(k)}\delta^{(k)\top}$ with its expected form $\sigma^2 I_p$ and expect that $\dot{\ell}(X_i, \hat{\theta})$ can be estimated by $\sigma^{-2}\{\ell(X_i, \hat{\theta} + \delta^{(k)}) - \ell(X_i, \hat{\theta})\}\delta^{(k)}$ with little bias. This motivates us to construct an initial estimator of Σ_1 as $\hat{\Sigma}_{1,int}^{(k)} = (2\sigma^{-4}N)^{-1}\sum_{i=1}^{N}\{\ell(X_i, \hat{\theta} + \delta^{(k)}) - \ell(X_i, \hat{\theta})\}^2(\delta^{(k)}\delta^{(k)\top})$. We then have

$$\begin{aligned}\hat{\Sigma}_{1,int}^{(k)} &= (2\sigma^{-4}N)^{-1}\sum_{i=1}^{N}\Big\{\ell(X_i, \hat{\theta} + \delta^{(k)}) - \ell(X_i, \hat{\theta})\Big\}^2(\delta^{(k)}\delta^{(k)\top}) \\ &\approx (2\sigma^{-4}N)^{-1}\sum_{i=1}^{N}\Big\{\delta^{(k)\top}\dot{\ell}(X_i, \hat{\theta})\dot{\ell}^{\top}(X_i, \hat{\theta})\delta^{(k)}\Big\}(\delta^{(k)}\delta^{(k)\top}).\end{aligned}$$

Unfortunately, the above initial estimator is biased. Its asymptotic bias is given by the following Proposition 1.

Proposition 1. *Define* $\hat{\Sigma}_{1,int}^{(0k)} = (2\sigma^{-4}N)^{-1}\sum_{i=1}^{N}\Big\{\delta^{(k)\top}\dot{\ell}(X_i, \theta_0)\dot{\ell}^{\top}(X_i, \theta_0)\delta^{(k)}\Big\}(\delta^{(k)}\delta^{(k)\top})$, *then we have* $E\{\hat{\Sigma}_{1,int}^{(0k)}\} = \Sigma_1 + 2^{-1}tr(\Sigma_1)I_p$.

Based on the above proposition, one can immediately have,

$$E\big(\hat{\Sigma}_{1,int}^{(k)}\big) \approx E\big\{\hat{\Sigma}_{1,int}^{(0k)}\big\} = \Sigma_1 + 2^{-1}tr(\Sigma_1)I_p, \tag{2.2}$$

which involves a non-negligible bias term $2^{-1}tr(\Sigma_1)I_p$. Therefore, bias-correction is necessary. From (2.2), we know that $E\{tr(\hat{\Sigma}_{1,int}^{(k)})\} \approx 2^{-1}(p+2)tr(\Sigma_1)$. Accordingly, $tr(\Sigma_1)$ can be approximated by $2(p+2)^{-1}tr(\hat{\Sigma}_{1,int}^{(k)})$. By plugging this expression into (2.2), we obtain the following bias-corrected estimator,

$$\hat{\Sigma}_1^{(k)} = \hat{\Sigma}_{1,int}^{(k)} - (p+2)^{-1}tr(\hat{\Sigma}_{1,int}^{(k)})I_p.$$

The bias-corrected estimator $\hat{\Sigma}_1^{(k)}$ is nearly unbiased. However, its variability is still large. To reduce the variability, we then average over different replications of $\hat{\Sigma}_1^{(k)}$ and obtain the final estimator of Σ_1 as

$$\hat{\Sigma}_1 = K^{-1}\sum_{k=1}^{K}\hat{\Sigma}_1^{(k)}. \tag{2.3}$$

This leads to a consistent estimator of Σ_1. See Theorem 1 proposed at the end of this subsection for rigorous theoretical justification.

We next consider how to estimate Σ_2. A natural estimator of Σ_2 is $N^{-1}\sum_{i=1}^{N}\{\ddot{\ell}(X_i,\hat{\theta})\}$. Similar to Σ_1, we employ the Taylor series expansion and obtain that

$$\ell(X_i,\hat{\theta}+\delta^{(k)})\approx\ell(X_i,\hat{\theta})+\delta^{(k)\top}\dot{\ell}(X_i,\hat{\theta})+\delta^{(k)\top}\ddot{\ell}(X_i,\hat{\theta})\delta^{(k)}/2. \tag{2.4}$$

Multiplying $\delta^{(k)}\delta^{(k)\top}$ on both sides of (2.4), we obtain $\delta^{(k)}\{\ell(X_i,\hat{\theta}+\delta^{(k)})-\ell(X_i,\hat{\theta})\}\delta^{(k)\top}\approx\delta^{(k)}\delta^{(k)\top}\{\delta^{(k)\top}\dot{\ell}(X_i,\hat{\theta})\}+\delta^{(k)}\delta^{(k)\top}\ddot{\ell}(X_i,\hat{\theta})\delta^{(k)}\delta^{(k)\top}/2$. Note that $E\{\dot{\ell}(X_i,\theta_0)\}=0$ by Condition (C2) below and $E(\delta^{(k)}\delta^{(k)\top})=\sigma^2 I_p$. We expect that $\ddot{\ell}(X_i,\hat{\theta})$ can be estimated by $2\sigma^{-4}\delta^{(k)}\{\ell(X_i,\hat{\theta}+\delta^{(k)})-\ell(X_i,\hat{\theta})\}\delta^{(k)\top}$. This motivates us to construct an initial estimator of Σ_2 as $\hat{\Sigma}_{2,int}^{(k)}=(\sigma^4N)^{-1}\sum_{i=1}^{N}\left\{\ell(X_i,\hat{\theta}+\delta^{(k)})-\ell(X_i,\hat{\theta})\right\}(\delta^{(k)}\delta^{(k)\top})$. Through Taylor series expansion, we then have,

$$\begin{aligned}\hat{\Sigma}_{2,int}^{(k)}&=(\sigma^4N)^{-1}\sum_{i=1}^{N}\left\{\ell(X_i,\hat{\theta}+\delta^{(k)})-\ell(X_i,\hat{\theta})\right\}(\delta^{(k)}\delta^{(k)\top})\\&\approx(\sigma^4N)^{-1}\sum_{i=1}^{N}\left\{\dot{\ell}^{\top}(X_i,\hat{\theta})\delta^{(k)}\right\}(\delta^{(k)}\delta^{(k)\top})+\frac{1}{2}(\sigma^4N)^{-1}\sum_{i=1}^{N}\left\{\delta^{(k)\top}\ddot{\ell}(X_i,\hat{\theta})\delta^{(k)}\right\}(\delta^{(k)}\delta^{(k)\top})\\&=Q_1+Q_2.\end{aligned}$$

One can easily verify that $Q_1=0$ because $\sum_{i=1}^{N}\dot{\ell}(X_i,\hat{\theta})=0$ by the definition of $\hat{\theta}$. Then, $\hat{\Sigma}_{2,int}^{(k)}\approx Q_2$. Similar to the discussion of Σ_1, Q_2 is biased, and its asymptotic bias is given by Proposition 2.

Proposition 2. *Define* $\hat{\Sigma}_{2,int}^{(0k)}=(2\sigma^4N)^{-1}\sum_{i=1}^{N}\left\{\delta^{(k)\top}\ddot{\ell}(X_i,\theta_0)\delta^{(k)}\right\}(\delta^{(k)}\delta^{(k)\top})$, *then we have* $E\{\hat{\Sigma}_{2,int}^{(0k)}\}=\Sigma_2+2^{-1}tr(\Sigma_2)I_p$.

Based on the results of Proposition 2, we have

$$E(Q_2)=E(\hat{\Sigma}_{2,int}^{(k)})\approx E(\hat{\Sigma}_{2,int}^{(0k)})=\Sigma_2+2^{-1}tr(\Sigma_2)I_p. \tag{2.5}$$

Similarly, to correct the bias term $2^{-1}tr(\Sigma_2)I_p$, we define the bias-corrected estimator as, $\hat{\Sigma}_2^{(k)}=\hat{\Sigma}_{2,int}^{(k)}-(p+2)^{-1}tr(\hat{\Sigma}_{2,int}^{(k)})I_p$. By averaging over all replications of $k=1,\cdots,K$, we then obtain the final estimator of Σ_2 as

$$\hat{\Sigma}_2=K^{-1}\sum_{k=1}^{K}\hat{\Sigma}_2^{(k)}. \tag{2.6}$$

We can theoretically and rigorously show that $\hat{\Sigma}_2$ is an asymptotically unbiased and consistent estimator for Σ_2 under appropriate regularity conditions. Similar to that of $\hat{\Sigma}_1$, the computation of $\hat{\Sigma}_2$ does not need to know its analytical formula, nor does it require the calculation of derivatives. Before providing the theoretical results of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$, we first consider the following regularity conditions.

(C1) Assume $\ell(X_i,\theta)$ has at least r-th order continuous derivatives for some $r\geq 3$. In addition, $\ddot{\ell}(X_i,\theta)$ and $\dddot{\ell}(X_i,\theta)$ are bounded uniformly for θ within a small neighborhood of θ_0. Further assume $\theta_0\in\Theta$ and Θ is an open set in $\mathbb{R}^p$.
(C2) Assume $E\{\dot{\ell}(X_i,\theta_0)\}=0$. In addition, both Σ_1 and Σ_2 are positive definite matrices with bounded eigenvalues for any N.

Both Conditions (C1) and (C2) are standard, and similar conditions are popularly used in the literature (e.g., Shao (2003)). Based on the two conditions, the theoretical results of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are given in Theorem 1.

Theorem 1. *Under conditions (C1)–(C2), we have* $\hat{\Sigma}_1\to_p\Sigma_1$ *and* $\hat{\Sigma}_2\to_p\Sigma_2$ *as* $\min\{N,K\}\to\infty$ *by setting* $\sigma\to 0$.

The above theorem implies that $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are both consistent by setting σ to be sufficiently small as $\min\{N,K\}\to\infty$. Our simulation results show that $\sigma=N^{-1}$ works satisfactorily.

Remark. It should be noted that the formulas of $\hat{\Sigma}_{1,int}^{(k)}$ and $\hat{\Sigma}_{2,int}^{(k)}$ can be expressed in vector forms. For example, $\hat{\Sigma}_{1,int}^{(k)}$ can be re-written as $(2\sigma^{-4}N)^{-1}W^{\top}W\delta^{(k)}\delta^{(k)\top}$, where $W=(W_1,\cdots,W_N)^{\top}$ and $W_i=\ell(X_i,\hat{\theta}+\delta^{(k)})-\ell(X_i,\hat{\theta})$ for $i=1,\cdots,N$. This vector form is very convenient for parallel computing (e.g., Vegh (2018); Maslian et al. (2019)). This is because a vector can be naturally decomposed into different elements. Those elements can then be separately processed by different computers simultaneously. That makes vector forms convenient for parallel computing.

2.2. Asymptotic variance for M-estimator with nuisance parameters

We next consider an M-estimator with nuisance parameters. The corresponding loss function is defined as $\ell(X_i, \theta, \gamma) \in \mathbb{R}$, where $\theta \subset \Theta \in \mathbb{R}^p$ $(p < \infty)$ is the target parameter of interest and $\gamma \subset \Gamma \in \mathbb{R}^q$ $(q < \infty)$ is the nuisance parameter with a consistent estimator of $\hat{\gamma}$. A typical example is the linear regression model with heteroscedastic error variances (Greene, 1997; Wooldridge, 2015). In this case, the parameters related to the variances are nuisance parameters. Another example is regression models with missing data in which the propensity function is assumed to be in a parametric form; see, for example, Huang et al. (2005) and Ibrahim and Molenberghs (2009). In this case, the parameters related to the propensity function are the nuisance parameters. Based on the preliminary estimator $\hat{\gamma}$, a two-step M-estimator is proposed as $\hat{\theta} = \text{argmin}_{\theta\in\Theta}\mathcal{L}(\theta, \hat{\gamma})$ with $\mathcal{L}(\theta, \hat{\gamma}) = N^{-1}\sum_i \ell(X_i, \theta, \hat{\gamma})$. Let θ_0 and γ_0 be the true parameters. Then, under Conditions (C3) and (C4) below, we have

$$\begin{aligned}\hat{\theta} - \theta_0 &= \Big\{N^{-1}\sum_{i=1}^{N}\frac{\partial^2\ell(X_i,\theta_0,\hat{\gamma})}{\partial\theta\partial\theta^\top}\Big\}^{-1}\Big\{N^{-1}\sum_{i=1}^{N}\frac{\partial\ell(X_i,\theta_0,\hat{\gamma})}{\partial\theta}\Big\}\{1+o_p(1)\}\\ &= \Big\{N^{-1}\sum_{i=1}^{N}\frac{\partial\ell^2(X_i,\theta_0,\gamma_0)}{\partial\theta\partial\theta^\top}\Big\}^{-1}\Big\{N^{-1}\sum_{i=1}^{N}\frac{\partial\ell(X_i,\theta_0,\gamma_0)}{\partial\theta}\Big\}\{1+o_p(1)\}. \qquad (2.7)\end{aligned}$$

Accordingly, one can obtain $\sqrt{N}(\hat{\theta} - \theta_0) \to_d N(0, \bar{\Sigma}_2^{-1}\bar{\Sigma}_1\bar{\Sigma}_2^{-1})$ with

$$\bar{\Sigma}_1 = E\Big[\frac{\partial\ell(X_i,\theta_0,\gamma_0)}{\partial\theta}\Big\{\frac{\partial\ell(X_i,\theta_0,\gamma_0)}{\partial\theta}\Big\}^\top\Big] \text{ and } \bar{\Sigma}_2 = E\Big\{\frac{\partial^2\ell(X_i,\theta_0,\gamma_0)}{\partial\theta\partial\theta^\top}\Big\}.$$

Similar to the discussions in subsection 2.1, $\bar{\Sigma}_1$ and $\bar{\Sigma}_2$ can be estimated by $\hat{\Sigma}_{n1} = K^{-1}\sum_{k=1}^{K}\Big\{\hat{\Sigma}_{n1,int}^{(k)} - (p+2)^{-1}tr(\hat{\Sigma}_{n1,int}^{(k)})I_p\Big\}$ with $\hat{\Sigma}_{n1,int}^{(k)} = (2\sigma^4 N)^{-1}\sum_{i=1}^{N}\Big\{\ell(X_i, \hat{\theta} + \delta^{(k)}, \hat{\gamma}) - \ell(X_i, \hat{\theta}, \hat{\gamma})\Big\}^2(\delta^{(k)}\delta^{(k)\top})$, and $\hat{\Sigma}_{n2} = K^{-1}\sum_{k=1}^{K}\Big\{\hat{\Sigma}_{n2,int}^{(k)} - (p+2)^{-1}\times tr(\hat{\Sigma}_{n2,int}^{(k)})I_p\Big\}$ with $\hat{\Sigma}_{n2,int}^{(k)} = (\sigma^4 N)^{-1}\sum_{i=1}^{N}\Big\{\ell(X_i, \hat{\theta} + \delta^{(k)}, \hat{\gamma}) - \ell(X_i, \hat{\theta}, \hat{\gamma})\Big\}(\delta^{(k)}\delta^{(k)\top})$. Here $\delta^{(k)}$s are the random perturbations that are defined in subsection 2.1. Before providing the theoretical results of $\bar{\Sigma}_1$ and $\bar{\Sigma}_2$, we first consider the following regularity conditions.

(C3) Assume $\ell(X_i, \theta, \gamma)$ has at least r-th order continuous derivatives with-respect to θ for some $r \geq 3$. In addition, $\ddot{\ell}(X_i, \theta, \gamma)$ and $\dddot{\ell}(X_i, \theta, \gamma)$ are bounded uniformly for θ within a small neighborhood of θ_0, and γ within a small neighborhood of γ_0. Further assume that $\theta_0 \in \Theta$ and Θ is an open set in $\mathbb{R}^p$.

(C4) Assume that $E\big\{\dot{\ell}(X_i, \theta_0, \gamma_0)\big\} = 0$ and $E\big\{\partial^2\ell(X_i, \theta_0, \gamma_0)/\partial\theta\partial\gamma\big\} = 0$. Additionally, both $\bar{\Sigma}_1$ and $\bar{\Sigma}_2$ are positive definite matrices with bounded eigenvalues for any N.

Theorem 2 shows that both $\hat{\Sigma}_{n1}$ and $\hat{\Sigma}_{n2}$ are consistent estimators of $\bar{\Sigma}_1$ and $\bar{\Sigma}_2$, respectively, by setting σ to be sufficiently small as $\min\{N, K\} \to \infty$.

Theorem 2. *Under conditions (C3)–(C4), we have* $\hat{\Sigma}_{n1} \to_p \bar{\Sigma}_1$ *and* $\hat{\Sigma}_{n2} \to_p \bar{\Sigma}_2$ *as* $\min\{N, K\} \to \infty$ *by setting* $\sigma \to 0$.

3. Numerical studies

To assess the finite sample performance of the proposed method, we conduct several simulation studies with four different settings. In the first setting, we consider a standard logistic regression model, where the estimators are traditional M-estimators. Second, we consider a linear regression model with heteroscedastic error variances, where the variances involve nuisance parameters. The third setting is similar to the second one, except that the variance of the error term is involved with both the target and nuisance parameters. Finally, to investigate a more complex loss function, we consider a multiplicative model in the last setting. It should be noted that, to evaluate the effectiveness of the proposed Gaussian random perturbation method, all the considered simulation examples have explicit asymptotic covariance.

3.1. Simulation models

Setting 1 *(Traditional M-estimator).* A standard logistic regression model is used to generate the data. It is given by,

$$P(Y_i = 1|X_i, \theta) = \frac{\exp(X_i^\top\theta)}{1 + \exp(X_i^\top\theta)},$$

where $X_i = (X_{i1}, X_{i2}, X_{i3})^\top \in \mathbb{R}^3$ is a three-dimensional multivariate normal random variable with 0 mean and unit variance. The regression coefficient is given by $\theta = (0.2, 1.0, 2.5)^\top$.

Setting 2 *(Asymptotic covariance with nuisance parameters).* In this example, we generate the data following a linear regression model with heteroscedastic error variances. It is given by,

$$Y_i = X_i^\top \theta + \epsilon_i,$$

where $X_i = (X_{i1}, X_{i2}, X_{i3})^\top \in \mathbb{R}^3$ is a three-dimensional multivariate normal random variable with 0 mean and unit variance. The regression coefficient is given by $\theta = (0.3, 1.5, 3)^\top$. The error term ϵ_i is a random variable generated from a normal distribution with 0 mean and covariance $\Sigma_\epsilon = \text{diag}\{\exp(Z_i^\top \gamma)\}$, where the sign "diag" indicates a diagonal matrix. Here, $Z_i = (Z_{i1}, Z_{i2})^\top \in \mathbb{R}^2$ is a two-dimensional multivariate normal random variable with 0 mean and unit variance, and the nuisance parameter γ is set to be $\gamma = (0.5, \sqrt{2})^\top$.

Setting 3 *(Asymptotic covariance with both target and nuisance parameters).* In this case, the simulation setup is the same as that of Setting 2 except for the generation mechanism of error term ϵ_i. In this example, error term ϵ_i is generated from a normal distribution with 0 mean and covariance $\Sigma_\epsilon = \text{diag}\{\exp(X_i^\top \theta + Z_i^\top \gamma)\}$. Accordingly, the asymptotic covariance matrix of $\hat{\theta}$ involves both target parameter θ and nuisance parameter γ. To make the coefficient values more diverse, we investigate another set of θ as $\theta = (0.1, 0.5, 1)^\top$.

Setting 4 *(Multiplicative model).* In the last case, we generate the data according to the following multiplicative model,

$$Y_i = \exp(X_i^\top \theta)\epsilon_i,$$

where $X_i = (X_{i1}, X_{i2}, X_{i3})^\top \in \mathbb{R}^3$ is a three-dimensional multivariate normal random variable with zero mean and unit variance. The error term ϵ_i is specified as $\log(\epsilon_i)$, and it follows a normal distribution within 0 mean unit variance. The regression coefficient is given by $\theta = (0.1, 0.5, 1)^\top$.

3.2. Performance measurements

For a reliable evaluation, a total of $M = 1,000$ simulation iterations are conducted for all the simulation studies. The sample sizes are set as $N = (5, 10, 20) \times 10^3$, and the true asymptotic covariance matrix with respect to $\sqrt{N}(\hat{\theta} - \theta_0)$ is defined as $\Sigma = \Sigma_2^{-1}\Sigma_1\Sigma_2^{-1} \in \mathbb{R}^{3\times 3}$. It should be noted that the matrices Σ_1, Σ_2 and Σ are usually unknown in practice. To obtain their true values, we adopt a simulation-based method. Specifically, Σ_1 is simulated by averaging 50,000 replications of $\dot{\ell}(X_i, \theta)\dot{\ell}^\top(X_i, \theta)$, and Σ_2 is simulated by averaging 50,000 replications of $\ddot{\ell}(X_i, \theta)$. Thereafter, Σ can be obtained by calculating a sandwich form of Σ_1 and Σ_2. It is remarkable that the loss functions in Settings 1-3 are standard and popularly used. As such, we omit their derivative forms in the main manuscript to save space. However, the loss function in Setting 4 is a little unusual and complicated. Following Chen et al. (2010), we provide a least squares relative errors (LSRE) criterion. Thus, loss function $\mathcal{L}(\theta)$ is defined as $\mathcal{L}(\theta) = \frac{1}{2}\sum_{i=1}^N \left[\left\{ \frac{Y_i - \exp(X_i^\top \theta)}{Y_i} \right\}^2 + \left\{ \frac{Y_i - \exp(X_i^\top \theta)}{\exp(X_i^\top \beta)} \right\}^2 \right]$, in which the parameters are involved in both the denominator and numerator of the second part of $\mathcal{L}(\theta)$. Accordingly, the first and second order derivatives of this loss function are complicated, so that we present them in the appendix.

Let $\hat{\Sigma}$ be the estimated covariance matrix of Σ. To assess the effectiveness of the proposed method, we provide here a comparative study. Specifically, in addition to the proposed Gaussian random perturbation (GRP) method, other estimation strategies for Σ are considered. They are, respectively, the bootstrap (BP) method and non-bootstrap (Non-BP) method for all the four settings. For the bootstrap method, we consider $R = 200$ bootstrap iterations. For each bootstrap iteration, we randomly select N samples with replacement as the bootstrap samples, and then evaluate the estimators $\tilde{\theta}^{(r)} = (\tilde{\theta}_1^{(r)}, \tilde{\theta}_2^{(r)}, \tilde{\theta}_3^{(r)})$ for $1 \le r \le R$. We then compute the sample covariance matrix of $\tilde{\theta}^{(1)}, \cdots, \tilde{\theta}^{(R)}$, leading to the bootstrap covariance estimate $\hat{\Sigma}_{\text{BP}}$. For the non-bootstrap method, the analytical forms of the covariance matrix for the four settings are obtained using different estimation methods. Particularly, the analytical form of $\hat{\Sigma}$ in Setting 1 is obtained using the maximum likelihood estimation, which is given by $\hat{\Sigma}_{\text{Non-BP}} = \left\{ \sum_{i=1}^N \exp(X_i^\top \hat{\theta}) X_i^\top X_i / (1 + \exp(X_i^\top \hat{\theta}))^2 \right\}^{-1}$. The analytical forms of $\hat{\Sigma}$ in Settings 2 and 3 are obtained using the weighed least squares estimation, which are given by $\hat{\Sigma}_{\text{Non-BP}} = \left\{ \sum_{i=1}^N X_i^\top X_i / \exp(Z_i^\top \hat{\gamma}) \right\}^{-1}$ and $\hat{\Sigma}_{\text{Non-BP}} = \left\{ \sum_{i=1}^N X_i^\top X_i / \exp(X_i^\top \hat{\theta} + Z_i^\top \hat{\gamma}) \right\}^{-1}$, respectively. Finally, the analytical form of $\hat{\Sigma}$ in Setting 4 is obtained using the traditional sandwich-typed estimation, which is given in the appendix.

We next employ the following two measures to gauge the performance of the proposed method. First, for each simulation iteration, let $\hat{\Sigma}_{\text{GRP}}^{(m)}$ be the estimated asymptotic covariance matrix obtained in the m-th replication using the proposed Gaussian random perturbation (GRP) method. To measure the estimation efficiency, we calculate the root mean square error for $\hat{\Sigma}_{\text{GRP}}$ as $\text{RMSE}_{\hat{\Sigma}_{\text{GRP}}} = M^{-1}\{||\hat{\Sigma}_{\text{GRP}}^{(m)}\Sigma^{-1} - I_3||^2\}^{1/2}$, where $I_3 \in \mathbb{R}^{3\times 3}$ is an identity matrix and $||\cdot||^2$ is the L_2 norm. Similarly, we obtain the RMSE values for $\hat{\Sigma}_{\text{Non-BP}}$ and $\hat{\Sigma}_{\text{BP}}$. Second, let $\hat{\theta}^{(m)} = (\hat{\theta}_k^{(m)})^\top = (\hat{\theta}_1^{(m)}, \hat{\theta}_2^{(m)}, \hat{\theta}_3^{(m)})^\top$ be the estimates obtained in the m-th replication using the proposed GRP method. Then, for a given parameter θ_k with $1 \le k \le 3$, a 95% confidence interval is constructed as $\text{CI}_k^{(m)} = (\hat{\theta}_k^{(m)} - z_{0.975}\widehat{\text{SE}}_k^{(m)}, \hat{\theta}_k^{(m)} + z_{0.975}\widehat{\text{SE}}_k^{(m)})$, where $\widehat{\text{SE}}_k^m$ is the kth diagonal element of $(\hat{\Sigma}_{\text{GRP}}^m/N)^{1/2}$ and z_α is the α-th quantile of a standard normal distribution. Consequently, the empirical coverage probability

Table 1
Simulation results of RMSE values using different kinds of covariance estimators. Specifically, $\hat{\Sigma}_{\text{GRP}}$ stands for using the proposed GRP method, $\hat{\Sigma}_{\text{Non-BP}}$ stands for using non-bootstrap methods, $\hat{\Sigma}_{\text{BP}}$ stands for using BP method.

Simulation	N	$\hat{\Sigma}_{\text{GRP}}$	$\hat{\Sigma}_{\text{Non-BP}}$	$\hat{\Sigma}_{\text{BP}}$
Setting 1	5000	0.216	0.094	0.294
	10000	0.207	0.076	0.282
	20000	0.196	0.060	0.268
Setting 2	5000	0.267	0.171	0.328
	10000	0.217	0.127	0.296
	20000	0.178	0.090	0.262
Setting 3	5000	0.620	0.344	0.476
	10000	0.501	0.258	0.407
	20000	0.425	0.190	0.351
Setting 4	5000	1.018	0.992	0.812
	10000	0.945	0.921	0.783
	20000	0.847	0.823	0.758

Table 2
Simulation results of CP (in %) for $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)^\top$. The values given outside the parentheses are computed using the proposed GRP method. The values given in the first column inside the parentheses are computed using Non-BP method. The values given in the second column inside the parentheses are computed using BP method.

Simulation	N	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
Setting 1	5000	94.5(94.6,95.1)	95.2(95.2,95.1)	95.2(95.4,95.1)
	10000	95.2(96.1,95.1)	96.1(95.1,95.1)	94.3(96.1,95.1)
	20000	94.5(94.8,94.8)	94.4(96.2,94.8)	95.4(95.2,94.8)
Setting 2	5000	94.7(94.7,94.5)	95.6(95.9,94.5)	94.0(94.0,94.5)
	10000	94.9(95.4,94.6)	94.4(94.2,94.6)	94.4(94.5,94.6)
	20000	94.6(94.0,94.8)	94.9(94.7,94.8)	96.3(96.3,94.8)
Setting 3	5000	96.1(96.3,95.7)	96.2(95.2,94.3)	94.6(94.9,94.2)
	10000	95.5(94.8,94.0)	96.2(95.6,94.7)	95.6(95.2,94.3)
	20000	96.1(95.6,95.2)	95.2(94.5,94.2)	95.5(95.1,94.5)
Setting 4	500	94.1(94.4,93.2)	93.3(94.1,93.2)	92.6(93.1,92.2)
	10000	93.4(94.0,92.8)	93.3(93.5,93.0)	92.4(93.4,91.8)
	20000	95.5(95.5,94.0)	96.0(94.4,93.2)	94.5(94.4,93.9)

(CP) is computed as $\text{CP}_k = M^{-1}\sum_{m=1}^{M} I(\theta_k \in \text{CI}_k^{(m)})$, where $I(\cdot)$ is the indicator function. Similarly, we obtain the CP values for $\hat{\Sigma}_{\text{Non-BP}}$ and $\hat{\Sigma}_{\text{BP}}$.

Remark. Since our proposed method does not need to calculate the derivative of the objective function, it is more suitable to those settings that the derivative calculations are complex or not computable. However, in order to compare with the traditional MLE, the simulation settings considered in this article all have differentiable objective function and the derivatives are computable. We did not consider the cases with more complex objective functions, which limits the simulation studies.

3.3. Simulation results

The detailed simulation results are summarized in Tables 1–2, from which we can draw a number of conclusions. First, the proposed asymptotic covariance matrix estimator (i.e., $\hat{\Sigma}_{\text{GRP}}$) is consistent with its mean root square error values decreasing towards 0 as $N \to \infty$. Particularly, we can see that in the cases of Settings 1-3, when the form of asymptotic variance is simple, the RMSE values of the proposed method are slightly worse to the Non-BP method. However, when the asymptotic variance becomes complicated, such as Setting 4, our method is comparable with the Non-BP method. Second, in all of the four settings, the reported coverage probabilities (i.e., CPs) obtained by GRP, BP and Non-BP are nearly the same. Moreover, the coverage probability values are fairly close to their nominal 95% level, which suggests that the estimated standard errors (i.e., $\widehat{\text{SE}}$) are well approximated. In summary, all of the results indicate that the proposed Gaussian random perturbation estimation for $\hat{\Sigma}$ is unbiased and consistent.

Table 3
Estimation results for the real data example.

Variable	Coefficient	$\widehat{\mathrm{SE}_1}$	$\widehat{\mathrm{SE}_2}$	P_1	P_2
Intercept	−5.090	0.078	0.090	<0.001	<0.001
Tenure	−0.316	0.068	0.059	<0.001	<0.001
Expense	−0.279	0.060	0.060	<0.001	<0.001
Degree	−0.873	0.141	0.193	<0.001	<0.001
Tightness	−0.287	0.046	0.052	<0.001	<0.001
Entropy	−0.328	0.083	0.095	<0.001	<0.001

3.4. A real data example

To illustrate the practical usefulness of our method, we present a real data example for analyzing customer churn. This study aims to understand what factors affect customer churn in the mobile communication industry. The data are drawn from a mobile communication company in China that contains a total of $N = 44,571$ customers with their call-record information. For each customer i, we define a binary response variable, where $Y_i = 1$ if customer i stops using the service, otherwise $Y_i = 0$. The corresponding churn rate is 1.17%. The literature suggests that other than traditional factors (e.g., tenure, expense, etc.), social factors can have an important influence on customer churn (Nitzan and Libai, 2011). To investigate the possible factors affecting customer churn, we consider five covariates, which are called *tenure, expense, degree, tightness*, and *entropy*. Specifically, *tenure* is defined as the length of time the customer uses the service. *Expense* is defined as the average cost to a customer of using a mobile phone over a period of time.

To proceed with the explanation for the next three variables, we need to define an auxiliary variable called the adjacency matrix. Particularly, we assume that the network structure of N customers is captured by the adjacency matrix $A = (a_{ij}) \in \mathbb{R}^{N\times N}$, where $a_{ij} = 1$ if node i calls (or is called by) node j $(i \neq j)$, and $a_{ij} = 0$ otherwise. We then define $a_{ii} = 0$ and $\sum_{i=1}^{N} a_{ij} > 0$ for completeness. Then, *degree* is defined as $D_i = \sum_{j\neq i} a_{ij}$, indicating the number of contacts involved with a focal customer in one's own network. *Tightness* is defined as $T_i = Time_i/D_i$, where $Time_i$ is the total communication time between customer i and their connected members. *Entropy* is defined as $E_i = -\sum_{a_{ij}=1} p_{ij}\log(p_{ij})$, where $p_{ij} = Commu_{i,j}/Time_i$ and $Commu_{i,j}$ is the total communication time between j and i. Typically, a large entropy indicates a more dispersed average communication time.

To investigate the influence of the proposed factors on the customer churn rate, we conduct a standard logistic regression. All the variables have been standardized so that the mean is 0 and variance is 1. Table 3 reports the coefficients estimated via maximum likelihood estimation, the standard errors (i.e., $\widehat{\mathrm{SE}_1}$, $\widehat{\mathrm{SE}_2}$), and corresponding p-values (i.e., P_1, P_2) estimated using both the traditional method and our proposed method. From the table, we can see that the standard errors estimated by the proposed random perturbation method are very similar to those obtained using the traditional method. This indicates that our method is robust in practice with real data. Furthermore, we find that all the proposed factors are significant at the 0.1% level, which means that they all have a significant effect in explaining customer churn behavior.

4. Discussion

We propose here a Gaussian random perturbation method for estimating the asymptotic covariance matrix of general M-estimators. The key idea is to generate a small random perturbation around the local M-estimator. By re-evaluating the loss function at the randomly perturbed M-estimator, we obtain the estimator of the first and second order derivatives of the loss function via Taylor series expansion. This leads to a novel estimator for the asymptotic covariance matrix. We then rigorously show that the resulting covariance estimator is statistically consistent under appropriate regularity conditions. The method does not require the computation of the derivatives of the loss function, and it is convenient for parallel computing. The practical usefulness of the method is further demonstrated via both simulation and real data analysis.

To generalize the usefulness of the proposed Gaussian random perturbation method, we provide here two possible future research directions. First, it is of great importance to generalize the proposed method to accommodate loss functions with discontinuous first order derivatives, such as the quantile regression. Second, our method can be extended to nonparametric regression models. We believe these efforts may increase the value of the concept of Gaussian random perturbation considerably.

Acknowledgements

Jing Zhou's research is supported in part by the National Natural Science Foundation of China (No. 72171226), the Beijing Municipal Social Science Foundation (No. 19GLC052) and the National Statistical Science Research Project (No. 2020LZ38). Wei Lan's research was supported by the National Natural Science Foundation of China (Nos. 71532001, 11931014, 12171395, 71991472) and the Joint Lab of Data Science and Business Intelligence at Southwestern University of Finance and Economics. Hansheng Wang's research is partially supported by National Natural Science Foundation of China (No. 11831008) and also partially supported by the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (KLATASDS-MOE-ECNU-KLATASDS2101).

Appendix

This Appendix includes four parts: Appendix A presents a useful lemma, which provides proof for Theorems 1–2. Appendix B demonstrates Theorem 1, and Appendix C demonstrates Theorem 2. Appendix D gives the analytical form of the sandwich- type covariance matrix of the model in Setting 4 in the simulation.

Appendix A. A useful lemma

Lemma 1. *Let $A \in \mathbb{R}^{p\times p}$ with $p < \infty$ be an arbitrary symmetric matrix of bounded eigenvalues, $\delta = (\delta_1, \cdots, \delta_p)^\top \in \mathbb{R}^p$, which is a random vector generated from a multivariate normal distribution with mean 0 and covariance matrix $\sigma^2 I_p \in \mathbb{R}^{p\times p}$. Then we have, (i). $E\{(\delta^\top A\delta)\delta\delta^\top\} = 2\sigma^4 A + \sigma^4 tr(A) I_p$; (ii). for any $j_1, j_2 = 1, \cdots, p$, we have $var\{(\delta^\top A\delta)\delta_{j_1}\delta_{j_2}\} < \infty$; (iii). for any $j_1, j_2 = 1, \cdots, p$, we have $E\{(\delta^\top A\delta)^2\delta_{j_1}\delta_{j_2}\} < \infty$; and (iv). for any $j_1, j_2 = 1, \cdots, p$, we have $var\{(\delta^\top A\delta)^2\delta_{j_1}\delta_{j_2}\} < \infty$.*

Proof. We first prove (i). Define $Z = \delta/\sigma$, then $Z \sim N(0, I_p)$, and we have

$$\sigma^{-4} E\left\{(\delta^\top A\delta)\delta\delta^\top\right\} = E\left\{(Z^\top AZ)ZZ^\top\right\}. \tag{A.1}$$

Write $A = U^\top DU$ for some orthogonal matrix U ($UU^\top = U^\top U = I_p$) and diagonal matrix $D = \text{diag}\{d_j\} \in \mathbb{R}^{p\times p}$. Define $\widetilde{Z} = UZ = (\widetilde{Z}_1, \cdots, \widetilde{Z}_p)^\top \in \mathbb{R}^p$, then we have $\widetilde{Z} \sim N(0, I_p)$. The equation (A.1) can be further written as,

$$\begin{aligned} E\left\{(Z^\top AZ)ZZ^\top\right\} &= E\left\{(\widetilde{Z}^\top D\widetilde{Z})(U^\top\widetilde{Z}\widetilde{Z}^\top U) = E\left\{(\sum_{j=1}^p \widetilde{Z}_j^2 d_j)(U^\top\widetilde{Z}\widetilde{Z}^\top U)\right\}\right. \\ &= U^\top E\left\{(\sum_{j=1}^p \widetilde{Z}_j^2 d_j)\widetilde{Z}\widetilde{Z}^\top\right\}U = U^\top\left[E\{\widetilde{Z}_{j1}\widetilde{Z}_{j2}(\sum_{j=1}^p \widetilde{Z}_j^2 d_j)\}\right]U. \end{aligned} \tag{A.2}$$

It should be noted that $E\{\widetilde{Z}_{j_1}\widetilde{Z}_{j_2}(\sum_{j=1}^p \widetilde{Z}_j^2 d_j)\} = 0$ if $j_1 \neq j_2$. Otherwise, when $j_1 = j_2$, the quantity of $E\{\widetilde{Z}_{j_1}\widetilde{Z}_{j_2}(\sum_{j=1}^p \widetilde{Z}_j^2 d_j)\}$ becomes $2d_j + \sum_j d_j$. Therefore, we have $E\left\{(Z^\top AZ)ZZ^\top\right\} = 2\sigma^4 A + \sigma^4 tr(A) I_p$. This completes the first part of the proof.

We next prove (ii). From the proof of (i), we have $(\delta^\top A\delta)\delta_{j_1}\delta_{j_2} = \widetilde{Z}_{j_1}\widetilde{Z}_{j_2}(\sum_{j=1}^p \widetilde{Z}_j^2 d_j)$. According to the Cauchy-Schwartz inequality, we have

$$\text{var}\{(\delta^\top A\delta)\delta_{j_1}\delta_{j_2}\} \le p\sum_j d_j^2 \text{var}\{\widetilde{Z}_j^2\widetilde{Z}_{j_1}\widetilde{Z}_{j_2}\}.$$

When $j_1 \neq j_2$, we have $E(\widetilde{Z}_j^2\widetilde{Z}_{j_1}\widetilde{Z}_{j_2}) = 0$. Then, $\text{var}\{\widetilde{Z}_j^2\widetilde{Z}_{j_1}\widetilde{Z}_{j_2}\} = E(\widetilde{Z}_j^4\widetilde{Z}_{j_1}^2\widetilde{Z}_{j_2}^2) < \infty$. In addition, when $j_1 = j_2$, then $\text{var}\{\widetilde{Z}_j^2\widetilde{Z}_{j_1}\widetilde{Z}_{j_2}\} = \text{var}\{\widetilde{Z}_j^2\widetilde{Z}_{j_1}^2\} \le E(\widetilde{Z}_j^4\widetilde{Z}_{j_1}^4) < \infty$. Combining the results above, we then have $\text{var}\{(\delta^\top A\delta)\delta_{j_1}\delta_{j_2}\} < \infty$, which completes the second part of this lemma. (iii) and (iv) can be proved similarly. This completes the entire proof.

Appendix B. Proof of Theorem 1

The consistency of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ can be proved separately in the following two steps as $\min\{K, N\} \to \infty$.

Step I. We first prove that $\hat{\Sigma}_2 \to_p \Sigma_2$. By definition, it suffices to prove $\hat{\Sigma}_{2,int} \to_p \Sigma_2 + 2^{-1}\text{tr}(\Sigma_2) I_p$, where $\hat{\Sigma}_{2,int} = K^{-1}\sum_{k=1}^K \hat{\Sigma}_{2,int}^{(k)}$. For any $k = 1, \cdots, K$, by the Taylor series expansion, we have

$$\begin{aligned} \hat{\Sigma}_{2,int}^{(k)} &= (\sigma^4 N)^{-1}\sum_{i=1}^N \left\{\dot{\ell}^\top(X_i, \hat{\theta})\delta^{(k)}\right\}\delta^{(k)}\delta^{(k)\top} + (\sigma^4 N)^{-1}\sum_{i=1}^N \left\{\delta^{(k)\top}\ddot{\ell}(X_i, \hat{\theta})\delta^{(k)}\right\}\delta^{(k)}\delta^{(k)\top}/2 \\ &+ (\sigma^4 N)^{-1}\sum_{i=1}^N \left\{\sum_{j_1=1}^p\sum_{j_2=1}^p\sum_{j_3=1}^p \delta_{j_1}^{(k)}\delta_{j_2}^{(k)}\delta_{j_3}^{(k)} \frac{\partial \ell^3(X_i, \theta_1^{*(k)})}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}}\right\}\delta^{(k)}\delta^{(k)\top}/6 \triangleq \Lambda_1^{(k)} + \Lambda_2^{(k)} + \Lambda_3^{(k)}, \end{aligned}$$

where $\theta_1^{*(k)}$ lies between $\hat{\theta}$ and $\hat{\theta} + \delta^{(k)}$. One can easily verify that $\Lambda_1^{(k)} = 0$ by the definition of the M-estimator; we then only need to consider the last two parts, $\Lambda_2^{(k)}$ and $\Lambda_3^{(k)}$, separately.

We first consider $\Lambda_2^{(k)}$. By the Taylor series expansion, we have

$$\Lambda_2^{(k)} = \frac{1}{2}(\sigma^4 N)^{-1}\sum_{i=1}^N \left\{\delta^{(k)\top}\ddot{\ell}(X_i, \theta_0)\delta^{(k)}\right\}\delta^{(k)}\delta^{(k)\top}$$

$$+\frac{1}{4}(\sigma^4 N)^{-1}\sum_{i=1}^{N}\Big\{\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\delta_{j_1}^{(k)}\delta_{j_2}^{(k)}(\hat{\theta}-\theta_0)^\top\frac{\partial\ell^3(X_i,\theta_2^{*(k)})}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta}\Big\}\delta^{(k)}\delta^{(k)\top}\doteq\Lambda_{21}^{(k)}+\Lambda_{22}^{(k)},$$

where $\theta_2^{*(k)}$ is located between $\hat{\theta}$ and θ_0. We next consider $\Lambda_{21}^{(k)}$ and $\Lambda_{22}^{(k)}$ separately. By Lemma 1(i), we have $E(\Lambda_{21}^{(k)})=\Sigma_2+2^{-1}\mathrm{tr}(\Sigma_2)I_p$, and by Lemma 1(ii) we have $\mathrm{var}(\Lambda_{21,j_1j_2}^{(k)})=O(1)$, where $\Lambda_{21}^{(k)}=(\Lambda_{21,j_1j_2}^{(k)})\in\mathbb{R}^{p\times p}$. This immediately leads to $K^{-1}\sum_{k=1}^{K}\Lambda_{21}^{(k)}\to_p\Sigma_2+2^{-1}\mathrm{tr}(\Sigma_2)I_p$ by Lemma 1(i) and (ii) again. In addition, by Condition (C1), $\dddot{\ell}(X_i,\theta^*)$ is uniformly bounded in a small neighborhood of θ_0. Then, we have $\Lambda_{22}^{(k)}=O_p\{\|\hat{\theta}-\theta_0\|\}=o_p(1)$ for any k. Combining the results above, we have $K^{-1}\sum_{k=1}^{K}\Lambda_2^{(k)}=\Sigma_2+2^{-1}\mathrm{tr}(\Sigma_2)I_p+o_p(1)$.

We next consider $\Lambda_3^{(k)}$. By Condition (C1) again, $\dddot{\ell}(X_i,\theta)$ is bounded uniformly for θ in a small neighborhood of θ_0. Then, there exists a finite positive constant C_1 such that

$$\Lambda_3^{(k)}\le(\sigma^4N)^{-1}C_1\sum_{i=1}^{N}\Big\{\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{j_3=1}^{p}|\delta_{j_1}^{(k)}\delta_{j_2}^{(k)}\delta_{j_3}^{(k)}|\Big\}\delta^{(k)}\delta^{(k)\top}/6=O_p(\sigma)=o_p(1).$$

Combining the results above, we have completed the first part of the proof.

Step II. We next prove $\hat{\Sigma}_1\xrightarrow{p}\Sigma_1$. Similarly, it suffices to prove $\hat{\Sigma}_{1,int}\to_p\Sigma_1+2^{-1}\mathrm{tr}(\Sigma_1)I_p$, where $\hat{\Sigma}_{1,int}=K^{-1}\sum_{k=1}^{K}\hat{\Sigma}_{1,int}^{(k)}$. For any $k=1,\cdots,K$, by the Taylor series expansion, we have

$$\ell(X_i,\hat{\theta}+\delta^{(k)})-\ell(X_i,\hat{\theta})=\delta^{(k)\top}\dot{\ell}(X_i,\hat{\theta})+\delta^{(k)\top}\ddot{\ell}(X_i,\hat{\theta}+\eta_i^{(k)}\delta^{(k)})\delta^{(k)}$$

for some constant $\eta_i^{(k)}$ lies between $(0,1)$. Accordingly, we obtain

$$\begin{aligned}\hat{\Sigma}_{1,int}^{(k)}&=\frac{1}{2}(\sigma^4N)^{-1}\sum_{i=1}^{N}\Big\{\dot{\ell}^\top(X_i,\hat{\theta})\delta^{(k)}+\delta^{(k)\top}\ddot{\ell}(X_i,\hat{\theta}+\eta_i^{(k)}\delta^{(k)})\delta^{(k)}\Big\}^2\delta^{(k)}\delta^{(k)\top}\\&=\frac{1}{2}(\sigma^4N)^{-1}\sum_{i=1}^{N}\Big\{\delta^{(k)\top}\dot{\ell}(X_i,\hat{\theta})\Big\}^2\delta^{(k)}\delta^{(k)\top}\\&+(\sigma^4N)^{-1}\sum_{i=1}^{N}\Big\{\dot{\ell}^\top(X_i,\hat{\theta})\delta^{(k)}\delta^{(k)\top}\ddot{\ell}(X_i,\hat{\theta}+\eta_i^{(k)}\delta^{(k)})\delta^{(k)}\Big\}\delta^{(k)}\delta^{(k)\top}\\&+\frac{1}{2}(\sigma^4N)^{-1}\sum_{i=1}^{N}\Big\{\delta^{(k)\top}\ddot{\ell}(X_i,\hat{\theta}+\eta_i^{(k)}\delta^{(k)})\delta^{(1)}\Big\}^2\delta^{(k)}\delta^{(1)\top}\triangleq\Delta_1^{(k)}+\Delta_2^{(k)}+\Delta_3^{(k)}.\end{aligned}$$

We next consider the above three parts separately. First, we consider $\Delta_1^{(k)}$. Similar to the proof of Step I, we have

$$\Delta_1^{(k)}=\frac{1}{2}(\sigma^4N)^{-1}\sum_{i=1}^{N}\Big\{\dot{\ell}^\top(X_i,\theta_0)\delta^{(k)}\Big\}^2\delta^{(k)}\delta^{(k)\top}\big\{1+o_p(1)\big\}.$$

By Lemma 1(i) and (ii), and using a similar argument to that in the proof of Step I, we have $K^{-1}\sum_{k=1}^{K}\Delta_1^{(k)}\to_p\Sigma_1+2^{-1}\mathrm{tr}(\Sigma_1)I_p$ as $\min\{N,K\}\to\infty$.

We next consider $\Delta_2^{(k)}$. By Condition (C1), $\ddot{\ell}(X_i,\theta)$ is bounded uniformly in a small neighborhood of θ_0 and $\mathrm{var}\big\{\dot{\ell}(X_i,\theta_0)\big\}<\infty$. Then, there exists a finite positive constant C_2 such that

$$\Delta_2^{(k)}\le(\sigma^4N)^{-1}\sum_{i=1}^{N}|\dot{\ell}^\top(X_i,\theta_0)\delta^{(k)}|\delta^{(k)}\delta^{(k)\top}\delta^{(k)}\delta^{(k)\top}=O_p(\sigma)=o_p(1).$$

Lastly, we consider $\Delta_3^{(k)}$. Similarly, we have

$$\Delta_3^{(k)}=\frac{1}{2}(\sigma^4N)^{-1}\sum_{i=1}^{N}\Big\{\delta^{(k)\top}\ddot{\ell}(X_i,\theta_0)\delta^{(k)}\Big\}^2\delta^{(k)}\delta^{(k)\top}\big\{1+o_p(1)\big\}.$$

By Lemma 1(iii) and (iv), we have

$$N^{-1}\sum_{i=1}^{N}\left\{\delta^{(k)\top}\ddot{\ell}(X_i,\theta_0)\delta^{(k)}\right\}^2\delta^{(k)}\delta^{(k)\top}=O_p(\sigma^6).$$

Then, we obtain $K^{-1}\sum_{k=1}^{K}\Delta_3^{(k)}=O_p(\sigma^2)=o_p(1)$ since $\delta^{(k)}$s are iid. Combining the results above, we then have

$$K^{-1}\sum_{k=1}^{K}\hat{\Sigma}_{1,int}^{(k)}=K^{-1}\sum_{k=1}^{K}\left\{\Delta_1^{(k)}+\Delta_2^{(k)}+\Delta_3^{(k)}\right\}=K^{-1}\sum_{k=1}^{K}\Delta_1^{(k)}+o_p(1)\to_p\Sigma_1+2^{-1}\mathrm{tr}(\Sigma_1)I_p,$$

which completes the entire proof.

Appendix C. Proof of Theorem 2

Similar to the proof of Theorem 1, we prove the consistency of $\hat{\Sigma}_{n1}$ and $\hat{\Sigma}_{n2}$ separately in the following two steps.

Step 1. We first consider $\hat{\Sigma}_{n2}$. By definition, it suffices to prove $\hat{\Sigma}_{n2,int}\to_p\bar{\Sigma}_2+2^{-1}\mathrm{tr}(\bar{\Sigma}_2)I_p$, where $\hat{\Sigma}_{n2,int}=K^{-1}\sum_{k=1}^{K}\hat{\Sigma}_{n2,int}^{(k)}$. By the Taylor series expansion, we have

$$\begin{aligned}\hat{\Sigma}_{n2,int}^{(k)}&=(\sigma^4N)^{-1}\sum_{i=1}^{N}\left\{\frac{\partial\ell^\top(X_i,\hat{\theta},\hat{\gamma})}{\partial\theta}\delta^{(k)}\right\}\delta^{(k)}\delta^{(k)\top}+(\sigma^4N)^{-1}\sum_{i=1}^{N}\left\{\delta^{(1)\top}\frac{\partial\ell^2(X_i,\hat{\theta},\hat{\gamma})}{\partial\theta\partial\theta^\top}\delta^{(k)}\right\}\delta^{(k)}\delta^{(k)\top}/2\\&\quad+(\sigma^4N)^{-1}\sum_{i=1}^{N}\left\{\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{j_3=1}^{p}\delta_{j_1}^{(k)}\delta_{j_2}^{(k)}\delta_{j_3}^{(k)}\frac{\partial\ell^3(X_i,\theta_{n1}^{*(k)},\hat{\gamma})}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}}\right\}\delta^{(k)}\delta^{(k)\top}/6\triangleq\Lambda_{n1}^{(k)}+\Lambda_{n2}^{(k)}+\Lambda_{n3}^{(k)},\end{aligned}$$

where $\theta_{n1}^{*(k)}$ lies between $\hat{\theta}$ and $\hat{\theta}+\delta^{(k)}$. One can easily verify that $\Lambda_{n1}^{(k)}=0$ by the definition of the M-estimator, we then only need to consider the last two parts, $\Lambda_{n2}^{(k)}$ and $\Lambda_{n3}^{(k)}$, separately. By Condition (C3), $p<\infty$, and the definition of $\delta^{(k)}$, there exists a finite positive constant C_{n1} such that

$$\Lambda_{n3}^{(k)}\le(\sigma^4N)^{-1}C_{n1}\sum_{i=1}^{N}\left\{\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{j_3=1}^{p}|\delta_{j_1}^{(k)}\delta_{j_2}^{(k)}\delta_{j_3}^{(k)}|\right\}\delta^{(k)}\delta^{(k)\top}/6=O_p(\sigma)=o_p(1),$$

which leads to $K^{-1}\sum_{k=1}^{K}\Lambda_{n3}^{(k)}=o_p(1)$ as $\delta^{(k)}$s are iid. We next consider $\Lambda_{n2}^{(k)}$. By the Taylor series expansion, we again have

$$\begin{aligned}\Lambda_{n2}^{(k)}&=\frac{1}{2}(\sigma^4N)^{-1}\sum_{i=1}^{N}\left\{\delta^{(k)\top}\frac{\partial\ell^2(X_i,\theta_0,\hat{\gamma})}{\partial\theta\partial\theta^\top}\delta^{(k)}\right\}\delta^{(k)}\delta^{(k)\top}\\&\quad+\frac{1}{4}(\sigma^4N)^{-1}\sum_{i=1}^{N}\left\{\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\delta_{j_1}^{(k)}\delta_{j_2}^{(k)}(\hat{\theta}-\theta_0)^\top\frac{\partial\ell^3(X_i,\theta_{n2}^{*(k)},\hat{\gamma})}{\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta}\right\}\delta^{(k)}\delta^{(k)\top}\doteq\Lambda_{n21}^{(k)}+\Lambda_{n22}^{(k)},\end{aligned}$$

where $\theta_{n2}^{*(k)}$ lies between $\hat{\theta}$ and θ_0. Similarly, by Condition (C3), we have $\Lambda_{n22}^{(k)}=O_p(\sigma)=o_p(1)$ and $K^{-1}\sum_{k=1}^{K}\Lambda_{n22}^{(k)}=o_p(1)$. In addition, by a similar technique for proving $\Lambda_{21}^{(k)}$, we can find that $\Lambda_{n21}^{(k)}=\frac{1}{2}(\sigma^4N)^{-1}\sum_{i=1}^{N}\left\{\delta^{(k)\top}\frac{\partial\ell^2(X_i,\theta_0,\gamma_0)}{\partial\theta\partial\theta^\top}\delta^{(k)}\right\}\delta^{(k)}\times\delta^{(k)\top}\{1+o_p(1)\}$ and $K^{-1}\sum_{k=1}^{K}\Lambda_{n21}^{(k)}\to_p\bar{\Sigma}_2+2^{-1}\mathrm{tr}(\bar{\Sigma}_2)I_p$. Combining the results above, we have competed the first part of the proof.

Step 2. We next consider $\hat{\Sigma}_{n1}$. By definition, it suffices to prove $\hat{\Sigma}_{n1,int}\to_p\bar{\Sigma}_1+2^{-1}\mathrm{tr}(\bar{\Sigma}_1)I_p$, where $\hat{\Sigma}_{n1,int}=K^{-1}\sum_{k=1}^{K}\hat{\Sigma}_{n1,int}^{(k)}$. For any $k=1,\cdots,K$, by the Taylor series expansion, we have

$$\ell(X_i,\hat{\theta}+\delta^{(k)},\hat{\gamma})-\ell(X_i,\hat{\theta},\hat{\gamma})=\delta^{(k)\top}\frac{\partial\ell(X_i,\hat{\theta},\hat{\gamma})}{\partial\theta}+\delta^{(k)\top}\frac{\ell^2(X_i,\hat{\theta}+\eta_i^{*(k)}\delta^{(k)},\hat{\gamma})}{\partial\theta\partial\theta^\top}\delta^{(k)}$$

for some constant $\eta_i^{*(k)}$ lies between $(0,1)$. Accordingly, we obtain

$$\begin{aligned}\hat{\Sigma}_{1,int}^{(k)} &= \frac{1}{2}(\sigma^4 N)^{-1}\sum_{i=1}^{N}\Big\{\frac{\partial \ell^\top(X_i,\hat{\theta},\hat{\gamma})}{\partial\theta}\delta^{(k)} + \delta^{(k)\top}\frac{\ell^2(X_i,\hat{\theta}+\eta_i^{*(k)}\delta^{(k)},\hat{\gamma})}{\partial\theta\partial\theta^\top}\delta^{(k)}\Big\}^2\delta^{(k)}\delta^{(k)\top}\\ &= \frac{1}{2}(\sigma^4 N)^{-1}\sum_{i=1}^{N}\Big\{\delta^{(k)\top}\frac{\partial \ell(X_i,\hat{\theta},\hat{\gamma})}{\partial\theta}\Big\}^2\delta^{(k)}\delta^{(k)\top}\\ &+ (\sigma^4 N)^{-1}\sum_{i=1}^{N}\Big\{\frac{\partial \ell^\top(X_i,\hat{\theta},\hat{\gamma})}{\partial\theta}\delta^{(k)}\delta^{(k)\top}\frac{\ell^2(X_i,\hat{\theta}+\eta_i^{*(k)}\delta^{(k)},\hat{\gamma})}{\partial\theta\partial\theta^\top}\delta^{(k)}\Big\}\delta^{(k)}\delta^{(k)\top}\\ &+ \frac{1}{2}(\sigma^4 N)^{-1}\sum_{i=1}^{N}\Big\{\delta^{(k)\top}\frac{\ell^2(X_i,\hat{\theta}+\eta_i^{*(k)}\delta^{(k)},\hat{\gamma})}{\partial\theta\partial\theta^\top}\delta^{(k)}\Big\}^2\delta^{(k)}\delta^{(k)\top} \triangleq \Delta_{n1}^{(k)} + \Delta_{n2}^{(k)} + \Delta_{n3}^{(k)}.\end{aligned}$$

We next consider the above three parts separately. Similar to the proofs for $\Delta_2^{(k)}$ and $\Delta_3^{(k)}$, by Conditions (C3) and (C4), one can verify that $K^{-1}\sum_{k=1}^{K}\Delta_{n2}^{(k)} = o_p(1)$ and $K^{-1}\sum_{k=1}^{K}\Delta_{n3}^{(k)} = o_p(1)$ because $\sigma \to 0$. Then, we only need to consider $\Delta_{n1}^{(k)}$. By Taylor series expansion, we can obtain that

$$\frac{\partial \ell(X_i,\theta_0,\hat{\gamma})}{\partial\theta} - \frac{\partial \ell(X_i,\theta_0,\hat{\gamma})}{\partial\theta} = \frac{\partial^2 \ell(X_i,\theta_0,\gamma_0)}{\partial\theta\partial\gamma}(\hat{\gamma}-\gamma_0)\{1+o_p(1)\}.$$

In addition, by Condition (C4), we have $E\big\{\partial^2\ell(X_i,\theta_0,\gamma_0)/\partial\theta\partial\gamma\big\} = 0$. Using this result, one can then verify that

$$\Delta_{n1}^{(k)} = \frac{1}{2}(\sigma^4 N)^{-1}\sum_{i=1}^{N}\Big\{\frac{\partial \ell^\top(X_i,\theta_0,\gamma_0)}{\partial\theta}\delta^{(k)}\Big\}^2\delta^{(k)}\delta^{(k)\top}\big\{1+o_p(1)\big\}.$$

By Lemma 1(i) and (ii), and using a similar argument to that in the proof of Theorem 1, we have $K^{-1}\sum_{k=1}^{K}\Delta_{n1}^{(k)} \to_p \bar{\Sigma}_1 + 2^{-1}\mathrm{tr}(\bar{\Sigma}_1)I_p$ as $\min\{N,K\}\to\infty$, which completes the entire proof.

Appendix D. Analytical form of the covariance matrix of the model in Setting 4

To obtain the analytical form of the covariance matrix of the multiplicative model in Setting 4, we need to know the analytical forms of Σ_1 and Σ_2. Accordingly, we know that $\Sigma_1 = E\big[\dot{\mathcal{L}}(\theta)\dot{\mathcal{L}}^\top(\theta)\big]/N$ and $\Sigma_2 = E\big[\ddot{\mathcal{L}}(\theta)\big]/N$. To this end, the key problem is to calculate $\dot{\mathcal{L}}(\theta)$ and $\ddot{\mathcal{L}}(\theta)$. Since the loss function $\mathcal{L}(\theta)$ in Setting 4 is defined as

$$\mathcal{L}(\theta) = \frac{1}{2}\sum_{i=1}^{N}\Big[\Big\{\frac{Y_i-\exp(X_i^\top\theta)}{Y_i}\Big\}^2 + \Big\{\frac{Y_i-\exp(X_i^\top\theta)}{\exp(X_i^\top\beta)}\Big\}^2\Big],$$

its first order derivative, $\dot{\mathcal{L}}(\theta)$, is derived as $\dot{\mathcal{L}}(\theta) = -\sum_{i=1}^{N}(z_{i1}-z_{i2}+z_{i3}-z_{i4})$, where $z_{i1} = \exp(X_i^\top\beta)X_i/Y_i$, $z_{i2} = \{\exp(X_i^\top\beta)\}^2X_i/Y_i^2$, $z_{i3} = X_iY_i^2/\{\exp(X_i^\top\beta)\}^2$, and $z_{i4} = X_iY_i/\exp(X_i^\top\beta)$. Then we have,

$$\begin{aligned}\Sigma_1 &= \frac{1}{N}\sum_{i=1}^{N}E\big\{(z_{i1}-z_{i2}+z_{i3}-z_{i4})(z_{i1}-z_{i2}+z_{i3}-z_{i4})^\top\big\}\\ &= \frac{1}{N}\sum_{i=1}^{N}E\big\{z_{i1}z_{i1}^\top - 2z_{i1}z_{i2}^\top + 2z_{i1}z_{i3}^\top - 2z_{i1}z_{i4}^\top + z_{i2}z_{i2}^\top - 2z_{i2}z_{i3}^\top\\ &+ 2z_{i2}z_{i4}^\top + z_{i3}z_{i3}^\top - 2z_{i3}z_{i4}^\top + z_{i4}z_{i4}^\top\big\}\\ &= \frac{1}{N}\sum_{i=1}^{N}E(z_{i1}z_{i1}^\top) - 2E(z_{i1}z_{i2}^\top) + 2E(z_{i1}z_{i3}^\top) - 2E(z_{i1}z_{i4}^\top) + E(z_{i2}z_{i2}^\top)\\ &- 2E(z_{i2}z_{i3}^\top) + 2E(z_{i2}z_{i4}^\top) + E(z_{i3}z_{i3}^\top) - 2E(z_{i3}z_{i4}^\top) + E(z_{i4}z_{i4}^\top).\end{aligned} \tag{A.3}$$

According to Setting 4, we know $\log(Y_i) = X_i^\top\beta + \log(\epsilon_i)$, where $\log(\epsilon_i) \sim N(0,1)$. We then have $\log(Y_i) \sim N(X_i^\top\beta, 1)$. Therefore, the conditional moments of Y_i can be obtained as $E(Y_i^t|X_i) = \exp(X_i^\top\beta t + t^2/2)$, where $t\in\mathbb{R}$. Then, we have the following calculations as

$$
\begin{aligned}
E(z_{i1}z_{i1}^\top) &= E(z_{i4}z_{i4}^\top) = E\Big[X_iX_i^\top \exp(2X_i^\top\beta)/Y_i^2\Big] = \exp(2)E(X_iX_i^\top)\\
E(z_{i1}z_{i2}^\top) &= E(z_{i3}z_{i4}^\top) = E\Big[X_iX_i^\top \exp(3X_i^\top\beta)/Y_i^3\Big] = \exp(9/2)E(X_iX_i^\top)\\
E(z_{i1}z_{i3}^\top) &= E(z_{i2}z_{i4}^\top) = E\Big[X_iX_i^\top Y_i/\exp(X_i^\top\beta)\Big] = \exp(1/2)E(X_iX_i^\top)\\
E(z_{i2}z_{i2}^\top) &= E(z_{i3}z_{i3}^\top) = E\Big[X_iX_i^\top \exp(4X_i^\top\beta)/Y_i^4\Big] = \exp(8)E(X_iX_i^\top)\\
E(z_{i1}z_{i4}^\top) &= E(z_{i2}z_{i3}^\top) = E(X_iX_i^\top).
\end{aligned}
\tag{A.4}
$$

By plugging (A.4) into (A.3), we can obtain the final analytical result of Σ_1. This completes the calculation of Σ_1. Next, we consider the calculation of Σ_2. The second order derivative $\ddot{\mathcal{L}}(\theta)$ is derived as $\ddot{\mathcal{L}}(\theta) = \sum_{i=1}^N (w_{i1} - w_{i2})$, where $w_{i1} = X_iX_i^\top Y_i\{2Y_i - \exp(X_i^\top\beta)\}/\exp(2X_i^\top\beta)$ and $w_{i2} = X_iX_i^\top \exp(X_i^\top\beta)\{Y_i - 2\exp(X_i^\top\beta)\}/Y_i^2$. Then, we have

$$
\Sigma_2 = \frac{1}{N}E\big[\ddot{\ell}(X_i,\theta)\big] = \frac{1}{N}\sum_{i=1}^N E(w_{i1} - w_{i2}) = \frac{1}{N}\sum_{i=1}^N \big\{E(w_{i1}) - E(w_{i2})\big\}. \tag{A.5}
$$

Accordingly, we have the following calculation as

$$
\begin{aligned}
E(w_{i1}) &= E\Big[X_iX_i^\top Y_i\{2Y_i - \exp(X_i^\top\beta)\}\Big]/E\big\{\exp(2X_i^\top\beta)\big\}\\
&= E\Big[X_iX_i^\top\big\{2E(Y_i^2|X_i) - \exp(X_i^\top\beta)E(Y_i|X_i)\big\}\Big]/E\big\{\exp(2X_i^\top\beta)\big\}\\
&= E\Big[X_iX_i^\top\{2\exp(2) - \exp(1/2)\}\Big].
\end{aligned}
\tag{A.6}
$$

We can calculate $E(w_{i2})$ in the same way. Given careful derivation, we find $E(w_{i2}) = E(w_{i1}) = E\Big[X_iX_i^\top\{2\exp(2) - \exp(1/2)\}\Big]$. Therefore, by plugging (A.6) into (A.5), and by the fact that $E(w_{i2}) = E(w_{i1})$, we can obtain the analytical form of Σ_2. This completes the calculation of Σ_2. With the derived analytical forms of Σ_1 and Σ_2, we can obtain the sandwich-type covariance matrix by calculating $\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}$.

References

Battey, H., Fan, J., Liu, H., Lu, J., Zhu, Z., 2018. Distributed testing and estimation under sparse high dimensional models. Ann. Stat. 46, 1352–1382.

Casella, G., Fienberg, S., Olkin, I., 2015. An Introduction to Statistical Learning with Applications in R, seventh edition. Springer, New York/Heidelberg/Dordrecht/London.

Chen, K., Guo, S., Lin, Y., Ying, Z., 2010. Least absolute relative error estimation. J. Am. Stat. Assoc. 105 (491), 1104–1112.

Chen, X., Wan, A., Zhou, Y., 2015. Efficient quantile regression analysis with missing observations. J. Am. Stat. Assoc. 110, 723–741.

Efron, B., Stein, C., 1981. The jackknife estimate of variance. Ann. Stat. 9, 586–596.

Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat. Sci. 1, 54–75.

Efron, B., Gong, G., 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am. Stat. 37, 36–48.

Efron, B., 1994. Missing data, imputation, and the bootstrap. J. Am. Stat. Assoc. 89, 463–475.

Fan, J., Wang, D., Wang, K., Zhu, Z., 2019. Distributed estimation of principal eigenspaces. Ann. Stat. 47, 3009–3031.

Greene, W.H., 1997. Econometric Analysis, third edition. Prentice-Hall, Inc.

Huang, R., Liang, Y., Carriere, K.C., 2005. The role of proxy information in missing data analysis. Stat. Methods Med. Res. 14 (5), 457.

Ibrahim, J., Molenberghs, G., 2009. Rejoinder on: missing data methods in longitudinal studies: a review. TEST 18, 68–75.

Jiao, J., Han, Y., 2020. Bias correction with jackknife, bootstrap, and Taylor series. IEEE Trans. Inf. Theory 66, 4392–4418.

Jordan, M.I., Lee, J.D., Yang, Y., 2018. Communication-efficient distributed statistical inference. J. Am. Stat. Assoc. 114, 1–14.

Lehmann, E.L., Casella, G., 1983. Theory of Point Estimation. Wiley.

Li, X., Li, R., Xia, Z., Xu, C., 2020. Distributed feature screening via component wise debiasing. J. Mach. Learn. Res. 21, 1–32.

Lin, H., Liu, W., Lan, W., 2021. Regression analysis with individual-specific patterns of missing covariates. J. Bus. Econ. Stat. 39, 179–188.

Maslian, I., Abdelfattah, A., Haidar, A., Tomov, S., Baboulin, M., Falcou, J., et al., 2019. Algorithms and optimization techniques for high-performance matrix-matrix multiplications of very small matrices. Parallel Comput. 81, 1–21.

Nitzan, I., Libai, B., 2011. Social effects on customer retention. J. Mark. 75, 24–38.

Shao, J., 2003. Mathematical Statistics, second edition. Springer-Verlag, New York.

Shao, J., Wang, H., 2002. Sample correlation coefficients based on survey data under regression imputation. J. Am. Stat. Assoc. 97, 544–552.

Vegh, J., 2018. Introducing the explicitly many-processor approach. Parallel Comput. 75, 28–40.

Wang, Q., Dai, P., 2008. Semiparametric model-based inference in the presence of missing responses. Biometrika 95, 721–734.

Wooldridge, J.M., 2001. Applications of generalized method of moments estimation. J. Econ. Perspect. 15, 87–100.

Wooldridge, J.M., 2015. Introductory Econometrics A Modern Approach, sixth edition. Cengage Learing.

Zhao, J., Shao, J., 2015. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. J. Am. Stat. Assoc. 110, 1577–1590.

Zhou, J., Liu, J., Wang, F., Wang, H., 2022. Autoregressive model with spatial dependence and missing data. J. Bus. Econ. Stat. 40, 28–34.