



Research article

Corrected optimal subsampling for a class of generalized linear measurement error models

Ruiyuan Chang, Xiuli Wang and Mingqiu Wang*

School of Statistics and Data Science, Qufu Normal University, Qufu 273100, China

* **Correspondence:** Email: mqwang@vip.126.com.

Abstract: Subsampling techniques have been promoted in massive data and can substantially reduce the computing time. However, existing subsampling techniques do not consider the case of dirty data, especially the inaccuracy of covariates due to measurement errors, which will lead to the inconsistent estimators of regression coefficients. Therefore, to eliminate the influence of measurement errors on parameter estimators for massive data, this paper combined the corrected score function with the subsampling technique. The consistency and asymptotic normality of the estimators in the general subsampling are also derived. In addition, optimal subsampling probabilities are obtained based on the general subsampling algorithm using the A-optimality and L-optimality criteria and the truncation method, and then an adaptive two-step algorithm is developed. The effectiveness of the proposed method is demonstrated through numerical simulations and two real data analyses.

Keywords: A-optimality criteria; generalized linear models; L-optimality criteria; massive data; measurement errors; optimal subsampling

Mathematics Subject Classification: 62F12, 62J12

1. Introduction

The generalized linear model (GLM), a generalization of the linear model with wide applications in many research areas, was proposed by Nelder and Wedderburn [1] in 1972 for discrete dependent variables, which cannot be dealt with by the ordinary linear regression model. The GLM allows the response variable to be nonnormal distributions, including binomial, Poisson, gamma, and inverse Gaussian distributions, whose means are linked with the predictors by a link function.

Nowadays, with the rapid development of science and technology, massive data is ubiquitous in many fields, including medicine, industry, and economics. Extracting effective information from massive data is the core challenge of big data analysis. However, the limited arithmetic power of computers tends to consume a lot of computing time. In order to deal with this challenge, parallel

computing and distributed computing are commonly used, and subsampling techniques have emerged as a result, i.e., a small number of representative samples are extracted from massive data. Imberg et al. [2] proposed a theory on optimal design in the context of general data subsampling issues. It includes and extends most existing methods, works out optimality conditions, and offers algorithms for finding optimal subsampling scheme designs, which introduces a new class of invariant linear optimality criteria. Chao et al. [3] presented an optimal subsampling approach for modal regression with big data. The estimators are obtained by means of a two-step algorithm based on the modal expectation maximization when the bandwidth is not related to the subsample size.

There has been a great deal of research on subsampling algorithms of specific models. Wang et al. [4] devised a rapid subsampling algorithm to approximate the maximum likelihood estimators in the context of logistic regression. Based on the previous study, Wang [5] presented an enhanced estimation method for logistic regression, which has a higher estimation efficiency. In the case that data are usually distributed in multiple distributed sites for storage, Zuo et al. [6] developed a distributed subsampling procedure to effectively approximate the maximum likelihood estimators of logistic regression. Ai et al. [7] focused on the optimal subsampling method under the A-optimality criteria based on the method developed by Wang [4] for generalized linear models to quickly approximate maximum likelihood estimators from massive data. Yao and Wang [8] examined optimal subsampling methods for various models, including logistic regression models, softmax regression models, generalized linear models, quantile regression models, and quasi-likelihood estimators. Yu et al. [9] proposed an efficient subsampling procedure for online data streams with a multinomial logistic model. Yu et al. [10] studied the subsampling technique for the Akaike information criterion (AIC) and the smoothed AIC model-averaging framework for generalized linear models. Yu et al. [11] reviewed several subsampling methods for massive datasets from the viewpoint of statistical design.

To the best of our knowledge, all the existing methods above assume that the covariates are fully observable. However, in practice, this assumption is not realistic, and covariates may be inaccurately observed owing to measurement errors, which will lead to biases in the estimators of the regression coefficients. This means that we may incorrectly determine some unimportant variables as significant, which in turn affects the model selection and interpretation. Therefore, it is necessary to consider measurement errors. Liang et al. [12], Li and Xue [13], and Liang and Li [14] investigated the partial linear measurement error models. Stefanski [15] and Nakamura [16] obtained the corrected score functions of the GLM, such as linear regression, gamma regression, inverse gamma regression, and Poisson regression. Yang et al. [17] proposed an empirical likelihood method based on the moment identity of the corrected score function to perform statistical inference for a class of generalized linear measurement error models. Fuller [18] estimated the variable error model using the maximum likelihood method and studied statistical inference. Hu and Cui [19] proposed a corrected error variance method to accurately estimate the error variance, which can effectively reduce the influence of measurement error and false correlation at the same time. Carroll et al. [20] summarized the measurement errors in linear regression and described some simple and universally applicable measurement error analysis methods. Yi et al. [21] presented a regression calibration method, which is one of the first statistical methods introduced to address measurement errors in the covariates. In addition, they presented an overview of the conditional score and corrected score approaches for measurement error correction. Regarding the measurement errors in different situations existing in actual data, extensive research has been carried out, and a variety of methods have been proposed,

see [22–25]. Recently, a class of variable selection procedures has been developed for measurement error models, see [26, 27]. More recently, Ju et al. [28] studied the optimal subsampling algorithm and the random perturbation subsampling algorithm for big data linear models with measurement errors. The aim of this paper is to estimate the parameters using a subsampling algorithm for a class of generalized linear measurement error models in the massive data analysis.

In this paper, we study a class of the GLM with measurement errors, such as logistic regression models and Poisson regression models. We combine the corrected score function method with subsampling techniques to investigate subsampling algorithms. The consistency and asymptotic normality of the estimators obtained in the general subsampling algorithm are derived. We optimize the subsampling probabilities based on the design of A-optimality and L-optimality criteria and incorporate a truncation method in the optimal subsampling probabilities to obtain the optimal estimators. In addition, we develop an adaptive two-step algorithm and obtain the consistency and asymptotic normality of the final subsampling estimators. Finally, the effectiveness of the proposed method is demonstrated through numerical simulations and real data analysis.

The remainder of this paper is organized as follows: Section 2 introduces the corrected score function under different distributions and derives the general subsampling algorithm and the adaptive two-step algorithm. Sections 3 and 4 verify the effectiveness of the proposed method by generating simulated experimental data and two real data sets, respectively. Section 5 provides conclusions.

2. Model and methodology

In the GLM, it is assumed that the conditional distribution of the response variable belongs to the exponential family

$$f(y; \theta) = \exp \left\{ \frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi) \right\},$$

where $a(\cdot), b(\cdot), c(\cdot, \cdot)$ are known functions, θ is called the natural parameter, and ϕ is called the dispersion parameter.

Let $\{(\mathbf{X}_i, Y_i)\}_{i=1}^N$ be independent and identically distributed random samples, $\mu_i = E(Y_i | \mathbf{X}_i)$, $V(\mu_i) = \text{Var}(Y_i | \mathbf{X}_i)$, where the covariate $\mathbf{X}_i \in \mathbb{R}^p$ and the response variable $Y_i \in \mathbb{R}$, $V(\cdot)$ is a known variance function. The conditional expectation of Y_i given \mathbf{X}_i is

$$g(\mu_i) = \mathbf{X}_i^T \boldsymbol{\beta}, \quad (2.1)$$

where $g(\cdot)$ is the canonical link function, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is a p -dimensional unknown regression parameter.

In practice, covariates are not always accurately observed, and there are measurement errors that cannot be ignored. Let \mathbf{W}_i be an accurate observation of the covariate \mathbf{X}_i . Assuming that the additive measurement error model is

$$\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i, \quad (2.2)$$

where $\mathbf{U}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_u)$, and it is independent of (\mathbf{X}_i, Y_i) . Combining (2.1) and (2.2) yields a generalized linear model with measurement errors.

Define the log-likelihood function as $\ell(\boldsymbol{\beta}; Y_i) = \sum_{i=1}^N \log f(Y_i; \boldsymbol{\beta})$. If \mathbf{X}_i is observable, the score function for $\boldsymbol{\beta}$ in (2.1) is

$$\sum_{i=1}^N \boldsymbol{\eta}_i(\boldsymbol{\beta}; \mathbf{X}_i, Y_i) = \sum_{i=1}^N \frac{\partial \ell(\boldsymbol{\beta}; Y_i)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N \frac{Y_i - \mu_i}{V(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \boldsymbol{\beta}},$$

and satisfies $E[\boldsymbol{\eta}_i(\boldsymbol{\beta}; \mathbf{X}_i, Y_i) \mid \mathbf{X}_i] = \mathbf{0}$. However, when there is an error in \mathbf{X}_i , directly replacing \mathbf{X}_i with \mathbf{W}_i to calculate $\boldsymbol{\eta}_i(\boldsymbol{\beta}; \mathbf{X}_i, Y_i)$ causes a bias, i.e., $E[\boldsymbol{\eta}_i(\boldsymbol{\beta}; \mathbf{X}_i, Y_i)] = \mathbf{0}$ will not always hold, hence a correction is needed. We define an unbiased score function $\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta}; \mathbf{W}_i, Y_i)$ for $\boldsymbol{\beta}$ satisfying $E[\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta}; \mathbf{W}_i, Y_i) \mid \mathbf{X}_i] = \mathbf{0}$ by the idea of [16]. The maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ of $\boldsymbol{\beta}$ is the solution of the estimating equation

$$\boldsymbol{Q}(\boldsymbol{\beta}) := \sum_{i=1}^N \boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta}; \mathbf{W}_i, Y_i) = \mathbf{0}. \quad (2.3)$$

Based on the following moment identities associated with the error model (2.2),

$$E(\mathbf{W}_i \mid \mathbf{X}_i) = \mathbf{X}_i,$$

$$E(\mathbf{W}_i \mathbf{W}_i^T \mid \mathbf{X}_i) = \mathbf{X}_i \mathbf{X}_i^T + \boldsymbol{\Sigma}_u,$$

$$E(\exp(\mathbf{W}_i^T \boldsymbol{\beta}) \mid \mathbf{X}_i) = \exp\left(\mathbf{X}_i^T \boldsymbol{\beta} + \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}\right),$$

$$E\left[\mathbf{W}_i \exp(\mathbf{W}_i^T \boldsymbol{\beta}) \mid \mathbf{X}_i\right] = (\mathbf{X}_i + \boldsymbol{\Sigma}_u \boldsymbol{\beta}) \exp\left(\mathbf{X}_i^T \boldsymbol{\beta} + \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}\right),$$

$$E\left[\mathbf{W}_i \exp(-\mathbf{W}_i^T \boldsymbol{\beta}) \mid \mathbf{X}_i\right] = (\mathbf{X}_i - \boldsymbol{\Sigma}_u \boldsymbol{\beta}) \exp\left[-\mathbf{X}_i^T \boldsymbol{\beta} + \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}\right],$$

$$E\left[\mathbf{W}_i \exp(-2\mathbf{W}_i^T \boldsymbol{\beta}) \mid \mathbf{X}_i\right] = (\mathbf{X}_i - 2\boldsymbol{\Sigma}_u \boldsymbol{\beta}) \exp\left[-2\mathbf{X}_i^T \boldsymbol{\beta} + 2\boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}\right],$$

then we can construct the unbiased score function for binary logistic measurement error regression models and Poisson measurement error regression models, which are widely used in practice.

(1) Binary logistic measurement error regression models.

We consider the logistic measurement error regression model

$$\begin{cases} P(Y_i = 1 \mid \mathbf{X}_i) = \frac{1}{1 + \exp(-\mathbf{X}_i^T \boldsymbol{\beta})}, \\ \mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i, \end{cases}$$

with mean $\mu_i = [1 + \exp(-\mathbf{X}_i^T \boldsymbol{\beta})]^{-1}$ and variance $\text{Var}(Y_i \mid \mathbf{X}_i) = \mu_i(1 - \mu_i)$. Followed by Huang and Wang [29], the corrected score function is

$$\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta}; \mathbf{W}_i, Y_i) = \mathbf{W}_i Y_i + (\mathbf{W}_i + \boldsymbol{\Sigma}_u \boldsymbol{\beta}) \exp\left(-\mathbf{W}_i^T \boldsymbol{\beta} - \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{\Sigma}_u \boldsymbol{\beta}\right) Y_i - \mathbf{W}_i,$$

and its first-order derivative is

$$\Omega_i^*(\Sigma_u, \beta; W_i, Y_i) = \frac{\partial \eta_i^*(\Sigma_u, \beta; W_i, Y_i)}{\partial \beta^T} = [\Sigma_u - (W_i + \Sigma_u \beta)(W_i + \Sigma_u \beta)^T] \exp\left(-W_i^T \beta - \frac{1}{2} \beta^T \Sigma_u \beta\right) Y_i.$$

(2) Poisson measurement error regression models.

Let Y_i follow the Poisson distribution with mean μ_i , $\text{Var}(Y_i | X_i) = \mu_i$. Consider the log linear measurement error model

$$\begin{cases} \log(\mu_i) = X_i^T \beta, \\ W_i = X_i + U_i, \end{cases}$$

then we have the corrected score function

$$\eta_i^*(\Sigma_u, \beta; W_i, Y_i) = W_i Y_i - (W_i - \Sigma_u \beta) \exp\left(W_i^T \beta - \frac{1}{2} \beta^T \Sigma_u \beta\right),$$

and its first-order derivative is

$$\Omega_i^*(\Sigma_u, \beta; W_i, Y_i) = \frac{\partial \eta_i^*(\Sigma_u, \beta; W_i, Y_i)}{\partial \beta^T} = [\Sigma_u - (W_i - \Sigma_u \beta)(W_i - \Sigma_u \beta)^T] \exp\left(W_i^T \beta - \frac{1}{2} \beta^T \Sigma_u \beta\right).$$

2.1. General subsampling algorithm

It is assumed that π_i is the probability of sampling the i -th sample (W_i, Y_i) , $i = 1, \dots, N$. Let S be the set of the subsamples $(\widetilde{W}_i, \widetilde{Y}_i)$ with corresponding sampling probabilities $\widetilde{\pi}_i$, i.e., $S = \{(\widetilde{W}_i, \widetilde{Y}_i, \widetilde{\pi}_i)\}$ with the subsample size r . The general subsampling algorithm is shown in Algorithm 1.

Algorithm 1 General subsampling algorithm.

Step 1. Given the subsampling probabilities π_i , $i = 1, \dots, N$ of all data points.

Step 2. Perform repeated sampling with replacement r times to form the subsample set $S = \{(\widetilde{W}_i, \widetilde{Y}_i, \widetilde{\pi}_i)\}$, where \widetilde{W}_i , \widetilde{Y}_i and $\widetilde{\pi}_i$ represent the covariate, response variable and subsampling probability in the subsample, respectively.

Step 3. Based on the subsample set S , solve the weighted estimation equation $Q^*(\beta)$ to obtain $\widetilde{\beta}$, where

$$Q^*(\beta) := \frac{1}{r} \sum_{i=1}^r \frac{1}{\widetilde{\pi}_i} \widetilde{\eta}_i^*(\Sigma_u, \beta; \widetilde{W}_i, \widetilde{Y}_i) = \mathbf{0}, \quad (2.4)$$

where $\widetilde{\eta}_i^*(\Sigma_u, \beta; \widetilde{W}_i, \widetilde{Y}_i)$ is the unbiased score function of i -th sample point in the subsample and its first order derivative is $\widetilde{\Omega}_i^*(\Sigma_u, \beta; \widetilde{W}_i, \widetilde{Y}_i)$.

To obtain the consistency and asymptotic normality of $\widetilde{\beta}$, the following assumptions should be made. For simplicity, denote $\eta_i^*(\Sigma_u, \beta; W_i, Y_i)$ and $\Omega_i^*(\Sigma_u, \beta; W_i, Y_i)$ as $\eta_i^*(\Sigma_u, \beta)$ and $\Omega_i^*(\Sigma_u, \beta)$.

A1: It is assumed that $W_i^T \beta$ is almost necessarily in the interior of a closed set $K \in \Theta$, Θ is a natural parameter space.

A2: The regression parameters are located in the ball $\Lambda = \{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq B\}$, β_t and $\hat{\beta}_{\text{MLE}}$ are true parameters and maximum likelihood estimators, which are interior points of Λ , and B is a constant, where $\|\cdot\|_1$ denotes ℓ_1 -norm.

A3: As $n \rightarrow \infty$, the observed information matrix $\mathbf{M}_X := \frac{1}{N} \sum_{i=1}^N \boldsymbol{\Omega}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})$ is a positive definite matrix in probability.

A4: Assume that for all $\boldsymbol{\beta} \in \Lambda$, $\frac{1}{N} \sum_{i=1}^N \|\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta})\|^4 = O_P(1)$, where $\|\cdot\|$ denotes the Euclidean norm.

A5: Suppose that the full sample covariates have finite 6th-order moments, i.e., $E\|\mathbf{W}_1\|^6 \leq \infty$.

A6: For any $\delta \geq 0$, we assume that

$$\frac{1}{N^{2+\delta}} \sum_{i=1}^N \frac{\|\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})\|^{2+\delta}}{\pi_i^{1+\delta}} = O_P(1), \quad \frac{1}{N^{2+\delta}} \sum_{i=1}^N \frac{|\boldsymbol{\Omega}_i^{*(j_1 j_2)}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})|^{2+\delta}}{\pi_i^{1+\delta}} = O_P(1),$$

where $\boldsymbol{\Omega}_i^{*(j_1 j_2)}$ represents the elements of the j_1 -th row and j_2 -th column of the matrix $\boldsymbol{\Omega}_i^*$.

A7: Assume that $\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta})$ and $\boldsymbol{\Omega}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta})$ are $m(\mathbf{W}_i)$ -Lipschitz continuous. For any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \Lambda$, there exist functions $m_1(\mathbf{W}_i)$ and $m_2(\mathbf{W}_i)$ such that $\|\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta}_1) - \boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta}_2)\| \leq m_1(\mathbf{W}_i) \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|$, $\|\boldsymbol{\Omega}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta}_1) - \boldsymbol{\Omega}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta}_2)\|_S \leq m_2(\mathbf{W}_i) \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|$, where $\|\mathbf{A}\|_S$ denotes the spectral norm of matrix \mathbf{A} . Further assume that $E\{m_1(\mathbf{W}_i)\} \leq \infty$ and $E\{m_2(\mathbf{W}_i)\} \leq \infty$.

Assumptions A1 and A2 are also used in Cléménçon et al. [30]. The set Λ in Assumption A2 is also known as the admissible set and is a prerequisite for consistency estimation for the GLM with full data [31]. Assumption A3 imposes a condition on the covariates to ensure that the MLE based on the full dataset is consistent. In order to obtain the Bahadur representation of the subsampling estimators, Assumptions A4 and A5 are required. Assumption A6 is a moment condition for the subsampling probability and is also required for the Lindberg-Feller central limit theorem. Assumption A7 adds a restriction on smoothing, which can be found in [32].

The following theorems show the consistency and asymptotic normality of the subsampling estimators.

Theorem 2.1. *If Assumptions A1–A7 hold, as $r \rightarrow \infty$ and $N \rightarrow \infty$, $\tilde{\boldsymbol{\beta}}$ converges to $\hat{\boldsymbol{\beta}}_{MLE}$ in conditional probability given \mathcal{F}_N , and the convergence rate is $r^{\frac{1}{2}}$. That is, for all $\varepsilon > 0$, there exist constants Δ_ε and r_ε such that*

$$P\left(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}\| \geq r^{-\frac{1}{2}} \Delta_\varepsilon \mid \mathcal{F}_N\right) < \varepsilon, \quad (2.5)$$

for all $r > r_\varepsilon$.

Theorem 2.2. *If Assumptions A1–A7 hold, as $r \rightarrow \infty$ and $N \rightarrow \infty$, conditional on \mathcal{F}_N , the estimator $\tilde{\boldsymbol{\beta}}$ obtained from Algorithm 1 satisfies*

$$\mathbf{V}^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}), \quad (2.6)$$

where $\mathbf{V} = \mathbf{M}_X^{-1} \mathbf{V}_C \mathbf{M}_X^{-1} = O_P(r^{-1})$, and

$$\mathbf{V}_C = \frac{1}{N^2 r} \sum_{i=1}^N \frac{\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE}) \boldsymbol{\eta}_i^{*\top}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})}{\pi_i}.$$

Remark 1. *In order to get the standard error of the corresponding estimator, we estimate the variance-covariance matrix of $\tilde{\boldsymbol{\beta}}$ by*

$$\widehat{\mathbf{V}} = \widehat{\mathbf{M}}_X^{-1} \widehat{\mathbf{V}}_C \widehat{\mathbf{M}}_X^{-1},$$

where

$$\widehat{\mathbf{M}}_X = \frac{1}{Nr} \sum_{i=1}^r \frac{\widetilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})}{\widetilde{\pi}_i},$$

$$\widehat{\mathbf{V}}_C = \frac{1}{N^2 r^2} \sum_{i=1}^r \frac{\widetilde{\boldsymbol{\eta}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE}) \widetilde{\boldsymbol{\eta}}_i^{*\top}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})}{\widetilde{\pi}_i^2}.$$

Based on the A-optimality criteria in the optimal design language, the optimal subsampling probabilities are obtained by minimizing the asymptotic mean square error of $\boldsymbol{\beta}$ in Theorem 2.2.

However, $\boldsymbol{\Sigma}_u$ is usually unknown in practice. Therefore, we need to estimate the covariance matrix $\boldsymbol{\Sigma}_u$ as suggested by [12]. We observe that the consistent, unbiased moment estimator of $\boldsymbol{\Sigma}_u$ is

$$\widehat{\boldsymbol{\Sigma}}_u = \frac{\sum_{i=1}^N \sum_{j=1}^{m_i} (\mathbf{W}_{ij} - \overline{\mathbf{W}}_i)(\mathbf{W}_{ij} - \overline{\mathbf{W}}_i)^\top}{\sum_{i=1}^N (m_i - 1)},$$

where $\overline{\mathbf{W}}_i$ is the sample mean of the replicates, and m_i is the number of repeated measurements of the i -th individual.

Theorem 2.3. Define $g_i^{mV} = \|\mathbf{M}_X^{-1} \boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})\|$, $i = 1, \dots, N$. The subsampling strategy is mV -optimal if the subsampling probability is chosen such that

$$\pi_i^{mV} = \frac{g_i^{mV}}{\sum_{j=1}^N g_j^{mV}}, \quad (2.7)$$

which is obtained by minimizing $\text{tr}(\mathbf{V})$.

Theorem 2.4. Define $g_i^{mVc} = \|\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})\|$, $i = 1, \dots, N$. The subsampling strategy is mVc -optimal if the subsampling probability is chosen such that

$$\pi_i^{mVc} = \frac{g_i^{mVc}}{\sum_{j=1}^N g_j^{mVc}}, \quad (2.8)$$

which is obtained by minimizing $\text{tr}(\mathbf{V}_C)$.

Remark 2. \mathbf{M}_X and \mathbf{V}_C are non-negative definite matrices, and $\mathbf{V} = \mathbf{M}_X^{-1} \mathbf{V}_C \mathbf{M}_X^{-1}$, then $\text{tr}(\mathbf{V}) = \text{tr}(\mathbf{M}_X^{-1} \mathbf{V}_C \mathbf{M}_X^{-1}) \leq \sigma_{\max}(\mathbf{M}_X^{-2}) \text{tr}(\mathbf{V}_C)$, where $\sigma_{\max}(\mathbf{A})$ represents the maximum eigenvalue of square matrix \mathbf{A} . As $\sigma_{\max}(\mathbf{M}_X^{-2})$ does not depend on $\boldsymbol{\pi}$, minimizing $\text{tr}(\mathbf{V}_C)$ means minimizing the upper bound of $\text{tr}(\mathbf{V})$. In fact, for two given subsampling probabilities $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$, $\mathbf{V}(\boldsymbol{\pi}_1) \leq \mathbf{V}(\boldsymbol{\pi}_2)$ if and only if $\mathbf{V}_C(\boldsymbol{\pi}_1) \leq \mathbf{V}_C(\boldsymbol{\pi}_2)$. Therefore, minimizing $\text{tr}(\mathbf{V}_C)$ reduces considerable computational time compared to minimizing $\text{tr}(\mathbf{V})$, and $\text{tr}(\mathbf{V}_C)$ does not take into account the structural information of the data.

2.2. Optimal subsampling algorithm

The optimal subsampling probabilities are defined as $\{\pi_i^{\text{op}}\}_{i=1}^N = \{\pi_i^{\text{mV}}\}_{i=1}^N$ or $\{\pi_i^{\text{mVc}}\}_{i=1}^N$. However, because π_i^{op} depends on $\hat{\beta}_{\text{MLE}}$, it cannot be used directly in applications. To calculate π_i^{op} , it is necessary to use a prior estimator $\tilde{\beta}_0$, which is obtained by the prior subsample of size r_0 .

We know π_i^{op} is proportional to $\|\eta_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}})\|$, however, in actual situations, there may be some data points that make $\eta_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}}) = \mathbf{0}$, which will never be included in a subsample, and some data points with $\eta_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}}) \approx \mathbf{0}$ also have small probabilities of being sampled. If these special data points are excluded, some sample information will be missed, but if these data points are included, the variance of the subsampling estimator may increase.

To avoid Eq (2.4) from being inflated by these special data points, this paper adopts a truncation method, setting a threshold ω for $\|\eta_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}})\|$, that is, replacing $\|\eta_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}})\|$ with $\max\{\|\eta_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}})\|, \omega\}$, where ω is a very small positive number, for example, 10^{-4} . In applications, the choice and design of the truncation weight function, which is a commonly used technique, are crucial to improving the robustness of the model and optimizing the performance.

We replace $\hat{\beta}_{\text{MLE}}$ in the matrix \mathbf{V} with $\tilde{\beta}_0$, denoted as $\tilde{\mathbf{V}}$, then $\text{tr}(\tilde{\mathbf{V}}) \leq \text{tr}(\tilde{\mathbf{V}}^\omega) \leq \text{tr}(\tilde{\mathbf{V}}) + \frac{\omega^2}{N^2 r} \sum_{i=1}^N \frac{\|\mathbf{M}_X^{-1}\|^2}{\pi_i^{\text{op}}}$. Therefore, when ω is sufficiently small, $\text{tr}(\tilde{\mathbf{V}}^\omega)$ approaches $\text{tr}(\tilde{\mathbf{V}})$. The threshold ω is set to make the subsample estimators more robust without sacrificing excessively estimation efficiency. $\tilde{\mathbf{M}}_X = \frac{1}{Nr_0} \sum_{i=1}^{r_0} \mathbf{\Omega}_i^*(\Sigma_u, \tilde{\beta}_0)$ based on the prior subsample can be used to approximate \mathbf{M}_X . The two-step algorithm is presented in Algorithm 2.

Algorithm 2 Optimal subsampling algorithm.

Step 1. Extract a prior subsample set S_{r_0} with a subsample size of r_0 from the full data, assuming that the subsampling probabilities of the prior subsample are $\pi^{\text{UNIF}} = \{\pi_i := \frac{1}{N}\}_{i=1}^N$. We use Algorithm 1 to obtain a prior estimator $\tilde{\beta}_0$, replace $\hat{\beta}_{\text{MLE}}$ with $\tilde{\beta}_0$ in Eqs (2.7) and (2.8) to get the optimal subsampling probabilities $\{\pi_i^{\text{opt}}\}_{i=1}^N$.

Step 2. Use the optimal subsample probabilities $\{\pi_i^{\text{opt}}\}_{i=1}^N$ computed in Step 1 to extract a subsample size of r with replacement. According to the step in Algorithm 1, combining the subsamples from Step 1 and solving the estimating Eq (2.4) to get the estimator $\check{\beta}$ based on a subsample of total size $r_0 + r$.

Remark 3. In Algorithm 2, $\tilde{\beta}_0$ in Step 1 satisfies

$$\mathbf{Q}_{\tilde{\beta}_0}^{*0}(\beta) = \frac{1}{r_0} \sum_{i=1}^{r_0} \frac{\tilde{\eta}_i^*(\Sigma_u, \beta)}{\pi_i^{\text{UNIF}}} = \mathbf{0}$$

with the prior subsample set S_{r_0} , and

$$\mathbf{M}_X^{\tilde{\beta}_0} = \frac{1}{Nr_0} \sum_{i=1}^{r_0} \frac{\tilde{\mathbf{\Omega}}_i^*(\Sigma_u, \tilde{\beta}_0)}{\pi_i^{\text{UNIF}}}.$$

In Step 2, the subsampling probabilities are $\{\pi_i^{\text{opt}}\}_{i=1}^N = \{\pi_i^{\text{mVt}}\}_{i=1}^N$ or $\{\pi_i^{\text{mVct}}\}_{i=1}^N$, let

$$g_i^{mVt} = \begin{cases} \|\mathbf{M}_X^{-1} \boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})\|, & \text{if } \|\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})\| > \omega \\ \omega \|\mathbf{M}_X^{-1}\|, & \text{if } \|\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})\| < \omega \end{cases}, \quad i = 1, \dots, N,$$

$$g_i^{mVct} = \max\{\|\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})\|, \omega\},$$

then

$$\pi_i^{mVt} = \frac{g_i^{mVt}}{\sum_{j=1}^N g_j^{mVt}} \quad \text{and} \quad \pi_i^{mVct} = \frac{g_i^{mVct}}{\sum_{j=1}^N g_j^{mVct}}.$$

The subsample set is $S_{r_0} \cup \{(\tilde{W}_i, \tilde{Y}_i, \tilde{\pi}_i^{opt}) \mid i = 1, \dots, r\}$ with a subsample size of $r + r_0$, and $\check{\boldsymbol{\beta}}$ is the solution to the corresponding estimating equation

$$\mathbf{Q}_{\check{\boldsymbol{\beta}_0}^{two-step}}(\boldsymbol{\beta}) = \frac{1}{r + r_0} \sum_{i=1}^{r+r_0} \frac{\tilde{\boldsymbol{\eta}}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta})}{\tilde{\pi}_i^{opt}} = \frac{r}{r + r_0} \mathbf{Q}_{\check{\boldsymbol{\beta}_0}^*}(\boldsymbol{\beta}) + \frac{r_0}{r + r_0} \mathbf{Q}_{\check{\boldsymbol{\beta}_0}^{*0}}(\boldsymbol{\beta}) = \mathbf{0},$$

where

$$\mathbf{Q}_{\check{\boldsymbol{\beta}_0}^*}(\boldsymbol{\beta}) = \frac{1}{r} \sum_{i=1}^r \frac{\tilde{\boldsymbol{\eta}}_i^*(\boldsymbol{\Sigma}_u, \boldsymbol{\beta})}{\tilde{\pi}_i^{opt}}.$$

Theorem 2.5. If Assumptions A1–A7 hold, as $r_0 r^{-1} \rightarrow 0$, $r_0 \rightarrow \infty$, $r \rightarrow \infty$ and $N \rightarrow \infty$, if $\check{\boldsymbol{\beta}_0}$ exists, then the estimator $\check{\boldsymbol{\beta}}$ obtained from Algorithm 2 converges to $\hat{\boldsymbol{\beta}}_{MLE}$ in conditional probability given \mathcal{F}_N , and its convergence rate is $r^{\frac{1}{2}}$. For all $\varepsilon > 0$, there exist finite Δ_ε and r_ε such that

$$P(\|\check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}\| \geq r^{-\frac{1}{2}} \Delta_\varepsilon \mid \mathcal{F}_N) < \varepsilon, \quad (2.9)$$

for all $r > r_\varepsilon$.

Theorem 2.6. If Assumptions A1–A7 hold, as $r_0 r^{-1} \rightarrow 0$, $r_0 \rightarrow \infty$, $r \rightarrow \infty$ and $N \rightarrow \infty$, conditional on \mathcal{F}_N , the estimator $\check{\boldsymbol{\beta}}$ obtained from Algorithm 2 satisfies

$$\mathbf{V}_{opt}^{-\frac{1}{2}}(\check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}), \quad (2.10)$$

where $\mathbf{V}_{opt} = \mathbf{M}_X^{-1} \mathbf{V}_C^{opt} \mathbf{M}_X^{-1} = O_P(r^{-1})$, and

$$\mathbf{V}_C^{opt} = \frac{1}{N^2 r} \sum_{i=1}^N \frac{\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE}) \boldsymbol{\eta}_i^{*\top}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})}{\pi_i^{opt}}.$$

Remark 4. We estimate the variance-covariance matrix of $\check{\boldsymbol{\beta}}$ by

$$\widehat{\mathbf{V}}_{opt} = \widehat{\mathbf{M}}_X^{-1} \widehat{\mathbf{V}}_C^{opt} \widehat{\mathbf{M}}_X^{-1},$$

where

$$\widehat{\mathbf{M}}_X = \frac{1}{N(r_0 + r)} \left[\sum_{i=1}^{r_0} \frac{\tilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})}{\tilde{\pi}_i^{UNIF}} + \sum_{i=1}^r \frac{\tilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})}{\tilde{\pi}_i^{opt}} \right],$$

$$\widehat{\mathbf{V}}_C^{opt} = \frac{1}{N^2(r_0 + r)^2} \left[\sum_{i=1}^{r_0} \frac{\tilde{\boldsymbol{\eta}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE}) \tilde{\boldsymbol{\eta}}_i^{*\top}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})}{\tilde{\pi}_i^{UNIF^2}} + \sum_{i=1}^r \frac{\tilde{\boldsymbol{\eta}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE}) \tilde{\boldsymbol{\eta}}_i^{*\top}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{MLE})}{\tilde{\pi}_i^{opt^2}} \right].$$

3. Simulation studies

In this section, we perform numerical simulations using synthetic data to evaluate the finite sample performance of the proposed method in Algorithm 2 (denoted as mV and mVc). For a fair comparison, we also give the results of the uniform subsampling method and set the size to be the same as that of Algorithm 2. The estimators of the above three subsampling methods, uniform—the uniform subsampling, mV—the mV probability subsampling, and mVc—the mVc probability subsampling, are compared with MLE—the maximum likelihood estimators for full data. In addition, we conduct simulation experiments using two models: the logistic regression model and the Poisson regression model.

3.1. Binary logistic measurement error regression model

Set the sample size $N = 100000$, the true value $\beta_t = (0.5, -0.6, 0.5)^T$, the covariate $X_i \sim N_3(\mathbf{0}, \Sigma)$, where $\Sigma = 0.5\mathbf{I} + 0.5\mathbf{1}\mathbf{1}^T$, \mathbf{I} is an identity matrix. The response Y_i follows a binomial distribution with $P(Y_i = 1 | X_i) = \left(1 + \exp(-X_i^T \beta_t)\right)^{-1}$. We consider the following three cases to generate the measurement error term U_i .

- Case 1: $U_i \sim N_3(\mathbf{0}, 0.4^2\mathbf{I})$;
- Case 2: $U_i \sim N_3(\mathbf{0}, 0.5^2\mathbf{I})$;
- Case 3: $U_i \sim N_3(\mathbf{0}, 0.6^2\mathbf{I})$.

The subsample size in Step 1 of Algorithm 2 is selected as $r_0 = 400$. The subsample size r is set to be 500, 1000, 1500, 2000, 2500, and 5000. In order to verify that $\check{\beta}$ can asymptotically approach β_t , we repeat $K = 1000$ and calculate $MSE = \frac{1}{K} \sum_{k=1}^K \|\check{\beta}^{(k)} - \beta_t\|^2$, where $\check{\beta}^{(k)}$ is the parameter estimator of the subsample generated by the k -th repetition.

The simulation results are shown in Figure 1, which can be seen that both mV and mVc always have smaller MSEs than uniform subsampling. The MSEs of all the subsampling methods decrease as an increase of r , which confirms the theoretical results of the consistency of the subsampling methods. As the variance of the error term increases, the MSEs of uniform, mV, and mVc also increase. The mV is better than the mVc because the subsampling probabilities of mV take the structural information of the data into account. A comparison between the corrected and uncorrected methods shows that the MSEs of the corrected methods are much smaller than those of the uncorrected methods, and the difference between the corrected and uncorrected methods increases as the error variance increases.

Now, we evaluate the statistical inference performance of the optimal subsampling method for different r and variances of U_i . The parameter β_1 is taken as an example, and a 95% confidence interval is constructed. Table 1 reports the empirical coverage probabilities and average lengths of three subsampling methods. It is evident that both mV and mVc have similar performance and consistently outperform the uniform subsampling method. As r increases, the length of the confidence interval uniformly decreases.

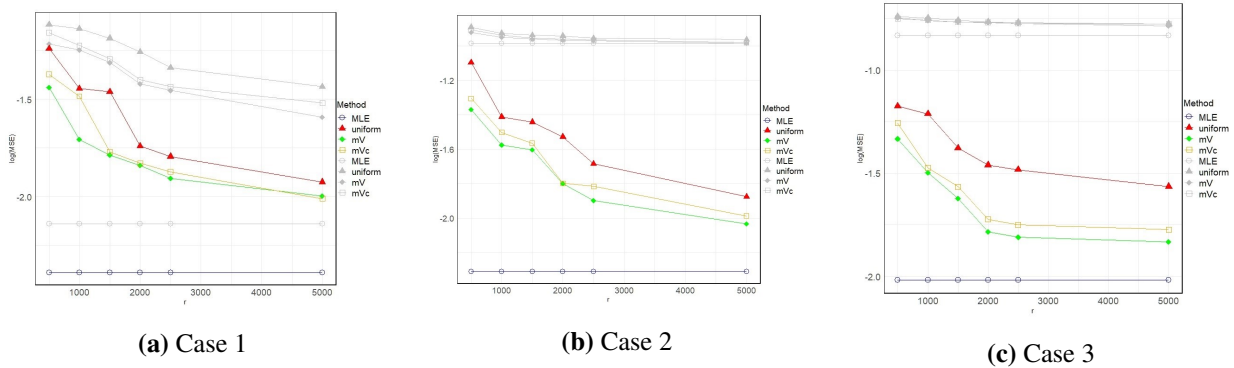


Figure 1. MSEs for $\check{\beta}$ with different second step subsample size r and $r_0 = 400$. The colorful icons and lines represent the corrected subsampling methods. The gray icons and lines represent the uncorrected subsampling methods.

Table 1. Empirical coverage probabilities and average lengths of confidence intervals for β_1 in the logistic regression models with different r and $r_0 = 500$.

Case	r	uniform		mV		mVc	
		Coverage	Length	Coverage	Length	Coverage	Length
Case 1	500	0.958	0.565	0.932	0.331	0.942	0.457
	1000	0.952	0.453	0.925	0.248	0.954	0.333
	1500	0.960	0.387	0.920	0.206	0.964	0.274
	2000	0.932	0.345	0.907	0.180	0.954	0.237
	2500	0.938	0.313	0.910	0.160	0.956	0.211
	5000	0.964	0.302	0.908	0.148	0.937	0.202
Case 2	500	0.956	0.634	0.946	0.602	0.962	0.613
	1000	0.946	0.621	0.934	0.586	0.946	0.593
	1500	0.927	0.597	0.954	0.551	0.962	0.561
	2000	0.943	0.543	0.956	0.524	0.921	0.518
	2500	0.970	0.475	0.958	0.453	0.944	0.462
	5000	0.963	0.438	0.932	0.417	0.947	0.441
Case 3	500	0.958	0.706	0.956	0.432	0.968	0.550
	1000	0.946	0.561	0.972	0.399	0.970	0.409
	1500	0.944	0.479	0.968	0.321	0.960	0.329
	2000	0.936	0.425	0.964	0.265	0.958	0.281
	2500	0.926	0.389	0.966	0.249	0.954	0.250
	5000	0.915	0.356	0.947	0.220	0.942	0.236

3.2. Poisson measurement error regression model

Let $\beta_t = (0.5, -0.6, 0.5)^T$, the covariate $X_i \sim N_3(\mathbf{0}, \Sigma)$, where $\Sigma = 0.3\mathbf{I} + 0.511\mathbf{1}\mathbf{1}^T$, \mathbf{I} is an identity matrix. We consider the following three cases to generate the measurement error term U_i .

- Case 1: $U_i \sim N_3(\mathbf{0}, 0.3^2\mathbf{I})$;
- Case 2: $U_i \sim N_3(\mathbf{0}, 0.4^2\mathbf{I})$;
- Case 3: $U_i \sim N_3(\mathbf{0}, 0.5^2\mathbf{I})$.

We also generate a sample of $N = 100000$ following $Poisson(\mu_i)$, where $\mu_i = \exp(X_i^T \beta_t)$. distribution, and summarize the MSEs with the number of simulations $K = 1000$ in Figure 2. The

other settings are the same as those in the logistic regression example.

In Figure 2, it can be seen that the MSEs of both the mV and mVc methods are smaller than those of the uniform subsampling, with the mV method being the optimal. In addition, the corrected method is obviously effective, which is consistent with Figure 1. Table 2 reports the empirical coverage probabilities and average lengths of 95% confidence interval of the parameter β_3 for three subsampling methods. The conclusions of Table 2 are consistent with those of Table 1, but the average lengths of the intervals for Poisson regression are significantly longer than those for logistic regression.

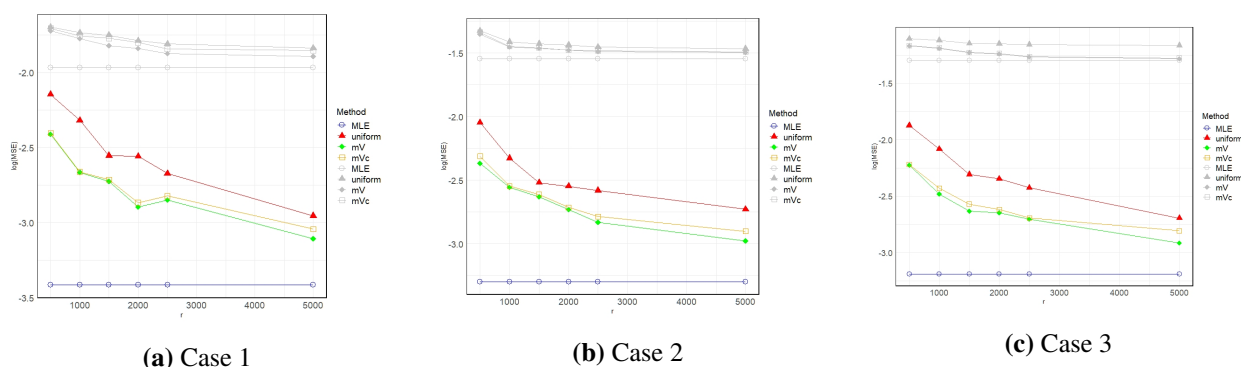


Figure 2. MSEs for β with different second step subsample size r and $r_0 = 400$. The colorful icons and lines represent the corrected subsampling methods. The gray icons and lines represent the uncorrected subsampling methods.

Table 2. Empirical coverage probabilities and average lengths of confidence intervals for β_3 in the Poisson regression models with different r and $r_0 = 500$.

Case	r	uniform		mV		mVc	
		Coverage	Length	Coverage	Length	Coverage	Length
Case 1	500	0.962	0.441	0.962	0.383	0.958	0.399
	1000	0.944	0.352	0.964	0.291	0.964	0.304
	1500	0.932	0.302	0.964	0.241	0.966	0.255
	2000	0.952	0.268	0.930	0.210	0.944	0.223
	2500	0.946	0.244	0.958	0.188	0.974	0.201
	5000	0.952	0.234	0.961	0.173	0.943	0.185
Case 2	500	0.938	0.127	0.936	0.108	0.948	0.109
	1000	0.936	0.102	0.946	0.082	0.934	0.082
	1500	0.942	0.087	0.934	0.069	0.936	0.068
	2000	0.952	0.078	0.956	0.060	0.952	0.059
	2500	0.946	0.071	0.932	0.053	0.944	0.053
	5000	0.935	0.068	0.965	0.045	0.971	0.047
Case 3	500	0.940	0.185	0.936	0.153	0.953	0.156
	1000	0.950	0.148	0.954	0.113	0.958	0.118
	1500	0.932	0.127	0.950	0.094	0.958	0.099
	2000	0.946	0.113	0.952	0.082	0.960	0.086
	2500	0.942	0.103	0.932	0.073	0.950	0.077
	5000	0.937	0.096	0.956	0.065	0.964	0.061

In order to explore the influence of different subsample size allocated in the two-step algorithm,

we calculate the MSEs for r_0 at different proportions under the condition that the total subsample size remains constant. Set the total subsample size $r_0 + r = 3000$, and the result is shown in Figure 3. It can be seen that the accuracy of the two-step algorithm will initially improve with the increase of r_0 . However, when r_0 increases to a certain extent, the accuracy of the algorithm begins to decrease. There are two reasons: (1) if r_0 is too small, the estimators in the first step will be biased, and it is difficult to ensure the accuracy; (2) if r_0 are too large, then the performances of mV and mVc are similar to that of the uniform subsampling. When $r_0/(r_0 + r)$ is around 0.25, the two-step algorithm performs the best.

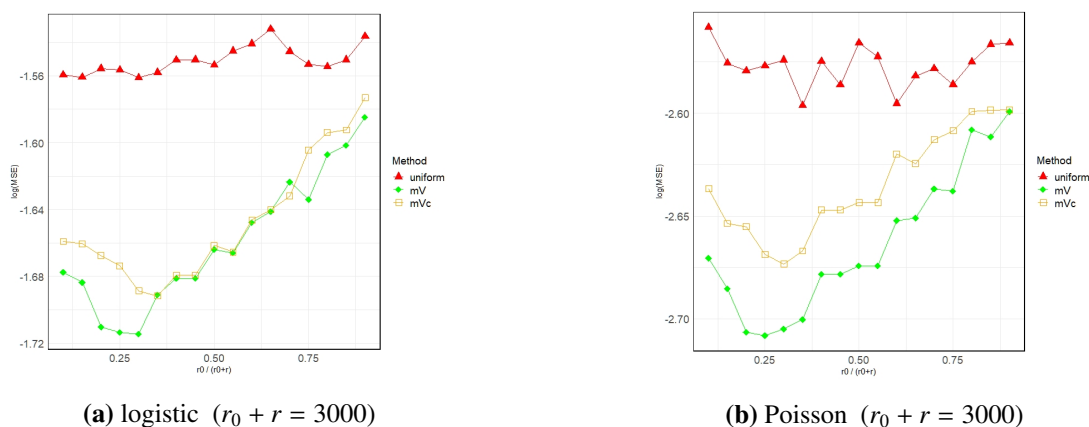


Figure 3. MSEs vs proportions of the first step subsample with fixed total subsample size for logistic and Poisson models with Case 1.

We use the Sys.time() function in R to calculate the running time of three subsampling methods and full data. We conduct 1000 repetitions, set $r_0 = 200$, and consider different r values in Case 1. The results are shown in Tables 3 and 4. It is easy to find that the uniform subsampling algorithm requires the least computation time. Because there is no need to calculate the subsampling probabilities. In addition, the mV method takes longer than the mVc method, and this result is consistent with the theoretical analysis in Section 2.

Table 3. Computing time (in seconds) for logistic regression with Case 1 for different r and fixed $r_0 = 200$.

Method	r					
	300	500	800	1200	1600	2000
uniform	0.2993	0.3337	0.4985	0.5632	0.8547	0.5083
mV	3.5461	3.6485	3.8623	4.1256	4.4325	5.2365
mVc	3.2852	3.3658	3.5463	3.8562	4.0235	4.4235
Full	45.9075					

Table 4. Computing time (in seconds) for Poisson regression with Case 1 for different r and fixed $r_0 = 200$.

Method	r					
	300	500	800	1200	1600	2000
uniform	0.4213	0.4868	0.5327	0.5932	0.7147	0.8883
mV	4.6723	4.8963	5.2369	5.6524	6.0128	6.3567
mVc	4.3521	4.6329	4.9658	5.2156	5.7652	5.9635
Full	51.2603					

4. Real data

4.1. Global census dataset

In this section, we apply the proposed method to analyze the 1994 global census data, which contains 42 countries, from the Machine Learning Database [33]. There are 5 covariates in the data: x_1 represents age; x_2 represents the population weight value, which is assigned by the Population Division of the Census Bureau and is related to socioeconomic characteristics; x_3 represents the highest level of education, that is, the highest level of education since primary school; x_4 represents capital loss, which refers to the loss of income from bad investment, which is the difference between the lower selling price and the higher purchase price of an individual's investment; x_5 represents weekly working hours. If an individual's annual income exceeds 50,000 dollars, it is expressed as $y_i = 1$ and $y_i = 0$ otherwise.

To verify the effectiveness of the proposed method, we add the measurement errors to the covariates x_2 , x_4 and x_5 in this dataset, and the covariance matrix of the measurement error is

$$\Sigma_u = \begin{bmatrix} 0 & & & & \\ & 0.04 & & & \\ & & 0 & & \\ & & & 0.04 & \\ & & & & 0.04 \end{bmatrix}.$$

We split the full dataset into a training set of 32561 observations and a test set of 16281 observations in a 2:1 ratio. We apply the proposed method to the training set and evaluate the classification performance with the test set. We calculate $LEMSE = \log\left(\frac{1}{K} \sum_{k=1}^K \|\check{\beta}^{(k)} - \hat{\beta}_{MLE}\|^2\right)$ based on 1000 bootstrap subsample estimators with $r = 500, 1000, 1500, 2200, 2500$, and $r_0 = 500$. The corrected MLE estimators for the training set are $\hat{\beta}_{MLE,0}^{err} = -1.6121$, $\hat{\beta}_{MLE,1}^{err} = 1.1992$, $\hat{\beta}_{MLE,2}^{err} = 0.0103$, $\hat{\beta}_{MLE,3}^{err} = 0.9142$, $\hat{\beta}_{MLE,4}^{err} = 0.2617$, $\hat{\beta}_{MLE,5}^{err} = 0.8694$.

Table 5 shows the average estimators and the corresponding standard errors based on the proposed method ($r_0 = 500$, $r = 2000$). It can be seen that the estimators from three subsampling methods are close to the estimators from the full data. In general, the mV and mVc subsampling methods produce small standard errors.

Table 5. Average estimators based on subsamples with measurement error and subsample size $r = 2000$. The numbers in parentheses are the standard errors of the average estimators.

	uniform	mV	mVc
Intercept	-1.6084(0.069)	-1.5998(0.055)	-1.3122(0.052)
$\check{\beta}_1^{err}$	1.2879(0.205)	1.1880(0.103)	1.2038(0.097)
$\check{\beta}_2^{err}$	0.0105(0.106)	0.0104(0.059)	0.0111(0.046)
$\check{\beta}_3^{err}$	1.0033(0.201)	0.9217(0.067)	0.9199(0.054)
$\check{\beta}_4^{err}$	0.2636(0.094)	0.2698(0.054)	0.2555(0.063)
$\check{\beta}_5^{err}$	0.9469(0.229)	0.8741(0.083)	0.8628(0.076)

All subsampling methods show that each variable has a positive impact on income, with age, highest education level, and weekly working hours having significant impacts on income. Interestingly, capital

losses have a significant positive impact on income because low-income people rarely invest. However, the population weight value has the smallest impact on income, the reason should be more inclined to reflect the overall distribution characteristics among groups rather than the specific economic performance of individuals. Income is a highly volatile variable, and the income gap between different groups may be large. Even under the same socioeconomic characteristics, the income distribution may have a large variance. This high variability weakens the overall impact of the population weight on income.

Fix $r_0 = 500$, Figure 4(a) shows the LEMSEs calculated for the subsample with measurement errors. We can see that the LEMSEs of the corrected methods are much smaller than those of the uncorrected methods. As r increases, the LEMSEs become increasingly small. The estimators of the subsampling methods are consistent and the mV method is the best. Figure 4(b) shows the proportion of responses in the test set being correctly classified for different subsample sizes. The mV performs slightly better than the mVc. It can also be seen that the prediction accuracy of the corrected subsampling methods is slightly greater compared with the correspondingly uncorrected methods.

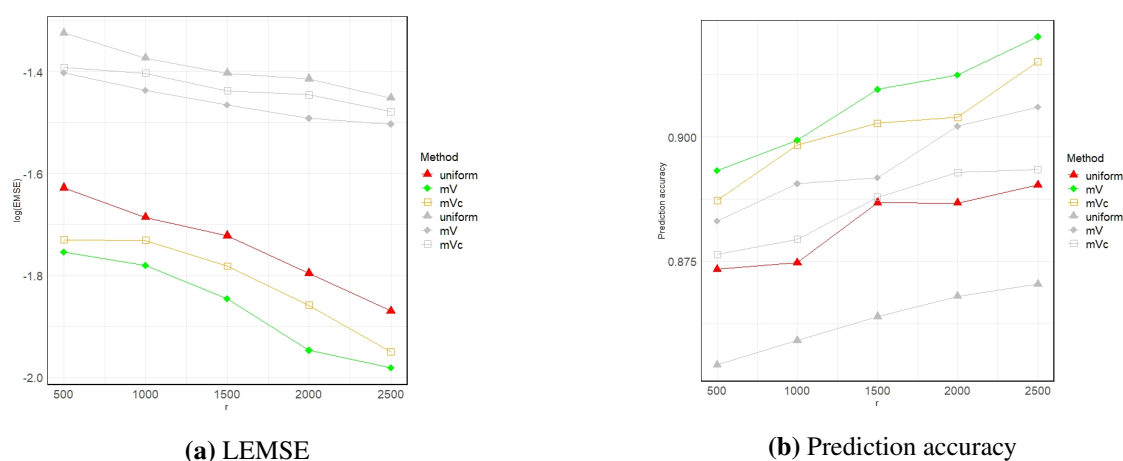


Figure 4. LEMSEs and model prediction accuracy (proportion of correctly classified models) for the subsample with measurement errors. The colorful icons and lines represent the corrected subsampling methods. The gray icons and lines represent the uncorrected subsampling methods.

4.2. Creditcard fraud dataset

This subsection applies the corrected subsampling method to creditcard fraud detection dataset from Kaggle *, and the dependent variable is whether an individual has committed creditcard fraud. There are 284,807 pieces of data in the dataset, with a total of 492 fraud cases. Since the data involves sensitive information, the covariates have all been processed by principal component analysis with a total of 28 principal components. Amount represents the consumption amount, class is the dependent variable, 1 represents fraud, and 0 means normal. The first four principal components and the consumption amount are selected as independent variables.

To verify the effectiveness of the proposed method, we add the measurement errors to the covariates, and the covariance matrix of the measurement error is $\Sigma_u = 0.16\mathbf{I}$. We split the dataset into the training

*<https://www.kaggle.com/datasets/creepycrap/creditcard-fraud-dataset>

set and the test set in a 3:1 ratio and summarize the LEMSEs based on the number of simulations $K = 1000$ with $r = 500, 1000, 1500, 2200, 2500, 5000$, and $r_0 = 500$.

The MLE estimators for the training set are $\hat{\beta}_{MLE,0}^{err} = -8.8016, \hat{\beta}_{MLE,1}^{err} = -0.6070, \hat{\beta}_{MLE,2}^{err} = 0.0737, \hat{\beta}_{MLE,3}^{err} = -0.9056, \hat{\beta}_{MLE,4}^{err} = 1.4553, \hat{\beta}_{MLE,5}^{err} = -0.1329$. Table 6 shows the average estimators and the corresponding standard errors ($r_0 = 500, r = 2000$). It can be seen that the estimators from three subsampling methods are close to the estimators from the full data. In general, the mV and mVc subsampling methods produce small standard errors. From Figure 5, we can obtain similar results as in Figure 4.

Table 6. Average estimators based on subsamples with measurement error and subsample size $r = 2000$. The numbers in parentheses are the standard errors of the average estimators.

	uniform	mV	mVc
Intercept	-8.7934(0.0678)	-8.8105(0.0562)	-8.8135(0.0543)
$\check{\beta}_1^{err}$	-0.6123(0.341)	-0.6047(0.142)	-0.6035(0.105)
$\check{\beta}_2^{err}$	0.0712(0.125)	0.0730(0.064)	0.0798(0.088)
$\check{\beta}_3^{err}$	-0.9321(0.245)	-0.9087(0.067)	-0.9123(0.057)
$\check{\beta}_4^{err}$	1.4618(0.198)	1.4580(0.054)	1.4603(0.075)
$\check{\beta}_5^{err}$	-0.1435(0.531)	-0.1347(0.242)	-0.1408(0.225)

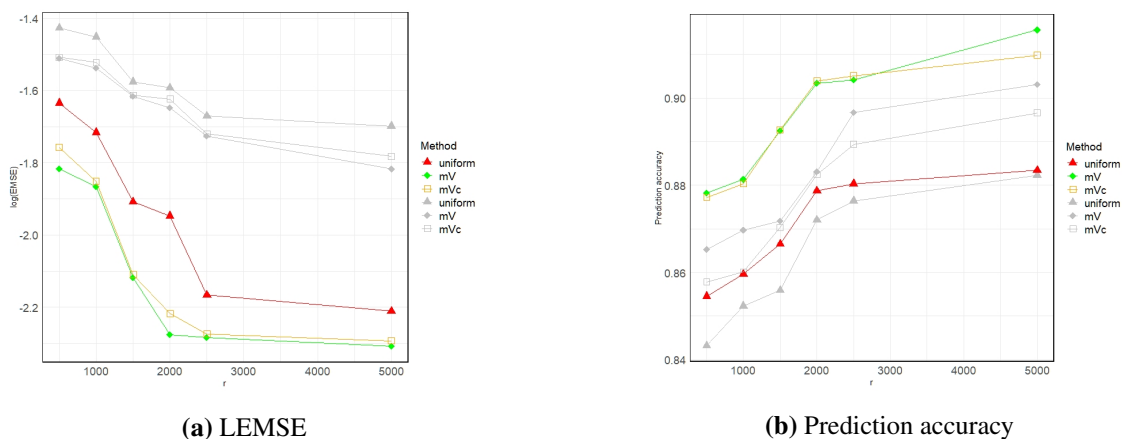


Figure 5. LEMSEs and model prediction accuracy (proportion of correctly classified models) for the subsample with measurement errors. The colorful icons and lines represent the corrected subsampling methods. The gray icons and lines represent the uncorrected subsampling methods.

5. Conclusions

In this paper, we not only combine the corrected score method with the subsampling technique, but also theoretically derive the consistency and asymptotic normality of the subsampling estimators. In addition, an adaptive two-step algorithm is developed based on optimal subsampling probabilities using A-optimality and L-optimality criteria and the truncation method. The theoretical results of the proposed method are tested with simulated and two real datasets, and the experimental results

demonstrate the effectiveness and good performance of the proposed method.

This paper merely assumes that the covariates are affected by the measurement error. However, in practical applications, the response variables can be influenced by measurement errors. The optimal subsampling probabilities are obtained by minimizing $tr(\mathbf{V})$ or $tr(\mathbf{V}_C)$ using the design ideas of the A-optimality and L-optimality criteria. In the future, the other optimality criteria for subsampling can be considered to develop more efficient algorithms.

Author contributions

Ruiyuan Chang: Furnished the algorithms and numerical results presented in the manuscript and composed the original draft of the manuscript; Xiuli Wang: Rendered explicit guidance regarding the proof of the theorem and refined the language of the entire manuscript; Mingqiu Wang: Rendered explicit guidance regarding the proof of theorems and the writing of codes and refined the language of the entire manuscript. All authors have read and consented to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (12271294) and the Natural Science Foundation of Shandong Province (ZR2024MA089).

Conflict of interest

The authors declare no conflict of interest.

References

1. J. A. Nelder, R. W. M. Wedderburn, Generalized linear models, *J. R. Statist. Soc. A*, **135** (1972), 370–384. <https://doi.org/10.2307/2344614>
2. H. Imberg, M. Axelson-Fisk, J. Jonasson, Optimal subsampling designs, arXiv: 2304.03019. <https://doi.org/10.48550/arXiv.2304.03019>
3. Y. Chao, L. Huang, X. Ma, J. Sun, Optimal subsampling for modal regression in massive data, *Metrika*, **87** (2024), 379–409. <https://doi.org/10.1007/s00184-023-00916-2>
4. H. Wang, R. Zhu, P. Ma, Optimal subsampling for large sample logistic regression, *J. Am. Stat. Assoc.*, **113** (2018), 829–844. <https://doi.org/10.1080/01621459.2017.1292914>
5. H. Wang, More efficient estimation for logistic regression with optimal subsamples, *J. Mach. Learn. Res.*, **20** (2019), 1–59.

6. L. Zuo, H. Zhang, H. Wang, L. Sun, Optimal subsample selection for massive logistic regression with distributed data, *Comput. Stat.*, **36** (2021), 2535–2562. <https://doi.org/10.1007/s00180-021-01089-0>
7. M. Ai, J. Yu, H. Zhang, H. Wang, Optimal subsampling algorithms for big data regression, *Stat. Sinica*, **31** (2021), 749–772. <https://doi.org/10.5705/ss.202018.0439>
8. Y. Yao, H. Wang, A review on optimal subsampling methods for massive datasets, *Journal of Data Science*, **19** (2021), 151–172. <https://doi.org/10.6339/21-JDS999>
9. J. Yu, Z. Ye, M. Ai, P. Ma, Optimal subsampling for data streams with measurement constrained categorical responses, *J. Comput. Graph. Stat.*, in press. <https://doi.org/10.1080/10618600.2024.2421990>
10. J. Yu, H. Wang, M. Ai, A subsampling strategy for AIC-based model averaging with generalized linear models, *Technometrics*, **67** (2025), 122–132. <https://doi.org/10.1080/00401706.2024.2407310>
11. J. Yu, M. Ai, Z. Ye, A review on design inspired subsampling for big data, *Stat. Papers*, **65** (2024), 467–510. <https://doi.org/10.1007/s00362-022-01386-w>
12. H. Liang, W. Härdle, R. J. Carroll, Estimation in a semiparametric partially linear errors-in-variables model, *Ann. Statist.*, **27** (1999), 1519–1535. <https://doi.org/10.1214/AOS/1017939140>
13. G. Li, L. Xue, Empirical likelihood confidence region for the parameter in a partially linear errors-in-variables model, *Commun. Stat.-Theor. M.*, **37** (2008), 1552–1564. <https://doi.org/10.1080/03610920801893913>
14. H. Liang, R. Li, Variable selection for partially linear models with measurement errors, *J. Am. Stat. Assoc.*, **104** (2009), 234–248. <https://doi.org/10.1198/jasa.2009.0127>
15. L. Stefanski, Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models, *Commun. Stat.-Theor. M.*, **18** (1989), 4335–4358. <https://doi.org/10.1080/03610928908830159>
16. T. Nakamura, Corrected score function for errors-in-variables models: methodology and application to generalized linear models, *Biometrika*, **77** (1990), 127–137. <https://doi.org/10.1093/biomet/77.1.127>
17. Y. Yang, G. Li, T. Tong, Corrected empirical likelihood for a class of generalized linear measurement error models, *Sci. China Math.*, **58** (2015), 1523–1536. <https://doi.org/10.1007/s11425-015-4976-6>
18. W. A. Fuller, *Measurement error models*, New York: John Wiley & Sons, 1987. <https://doi.org/10.1002/9780470316665>
19. H. Cui, F. Zou, L. Li, Feature screening and error variance estimation for ultrahigh-dimensional linear model with measurement errors, *Commun. Math. Stat.*, **13** (2025), 139–171. <https://doi.org/10.1007/s40304-022-00317-3>
20. R. J. Carroll, D. Ruppert, L. A. Stefanski, C. M. Crainiceanu, *Measurement error in nonlinear models: a modern perspective*, New York: Chapman and Hall/CRC, 2006. <https://doi.org/10.1201/9781420010138>

21. G. Y. Yi, A. Delaigle, P. Gustafson, *Handbook of measurement error models*, New York: Chapman and Hall/CRC, 2021. <https://doi.org/10.1201/9781315101279>
22. S. Lipovetsky, Book review: handbook of measurement error models, *Technometrics*, **65** (2023), 302–304. <https://doi.org/10.1080/00401706.2023.2201134>
23. T. Wang, H. Zhan, Optimal subsampling for multiplicative regression with massive data, *Stat. Neerl.*, **76** (2022), 418–449. <https://doi.org/10.1111/stan.12266>
24. X. Lin, R. J. Carroll, Nonparametric function estimation for clustered data when the predictor is measured without/with error, *J. Am. Stat. Assoc.*, **95** (2000), 520–534. <https://doi.org/10.1080/01621459.2000.10474229>
25. Y. Wei, R. J. Carroll, Quantile regression with measurement error, *J. Am. Stat. Assoc.*, **104** (2009), 1129–1143. <https://doi.org/10.1198/jasa.2009.tm08420>
26. P. Zhao, L. Xue, Variable selection for semiparametric varying coefficient partially linear errors-in-variables models, *J. Multivariate Anal.*, **101** (2010), 1872–1883. <https://doi.org/10.1016/j.jmva.2010.03.005>
27. Y. Ma, R. Li, Variable selection in measurement error models, *Bernoulli*, **16** (2010), 274–300. <https://doi.org/10.3150/09-bej205>
28. J. Ju, M. Wang, S. Zhao, Subsampling for big data linear models with measurement errors, arXiv: 2403.04361. <https://doi.org/10.48550/arXiv.2403.04361>
29. Y. Huang, C. Y. Wang, Consistent functional methods for logistic regression with errors in covariates, *J. Am. Stat. Assoc.*, **96** (2001), 1469–1482. <https://doi.org/10.1198/016214501753382372>
30. S. Cléménçon, P. Bertail, E. Chautru, Scaling up M-estimation via sampling designs: the Horvitz-Thompson stochastic gradient descent, *Proceedings of IEEE International Conference on Big Data*, 2014, 25–30. <https://doi.org/10.1109/BigData.2014.7004208>
31. L. Fahrmeir, H. Kaufmann, Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, *Ann. Statist.*, **13** (1985), 342–368. <https://doi.org/10.1214/aos/1176346597>
32. A. Datta, H. Zou, CoCoLasso for high-dimensional error-in-variables regression, *Ann. Statist.*, **45** (2017), 2400–2426. <https://doi.org/10.1214/16-AOS1527>
33. K. Nottingham, R. Longjohn, M. Kelly, *UC Irvine Machine Learning Repository*, University of California, School of Information and Computer Science, 2013. Available from: <http://archive.ics.uci.edu/ml>.
34. J. Yu, H. Wang, M. Ai, H. Zhang, Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data, *J. Am. Stat. Assoc.*, **117** (2022), 265–276. <https://doi.org/10.1080/01621459.2020.1773832>

Appendix

The proofs of the following lemmas and theorems are primarily based on Wang et al. [5], Ai et al. [7] and Yu et al. [34].

Lemma 1. If Assumptions A1–A4 and A6 hold, as $r \rightarrow \infty$ and $N \rightarrow \infty$, conditional on \mathcal{F}_N , we have

$$\widetilde{\mathbf{M}}_X - \mathbf{M}_X = O_{P|\mathcal{F}_N}(r^{-\frac{1}{2}}), \quad (\text{A.1})$$

$$\frac{1}{N} \mathbf{Q}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) - \frac{1}{N} \mathbf{Q}(\hat{\boldsymbol{\beta}}_{\text{MLE}}) = O_{P|\mathcal{F}_N}(r^{-\frac{1}{2}}), \quad (\text{A.2})$$

$$\frac{1}{N} \mathbf{V}_C^{-\frac{1}{2}} \mathbf{Q}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}), \quad (\text{A.3})$$

where

$$\widetilde{\mathbf{M}}_X = \frac{1}{Nr} \sum_{i=1}^r \frac{\widetilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\widetilde{\pi}_i},$$

and

$$\mathbf{V}_C = \frac{1}{N^2 r} \sum_{i=1}^N \boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}) \boldsymbol{\eta}_i^{*\text{T}}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}).$$

Proof.

$$\begin{aligned} E\left(\widetilde{\mathbf{M}}_X \mid \mathcal{F}_N\right) &= E\left(\frac{1}{Nr} \sum_{i=1}^r \frac{\widetilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\widetilde{\pi}_i} \mid \mathcal{F}_N\right) \\ &= \frac{1}{Nr} \sum_{i=1}^r \sum_{j=1}^N \pi_j \frac{\boldsymbol{\Omega}_j^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\pi_j} \\ &= \frac{1}{N} \sum_{i=1}^N \boldsymbol{\Omega}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}) \\ &= \mathbf{M}_X. \end{aligned}$$

By Assumption A6, we have

$$\begin{aligned} &E\left[\left(\widetilde{\mathbf{M}}_X^{j_1 j_2} - \mathbf{M}_X^{j_1 j_2}\right)^2 \mid \mathcal{F}_N\right] \\ &= E\left[\left(\frac{1}{Nr} \sum_{i=1}^r \frac{\widetilde{\boldsymbol{\Omega}}_i^{*(j_1 j_2)}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\widetilde{\pi}_i} - \frac{1}{N} \sum_{i=1}^N \boldsymbol{\Omega}_i^{*(j_1 j_2)}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})\right)^2 \mid \mathcal{F}_N\right] \\ &= \frac{1}{r} \sum_{i=1}^N \pi_i \left(\frac{\boldsymbol{\Omega}_i^{*(j_1 j_2)}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{N\pi_i} - \mathbf{M}_X^{j_1 j_2}\right)^2 \\ &= \frac{1}{r} \sum_{i=1}^N \pi_i \left(\frac{\boldsymbol{\Omega}_i^{*(j_1 j_2)}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{N\pi_i}\right)^2 - \frac{1}{r} \left(\mathbf{M}_X^{j_1 j_2}\right)^2 \\ &\leq \frac{1}{r} \sum_{i=1}^N \pi_i \left(\frac{\boldsymbol{\Omega}_i^{*(j_1 j_2)}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{N\pi_i}\right)^2 \\ &= O_P(r^{-1}). \end{aligned}$$

It follows from Chebyshev's inequality that (A.1) holds.

$$\begin{aligned}
 E\left(\frac{1}{N}\mathbf{Q}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})\middle|\mathcal{F}_N\right) &= E\left(\frac{1}{N}\frac{1}{r}\sum_{i=1}^r\frac{\tilde{\boldsymbol{\eta}}_i^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\tilde{\pi}_i}\middle|\mathcal{F}_N\right) \\
 &= \frac{1}{Nr}\sum_{i=1}^r\sum_{j=1}^N\pi_j\frac{\boldsymbol{\eta}_j^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\pi_j} \\
 &= \frac{1}{N}\sum_{i=1}^N\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}}) \\
 &= \mathbf{0}.
 \end{aligned}$$

By Assumption A4, we have

$$\begin{aligned}
 \text{Var}\left(\frac{1}{N}\mathbf{Q}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})\middle|\mathcal{F}_N\right) &= \text{Var}\left[\left(\frac{1}{N}\frac{1}{r}\sum_{i=1}^r\frac{\tilde{\boldsymbol{\eta}}_i^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\tilde{\pi}_i}\right)\middle|\mathcal{F}_N\right] \\
 &= \frac{1}{N^2r^2}\sum_{i=1}^r\sum_{j=1}^N\pi_j\frac{\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}})\boldsymbol{\eta}_i^{*\text{T}}(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\pi_j^2} \\
 &= \frac{1}{N^2r}\sum_{i=1}^N\frac{\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}})\boldsymbol{\eta}_i^{*\text{T}}(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}})}{\pi_i} \\
 &= O_p(r^{-1}).
 \end{aligned}$$

Now (A.2) follows from Markov's Inequality.

Let $\boldsymbol{\gamma}_i^* = (N\pi_i)^{-1}\tilde{\boldsymbol{\eta}}_i^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}})$, then $N^{-1}\mathbf{Q}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) = r^{-1}\sum_{i=1}^r\boldsymbol{\gamma}_i^*$ holds. Based on Assumption A6, for all $\varepsilon > 0$, we have

$$\begin{aligned}
 &\sum_{i=1}^r E\left\{\|r^{-\frac{1}{2}}\boldsymbol{\gamma}_i^*\|^2 I(\|\boldsymbol{\gamma}_i^*\| > r^{\frac{1}{2}}\varepsilon)\middle|\mathcal{F}_N\right\} \\
 &= \frac{1}{r}\sum_{i=1}^r E\left\{\|\boldsymbol{\gamma}_i^*\|^2 I(\|\boldsymbol{\gamma}_i^*\| > r^{\frac{1}{2}}\varepsilon)\middle|\mathcal{F}_N\right\} \\
 &\leq \frac{1}{r^{\frac{3}{2}}\varepsilon}\sum_{i=1}^r E\left\{\|\boldsymbol{\gamma}_i^*\|^3\middle|\mathcal{F}_N\right\} \\
 &= \frac{1}{r^{\frac{1}{2}}\varepsilon}\frac{1}{N^3}\sum_{i=1}^N\frac{\|\boldsymbol{\gamma}_i^*\|^3}{\pi_i^2} \\
 &= O_p(r^{-\frac{1}{2}}) = o_p(1).
 \end{aligned}$$

This shows that the Lindeberg-Feller conditions are satisfied in probability. Therefore (A.3) is true. \square

Lemma 2. If Assumptions A1–A7 hold, as $r \rightarrow \infty$ and $N \rightarrow \infty$, conditional on \mathcal{F}_N , for all $\mathbf{s}_r \rightarrow \mathbf{0}$, we have

$$\frac{1}{Nr}\sum_{i=1}^N\frac{\tilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}}+\mathbf{s}_r)}{\tilde{\pi}_i}-\frac{1}{N}\sum_{i=1}^N\boldsymbol{\Omega}_i^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}})=o_{P|\mathcal{F}_N}(1). \quad (\text{A.4})$$

Proof. The Eq (A.4) can be written as

$$\frac{1}{Nr} \sum_{i=1}^N \frac{\tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{MLE} + s_r)}{\tilde{\pi}_i} - \frac{1}{Nr} \sum_{i=1}^N \frac{\tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{MLE})}{\tilde{\pi}_i} + \frac{1}{Nr} \sum_{i=1}^N \frac{\tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{MLE})}{\tilde{\pi}_i} - \frac{1}{N} \sum_{i=1}^N \Omega_i^*(\Sigma_u, \hat{\beta}_{MLE}).$$

Let

$$\tau_1 := \frac{1}{Nr} \sum_{i=1}^N \frac{\tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{MLE} + s_r)}{\tilde{\pi}_i} - \frac{1}{Nr} \sum_{i=1}^N \frac{\tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{MLE})}{\tilde{\pi}_i},$$

then by Assumption A7, we have

$$\begin{aligned} E(\|\tau_1\|_S \mid \mathcal{F}_N) &= E \left\{ \frac{1}{Nr} \sum_{i=1}^N \frac{1}{\tilde{\pi}_i} \left\| \tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{MLE} + s_r) - \tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{MLE}) \right\|_S \mid \mathcal{F}_N \right\} \\ &= \frac{1}{Nr} \sum_{i=1}^r \sum_{j=1}^N \pi_j \frac{1}{\pi_j} \left\| \Omega_i^*(\Sigma_u, \hat{\beta}_{MLE} + s_r) - \Omega_i^*(\Sigma_u, \hat{\beta}_{MLE}) \right\|_S \\ &\leq \frac{1}{N} \sum_{i=1}^N m_2(W_i) \|s_r\| \\ &= o_P(1). \end{aligned}$$

It follows from Markov's inequality that $\tau_1 = o_{P|\mathcal{F}_N}(1)$.

Let

$$\tau_2 := \frac{1}{Nr} \sum_{i=1}^N \frac{\tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{MLE})}{\tilde{\pi}_i} - \frac{1}{N} \sum_{i=1}^N \Omega_i^*(\Sigma_u, \hat{\beta}_{MLE}),$$

then

$$E \left\{ \frac{1}{Nr} \sum_{i=1}^N \frac{\tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{MLE})}{\tilde{\pi}_i} \mid \mathcal{F}_N \right\} = \frac{1}{N} \sum_{i=1}^N \Omega_i^*(\Sigma_u, \hat{\beta}_{MLE}).$$

From the proof of Lemma 1, it follows that

$$E \left[\left(\tilde{\mathbf{M}}_X^{j_1 j_2} - \mathbf{M}_X^{j_1 j_2} \right)^2 \mid \mathcal{F}_N \right] = O_P(r^{-1}) = o_P(1).$$

Therefore $\tau_2 = o_{P|\mathcal{F}_N}(1)$, and (A.4) holds

□

Next, we will prove Theorems 2.1 and 2.2.

Proof of Theorem 2.1. $\tilde{\beta}$ is the solution of $\mathbf{Q}^*(\beta) = \frac{1}{r} \sum_{i=1}^r \frac{1}{\tilde{\pi}_i} \tilde{\eta}_i^*(\Sigma_u, \beta) = \mathbf{0}$, then

$$E \left(\frac{1}{N} \mathbf{Q}^*(\beta) \mid \mathcal{F}_N \right) = \frac{1}{Nr} \sum_{i=1}^r \sum_{j=1}^N \pi_j \frac{\eta_j^*(\Sigma_u, \beta)}{\pi_j} = \frac{1}{N} \sum_{i=1}^N \eta_i^*(\Sigma_u, \beta) = \frac{1}{N} \mathbf{Q}(\beta).$$

By Assumption A6, we have

$$\begin{aligned}
\text{Var}\left(\frac{1}{N}\mathbf{Q}^*(\boldsymbol{\beta})\mid\mathcal{F}_N\right) &= \text{Var}\left(\frac{1}{N}\frac{1}{r}\sum_{i=1}^r\frac{\tilde{\boldsymbol{\eta}}_i^*(\boldsymbol{\Sigma}_u,\boldsymbol{\beta})}{\tilde{\pi}_i}\mid\mathcal{F}_N\right) \\
&= \frac{1}{N^2r^2}\sum_{i=1}^r\sum_{j=1}^N\pi_j\frac{\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u,\boldsymbol{\beta})\boldsymbol{\eta}_i^{*\text{T}}(\boldsymbol{\Sigma}_u,\boldsymbol{\beta})}{\pi_j^2} \\
&= \frac{1}{N^2r}\sum_{i=1}^N\frac{\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u,\boldsymbol{\beta})\boldsymbol{\eta}_i^{*\text{T}}(\boldsymbol{\Sigma}_u,\boldsymbol{\beta})}{\pi_i} \\
&= O_P(r^{-1}).
\end{aligned}$$

Therefore, as $r \rightarrow \infty$, $N^{-1}\mathbf{Q}^*(\boldsymbol{\beta}) - N^{-1}\mathbf{Q}(\boldsymbol{\beta}) \rightarrow 0$ for all $\boldsymbol{\beta} \in \Lambda$ in conditional probability given \mathcal{F}_N . Thus, from Theorem 5.9 in [32], we have $\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\| = o_{P|\mathcal{F}_N}(1)$. By Taylor expansion,

$$\frac{1}{N}\mathbf{Q}^*(\tilde{\boldsymbol{\beta}}) = \mathbf{0} = \frac{1}{N}\mathbf{Q}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) + \frac{1}{Nr}\sum_{i=1}^r\frac{\tilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}} + \mathbf{s}_r)}{\tilde{\pi}_i}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}).$$

By Lemma 2, it follows that

$$\frac{1}{Nr}\sum_{i=1}^N\frac{\tilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}} + \mathbf{s}_r)}{\tilde{\pi}_i} - \frac{1}{N}\sum_{i=1}^N\boldsymbol{\Omega}_i^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}}) = o_{P|\mathcal{F}_N}(1),$$

then

$$\mathbf{0} = \frac{1}{N}\mathbf{Q}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) + \frac{1}{N}\sum_{i=1}^N\boldsymbol{\Omega}_i^*(\boldsymbol{\Sigma}_u,\hat{\boldsymbol{\beta}}_{\text{MLE}})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) + o_{P|\mathcal{F}_N}(1)(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}).$$

Here is

$$\frac{1}{N}\mathbf{Q}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) + \mathbf{M}_X(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) + o_{P|\mathcal{F}_N}\left(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|\right) = \mathbf{0},$$

we have

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}} &= -\mathbf{M}_X^{-1}\left\{\frac{1}{N}\mathbf{Q}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) + o_{P|\mathcal{F}_N}\left(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|\right)\right\} \\
&= -\mathbf{M}_X^{-1}\mathbf{V}_C^{\frac{1}{2}}\mathbf{V}_C^{-\frac{1}{2}}\frac{1}{N}\mathbf{Q}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) + \mathbf{M}_X^{-1}o_{P|\mathcal{F}_N}\left(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|\right) \\
&= O_{P|\mathcal{F}_N}(r^{-\frac{1}{2}}) + o_{P|\mathcal{F}_N}\left(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|\right).
\end{aligned} \tag{A.5}$$

By Lemma 1 and Assumption A3, $\mathbf{M}_X^{-1} = O_{P|\mathcal{F}_N}(1)$, we have $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}} = O_{P|\mathcal{F}_N}(r^{-\frac{1}{2}})$. □

Proof of Theorem 2.2. By Lemma 1 and (A.5), as $r \rightarrow \infty$, conditional on \mathcal{F}_N , it holds that

$$\mathbf{V}^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) = -\mathbf{V}^{-\frac{1}{2}}\mathbf{M}_X^{-1}\mathbf{V}_C^{\frac{1}{2}}\mathbf{V}_C^{-\frac{1}{2}}\frac{1}{N}\mathbf{Q}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) + o_{P|\mathcal{F}_N}(1).$$

By Lemma 1 and Slutsky's theorem, it follows that

$$\mathbf{V}^{-\frac{1}{2}}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}).$$

□

Proof of Theorem 2.3. To minimize the asymptotic variance $tr(\mathbf{V})$ of $\widetilde{\hat{\boldsymbol{\beta}}}$, the optimization problem is

$$\begin{cases} \min tr(\mathbf{V}) = \min \frac{1}{N^2 r} \sum_{i=1}^N \left[\frac{1}{\pi_i} \left\| \mathbf{M}_X^{-1} \boldsymbol{\eta}_i^* (\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}) \right\|^2 \right], \\ \text{s.t. } \sum_{i=1}^N \pi_i = 1, \quad 0 \leq \pi_i \leq 1, \quad i = 1, \dots, N. \end{cases} \quad (\text{A.6})$$

Define $g_i^{\text{mV}} = \left\| \mathbf{M}_X^{-1} \boldsymbol{\eta}_i^* (\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}) \right\|$, $i = 1, \dots, N$, it follows from Cauchy's inequality that

$$\begin{aligned} tr(\mathbf{V}) &= \frac{1}{N^2 r} \sum_{i=1}^N \left[\frac{1}{\pi_i} \left\| \mathbf{M}_X^{-1} \boldsymbol{\eta}_i^* (\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}) \right\|^2 \right] \\ &= \frac{1}{N^2 r} \left(\sum_{i=1}^N \pi_i \right) \left\{ \sum_{i=1}^N \left[\frac{1}{\pi_i} \left\| \mathbf{M}_X^{-1} \boldsymbol{\eta}_i^* (\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}) \right\|^2 \right] \right\} \\ &\geq \frac{1}{N^2 r} \left[\sum_{i=1}^N \left\| \mathbf{M}_X^{-1} \boldsymbol{\eta}_i^* (\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}) \right\| \right]^2 \\ &= \frac{1}{N^2 r} \left[\sum_{i=1}^N g_i^{\text{mV}} \right]^2. \end{aligned}$$

The equality sign holds if and only if $\pi_i \propto g_i^{\text{mV}}$, therefore

$$\pi_i^{\text{mV}} = \frac{g_i^{\text{mV}}}{\sum_{j=1}^N g_j^{\text{mV}}}$$

is the optimal solution. □

The proof of Theorem 2.4 is similar to Theorem 2.3.

Lemma 3. If Assumptions A1–A4 and A6 hold, as $r_0 \rightarrow \infty$, $r \rightarrow \infty$ and $N \rightarrow \infty$, conditional on \mathcal{F}_N , we have

$$\widetilde{\mathbf{M}}_X^{\hat{\beta}_0} - \mathbf{M}_X = O_{P|\mathcal{F}_N}(r^{-\frac{1}{2}}), \quad (\text{A.7})$$

$$\mathbf{M}_X^0 - \mathbf{M}_X = O_{P|\mathcal{F}_N}(r_0^{-\frac{1}{2}}), \quad (\text{A.8})$$

$$\frac{1}{N} \mathbf{Q}_{\hat{\beta}_0}^* (\hat{\boldsymbol{\beta}}_{\text{MLE}}) = O_{P|\mathcal{F}_N}(r^{-\frac{1}{2}}), \quad (\text{A.9})$$

$$\frac{1}{N} \mathbf{Q}_{\hat{\beta}_0}^{*0} (\hat{\boldsymbol{\beta}}_{\text{MLE}}) = O_{P|\mathcal{F}_N}(r_0^{-\frac{1}{2}}), \quad (\text{A.10})$$

$$\frac{1}{N} \mathbf{V}_C^{\text{opt}-\frac{1}{2}} \mathbf{Q}_{\hat{\beta}_0}^* (\hat{\boldsymbol{\beta}}_{\text{MLE}}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}), \quad (\text{A.11})$$

where

$$\widetilde{\mathbf{M}}_X^{\hat{\beta}_0} = \frac{1}{Nr} \sum_{i=1}^r \frac{\widetilde{\boldsymbol{\Omega}}_i^* (\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\widetilde{\pi}_i^{\text{opt}}},$$

$$\mathbf{M}_X^0 = \frac{1}{Nr_0} \sum_{i=1}^{r_0} \frac{\widetilde{\boldsymbol{\Omega}}_i^* (\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\widetilde{\pi}_i^{\text{UNIF}}}.$$

Proof.

$$\begin{aligned}
 E\left(\tilde{\mathbf{M}}_X^{\tilde{\beta}_0} \mid \mathcal{F}_N\right) &= E_{\tilde{\beta}_0} \left[E\left(\tilde{\mathbf{M}}_X^{\tilde{\beta}_0} \mid \mathcal{F}_N, \tilde{\beta}_0\right) \right] \\
 &= E_{\tilde{\beta}_0} \left[E\left(\frac{1}{Nr} \sum_{i=1}^r \frac{\tilde{\Omega}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\tilde{\pi}_i^{\text{opt}}} \mid \mathcal{F}_N, \tilde{\beta}_0 \right) \right] \\
 &= E_{\tilde{\beta}_0} \left[E\left(\mathbf{M}_X \mid \mathcal{F}_N, \tilde{\beta}_0\right) \right] \\
 &= \mathbf{M}_X.
 \end{aligned}$$

By Assumption A6, we have

$$\begin{aligned}
 &E\left[\left(\tilde{\mathbf{M}}_X^{\tilde{\beta}_0, j_1 j_2} - \mathbf{M}_X^{j_1 j_2}\right)^2 \mid \mathcal{F}_N\right] \\
 &= E_{\tilde{\beta}_0} \left\{ E\left[\left(\tilde{\mathbf{M}}_X^{\tilde{\beta}_0, j_1 j_2} - \mathbf{M}_X^{j_1 j_2}\right)^2 \mid \mathcal{F}_N, \tilde{\beta}_0\right] \right\} \\
 &= E_{\tilde{\beta}_0} \left[\frac{1}{r} \sum_{i=1}^N \pi_i^{\text{opt}} \left(\frac{\tilde{\Omega}_i^{*j_1 j_2}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{N\pi_i^{\text{opt}}} - \mathbf{M}_X^{j_1 j_2} \right)^2 \mid \mathcal{F}_N, \tilde{\beta}_0 \right] \\
 &= E_{\tilde{\beta}_0} \left[\frac{1}{r} \sum_{i=1}^N \pi_i^{\text{opt}} \left(\frac{\tilde{\Omega}_i^{*j_1 j_2}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{N\pi_i^{\text{opt}}} \right)^2 - \frac{1}{r} \left(\mathbf{M}_X^{j_1 j_2}\right)^2 \mid \mathcal{F}_N, \tilde{\beta}_0 \right] \\
 &\leq E_{\tilde{\beta}_0} \left[\frac{1}{r} \sum_{i=1}^N \pi_i^{\text{opt}} \left(\frac{\tilde{\Omega}_i^{*j_1 j_2}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{N\pi_i^{\text{opt}}} \right)^2 \mid \mathcal{F}_N, \tilde{\beta}_0 \right] \\
 &= \frac{1}{r} \sum_{i=1}^N \frac{\left(\tilde{\Omega}_i^{*j_1 j_2}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})\right)^2}{N^2 \pi_i^{\text{opt}}} \\
 &= O_P(r^{-1}).
 \end{aligned}$$

It follows from Chebyshev's inequality that (A.7) holds. Similarly, (A.8) also holds.

$$E\left(\frac{1}{N} \mathbf{Q}_{\tilde{\beta}_0}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) \mid \mathcal{F}_N\right) = E_{\tilde{\beta}_0} \left[E\left(\frac{1}{N} \frac{1}{r} \sum_{i=1}^r \frac{\tilde{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\tilde{\pi}_i^{\text{opt}}} \mid \mathcal{F}_N, \tilde{\beta}_0\right) \right] = \frac{1}{N} \sum_{i=1}^N \boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}) = \mathbf{0}.$$

By Assumption A6, we have

$$\begin{aligned}
 \text{Var}\left(\frac{1}{N} \mathbf{Q}_{\tilde{\beta}_0}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) \mid \mathcal{F}_N\right) &= E_{\tilde{\beta}_0} \left\{ \text{Var}\left[\left(\frac{1}{N} \frac{1}{r} \sum_{i=1}^r \frac{\tilde{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\tilde{\pi}_i^{\text{opt}}}\right) \mid \mathcal{F}_N, \tilde{\beta}_0\right] \right\} \\
 &= \frac{1}{N^2 r} \sum_{i=1}^N \frac{\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}) \boldsymbol{\eta}_i^{*\text{T}}(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\pi_i^{\text{opt}}} \\
 &= O_P(r^{-1}).
 \end{aligned}$$

Therefore, the (A.9) and (A.10) follow from Markov's Inequality.

Let

$$\boldsymbol{\gamma}_{i,\hat{\boldsymbol{\beta}}_0}^* = \frac{\tilde{\boldsymbol{\eta}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{N\bar{\pi}_i^{\text{opt}}},$$

for all $\varepsilon > 0$, it follows that $N^{-1}\mathbf{Q}_{\hat{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}}) = r^{-1} \sum_{i=1}^r \boldsymbol{\gamma}_{i,\hat{\boldsymbol{\beta}}_0}^*$,

$$\begin{aligned} & \sum_{i=1}^r E_{\tilde{\boldsymbol{\beta}}_0} \left\{ E \left[\left\| r^{-\frac{1}{2}} \boldsymbol{\gamma}_{i,\tilde{\boldsymbol{\beta}}_0}^* \right\|^2 I \left(\left\| \boldsymbol{\gamma}_{i,\tilde{\boldsymbol{\beta}}_0}^* \right\| > r^{\frac{1}{2}} \varepsilon \right) \middle| \mathcal{F}_N, \tilde{\boldsymbol{\beta}}_0 \right] \right\} \\ &= \frac{1}{r} \sum_{i=1}^r E_{\tilde{\boldsymbol{\beta}}_0} \left\{ E \left[\left\| \boldsymbol{\gamma}_{i,\tilde{\boldsymbol{\beta}}_0}^* \right\|^2 I \left(\left\| \boldsymbol{\gamma}_{i,\tilde{\boldsymbol{\beta}}_0}^* \right\| > r^{\frac{1}{2}} \varepsilon \right) \middle| \mathcal{F}_N, \tilde{\boldsymbol{\beta}}_0 \right] \right\} \\ &\leq \frac{1}{r^{\frac{3}{2}} \varepsilon} \sum_{i=1}^r E_{\tilde{\boldsymbol{\beta}}_0} \left[E \left(\left\| \boldsymbol{\gamma}_{i,\tilde{\boldsymbol{\beta}}_0}^* \right\|^3 \middle| \mathcal{F}_N, \tilde{\boldsymbol{\beta}}_0 \right) \right] \\ &= \frac{1}{r^{\frac{1}{2}} \varepsilon} \frac{1}{N^3} \sum_{i=1}^N \frac{\left\| \tilde{\boldsymbol{\eta}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}) \right\|^3}{\pi_i^{\text{opt}^2}} \\ &= O_P(r^{-\frac{1}{2}}) = o_P(1). \end{aligned}$$

This shows that the Lindeberg-Feller conditions are satisfied in probability. Therefore (A.11) is true. \square

Lemma 4. If Assumptions A1–A7 hold, as $r_0 \rightarrow \infty$, $r \rightarrow \infty$ and $N \rightarrow \infty$, for all $\mathbf{s}_{r_0} \rightarrow \mathbf{0}$ and $\mathbf{s}_r \rightarrow \mathbf{0}$, conditional on \mathcal{F}_N , we have

$$\frac{1}{Nr_0} \sum_{i=1}^{r_0} \frac{\tilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}} + \mathbf{s}_{r_0})}{\bar{\pi}_i^{\text{opt}}} - \frac{1}{N} \sum_{i=1}^N \boldsymbol{\Omega}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}) = o_{P|\mathcal{F}_N}(1), \quad (\text{A.12})$$

$$\frac{1}{Nr} \sum_{i=1}^r \frac{\tilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}} + \mathbf{s}_r)}{\bar{\pi}_i^{\text{opt}}} - \frac{1}{N} \sum_{i=1}^N \boldsymbol{\Omega}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}) = o_{P|\mathcal{F}_N}(1). \quad (\text{A.13})$$

Proof. The Eq (A.12) can be written as

$$\frac{1}{Nr_0} \sum_{i=1}^{r_0} \frac{\tilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}} + \mathbf{s}_{r_0})}{\bar{\pi}_i^{\text{opt}}} - \frac{1}{Nr_0} \sum_{i=1}^{r_0} \frac{\tilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\bar{\pi}_i^{\text{opt}}} + \frac{1}{Nr_0} \sum_{i=1}^{r_0} \frac{\tilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\bar{\pi}_i^{\text{opt}}} - \frac{1}{N} \sum_{i=1}^N \boldsymbol{\Omega}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}}).$$

Let

$$\boldsymbol{\tau}_1^0 := \frac{1}{Nr_0} \sum_{i=1}^{r_0} \frac{\tilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}} + \mathbf{s}_{r_0})}{\bar{\pi}_i^{\text{opt}}} - \frac{1}{Nr_0} \sum_{i=1}^{r_0} \frac{\tilde{\boldsymbol{\Omega}}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})}{\bar{\pi}_i^{\text{opt}}},$$

then by Assumption A7, we have

$$\begin{aligned}
 E\left(\|\tau_1^0\|_S \mid \mathcal{F}_N\right) &= E_{\tilde{\beta}_0} \left\{ E \left[\frac{1}{Nr_0} \sum_{i=1}^{r_0} \frac{1}{\tilde{\pi}_i^{\text{opt}}} \left\| \tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}} + s_{r_0}) - \tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}}) \right\|_S \mid \mathcal{F}_N, \tilde{\beta}_0 \right] \right\} \\
 &= \frac{1}{N} \sum_{i=1}^N \left\| \Omega_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}} + s_{r_0}) - \Omega_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}}) \right\|_S \\
 &\leq \frac{1}{N} \sum_{i=1}^N m_2(\mathbf{W}_i) \|s_{r_0}\| \\
 &= o_P(1).
 \end{aligned}$$

It follows from Markov’s inequality that $\tau_1^0 = o_{P|\mathcal{F}_N}(1)$.

Let

$$\tau_2^0 := \frac{1}{Nr_0} \sum_{i=1}^{r_0} \frac{\tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}})}{\tilde{\pi}_i^{\text{opt}}} - \frac{1}{N} \sum_{i=1}^N \Omega_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}}),$$

then

$$E_{\tilde{\beta}_0} \left\{ E \left[\frac{1}{Nr_0} \sum_{i=1}^{r_0} \frac{\tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}} + s_{r_0})}{\tilde{\pi}_i^{\text{opt}}} \mid \mathcal{F}_N, \tilde{\beta}_0 \right] \right\} = \frac{1}{N} \sum_{i=1}^N \Omega_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}}).$$

By the proof of Lemma 3, it follows that

$$E \left[\left(\tilde{\mathbf{M}}_X^{\tilde{\beta}_0, j_1 j_2} - \mathbf{M}_X^{j_1 j_2} \right)^2 \mid \mathcal{F}_N \right] = O_P(r_0^{-1}) = o_P(1),$$

we have $\tau_2^0 = o_{P|\mathcal{F}_N}(1)$. Therefore (A.12) holds. Similarly, (A.13) is also true. □

Next, we will prove Theorems 2.5 and 2.6.

Proof of Theorem 2.5.

$$E\left(\frac{1}{N} \mathbf{Q}_{\tilde{\beta}_0}^*(\beta) \mid \mathcal{F}_N\right) = E_{\tilde{\beta}_0} \left[E \left(\frac{1}{N} \frac{1}{r} \sum_{i=1}^r \frac{\tilde{\eta}_i^*(\Sigma_u, \beta)}{\tilde{\pi}_i^{\text{opt}}} \mid \mathcal{F}_N, \tilde{\beta}_0 \right) \right] = \frac{1}{N} \sum_{i=1}^N \eta_i^*(\Sigma_u, \beta) = \frac{1}{N} \mathbf{Q}(\beta).$$

By Assumption A6, we have

$$\text{Var} \left(\frac{1}{N} \mathbf{Q}_{\tilde{\beta}_0}^*(\beta) \mid \mathcal{F}_N \right) = E_{\tilde{\beta}_0} \left\{ \text{Var} \left(\frac{1}{N} \frac{1}{r} \sum_{i=1}^r \frac{\tilde{\eta}_i^*(\Sigma_u, \beta)}{\tilde{\pi}_i^{\text{opt}}} \mid \mathcal{F}_N, \tilde{\beta}_0 \right) \right\} = \frac{1}{N^2 r} \sum_{i=1}^N \frac{\eta_i^*(\Sigma_u, \beta) \eta_i^{*\text{T}}(\Sigma_u, \beta)}{\pi_i^{\text{opt}}} = O_P(r^{-1}).$$

Hence, as $r \rightarrow \infty$, $N^{-1} \mathbf{Q}_{\tilde{\beta}_0}^*(\beta) - N^{-1} \mathbf{Q}(\beta) \rightarrow \mathbf{0}$ for all $\beta \in \Lambda$ in conditional probability given \mathcal{F}_N .

$\check{\beta}$ is the solution of $\mathbf{Q}_{\tilde{\beta}_0}^{\text{two-step}}(\beta) = \mathbf{0}$, we have

$$\mathbf{0} = \frac{1}{N} \mathbf{Q}_{\tilde{\beta}_0}^{\text{two-step}}(\check{\beta}) = \frac{r}{r+r_0} \frac{1}{N} \mathbf{Q}_{\tilde{\beta}_0}^*(\check{\beta}) + \frac{r_0}{r+r_0} \frac{1}{N} \mathbf{Q}_{\tilde{\beta}_0}^{*0}(\check{\beta}). \tag{A.14}$$

By Lemma 4, we have

$$\frac{1}{Nr_0} \sum_{i=1}^{r_0} \frac{\tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}} + s_{r_0})}{\tilde{\pi}_i^{\text{opt}}} = \frac{1}{N} \sum_{i=1}^N \Omega_i^*(\Sigma_u, \hat{\beta}_{\text{MLE}}) + o_{P|\mathcal{F}_N}(1) = \mathbf{M}_X + o_{P|\mathcal{F}_N}(1),$$

and

$$\frac{1}{Nr} \sum_{i=1}^r \frac{\tilde{\Omega}_i^*(\Sigma_u, \hat{\beta}_{MLE} + s_r)}{\tilde{\pi}_i^{\text{opt}}} = \mathbf{M}_X + o_{P|\mathcal{F}_N}(1).$$

By Taylor expansion, we have

$$\begin{aligned} \frac{1}{N} \mathbf{Q}_{\check{\beta}_0}^*(\check{\beta}) &= \frac{1}{N} \mathbf{Q}_{\check{\beta}_0}^*(\hat{\beta}_{MLE}) + \frac{1}{Nr} \sum_{i=1}^r \frac{\Omega_i^*(\Sigma_u, \hat{\beta}_{MLE} + s_r)}{\tilde{\pi}_i^{\text{opt}}} (\check{\beta} - \hat{\beta}_{MLE}) \\ &= \frac{1}{N} \mathbf{Q}_{\check{\beta}_0}^*(\hat{\beta}_{MLE}) + \mathbf{M}_X (\check{\beta} - \hat{\beta}_{MLE}) + o_{P|\mathcal{F}_N}(1) (\check{\beta} - \hat{\beta}_{MLE}). \end{aligned} \quad (\text{A.15})$$

Similarly,

$$\frac{1}{N} \mathbf{Q}_{\check{\beta}_0}^{*0}(\check{\beta}) = \frac{1}{N} \mathbf{Q}_{\check{\beta}_0}^{*0}(\hat{\beta}_{MLE}) + \mathbf{M}_X (\check{\beta} - \hat{\beta}_{MLE}) + o_{P|\mathcal{F}_N}(1) (\check{\beta} - \hat{\beta}_{MLE}). \quad (\text{A.16})$$

As $r_0 r^{-1} \rightarrow 0$, $N^{-1} \mathbf{Q}_{\check{\beta}_0}^{*0}(\hat{\beta}_{MLE}) = O_{P|\mathcal{F}_N}(r_0^{-\frac{1}{2}})$, then

$$\frac{r_0}{r + r_0} \frac{1}{N} \mathbf{Q}_{\check{\beta}_0}^{*0}(\hat{\beta}_{MLE}) = \frac{r_0}{r + r_0} O_{P|\mathcal{F}_N}(r_0^{-\frac{1}{2}}) = o_{P|\mathcal{F}_N}(r^{-\frac{1}{2}}).$$

Combining this with (A.14)–(A.16), we have

$$\check{\beta} - \hat{\beta}_{MLE} = O_{P|\mathcal{F}_N}(r^{-\frac{1}{2}}) + o_{P|\mathcal{F}_N}(\|\check{\beta} - \hat{\beta}_{MLE}\|), \quad (\text{A.17})$$

which implies that $\check{\beta} - \hat{\beta}_{MLE} = O_{P|\mathcal{F}_N}(r^{-\frac{1}{2}})$. \square

Proof of Theorem 2.6. By Lemma 3, $\frac{1}{N} \mathbf{V}_C^{-\frac{1}{2}} \mathbf{Q}_{\check{\beta}_0}^*(\hat{\beta}_{MLE}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I})$, we have

$$\begin{aligned} \|\mathbf{V}_C - \mathbf{V}_C^{\text{opt}}\|_S &= \left\| \frac{1}{N^2 r} \sum_{i=1}^N \frac{\boldsymbol{\eta}_i^*(\Sigma_u, \hat{\beta}_{MLE}) \boldsymbol{\eta}_i^{*\text{T}}(\Sigma_u, \hat{\beta}_{MLE})}{\pi_i^{\text{op}}} - \frac{1}{N^2 r} \sum_{i=1}^N \frac{\boldsymbol{\eta}_i^*(\Sigma_u, \hat{\beta}_{MLE}) \boldsymbol{\eta}_i^{*\text{T}}(\Sigma_u, \hat{\beta}_{MLE})}{\pi_i^{\text{opt}}} \right\|_S \\ &\leq \frac{1}{N^2 r} \sum_{i=1}^N \left\| \frac{1}{\pi_i^{\text{op}}} - \frac{1}{\pi_i^{\text{opt}}} \right\| \|\boldsymbol{\eta}_i^*(\Sigma_u, \hat{\beta}_{MLE})\|^2 \\ &= \frac{1}{r} \sum_{i=1}^N \left\| 1 - \frac{\pi_i^{\text{op}}}{\pi_i^{\text{opt}}} \right\| \frac{\|\boldsymbol{\eta}_i^*(\Sigma_u, \hat{\beta}_{MLE})\|^2}{N^2 \pi_i^{\text{op}}}. \end{aligned}$$

Taking π_i^{mVc} as an example, by Assumption A4, the above equation can be summarized as

$$\begin{aligned}
& \frac{1}{r} \sum_{i=1}^N \left\| 1 - \frac{\pi_i^{\text{mVc}}}{\pi_i^{\text{mVct}}} \left\| \frac{\|\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})\|^2}{N^2} \frac{\sum_{j=1}^N g_j^{\text{mVc}}}{g_i^{\text{mVc}}} \right\| \right\| \\
&= \frac{1}{r} \sum_{i=1}^N \left\| 1 - \frac{\pi_i^{\text{mVc}}}{\pi_i^{\text{mVct}}} \left\| \frac{\|\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})\|^2}{N^2} \frac{\sum_{j=1}^N \|\boldsymbol{\eta}_j^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})\|}{\|\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})\|} \right\| \right\| \\
&\leq \frac{1}{r} \sum_{i=1}^N \left\| 1 - \frac{\pi_i^{\text{mVc}}}{\pi_i^{\text{mVct}}} \left\| \frac{\|\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})\|}{N} \frac{\sum_{j=1}^N \|\boldsymbol{\eta}_j^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})\|}{N} \right\| \right\| \\
&\leq \frac{1}{r} \left(\frac{1}{N} \sum_{i=1}^N \left\| 1 - \frac{\pi_i^{\text{mVc}}}{\pi_i^{\text{mVct}}} \right\|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^N \frac{\|\boldsymbol{\eta}_i^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})\|^2}{N} \right)^{\frac{1}{2}} \left(\frac{\sum_{j=1}^N \|\boldsymbol{\eta}_j^*(\boldsymbol{\Sigma}_u, \hat{\boldsymbol{\beta}}_{\text{MLE}})\|}{N} \right) \\
&= o_{P|\mathcal{F}_N}(r^{-1}).
\end{aligned}$$

Therefore $\|\mathbf{V}_C - \mathbf{V}_C^{\text{opt}}\|_S = o_{P|\mathcal{F}_N}(r^{-1})$, and

$$\begin{aligned}
\mathbf{V}_{\text{opt}}^{-\frac{1}{2}}(\check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) &= -\mathbf{V}_{\text{opt}}^{-\frac{1}{2}} \mathbf{M}_X^{-1} \frac{1}{N} \mathbf{Q}_{\check{\boldsymbol{\beta}}_0}^{\text{two-step}}(\hat{\boldsymbol{\beta}}_{\text{MLE}}) + o_{P|\mathcal{F}_N}(1) \\
&= -\mathbf{V}_{\text{opt}}^{-\frac{1}{2}} \mathbf{M}_X^{-1} \left[\frac{r}{r+r_0} \frac{1}{N} \mathbf{Q}_{\check{\boldsymbol{\beta}}_0}^*(\check{\boldsymbol{\beta}}) + \frac{r_0}{r+r_0} \frac{1}{N} \mathbf{Q}_{\check{\boldsymbol{\beta}}_0}^{*0}(\check{\boldsymbol{\beta}}) \right] + o_{P|\mathcal{F}_N}(1) \\
&= -\mathbf{V}_{\text{opt}}^{-\frac{1}{2}} \mathbf{M}_X^{-1} (\mathbf{V}_C^{\text{opt}})^{\frac{1}{2}} (\mathbf{V}_C^{\text{opt}})^{-\frac{1}{2}} \frac{1}{N} \mathbf{Q}_{\check{\boldsymbol{\beta}}_0}^*(\check{\boldsymbol{\beta}}) + o_{P|\mathcal{F}_N}(1),
\end{aligned}$$

which implies that

$$\mathbf{V}_{\text{opt}}^{-\frac{1}{2}} \mathbf{M}_X^{-1} (\mathbf{V}_C^{\text{opt}})^{\frac{1}{2}} \left\{ \mathbf{V}_{\text{opt}}^{-\frac{1}{2}} \mathbf{M}_X^{-1} (\mathbf{V}_C^{\text{opt}})^{\frac{1}{2}} \right\}^{\text{T}} = \mathbf{V}_{\text{opt}}^{-\frac{1}{2}} \mathbf{M}_X^{-1} (\mathbf{V}_C^{\text{opt}})^{\frac{1}{2}} (\mathbf{V}_C^{\text{opt}})^{\frac{1}{2}} \mathbf{M}_X^{-1} \mathbf{V}_{\text{opt}}^{-\frac{1}{2}} = \mathbf{I}.$$

Therefore

$$\mathbf{V}_{\text{opt}}^{-\frac{1}{2}}(\check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}).$$

□



AIMS Press

© 2025 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)