Subsampling in Longitudinal Models

Ziyang Wang[1*], HaiYing Wang[1] and Nalini Ravishanker[1]

[1*]Department of Statistics, University of Connecticut, 215 Glenbrook Road, Storrs, 06268, CT, USA.

*Corresponding author(s). E-mail(s): ziyang.wang@uconn.edu;
Contributing authors: haiying.wang@uconn.edu;
nalini.ravishanker@uconn.edu;

Abstract

For large scale data, subsampling methods are often used to approximate the full-data parameter estimates. An ideal subsampling method picks a small proportion of informative observations from the full data and produces an accurate approximate to the full-data estimate using much less computing power. Existing studies on subsampling methods focus on independent responses. This paper discusses subsampling methods for longitudinal data where observations within a block are correlated, and develops optimal subsampling methods to approximate the full-data maximum likelihood estimators of the model parameters. We first establish the conditional asymptotic distribution of the subsample estimator with general subsampling probabilities, and then derive the *optimal* subsampling method that minimizes the asymptotic mean square error of the subsample estimator. To evaluate the finite sample performance of the proposed method, we provide results based on numerical experiments with simulated data.

Keywords: Large data; Fisher scoring; Optimal subsampling

1 Introduction

In the big data era, huge amounts of data are being generated every day. While this greatly extends the possibility of getting more information, it makes certain standard statistical tools unfeasible because of the much more complex computation due to high data volume.

To reduce the computational burden, one possible solution is to perform the calculations on a smaller subset of the full data. This sacrifices a certain amount of information in exchange for easier computation. To extract the maximum amount of information from the full data, subsampling designs are developed so that more informative data points have higher chance of being selected. In the context of linear regression, statistical leverage scores and their variants are widely used in identifying influential rows in the covariate matrix (Drineas et al, 2006, 2010; Yang et al, 2015). This approach is called *algorithmic leveraging* (Ma et al, 2015) and it has been shown to perform well with limited computing power (Avron et al, 2010; Meng et al, 2020). Besides these, Ma et al (2015) provided an analysis of this method from a statistical perspective. However, the algorithmic leveraging approach does not use the information from the responses when assigning subsampling probabilities and this sampling scheme is referred to as non-informative subsampling. Zhu (2018) proposed agradient-based subsampling method where the subsampling probabilities depend on the responses as well as the covariates. Wang et al (2019) developed an information-based optimal subdataselection method that has high estimation efficiency. Li and Meng (2021) provided a review of these subsampling methods and evaluated their performance using real data. Meng et al (2020) developed the "LowCon" subsampling method to handle the cases when the models are misspecified. For logistic regression, Fithian and Hastie (2014) proposed the local case-control (LCC) subsampling method for imbalanced data. Using the A-optimality criterion in design of experiments, Wang et al (2018) proposed an optimal subsampling method that minimizes the asymptotic mean squared error of the subsampling estimator. The optimal subsampling method based on the A-optimality criterion has been extended to include multi-class logistic regression model (Yao and Wang, 2019), generalized linear models (Ai et al, 2021b), quantile regressions (Ai et al, 2021a; Wang and Ma, 2021) and quasi-likelihood models Yu et al (2021). The aforementioned subsampling methods are all related to independent data. In this paper, we focus on longitudinal data where observations within each block are assumed to be correlated.

In longitudinal data analysis, multivariate linear models with dependent covariance structures within the response vector are commonly used (Chapter 4 of Diggle et al, 2013). We consider the case of longitudinal data under a balanced design, i.e. the number of measurements for each subject are the same. The corresponding multivariate linear model has the following form:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad \text{for} \quad i = 1, \ldots, m, \tag{1}$$

where m denotes the total number of subjects; n is the number of measurements on each subject; $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,n})'$ is the $n \times 1$ response vector for the i-th subject; $\mathbf{X}_i = (\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,d})$ is the $n \times d$ design matrix with $\mathbf{x}_{i,j}$ being the $n \times 1$ vectors; $\boldsymbol{\beta}$ is the $d \times 1$ unknown parameter vector; and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ is the $n \times 1$ unobserved normal random error with $\mathbf{V}$ being its variance-covariance matrix whose structure depends on unknown covariance parameter $\boldsymbol{\xi} \in \mathbb{R}^q$.

In this paper, we assume that $\mathbf{X}_i$'s are nonrandom and $\boldsymbol{\varepsilon}_i$'s are independent. Since most models include an intercept term, we assume that $\mathbf{x}_{i,1} = \mathbf{1}$. We denote the full data matrix as $\mathcal{D}_m = (\mathcal{X}, \mathcal{Y})$, where $\mathcal{X} = (\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_m')'$, $\mathcal{Y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_m')'$.

In the model specified in (1), an explicit parametric model for $\mathbf{V}$ is often assumed. This represents the belief we may have about possible associations among observations within each subject, usually from subject expert knowledge or past experience. When such a model is not available, using (1) requires us to estimate $n(n+1)/2$ covariance parameters, which can be complicated, especially when n is large. One alternative is to use the random effects model (Laird and Ware, 1982). The idea is to model the correlations among observations within each subject parsimoniously by introducing random effects. To connect to model (1), the random effect model decomposes the overall aggregated variation of $\boldsymbol{\varepsilon}_i$ into two parts: the random effect and the pure error, e.g., error that may be caused by the measurement mechanism. In this paper, we assume that a parametric model for $\mathbf{V}$ is given and we focus on the compound symmetric correlation structure, i.e., the variance-covariance matrix of $\boldsymbol{\varepsilon}_i$, $\mathbf{V}$, has the form:

$$\mathbf{V}(\boldsymbol{\xi}) = \sigma^2 \left((1-\rho)\mathbf{I} + \rho\mathbf{J}\right), \boldsymbol{\xi} = (\sigma^2, \rho)', \tag{2}$$

where $-1/(n-1) < \rho < 1$, $\mathbf{I}$ is the $n \times n$ identity matrix, and $\mathbf{J}$ is the $n \times n$ matrix with each element being equal to unity. The inverse of the matrix $\mathbf{V}$ has the form:

$$\mathbf{V}^{-1} = \frac{1}{\sigma^2(1-\rho)}\left(\mathbf{I} - \frac{\rho}{1+(n-1)\rho}\mathbf{J}\right). \tag{3}$$

This correlation structure is the default option for many software packages and it is widely used in many scientific fields such as social sciences and medical studies (Zhao et al, 2019; Pusponegoro et al, 2017; Hong and Shyr, 2007; Kaplan et al, 2004).

The rest of the paper is organized as follows. In section 2, we describe the Fisher scoring method that is used to obtain parameter estimates in longitudinal models. In section 3, we first provide the general subsampling algorithm in longitudinal models. We then discuss several methods for specifying the subsampling probabilities when selecting subsamples. We also provide a two-step subsampling method for practical implementation. In section 4, we compare the empirical performance of all the subsampling procedures and discuss the implications of the numerical results. Section 5 summarizes the paper. Technical proofs for the theoretical results are given in the appendix.

2 Fisher scoring in longitudinal models

To facilitate the presentation, we use $\boldsymbol{\theta}$ to denote the full vector of unknown parameters, i.e., $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\xi}')' = (\boldsymbol{\beta}', \sigma^2, \rho)'$. To estimate the unknown parameters, the maximum likelihood estimator (MLE) obtained through maximizing

the log-likelihood function is

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\xi}}')' = \arg\max_{\boldsymbol{\beta},\boldsymbol{\xi}} \mathrm{M}_m(\boldsymbol{\beta}, \boldsymbol{\xi}), \tag{4}$$

where

$$\mathrm{M}_m(\boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{p}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi}), \tag{5}$$

and

$$\mathrm{p}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi}) = \log\left[\frac{1}{(2\pi)^{\frac{n}{2}}}\frac{1}{\sqrt{det(\mathbf{V})}}\exp\left(-\frac{(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})}{2}\right)\right].$$

There is no general closed form solution to equation (4), so an iterative method is required to solve it numerically. In this paper, we resort to the Fisher scoring method (Jennrich and Schluchter, 1986) to obtain the MLEs.

To facilitate the presentation, we first introduce some notations. Let $\dot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})$ and $\ddot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})$ be the gradient vector and Hessian matrix of $\mathrm{p}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})$, respectively, i.e.,

$$\dot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{\partial \mathrm{p}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \boldsymbol{\theta}} \quad \text{and} \quad \ddot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{\partial^2 \mathrm{p}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}.$$

Appendix A gives the explicit expressions for $\dot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})$ and $\ddot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})$. We use $\dot{\mathbf{p}}_{\boldsymbol{\beta}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})$ and $\dot{\mathbf{p}}_{\boldsymbol{\xi}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})$ to denote the sub-vectors of $\dot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})$ corresponding to $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$, respectively. Similarly, we define $\ddot{\mathbf{p}}_{\boldsymbol{\beta}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})$ and $\ddot{\mathbf{p}}_{\boldsymbol{\xi}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})$ as the upper-left and lower-right sub-matrices corresponding to $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$, respectively.

The Fisher scoring method obtains the MLEs by iterating the following two steps until $(\boldsymbol{\beta}^{(k+1)\prime}, \boldsymbol{\xi}^{(k+1)\prime})'$ converges:

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= \left(\sum_{i=1}^{m} \mathbf{X}_i'(\mathbf{V}^{(k)})^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i'(\mathbf{V}^{(k)})^{-1}\mathbf{y}_i, \\
\boldsymbol{\xi}^{(k+1)} &= \boldsymbol{\xi}^{(k)} + (\mathcal{I}^{(k)})^{-1}\mathbf{G}_{\boldsymbol{\xi}}^{(k)},
\end{aligned}$$

where $\mathbf{V}^{(k)}$ is $\mathbf{V}$ evaluated at $(\boldsymbol{\beta}^{(k)\prime}, \boldsymbol{\xi}^{(k)\prime})'$,

$$\mathbf{G}_{\boldsymbol{\xi}}^{(k)} = \frac{1}{m}\sum_{i=1}^{m} \dot{\mathbf{p}}_{\boldsymbol{\xi}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}^{(k)}, \boldsymbol{\xi}^{(k)}),$$

and $\mathcal{I}^{(k)}$ is the sub Fisher information matrix for $\boldsymbol{\xi}$ evaluated at the k-th step. Here, the sub Fisher information matrix $\mathcal{I}$ is defined as

$$\mathcal{I} = -\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\{\ddot{\mathrm{p}}_{\boldsymbol{\xi}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi})\},$$

where the expectation is taken with respect to the distribution of $\mathbf{y}_i$. Each component of $\mathcal{I}$ has the following explicit expression

$$\mathcal{I}_{u,v} = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \xi_u}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \xi_v}\right),\ \text{for } u, v = 1, 2,$$

where

$$\frac{\partial \mathbf{V}}{\partial \xi_1} = \frac{\partial \mathbf{V}}{\partial \sigma^2} = (1-\rho)\mathbf{I} + \rho\mathbf{J} \quad \text{and} \quad \frac{\partial \mathbf{V}}{\partial \xi_2} = \frac{\partial \mathbf{V}}{\partial \rho} = \sigma^2(\mathbf{J} - \mathbf{I}).$$

The aforementioned Fisher scoring method requires iterative calculations. If it is applied to the full data, the computing time for each iteration is $\mathcal{O}(m(nd^2+n^2d))$. Thus, the whole time for the Fisher scoring algorithm to converge is $\mathcal{O}(\zeta m(nd^2 + n^2d))$, where ζ is the number of iterations. With a large number of subjects m, this algorithm becomes computationally expensive, and a subsampling algorithm helps reduce the computational cost.

3 Subsampling algorithms for longitudinal models

We first describe a general subsampling procedure. Let $\eta_i > 0$ be the subsampling probability for the i-th subject if one data point is taken from the full data $\mathcal{D}_m$. Assume that $\sum_{i=1}^{m}\eta_i = 1$, so that $\{\eta_i\}_{i=1}^{m}$ is a sampling distribution. Take a random subsample of size r with replacement from the full data according to $\{\eta_i\}_{i=1}^{m}$ and denote the subsampled data as $\{\mathbf{X}_i^*, \mathbf{y}_i^*, \eta_i^*\}_{i=1}^{r}$. For example, if $m = 10$, $r = 2$, and the selected data points are $\{\mathbf{X}_1, \mathbf{y}_1\}$ and $\{\mathbf{X}_7, \mathbf{y}_7\}$, then $\eta_1^* = \eta_1$ and $\eta_2^* = \eta_7$.

The subsample estimators $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$ are obtained by maximizing the following target function:

$$\mathrm{M}_r^*(\boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{1}{rm}\sum_{i=1}^{r}\frac{\mathrm{p}(\mathbf{X}_i^*, \mathbf{y}_i^*;\ \boldsymbol{\beta}, \boldsymbol{\xi})}{\eta_i^*}.$$

The maximization can be implemented by the Fisher scoring method described in section 2, where we iterate the following two steps until $(\tilde{\boldsymbol{\beta}}^{(k+1)\prime}, \tilde{\boldsymbol{\xi}}^{(k+1)\prime})'$

converges:

$$\tilde{\boldsymbol{\beta}}^{(k+1)} = \left(\sum_{i=1}^{r} \frac{\mathbf{X}_i^{*\prime} (\tilde{\mathbf{V}}^{(k)})^{-1} \mathbf{X}_i^{*}}{\eta_i^{*}} \right)^{-1} \sum_{i=1}^{r} \frac{\mathbf{X}_i^{*\prime} (\tilde{\mathbf{V}}^{(k)})^{-1} \mathbf{y}_i^{*}}{\eta_i^{*}},$$
$$\tilde{\boldsymbol{\xi}}^{(k+1)} = \tilde{\boldsymbol{\xi}}^{(k)} + (\mathcal{I}^{*(k)})^{-1} \mathbf{G}_{\boldsymbol{\xi}}^{*(k)},$$

where $\tilde{\mathbf{V}}^{(k)}$ is $\mathbf{V}$ evaluated at $(\tilde{\boldsymbol{\beta}}^{(k)'}, \tilde{\boldsymbol{\xi}}^{(k)'})'$,

$$\mathbf{G}_{\boldsymbol{\xi}}^{*(k)} = \frac{1}{rm} \sum_{i=1}^{r} \frac{\dot{\mathbf{p}}_{\boldsymbol{\xi}}(\mathbf{X}_i^{*}, \mathbf{y}_i^{*};\ \tilde{\boldsymbol{\beta}}^{(k)}, \tilde{\boldsymbol{\xi}}^{(k)})}{\eta_i^{*}},$$

and $\tilde{\mathcal{I}}^{(k)}$ is the approximated sub-matrix of the Fisher information matrix for $\boldsymbol{\xi}$ evaluated at the k-th step based on the subsample. The (u, v)-th element of $\tilde{\mathcal{I}}^{(k)}$ is

$$\tilde{\mathcal{I}}_{u,v}^{(k)} = \frac{1}{2rm} \mathrm{tr} \left((\tilde{\mathbf{V}}^{(k)})^{-1} \frac{\partial \tilde{\mathbf{V}}^{(k)}}{\partial \xi_u} (\tilde{\mathbf{V}}^{(k)})^{-1} \frac{\partial \tilde{\mathbf{V}}^{(k)}}{\partial \xi_v} \right) \sum_{i=1}^{r} \frac{1}{\eta_i^{*}}, \text{ for } u, v = 1, 2,$$

with

$$\frac{\partial \tilde{\mathbf{V}}^{(k)}}{\partial \xi_1} = (1 - \rho^{(k)}) \mathbf{I} + \rho^{(k)} \mathbf{J} \quad \text{and} \quad \frac{\partial \tilde{\mathbf{V}}^{(k)}}{\partial \xi_2} = {\sigma^{(k)}}^2 (\mathbf{J} - \mathbf{I}).$$

The subsampling probabilities play an important role in selecting informative subsamples, and are crucial for the subsample estimator $(\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$ to better approximate the full data estimator. In the following, we discuss different approaches of assigning subsampling probabilities.

1. **Uniform subsampling method**: The subsampling probabilities η_i's are equal, i.e., $\eta_i^{\text{uni}} = 1/m$ for all the subjects. This is the simplest approach to specifying the sampling distribution.
2. **Leverage-based subsampling method**: Set $\eta_i^{\text{lev}} = \sqrt{\mathrm{tr}(\mathbf{H}_{ii})} / \sum_{i=1}^{m} \sqrt{\mathrm{tr}(\mathbf{H}_{ii})}$, where $\mathbf{H}_{ii} = \mathbf{X}_i (\mathcal{X}'\mathcal{X}) \mathbf{X}_i'$ is the diagonal sub-matrix of the hat matrix $\mathcal{X}(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'$ corresponding to $\mathbf{X}_i$. This method uses the information from the covariates matrix $\mathcal{X}$ when assigning the subsampling probabilities.
3. **Gradient-based subsampling method**: This method is a two-step procedure, where we first use the uniform subsampling method to select a subsample of size r_0 and obtain pilot estimates $\boldsymbol{\beta}_0$ and $\boldsymbol{\xi}_0$ for $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$, respectively. With the pilot estimates, we calculate the gradient vector for the i-th subject as

$$\dot{\mathbf{p}}_i^0 = \dot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}_0, \boldsymbol{\xi}_0).$$

The subsampling probabilities in the second step are assigned as $\eta_i^{\text{gradient}} = \|\dot{\mathbf{p}}_i^0\| / \sum_{i=1}^{m} \|\dot{\mathbf{p}}_i^0\|$ for $i = 1, ..., m$. When calculating the subsample estimator $(\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$, we combine the subsamples from both steps. Besides using

the information from the covariates, this method also includes information from the responses.

The leverage-based and gradient-based subsampling probabilities described above reduce to the widely used subsampling probabilities in least square problems with independent univariate responses Drineas et al (2006, 2010, 2012); Yang et al (2015), if $n = 1$ in model (1) of section 1.

In addition to the aforementioned methods, motivated by the idea of optimal subsampling using the A-optimality criterion introduced in Wang et al (2018), we also derive the optimal subsampling probabilities. This method assigns subsampling probabilities that minimize the asymptomatic mean squared error of $(\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$ in approximating $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\xi}}')'$.

Before providing the optimal subsampling probabilities, we first give the asymptotic distribution of the subsample estimator $(\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$ under a general subsampling distribution $\{\eta_i\}_{i=1}^m$. We need the following assumptions on the full data and the general subsampling probabilities η_i's.

Assumption 1 $m^{-2}\sum_{i=1}^m \eta_i^{-1}\|\mathbf{x}_{i,j}\|^4 = \mathcal{O}_p(1)$, for $1 \le j \le d$.

Assumption 2 $m^{-2}\sum_{i=1}^m \eta_i^{-1}\|\boldsymbol{e}_i\|^4 = \mathcal{O}_p(1)$, where $\boldsymbol{e}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$.

Assumption 3 There exists $\delta > 0$ such that $m^{-(2+\delta)}\sum_{i=1}^m \eta_i^{-1-\delta}|\mathbf{x}'_{i,j}\boldsymbol{e}_i|^{2+\delta} = \mathcal{O}_p(1)$, for $1 \le j \le d$.

Assumption 4 There exists $\delta > 0$ such that $m^{-(2+\delta)}\sum_{i=1}^m \eta_i^{-1-\delta}\|\boldsymbol{e}_i\|^{4+2\delta} = \mathcal{O}_p(1)$.

Assumption 5 $\ddot{\mathbf{M}}_m(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})$ approaches a positive-definite matrix in probability as $m \to \infty$.

Remark 1: If the design matrix $\mathbf{X}_i$ in equation (1) does not include an intercept term, we will need an additional assumption that $m^{-1}\sum_{i=1}^m (m\eta_i)^{-1} = \mathcal{O}_p(1)$.

Theorem 1 *Under Assumptions 1-5, as $m \to \infty$ and $r \to \infty$, conditionally on the full data $\mathcal{D}_m$, the subsample estimator $(\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$ satisfies that,*

$$\boldsymbol{\Sigma}^{-1/2}\begin{pmatrix}\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{\xi}} - \hat{\boldsymbol{\xi}}\end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{6}$$

where $\xrightarrow{\mathcal{L}}$ means converges in law,

$$\boldsymbol{\Sigma} = \left(\ddot{\mathbf{M}}_m(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})\right)^{-1} \boldsymbol{\Sigma}_c \left(\ddot{\mathbf{M}}_m(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})\right)^{-1},$$

$$\ddot{\mathbf{M}}_m(\boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{1}{m}\sum_{i=1}^m \ddot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi}),$$

$$\boldsymbol{\Sigma}_c = \frac{1}{rm^2}\sum_{i=1}^{m}\frac{1}{\eta_i}\dot{\mathbf{p}}_i\dot{\mathbf{p}}_i',$$

and $\dot{\mathbf{p}}_i = \dot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \hat{\beta}, \hat{\boldsymbol{\xi}})$.

Theorem 1 indicates that the conditional asymptotic mean squared error (AMSE) of $(\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$, given the full data, is equal to $\mathrm{tr}(\boldsymbol{\Sigma})$. Since $\boldsymbol{\Sigma}$ depends on the subsampling distribution $\{\eta_i\}_{i=1}^m$, we want to find the subsampling probability distribution that minimizes the AMSE $\mathrm{tr}(\boldsymbol{\Sigma})$. This corresponds to the A-optimality criterion in the optimal design of experiments. We call the subsampling probabilities obtained from using this criterion the optimal subsampling probabilities.

Theorem 2 *The optimal subsampling probabilities that minimize the AMSE* $tr(\boldsymbol{\Sigma})$ *are*

$$\eta_i^{\mathrm{opt}} = \frac{\left\|\ddot{\mathbf{M}}_m^{-1}(\hat{\beta}, \hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\right\|}{\sum_{j=1}^{m}\left\|\ddot{\mathbf{M}}_m^{-1}(\hat{\beta}, \hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_j\right\|},\ \textit{for } i = 1, ..., m. \tag{7}$$

In addition to minimizing the conditional AMSE of the full parameter vector $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$, we also consider a more general setting. Suppose we are interested in approximating a linear transformation, say $\mathbf{T}$, of the full data MLE $\hat{\boldsymbol{\theta}}$, i.e., $\mathbf{T}\hat{\boldsymbol{\theta}}$. Since equation (6) implies that the conditional asymptomatic variance-covariance matrix of $\tilde{\boldsymbol{\theta}}$ is $\boldsymbol{\Sigma}$, so the conditional asymptomatic variance-covariance matrix of $\mathbf{T}\tilde{\boldsymbol{\theta}}$ is $\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}'$, indicating that the conditional AMSE of $\mathbf{T}\tilde{\boldsymbol{\theta}}$ given the full data is $\mathrm{tr}(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}')$. To minimize $\mathrm{tr}(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}')$, we need to adjust our optimal subsampling probabilities in equation (7). This alternative criterion corresponds to the L-optimality in optimal design of experiments. We call the subsampling probabilities obtained from using this criterion the L-optimal subsampling probabilities, and their expressions are presented in the following Theorem.

Theorem 3 *The L-optimal subsampling probabilities that minimize the AMSE* $tr(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}')$ *are*

$$\eta_i^{\mathrm{Lopt}} = \frac{\left\|\mathbf{T}\ddot{\mathbf{M}}_m^{-1}(\hat{\beta}, \hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\right\|}{\sum_{j=1}^{m}\left\|\mathbf{T}\ddot{\mathbf{M}}_m^{-1}(\hat{\beta}, \hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_j\right\|},\ \textit{for } i = 1, ..., m. \tag{8}$$

The L-optimal subsampling probabilities are particularly useful when our primary interests involve only part of $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\xi}}')'$. For example, if $\hat{\boldsymbol{\xi}}$ is the only parameter of interest, we can use the above L-optimal subsampling probabilities by setting $\mathbf{T} = (\mathbf{c}_{d+1}, \mathbf{c}_{d+2})'$, where $\mathbf{c}_i$ is the i-th canonical basis vector.

Note that η_i^{opt} and η_i^{Lopt} in equations (7) and (8) depend on the full data MLE, so they are not directly available. In practice, we can approximate η_i^{opt}

and η_i^{Lopt} by $\tilde{\eta}_i^{\text{opt}}$ and $\tilde{\eta}_i^{\text{opt}}$ using a pilot sample as given in equations (9) and (10).

$$\tilde{\eta}_i^{\text{opt}} = \frac{\left\|\ddot{\mathbf{M}}_m^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\xi}_0)\dot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}_0, \boldsymbol{\xi}_0)\right\|}{\sum_{j=1}^m \left\|\ddot{\mathbf{M}}_m^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\xi}_0)\dot{\mathbf{p}}(\mathbf{X}_j, \mathbf{y}_j;\ \boldsymbol{\beta}_0, \boldsymbol{\xi}_0)\right\|},\ \text{for } i = 1, ..., m, \tag{9}$$

$$\tilde{\eta}_i^{\text{Lopt}} = \frac{\left\|\mathbf{T}\ddot{\mathbf{M}}_m^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\xi}_0)\dot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}_0, \boldsymbol{\xi}_0)\right\|}{\sum_{j=1}^m \left\|\mathbf{T}\ddot{\mathbf{M}}_m^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\xi}_0)\dot{\mathbf{p}}(\mathbf{X}_j, \mathbf{y}_j;\ \boldsymbol{\beta}_0, \boldsymbol{\xi}_0)\right\|},\ \text{for } i = 1, ..., m, \tag{10}$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\xi}_0$ are estimates based on the pilot sample.

The approximated subsampling probabilities $\tilde{\eta}_i^{\text{opt}}$ and $\tilde{\eta}_i^{\text{Lopt}}$ are subject to additional disturbance, which may inflate the asymptotic variance of the resulting estimator especially for small η_i^{opt} or η_i^{Lopt} (they are in the denominator of $\boldsymbol{\Sigma}_c$). To handle this, we propose a more practical approach and mix the approximated optimal subsampling probabilities with the uniform subsampling probability to protect the estimator from these data points. Specifically, use

$$\tilde{\eta}_{\alpha,i}^{\text{opt}} = (1-\alpha)\tilde{\eta}_i^{\text{opt}} + \alpha\frac{1}{m};\ \ \tilde{\eta}_{\alpha,i}^{\text{Lopt}} = (1-\alpha)\tilde{\eta}_i^{\text{Lopt}} + \alpha\frac{1}{m}, \tag{11}$$

where $\alpha \in (0, 1)$.

Remark 2: For subsampling the probabilities given in equation (11) to satisfy Assumption 1-4, we only need the condition given in equation (12) below. The proofs are in Appendix E.

$$\text{For some } \delta > 0, \frac{1}{m}\sum_{i=1}^m \|\mathbf{x}_{i,j}\|^{4+\delta} = \mathcal{O}_p(1),\ \text{for } j = 1, \ldots, d. \tag{12}$$

We summarize the practical implementation in a two-step procedure below.

4. **Optimal subsampling methods**: In Step 1, use the uniform subsampling method to select a pilot subsample of size r_0 and use it to obtain pilot estimates $\boldsymbol{\beta}_0$ and $\boldsymbol{\xi}_0$. Use equation (11) to get the approximated optimal subsampling distribution $\{\tilde{\eta}_{\alpha,i}^{\text{opt}}\}_{i=1}^m$ or $\{\tilde{\eta}_{\alpha,i}^{\text{Lopt}}\}_{i=1}^m$, respectively. In Step 2, sample with replacement to obtain another subsample of size r using $\{\tilde{\eta}_{\alpha,i}^{\text{opt}}\}_{i=1}^m$ or $\{\tilde{\eta}_{\alpha,i}^{\text{Lopt}}\}_{i=1}^m$, obtained in Step 1. The subsample estimators $(\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$ are calculated using the combined subsamples from both steps.

4 Simulation Results and Analysis

In this section, we use simulated data to evaluate and compare the performance of different subsampling methods. We set the total number of subjects to be $m = 5000$, the number of measurements for each subject to be $n = 4$. We generated data from model (1) by setting the true value of $\boldsymbol{\beta}$ to be a 5×1 vector such that $\boldsymbol{\beta} = (50, 75, 100, 125, 150)'$. The random error vectors

$\boldsymbol{\varepsilon}_i$'s were generated from a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{V} = \sigma^2 ((1-\rho)\mathbf{I}_4 + \rho\mathbf{J}_4)$ with $\sigma^2 = 2$ and $\rho = 0.6$.

To evaluate the effect of the covariates on the subsampling methods, we considered four different distributions when generating $\mathbf{X}_i$'s. Let $\boldsymbol{\Omega}_\mathbf{x} = \mathbf{I}_5 \otimes \boldsymbol{\Omega}$, where $\otimes$ is the Kronecker product and $\boldsymbol{\Omega}$ is a square matrix with $\Omega_{u,v} = 2 \times 0.6^{I(u \neq v)}$ for $u, v = 1, ..., 4$. The following distributions of vec$(\mathbf{X}_i)$'s were used to generate $\mathbf{X}_i$'s.

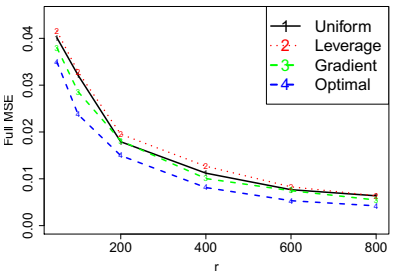
1. **Multivariate Normal**: $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_\mathbf{x})$.
2. **Mixture Normal**: $1/2\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_\mathbf{x}) + 1/2\mathcal{N}(\mathbf{0}, 4\boldsymbol{\Omega}_\mathbf{x})$.
3. $\mathbf{T_3}$: A multivariate t distribution with degrees of freedom 3, $\mathbf{T_3}(\mathbf{0}, \boldsymbol{\Omega}_\mathbf{x})$.
4. $\mathbf{T_2}$: A multivariate t distribution with degrees of freedom 2, $\mathbf{T_2}(\mathbf{0}, \boldsymbol{\Omega}_\mathbf{x})$.

To evaluate the performance of different subsampling methods, we first calculated the full data MLE, denoted as $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\xi}}')'$. For each subsampling procedure, we calculated the empirical MSEs of $(\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$ from 500 repetitions of the simulation using

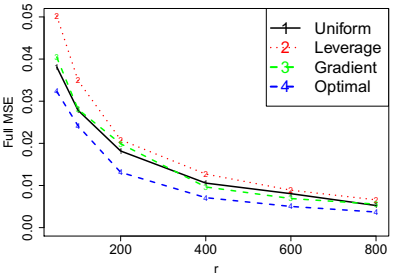
$$\text{MSE} = \frac{1}{500} \sum_{s=1}^{500} \left\| (\tilde{\boldsymbol{\beta}}^{(s)\prime}, \tilde{\boldsymbol{\xi}}^{(s)\prime})' - (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\xi}}')' \right\|^2,$$

where $(\tilde{\boldsymbol{\beta}}^{(s)\prime}, \tilde{\boldsymbol{\xi}}^{(s)\prime})'$ is the subsample estimate in the s-th repetition. For the gradient-based subsampling and the optimal subsampling methods, we set the first step sample size $r_0 = 80$, and set the second step subsample sizes to $r = 50, 100, 200, 400, 600$, and 800. For fair comparisons, the subsample size for uniform and leverage subsampling methods were set to be $r_0 + r$.

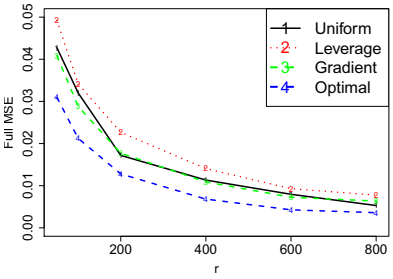
Figure 1 presents the MSEs of $(\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$ from different subsampling methods. The MSEs from the optimal subsampling method are the smallest among all the subsampling methods. This agrees with the theoretical result which shows that the optimal subsampling method minimizes the AMSE of $(\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$.



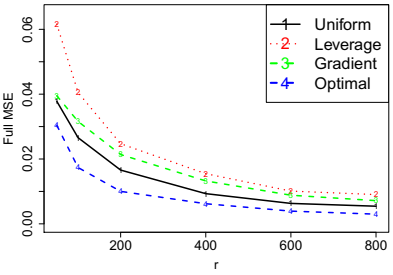
(a) Multivariate Normal



(b) Mixture Normal



(c) T_3

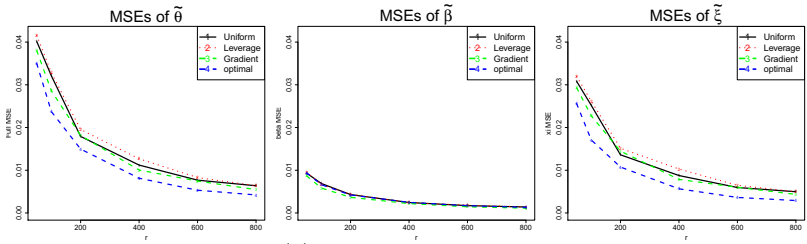


(d) T_2

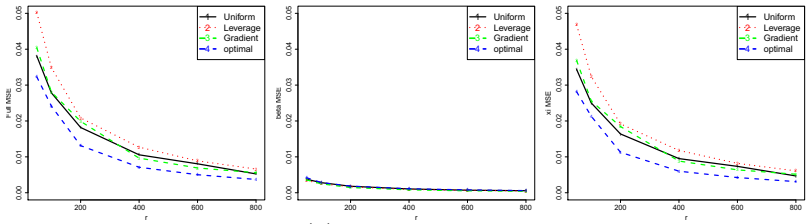
Fig. 1: MSEs of $(\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$ for different second step subsample sizes r, with the first step subsample size being fixed at $r_0 = 80$.

In addition to the results shown in Figure 1 where the MSEs are for the entire parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\xi}')'$, we also calculated the MSEs for $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\xi}}$ separately. The three columns in Figure 2 are for the MSEs of $\tilde{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\beta}}$, and $\tilde{\boldsymbol{\xi}}$, respectively. For a fair comparison, we use the same scale for their y-axes. To make the difference in $\tilde{\boldsymbol{\beta}}$ more clear, we have also created Figure 3 that contains MSEs for $\tilde{\boldsymbol{\beta}}$ only.

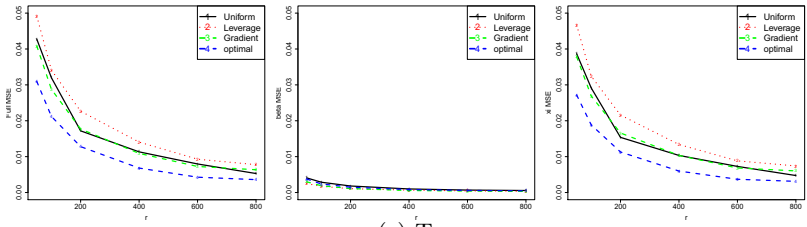
It is seen that the MSEs of $\tilde{\boldsymbol{\theta}}$ are dominated by the contribution from the covariance parameter estimator $\tilde{\boldsymbol{\xi}}$. Thus, a method that performs well in approximating $\hat{\boldsymbol{\xi}}$ usually has an overall better performance. Also, there appears to be a trade-off between the MSEs of $\tilde{\boldsymbol{\beta}}$ and the MSEs of $\tilde{\boldsymbol{\xi}}$; subsampling methods yielding smaller MSEs of $\tilde{\boldsymbol{\beta}}$ (leverage-based and gradient-based methods) tend to have larger MSEs of $\tilde{\boldsymbol{\xi}}$.



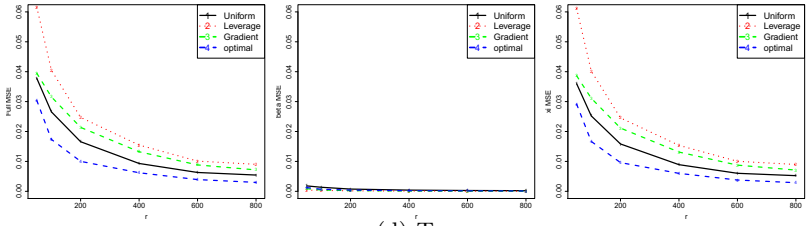
(a) Multivariate Normal



(b) Mixture Normal

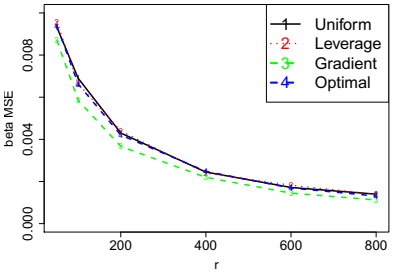


(c) T_3

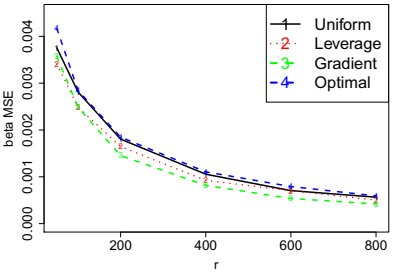


(d) T_2

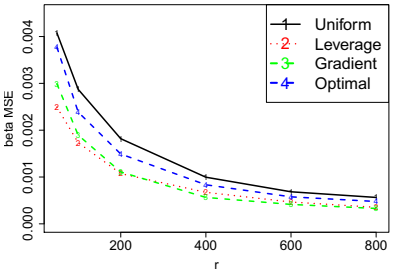
Fig. 2: Decomposition of MSEs of $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\xi}}')'$.



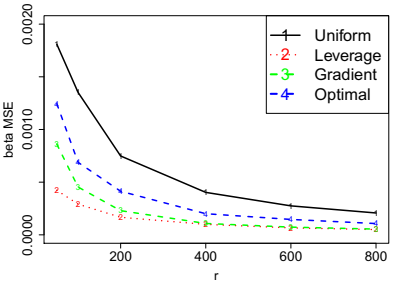
(a) Multivariate Normal



(b) Mixture Normal



(c) T_3

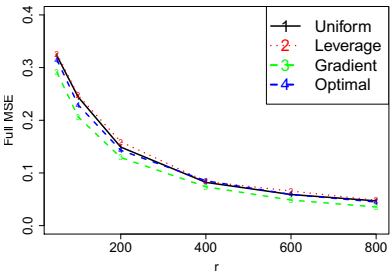


(d) T_2

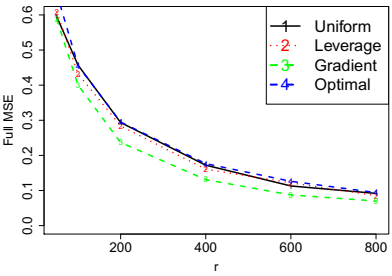
Fig. 3: MSEs of $\tilde{\boldsymbol{\beta}}$ for different second step subsample sizes r, with the first step subsample size being fixed at $r_0 = 80$.

The trade-off between the MSEs of $\tilde{\boldsymbol{\beta}}$ and the MSEs of $\tilde{\boldsymbol{\xi}}$ can help us choose which subsampling method to use in practice. For example, if the primary interest is to approximate $\hat{\boldsymbol{\beta}}$ as precisely as possible and $\hat{\boldsymbol{\xi}}$ is not of interest, then the widely used leverage-based subsampling method or the gradient-based subsampling method are appropriate. However, they are not recommended if better approximation of $\hat{\boldsymbol{\xi}}$ is of interest.

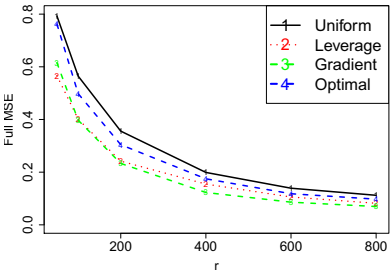
Also, from Theorem 3, we see that the gradient subsampling method minimizes the AMSE of $\mathbf{T}\tilde{\boldsymbol{\theta}}$ with $\mathbf{T} = \ddot{\mathbf{M}}_m(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})$. To verify this, we have also calculated the MSEs of $\ddot{\mathbf{M}}_m(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})\tilde{\boldsymbol{\theta}}$ in Figure 4. It is seen that the gradient method has the smallest MSEs for all the four cases.



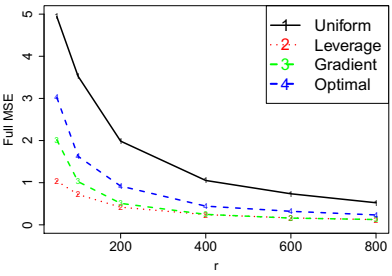
(a) Multivariate Normal



(b) Mixture Normal



(c) T_3



(d) T_2

Fig. 4: MSEs of $\ddot{\mathbf{M}}_m\tilde{\boldsymbol{\theta}}$ with $r_0 = 80$ and different r.

5 Discussion

In this paper, we discussed several subsampling algorithms in longitudinal models with a balanced design. In order to get a better subsample estimators of the vector of the model parameters, we proposed the optimal subsampling probabilities and provided a two-step procedure for practical implementation. Furthermore, we conducted numerical experiments to compare the performances of different subsampling methods, which confirmed the theoretical result that the optimal subsampling method should yield a better approximation to the full data MLE. We also discussed how to adjust the optimal subsampling probabilities if the primary interest of is a linear transformation of the entire parameter vector.

In this paper, we have assumed a balanced design. However, because of the covariance structure we imposed on the error terms, our subsampling algorithms can be easily extended to the scenario with an unbalanced design or missing values. For the compound symmetric correlation structure, as we can see from (2), the correlations between any two observations on the same subject are the same. With this property, although unbalanced designs or missing

values will lead to different correlation matrices $\mathbf{V}_i$'s for different subjects, the difference depends on i only through n_i. Specifically, for a subject with n_i observations, the covariance is

$$\mathbf{V}_i(\boldsymbol{\xi}) = \sigma^2 \left((1-\rho)\mathbf{I}_{n_i} + \rho \mathbf{J}_{n_i}\right), \boldsymbol{\xi} = (\sigma^2, \rho)'.$$

In this case, to calculate the optimal subsampling probabilities and perform the Fisher scoring algorithm, we can simply replace $\mathbf{V}$ and n by $\mathbf{V}_i$ and n_i in the corresponding equations.

The above extensions to unbalanced designs and the situation with missing values rely on the special correlation structure we have assumed. In general, extensions to more complex designs require more assumptions. For example, if we assume the covariance structure for the error terms to be auto-regressive of order one (AR(1)), we will need to assume that the observation times for each subject are discrete and equally-spaced time points. For more complicated designs, for example, when different subjects have different measurement times, the model we have specified in (1) may not be appropriate. Since these designs are also common in practice, future research to extend the optimal subsampling method to account for more complex designs is important.

Acknowledgments. The authors are grateful to an anonymous referee whose insightful comments greatly helped enhancing the paper. HaiYing Wang's research was supported by NSF grant CCF-2105571.

Declarations

- Competing interests: The authors have no competing interests to declare that are relevant to the content of this article.
- Availability of data and materials: The data that support the findings of this study are available on request from the corresponding author.

Appendix A Expressions of $\dot{\mathbf{p}}$ and $\ddot{\mathbf{p}}$

$$\dot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi}) = \begin{bmatrix} \frac{\mathbf{X}_i'\left(\mathbf{I} - \frac{\rho}{1+(n-1)\rho}\mathbf{J}\right)(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})}{\sigma^2(1-\rho)} \\ \frac{(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\left(\mathbf{I} - \frac{\rho}{1+(n-1)\rho}\mathbf{J}\right)(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})}{2\sigma^4(1-\rho)} - \frac{n}{2\sigma^2} \\ \frac{(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\left(\frac{1+(n-1)\rho^2}{(1+(n-1)\rho)^2}\mathbf{J} - \mathbf{I}\right)(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})}{2\sigma^2(1-\rho)^2} + \frac{n(n-1)\rho}{2(1-\rho)(1+(n-1)\rho)} \end{bmatrix}.$$

From the above, we have

$$\|\dot{\mathbf{p}}_i\| = \left(\frac{\left\| \mathbf{X}_i'\left(\mathbf{I} - \frac{\hat{\rho}}{1+(n-1)\hat{\rho}}\mathbf{J}\right)\boldsymbol{e}_i \right\|^2}{\hat{\sigma}^4(1-\hat{\rho})^2} + \frac{\left(\boldsymbol{e}_i'\left(\mathbf{I} - \frac{\hat{\rho}}{1+(n-1)\hat{\rho}}\mathbf{J}\right)\boldsymbol{e}_i\right)^2}{4\hat{\sigma}^8(1-\hat{\rho})^2} \right.$$

$$
\begin{aligned}
&- \frac{n\left(\boldsymbol{e}_i'\left(\mathbf{I}-\frac{\hat{\rho}}{1+(n-1)\hat{\rho}}\mathbf{J}\right)\boldsymbol{e}_i\right)}{2\hat{\sigma}^6(1-\hat{\rho})} + \frac{n^2}{4\hat{\sigma}^4} \\
&+ \frac{\left(\boldsymbol{e}_i'\left(\frac{1+\rho^2(n-1)}{(1+(n-1)\rho)^2}\mathbf{J}-\mathbf{I}\right)\boldsymbol{e}_i\right)^2}{4\hat{\sigma}^4(1-\hat{\rho})^4} + \frac{n(n-1)\hat{\rho}\left(\boldsymbol{e}_i'\left(\frac{1+\hat{\rho}^2(n-1)}{(1+(n-1)\rho)^2}\mathbf{J}-\mathbf{I}\right)\boldsymbol{e}_i\right)}{2\hat{\sigma}^2(1-\hat{\rho})^3(1+(n-1)\hat{\rho})} \\
&+ \frac{n^2(n-1)^2\hat{\rho}}{4(1-\hat{\rho})^2(1+(n-1)\hat{\rho})^2}\Bigg)^{1/2}.
\end{aligned}
$$

Hence, $\|\dot{\mathbf{p}}_i\|$ is bounded above and below such that

$$
\|\dot{\mathbf{p}}_i\| \geq \frac{\left\|\mathbf{X}_i'\left(\mathbf{I}-\frac{\hat{\rho}}{1+(n-1)\hat{\rho}}\mathbf{J}\right)\boldsymbol{e}_i\right\|}{\hat{\sigma}^2(1-\hat{\rho})},
$$

and

$$
\begin{aligned}
\|\dot{\mathbf{p}}_i\| \leq \Bigg(&\frac{\left\|\mathbf{X}_i'\left(\mathbf{I}-\frac{\hat{\rho}}{1+(n-1)\hat{\rho}}\mathbf{J}\right)\boldsymbol{e}_i\right\|^2}{\hat{\sigma}^4(1-\hat{\rho})^2} + \frac{\left(\boldsymbol{e}_i'\left(\mathbf{I}-\frac{\hat{\rho}}{1+(n-1)\hat{\rho}}\mathbf{J}\right)\boldsymbol{e}_i\right)^2}{4\hat{\sigma}^8(1-\hat{\rho})^2} \\
&+ \frac{n^2}{4\hat{\sigma}^4} + \frac{\left(\boldsymbol{e}_i'\left(\frac{1+\rho^2(n-1)}{(1+(n-1)\rho)^2}\mathbf{J}-\mathbf{I}\right)\boldsymbol{e}_i\right)^2}{4\hat{\sigma}^4(1-\hat{\rho})^4} + \frac{n^2(n-1)^2\hat{\rho}}{4(1-\hat{\rho})^2(1+(n-1)\hat{\rho})^2}\Bigg)^{1/2}.
\end{aligned} \tag{A1}
$$

$$
\ddot{\mathbf{p}}(\mathbf{X}_i, \mathbf{y}_i;\ \boldsymbol{\beta}, \boldsymbol{\xi}) = \begin{bmatrix} \frac{\partial^2 \mathrm{p}_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \frac{\partial^2 \mathrm{p}_i}{\partial\boldsymbol{\beta}\partial\sigma^2} & \frac{\partial^2 \mathrm{p}_i}{\partial\boldsymbol{\beta}\partial\rho} \\ \frac{\partial^2 \mathrm{p}_i}{\partial\sigma^2\partial\boldsymbol{\beta}'} & \frac{\partial^2 \mathrm{p}_i}{\partial\sigma^4} & \frac{\partial^2 \mathrm{p}_i}{\partial\sigma^2\partial\rho} \\ \frac{\partial^2 \mathrm{p}_i}{\partial\rho\partial\boldsymbol{\beta}'} & \frac{\partial^2 \mathrm{p}_i}{\partial\rho\partial\sigma^2} & \frac{\partial^2 \mathrm{p}_i}{\partial\rho^2} \end{bmatrix},
$$

where

$$
\begin{aligned}
&\frac{\partial^2 \mathrm{p}_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = -\frac{\mathbf{X}_i'\left(\mathbf{I}-\frac{\rho}{1+(n-1)\rho}\mathbf{J}\right)\mathbf{X}_i}{\sigma^2(1-\rho)}, \quad \frac{\partial^2 \mathrm{p}_i}{\partial\boldsymbol{\beta}\partial\sigma^2} = -\frac{\mathbf{X}_i'\left(\mathbf{I}-\frac{\rho}{1+(n-1)\rho}\mathbf{J}\right)(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})}{\sigma^4(1-\rho)}, \\
&\frac{\partial^2 \mathrm{p}_i}{\partial\boldsymbol{\beta}\partial\rho} = -\frac{\mathbf{X}_i'\left(\frac{1+(n-1)\rho^2}{(1+(n-1)\rho)^2}\mathbf{J}-\mathbf{I}\right)(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})}{\sigma^2(1-\rho)^2}, \\
&\frac{\partial^2 \mathrm{p}_i}{\partial\sigma^4} = -\frac{(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\left(\mathbf{I}-\frac{\rho}{1+(n-1)\rho}\mathbf{J}\right)(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})}{\sigma^6(1-\rho)} + \frac{n}{2\sigma^4}, \\
&\frac{\partial^2 \mathrm{p}_i}{\partial\sigma^2\partial\rho} = \frac{(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\left(\frac{1+(n-1)\rho^2}{(1+(n-1)\rho)^2}\mathbf{J}-\mathbf{I}\right)(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})}{2\sigma^4(1-\rho)^2}, \\
&\frac{\partial^2 \mathrm{p}_i}{\partial\rho^2} = \frac{(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\left(\mathbf{I}-\left[\frac{n-2\rho}{(1+(n-1)\rho)^2}-\frac{2-\rho}{1+(n-1)\rho}-\frac{\rho n^2}{(1+(n-1)\rho)^3}\right]\mathbf{J}\right)(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})}{\sigma^2(1-\rho)^3}
\end{aligned}
$$

$$- \frac{n(n-1)(1+(n-1)\rho^2)}{2(1-\rho)^2((1+(n-1)\rho))^2}\,.$$

Appendix B Proof of Theorem 1

We first introduce additional notations that will be used later:

$$\dot{\mathbf{M}}_m(\boldsymbol{\beta},\boldsymbol{\xi}) = \frac{1}{m}\sum_{i=1}^{m} \dot{\mathbf{p}}(\mathbf{X}_i,\mathbf{y}_i;\ \boldsymbol{\beta},\boldsymbol{\xi}),$$

$$\dot{\mathbf{M}}^*_r(\boldsymbol{\beta},\boldsymbol{\xi}) = \frac{1}{r}\sum_{i=1}^{r} \frac{\dot{\mathbf{p}}(\mathbf{X}^*_i,\mathbf{y}^*_i,\boldsymbol{\beta},\boldsymbol{\xi})}{m\eta^*_i},\quad \ddot{\mathbf{M}}^*_r(\boldsymbol{\beta},\boldsymbol{\xi}) = \frac{1}{r}\sum_{i=1}^{r} \frac{\ddot{\mathbf{p}}(\mathbf{X}^*_i,\mathbf{y}^*_i,\boldsymbol{\beta},\boldsymbol{\xi})}{m\eta^*_i}\,.$$

We begin by establishing a lemma that will be used in the proof of Theorem 1. Recall that $\mathcal{D}_m$ denotes the full data.

Lemma 1

$$\ddot{\mathbf{M}}^*_r(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) - \ddot{\mathbf{M}}_m(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) = \mathcal{O}_{p|\mathcal{D}_m}(r^{-1/2}) = o_{p|\mathcal{D}_m}(1)\,.$$

Proof By direct calculation, we have

$$\mathbb{E}[\ddot{\mathbf{M}}^*_r(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) \mid \mathcal{D}_m] = \ddot{\mathbf{M}}_m(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\,.$$

Let $\ddot{\mathrm{M}}^{*j_1j_2}_r(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})$ be the (j_1,j_2)-th entry of the matrix $\ddot{\mathbf{M}}^*_r(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})$, $\ddot{\mathrm{M}}^{j_1j_2}_m(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})$ be the (j_1,j_2)-th entry of the matrix $\ddot{\mathbf{M}}_m(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})$ and $\hat{\mathbf{V}}$ be $\mathbf{V}$ evaluated at $\hat{\boldsymbol{\xi}}$.

For $1 \le j_1, j_2 \le d$,

$$\begin{aligned}
\mathbb{V}(\ddot{\mathrm{M}}^{*j_1,j_2}_r \mid \mathcal{D}_m) &= \frac{1}{r}\sum_{i=1}^{m}\eta_i\left(-\frac{\mathbf{x}'_{i,j_1}\hat{\mathbf{V}}^{-1}\mathbf{x}_{i,j_2}}{m\eta_i} - \ddot{\mathbf{M}}^{j_1,j_2}_m\right)^2 \\
&= \frac{1}{rm^2}\sum_{i=1}^{m}\frac{\left(-\mathbf{x}'_{i,j_1}\hat{\mathbf{V}}^{-1}\mathbf{x}_{i,j_2}\right)^2}{\eta_i} - \frac{1}{r}\left(\ddot{\mathbf{M}}^{j_1,j_2}_m\right)^2 \\
&\le \frac{1}{rm^2}\sum_{i=1}^{m}\frac{\left(\mathbf{x}'_{i,j_1}\hat{\mathbf{V}}^{-1}\mathbf{x}_{i,j_2}\right)^2}{\eta_i} \\
&= \mathcal{O}_p(r^{-1})\,,
\end{aligned}$$

where the last equality is from Assumption 1.

For $1 \le j_1 \le d$, $j_2 = d+1$,

$$\begin{aligned}
\mathbb{V}(\ddot{\mathbf{M}}^{*j_1j_2}_r(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) \mid \mathcal{D}_m) &= \frac{1}{r}\sum_{i=1}^{m}\eta_i\left(-\frac{\mathbf{x}'_{i,j_1}\hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}}{\partial\sigma^2}\hat{\mathbf{V}}^{-1}\boldsymbol{e}_i}{m\eta_i} - \ddot{\mathbf{M}}^{j_1,j_2}_m(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\right)^2 \\
&= \frac{1}{rm^2}\sum_{i=1}^{m}\frac{\left(-\mathbf{x}'_{i,j_1}\hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}}{\partial\sigma^2}\hat{\mathbf{V}}^{-1}\boldsymbol{e}_i\right)^2}{\eta_i} - \frac{1}{r}\left(\ddot{\mathbf{M}}^{j_1,j_2}_m(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\right)^2
\end{aligned}$$

$$\leq \frac{1}{rm^2}\sum_{i=1}^{m}\frac{\left(\mathbf{x}'_{i,j_1}\hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}}{\partial\sigma^2}\hat{\mathbf{V}}^{-1}\boldsymbol{e}_i\right)^2}{\eta_i}$$
$$= \mathcal{O}_p(r^{-1}),$$

where the last equality is from Assumptions 1 and 2, and by using the Cauchy-Schwarz inequality.

For $1 \leq j_1 \leq d$, $j_2 = d+2$,

$$\mathbb{V}(\ddot{\mathbf{M}}_r^{*j_1j_2}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) \mid \mathcal{D}_m) = \frac{1}{r}\sum_{i=1}^{m}\eta_i\left(-\frac{\mathbf{x}'_{i,j_1}\hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}}{\partial\rho}\hat{\mathbf{V}}^{-1}\boldsymbol{e}_i}{m\eta_i} - \ddot{\mathbf{M}}_m^{j_1,j_2}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\right)^2$$
$$= \frac{1}{rm^2}\sum_{i=1}^{m}\frac{\left(-\mathbf{x}'_{i,j_1}\hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}}{\partial\rho}\hat{\mathbf{V}}^{-1}\boldsymbol{e}_i\right)^2}{\eta_i} - \frac{1}{r}\left(\ddot{\mathbf{M}}_m^{j_1,j_2}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\right)^2$$
$$\leq \frac{1}{rm^2}\sum_{i=1}^{m}\frac{\left(\mathbf{x}'_{i,j_1}\hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}}{\partial\rho}\hat{\mathbf{V}}^{-1}\boldsymbol{e}_i\right)^2}{\eta_i}$$
$$= \mathcal{O}_p(r^{-1}),$$

where the last equality is from Assumptions 1 and 2, and by using the Cauchy-Schwarz inequality. Since $\ddot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})$ is symmetric, its (j_1,j_2)-th entry when $d+1 \leq j_1 \leq d+2$, $1 \leq j_2 \leq d$ is also $O_p(r^{-1})$.

For $d+1 \leq j_1, j_2 \leq d+2$,

$$\mathbb{V}(\ddot{\mathbf{M}}_r^{*j_1j_2}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) \mid \mathcal{D}_m)$$
$$= \frac{1}{r}\sum_{i=1}^{m}\eta_i\left(-\frac{2\boldsymbol{e}_i'(\mathbf{D}_{j_2}\mathbf{D}_{j_1}-\dot{\mathbf{D}})\hat{\mathbf{V}}^{-1}\boldsymbol{e}_i - \mathrm{tr}\left(\mathbf{D}_{j_1}\mathbf{D}_{j_2}-\dot{\mathbf{D}}\right)}{2m\eta_i} - \ddot{\mathbf{M}}_m^{j_1,j_2}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\right)^2$$
$$= \frac{1}{r}\sum_{i=1}^{m}\eta_i\left(-\frac{2\boldsymbol{e}_i'(\mathbf{D}_{j_2}\mathbf{D}_{j_1}-\dot{\mathbf{D}})\hat{\mathbf{V}}^{-1}\boldsymbol{e}_i - \mathrm{tr}\left(\mathbf{D}_{j_1}\mathbf{D}_{j_2}-\dot{\mathbf{D}}\right)}{2m\eta_i}\right)^2 - \frac{1}{r}\left(\ddot{\mathbf{M}}_m^{j_1,j_2}\right)^2$$
$$\leq \frac{1}{rm^2}\sum_{i=1}^{m}\eta_i\left(-\frac{2\boldsymbol{e}_i'(\mathbf{D}_{j_2}\mathbf{D}_{j_1}-\dot{\mathbf{D}})\hat{\mathbf{V}}^{-1}\boldsymbol{e}_i - \mathrm{tr}\left(\mathbf{D}_{j_1}\mathbf{D}_{j_2}-\dot{\mathbf{D}}\right)}{2\eta_i}\right)^2$$
$$= \mathcal{O}_p(r^{-1}),$$

where $\mathbf{D}_{j_1} = \hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}}{\partial\boldsymbol{\xi}_{j_1-d}}$, $\mathbf{D}_{j_2} = \hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}}{\partial\boldsymbol{\xi}_{j_2-d}}$, $\dot{\mathbf{D}} = \hat{\mathbf{V}}^{-1}\frac{\partial^2\hat{\mathbf{V}}}{\partial\boldsymbol{\xi}_{j_1-d}\partial\boldsymbol{\xi}_{j_2-d}}$ and the last equality is from Assumption 2.

Then, by Markov's inequality, for any $\varepsilon > 0$

$$\ddot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) - \ddot{\mathbf{M}}_m(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) = O_{p|\mathcal{D}_m}(r^{-1/2}) = o_{p|\mathcal{D}_m}(1).$$

□

Now, we prove Theorem 1. By direct calculation, we have for any $(\boldsymbol{\beta}',\boldsymbol{\xi}')'$,

$$\mathbb{E}\left[\mathrm{M}_r^*(\boldsymbol{\beta},\boldsymbol{\xi}) \mid \mathcal{D}_m\right] = \mathrm{M}_m(\boldsymbol{\beta},\boldsymbol{\xi}).$$

Also, we have

$$
\begin{aligned}
&P\left(|\mathrm{M}_r^*(\boldsymbol{\beta},\boldsymbol{\xi}) - \mathrm{M}_m(\boldsymbol{\beta},\boldsymbol{\xi})| \geq \epsilon \mid \mathcal{D}_m\right) \leq \frac{\mathbb{V}\left[\mathrm{M}_r^*(\boldsymbol{\beta},\boldsymbol{\xi})\right]}{\epsilon^2} \\
&= \frac{1}{\epsilon^2 r}\left\{\frac{1}{m^2}\sum_{i=1}^{m}\frac{\mathrm{p}^2(\mathbf{X}_i,\mathbf{y}_i,\boldsymbol{\beta},\boldsymbol{\xi})}{\eta_i} - \left(\frac{1}{m}\sum_{i=1}^{m}\mathrm{p}(\mathbf{X}_i,\mathbf{y}_i,\boldsymbol{\beta},\boldsymbol{\xi})\right)^2\right\} \\
&\leq \frac{1}{rm^2\epsilon^2}\sum_{i=1}^{m}\frac{\mathrm{p}^2(\mathbf{X}_i,\mathbf{y}_i,\boldsymbol{\beta},\boldsymbol{\xi})}{\eta_i} = \frac{1}{rm^2\epsilon^2}\sum_{i=1}^{m}\frac{\left(C + \frac{(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})}{2}\right)^2}{\eta_i} \\
&= \frac{1}{rm^2\epsilon^2}\sum_{i=1}^{m}\frac{\left(C + \frac{(\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))'\mathbf{V}^{-1}(\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))}{2}\right)^2}{\eta_i} \\
&= \frac{1}{r}\Bigg(\frac{1}{m^2\epsilon^2}\sum_{i=1}^{m}\frac{C^2}{\eta_i} + \frac{1}{m^2\epsilon^2}\sum_{i=1}^{m}\frac{C\left((\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))'\mathbf{V}^{-1}(\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))\right)}{\eta_i} \\
&\qquad + \frac{1}{m^2\epsilon^2}\sum_{i=1}^{m}\frac{\left((\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))'\mathbf{V}^{-1}(\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))\right)^2}{4\eta_i}\Bigg),
\end{aligned}
$$

where

$$
C = \frac{n}{2}\log 2\pi\sigma^2 + \frac{1}{2}\log\left[(1+(n-1)\rho)\right] + \frac{n-1}{2}\log\left(1-\rho\right).
$$

From Assumption 1 and the fact that $\mathbf{x}_{i,1} = \mathbf{1}$, we have

$$
\frac{1}{m^2}\sum_{i=1}^{m}\frac{\|\mathbf{x}_{i,1}\|^4}{\eta_i} = \frac{n^2}{m}\sum_{i=1}^{m}\frac{1}{m\eta_i} = \mathcal{O}_p(1). \tag{B2}
$$

As discussed in Remark 1, if the model does not include an intercept term, we need to assume that

$$
\frac{1}{m}\sum_{i=1}^{m}\frac{1}{m\eta_i} = \mathcal{O}_p(1).
$$

Also,

$$
\begin{aligned}
&\frac{1}{m^2\epsilon^2}\sum_{i=1}^{m}\frac{C\left((\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))'\mathbf{V}^{-1}(\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))\right)}{\eta_i} \\
&= \frac{1}{\epsilon^2}\Bigg(\frac{1}{m^2}\sum_{i=1}^{m}\frac{C\mathbf{e}_i'\mathbf{V}^{-1}\mathbf{e}_i}{\eta_i} + \frac{1}{m^2}\sum_{i=1}^{m}\frac{2C\mathbf{e}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\eta_i} \\
&\qquad + \frac{1}{m^2}\sum_{i=1}^{m}\frac{C(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\eta_i}\Bigg) = \mathcal{O}_P(1),
\end{aligned}
$$

From Assumption 1 and 2, we have

$$\frac{1}{m^2}\sum_{i=1}^{m}\frac{C\mathbf{e}_i'\mathbf{V}^{-1}\mathbf{e}_i}{\eta_i}=\mathcal{O}_P(1);\quad \frac{1}{m^2}\sum_{i=1}^{m}\frac{C(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\eta_i}=\mathcal{O}_P(1)\,. \tag{B3}$$

Using Assumption 1 and 2 and by using the Cauchy-Schwarz inequality, we have

$$\frac{1}{m^2}\sum_{i=1}^{m}\frac{\mathbf{e}_i'\mathbf{x}_{i,j}}{\eta_i}\le\frac{1}{m^2}\sum_{i=1}^{m}\frac{|\mathbf{e}_i'\mathbf{x}_{i,j}|}{\eta_i}\le\sqrt{\frac{1}{m^2}\sum_{i=1}^{m}\frac{\|\mathbf{e}_i\|^2}{\eta_i}\frac{1}{m^2}\sum_{i=1}^{m}\frac{\|\mathbf{x}_{i,j}\|^2}{\eta_i}}=\mathcal{O}_P(1)\,.$$

So,

$$\frac{1}{m^2}\sum_{i=1}^{m}\frac{2C\mathbf{e}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\eta_i}=\mathcal{O}_P(1)\,. \tag{B4}$$

Combing (B3) and (B4), we have

$$\frac{1}{m^2\epsilon^2}\sum_{i=1}^{m}\frac{C\left((\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))'\mathbf{V}^{-1}(\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))\right)}{\eta_i}=\mathcal{O}_P(1)\,. \tag{B5}$$

Similarly,

$$\begin{aligned}
&\frac{1}{m^2\epsilon^2}\sum_{i=1}^{m}\frac{\left((\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))'\mathbf{V}^{-1}(\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))\right)^2}{4\eta_i}\\
&=\frac{1}{4\epsilon^2}\Bigg(\frac{1}{m^2}\sum_{i=1}^{m}\frac{\left(\boldsymbol{e}_i'\mathbf{V}^{-1}\boldsymbol{e}_i\right)^2}{\eta_i}+\frac{1}{m^2}\sum_{i=1}^{m}\frac{4\left(\boldsymbol{e}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)^2}{\eta_i}\\
&\quad+\frac{1}{m^2}\sum_{i=1}^{m}\frac{\left((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)^2}{\eta_i}+\frac{1}{m^2}\sum_{i=1}^{m}\frac{4\boldsymbol{e}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\boldsymbol{e}_i'\mathbf{V}^{-1}\boldsymbol{e}_i}{\eta_i}\\
&\quad+\frac{1}{m^2}\sum_{i=1}^{m}\frac{4(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\boldsymbol{e}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\eta_i}\\
&\quad+\frac{1}{m^2}\sum_{i=1}^{m}\frac{4(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\boldsymbol{e}_i'\mathbf{V}^{-1}\boldsymbol{e}_i}{\eta_i}\Bigg)=\mathcal{O}_P(1)\,,
\end{aligned}$$

From Assumption 1 and 2, we have

$$\frac{1}{m^2}\sum_{i=1}^{m}\frac{\left(\boldsymbol{e}_i'\mathbf{V}^{-1}\boldsymbol{e}_i\right)^2}{\eta_i}=\mathcal{O}_P(1);\quad \frac{1}{m^2}\sum_{i=1}^{m}\frac{\left((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)^2}{\eta_i}=\mathcal{O}_P(1)\,. \tag{B6}$$

Using Assumption 1 and 2 and by using the Cauchy-Schwarz inequality, we have

$$\frac{1}{m^2}\sum_{i=1}^{m}\frac{(\mathbf{e}_i'\mathbf{x}_{i,j})^2}{\eta_i} \le \frac{1}{m^2}\sum_{i=1}^{m}\frac{\|\mathbf{e}_i\|^2\|\mathbf{x}_{i,j}\|^2}{\eta_i} \le \sqrt{\frac{1}{m^2}\sum_{i=1}^{m}\frac{\|\mathbf{e}_i\|^4}{\eta_i}\frac{1}{m^2}\sum_{i=1}^{m}\frac{\|\mathbf{x}_{i,j}\|^4}{\eta_i}} = \mathcal{O}_P(1)\,.$$

So,

$$\frac{1}{m^2}\sum_{i=1}^{m}\frac{4\left(\boldsymbol{e}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)^2}{\eta_i} = \mathcal{O}_P(1)\,. \tag{B7}$$

Then, using (B6), (B7) and Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\frac{1}{m^2}\sum_{i=1}^{m}\frac{4\boldsymbol{e}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\boldsymbol{e}_i'\mathbf{V}^{-1}\boldsymbol{e}_i}{\eta_i}\\ &\quad\le 4\sqrt{\frac{1}{m^2}\sum_{i=1}^{m}\frac{\left(\boldsymbol{e}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)^2}{\eta_i}\frac{1}{m^2}\sum_{i=1}^{m}\frac{(\boldsymbol{e}_i'\mathbf{V}^{-1}\boldsymbol{e}_i)^2}{\eta_i}}\\ &\quad= \mathcal{O}_P(1)\,,\end{aligned} \tag{B8}$$

$$\begin{aligned}&\frac{1}{m^2}\sum_{i=1}^{m}\frac{4(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\boldsymbol{e}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\eta_i}\\ &\quad\le 4\sqrt{\frac{1}{m^2}\sum_{i=1}^{m}\frac{\left((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)^2}{\eta_i}\frac{1}{m^2}\sum_{i=1}^{m}\frac{\left(\boldsymbol{e}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)^2}{\eta_i}}\\ &= \mathcal{O}_P(1)\,,\end{aligned} \tag{B9}$$

$$\begin{aligned}&\frac{1}{m^2}\sum_{i=1}^{m}\frac{4(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\boldsymbol{e}_i'\mathbf{V}^{-1}\boldsymbol{e}_i}{\eta_i}\\ &\quad\le 4\sqrt{\frac{1}{m^2}\sum_{i=1}^{m}\frac{\left((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}_i'\mathbf{V}^{-1}\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)^2}{\eta_i}\frac{1}{m^2}\sum_{i=1}^{m}\frac{(\boldsymbol{e}_i'\mathbf{V}^{-1}\boldsymbol{e}_i)^2}{\eta_i}}\\ &= \mathcal{O}_P(1)\,,\end{aligned} \tag{B10}$$

Combing (B6), (B7), (B8), (B9) and (B10), we have

$$\frac{1}{m^2\epsilon^2}\sum_{i=1}^{m}\frac{\left((\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))'\mathbf{V}^{-1}(\mathbf{e}_i+\mathbf{X}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))\right)^2}{4\eta_i}=\mathcal{O}_P(1)\,. \tag{B11}$$

So, combing (B2), (B5), and (B11) the above, we have

$$P\left(|\mathrm{M}_r^*(\boldsymbol{\beta},\boldsymbol{\xi})-\mathrm{M}_m(\boldsymbol{\beta},\boldsymbol{\xi})|\geq\epsilon\mid\mathcal{D}_m\right)=\mathcal{O}_{p|\mathcal{D}_m}(r^{-1})=o_{p|\mathcal{D}_m}(1)\,.$$

Therefore, $\mathrm{M}_r^*(\boldsymbol{\beta},\boldsymbol{\xi})-\mathrm{M}_m(\boldsymbol{\beta},\boldsymbol{\xi})\to 0$ in conditional probability given $\mathcal{D}_m$. Note that the parameter space is compact and $(\hat{\boldsymbol{\beta}}',\hat{\boldsymbol{\xi}}')'$ is the unique global maximum of the continuous function $\ell(\boldsymbol{\beta},\boldsymbol{\xi})$. Thus, from Theorem 5.9 and corresponding remark in van der Vaart (1998), conditionally on $\mathcal{D}_m$ in probability,

$$\left\|\begin{matrix}\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\xi}}-\hat{\boldsymbol{\xi}}\end{matrix}\right\|=o_{p|\mathcal{D}_m}(1)\,.$$

The consistency ensures that $(\tilde{\boldsymbol{\beta}}',\tilde{\boldsymbol{\xi}}')'$ is close to $(\hat{\boldsymbol{\beta}}',\hat{\boldsymbol{\xi}}')'$ provided r is large. Applying Taylor expansion on $\mathrm{M}_r^*(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\xi}})$ at $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})$, we have

$$\mathbf{0}=\dot{\mathbf{M}}_r^*(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\xi}})=\dot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})+\ddot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\left(\begin{bmatrix}\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\xi}}-\hat{\boldsymbol{\xi}}\end{bmatrix}\right)+\mathbf{R}\,.$$

Letting $\dot{\mathrm{M}}_r^{*j}(\boldsymbol{\beta},\boldsymbol{\xi})$ be the j-th component of $\dot{\mathbf{M}}_r^*(\boldsymbol{\beta},\boldsymbol{\xi})$, we have:

$$0=\dot{\mathrm{M}}_r^{*j}(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\xi}})=\dot{\mathrm{M}}_r^{*j}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})+\ddot{\mathbf{M}}_r^{*j}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\left(\begin{bmatrix}\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\xi}}-\hat{\boldsymbol{\xi}}\end{bmatrix}\right)+R^j\,,$$

where

$$\begin{aligned}R^j&=\left(\begin{bmatrix}\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\xi}}-\hat{\boldsymbol{\xi}}\end{bmatrix}\right)'\int_0^1\int_0^1\frac{\partial^2\dot{\mathrm{M}}_r^{*j}(\hat{\boldsymbol{\theta}}+uv(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}vdudv\left(\begin{bmatrix}\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\xi}}-\hat{\boldsymbol{\xi}}\end{bmatrix}\right)\\&=\left(\begin{bmatrix}\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\xi}}-\hat{\boldsymbol{\xi}}\end{bmatrix}\right)'\int_0^1\int_0^1\frac{1}{r}\sum_{i=1}^{r}\frac{1}{m\eta_i^*}\frac{\partial^2\dot{\mathrm{p}}^j(\mathbf{X}_i^*,\mathbf{y}_i^*,,\hat{\boldsymbol{\theta}}+uv(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}vdudv\left(\begin{bmatrix}\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\xi}}-\hat{\boldsymbol{\xi}}\end{bmatrix}\right),\end{aligned}$$

$\hat{\boldsymbol{\theta}}=(\hat{\boldsymbol{\beta}}',\hat{\boldsymbol{\xi}}')'$, $\tilde{\boldsymbol{\theta}}=(\tilde{\boldsymbol{\beta}}',\tilde{\boldsymbol{\xi}}')'$, and

$$\frac{\partial^2\dot{\mathrm{p}}^j(\mathbf{x}_i^*,\mathbf{y}_i^*,\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}=\begin{bmatrix}\frac{\partial^2\dot{\mathrm{p}}^j}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \frac{\partial^2\dot{\mathrm{p}}^j}{\partial\boldsymbol{\beta}\partial\boldsymbol{\xi}'}\\ \frac{\partial^2\dot{\mathrm{p}}^j}{\partial\boldsymbol{\xi}\partial\boldsymbol{\beta}'} & \frac{\partial^2\dot{\mathrm{p}}^j}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}'}\end{bmatrix}\,.$$

For $1 \leq j \leq d$,

$$\dot{\mathrm{M}}_r^{*j}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) = \frac{1}{r} \sum_{i=1}^{r} \frac{\mathbf{x}_{i,j}^{*\prime} \hat{\mathbf{V}}^{-1} (\mathbf{y}_i^* - \mathbf{x}_i^* \hat{\boldsymbol{\beta}})}{\eta_i^*} \,.$$

For $j = d+1$,

$$\dot{\mathrm{M}}_r^{*j}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) = \frac{1}{r} \sum_{i=1}^{r} \frac{\mathrm{tr}\left[\hat{\mathbf{V}}^{-1} \left((\mathbf{y}_i^* - \mathbf{x}_i^* \hat{\boldsymbol{\beta}})(\mathbf{y}_i^* - \mathbf{x}_i^* \hat{\boldsymbol{\beta}})' - \mathbf{V}(\hat{\boldsymbol{\xi}}) \right) \hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}}{\partial \sigma^2} \right]}{2\eta_i^*} \,.$$

For $j = d+2$,

$$\dot{\mathrm{M}}_r^{*j}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) = \frac{1}{r} \sum_{i=1}^{r} \frac{\mathrm{tr}\left[\hat{\mathbf{V}}^{-1} \left((\mathbf{y}_i^* - \mathbf{x}_i^* \hat{\boldsymbol{\beta}})(\mathbf{y}_i^* - \mathbf{x}_i^* \hat{\boldsymbol{\beta}})' - \hat{\mathbf{V}} \right) \hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}}{\partial \rho} \right]}{2\eta_i^*} \,.$$

Define:

$$\mathbf{A} = -\hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}}{\partial \sigma^2} \hat{\mathbf{V}}^{-1} \,, \qquad \mathbf{B} = -\hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}}{\partial \rho} \hat{\mathbf{V}}^{-1} \,,$$

$$\dot{\mathbf{A}}_1 = \frac{\partial}{\partial \sigma^2} \mathbf{A} = -\left(2\mathbf{A} \frac{\partial \hat{\mathbf{V}}}{\partial \sigma^2} \hat{\mathbf{V}}^{-1} + \hat{\mathbf{V}}^{-1} \frac{\partial^2 \hat{\mathbf{V}}}{\partial \sigma^4} \hat{\mathbf{V}}^{-1} \right) \,,$$

$$\dot{\mathbf{A}}_2 = \frac{\partial}{\partial \rho} \mathbf{A} = -\left(\mathbf{B} \frac{\partial \hat{\mathbf{V}}}{\partial \sigma^2} \hat{\mathbf{V}}^{-1} + \hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}^2}{\partial \sigma^2 \rho} \hat{\mathbf{V}}^{-1} + \hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}}{\partial \sigma^2} \mathbf{B} \right) \,,$$

$$\dot{\mathbf{B}}_1 = \frac{\partial}{\partial \sigma^2} \mathbf{B} = \dot{\mathbf{A}}_2 \,, \; \dot{\mathbf{B}}_2 = \frac{\partial}{\partial \rho} \mathbf{B} = -\left(2\mathbf{B} \frac{\partial \hat{\mathbf{V}}}{\partial \rho} \hat{\mathbf{V}}^{-1} + \hat{\mathbf{V}}^{-1} \frac{\partial^2 \hat{\mathbf{V}}}{\partial \rho^2} \hat{\mathbf{V}}^{-1} \right) \,,$$

$$\ddot{\mathbf{A}}_{11} = \frac{\partial}{\partial \sigma^2} \dot{\mathbf{A}}_1 = -\left(2\dot{\mathbf{A}}_1 \frac{\partial \hat{\mathbf{V}}}{\partial \sigma^2} \hat{\mathbf{V}}^{-1} + 3\mathbf{A} \frac{\partial^2 \hat{\mathbf{V}}}{\partial \sigma^4} \hat{\mathbf{V}}^{-1} + 2\mathbf{A} \frac{\partial \hat{\mathbf{V}}}{\partial \sigma^2} \mathbf{A} \right.$$
$$\left. + \hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}^3}{\partial \sigma^6} \hat{\mathbf{V}}^{-1} + \hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}^2}{\partial \sigma^4} \mathbf{A} \right) \,,$$

$$\ddot{\mathbf{A}}_{12} = \frac{\partial}{\partial \rho} \dot{\mathbf{A}}_1 = -\left(2\dot{\mathbf{A}}_2 \frac{\partial \hat{\mathbf{V}}}{\partial \sigma^2} \hat{\mathbf{V}}^{-1} + 2\mathbf{A} \frac{\partial^2 \hat{\mathbf{V}}}{\partial \sigma^2 \partial \rho} \hat{\mathbf{V}}^{-1} + 2\mathbf{A} \frac{\partial \hat{\mathbf{V}}}{\partial \sigma^2} \mathbf{B} \right.$$
$$\left. + \mathbf{B} \frac{\partial \hat{\mathbf{V}}^2}{\partial \sigma^4} \hat{\mathbf{V}}^{-1} + \hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}^3}{\partial \sigma^4 \partial \rho} \hat{\mathbf{V}}^{-1} + \hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}^2}{\partial \sigma^4} \mathbf{B} \right) \,,$$

$$\ddot{\mathbf{A}}_{21} = \frac{\partial}{\partial\sigma^2}\dot{\mathbf{A}}_2 = \ddot{\mathbf{A}}_{12}\,,\ \ddot{\mathbf{B}}_{12} = \frac{\partial}{\partial\rho}\dot{\mathbf{B}}_1 = \ddot{\mathbf{B}}_{21}\,,$$

$$\ddot{\mathbf{A}}_{22} = \frac{\partial}{\partial\rho}\dot{\mathbf{A}}_2 = -\left(\dot{\mathbf{B}}_2\frac{\partial\hat{\mathbf{V}}}{\partial\sigma^2}\hat{\mathbf{V}}^{-1} + 2\mathbf{B}\frac{\partial^2\hat{\mathbf{V}}}{\partial\sigma^2\partial\rho}\hat{\mathbf{V}}^{-1} + 2\mathbf{B}\frac{\partial\hat{\mathbf{V}}}{\partial\sigma^2}\mathbf{B} + \hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}^3}{\partial\sigma^4\partial\rho}\hat{\mathbf{V}}^{-1} + 2\hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}^2}{\partial\sigma^2\partial\rho}\mathbf{B} + \hat{\mathbf{V}}^{-1}\frac{\partial}{\partial\sigma^2}\dot{\mathbf{B}}_2\right),$$

$$\ddot{\mathbf{B}}_{11} = \frac{\partial}{\partial\sigma^2}\dot{\mathbf{B}}_1 = -\left(\dot{\mathbf{A}}_1\frac{\partial\hat{\mathbf{V}}}{\partial\rho}\hat{\mathbf{V}}^{-1} + 2\mathbf{A}\frac{\partial^2\hat{\mathbf{V}}}{\partial\rho\partial\sigma^2}\hat{\mathbf{V}}^{-1} + 2\mathbf{A}\frac{\partial\hat{\mathbf{V}}}{\partial\rho}\mathbf{A} + \hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}^3}{\partial\rho\partial\sigma^4}\hat{\mathbf{V}}^{-1} + 2\hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}^2}{\partial\rho\partial\sigma^2}\mathbf{A} + \hat{\mathbf{V}}^{-1}\frac{\partial}{\partial\rho}\dot{\mathbf{A}}_1\right),$$

$$\ddot{\mathbf{B}}_{21} = \frac{\partial}{\partial\sigma^2}\dot{\mathbf{B}}_2 = -\left(2\dot{\mathbf{B}}_1\frac{\partial\hat{\mathbf{V}}}{\partial\rho}\hat{\mathbf{V}}^{-1} + 2\mathbf{B}\frac{\partial^2\hat{\mathbf{V}}}{\partial\rho\partial\sigma^2}\hat{\mathbf{V}}^{-1} + 2\mathbf{B}\frac{\partial\hat{\mathbf{V}}}{\partial\rho}\mathbf{A} + \mathbf{A}\frac{\partial\hat{\mathbf{V}}^2}{\partial\rho^2}\hat{\mathbf{V}}^{-1} + \hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}^3}{\partial\rho^2\partial\sigma^2}\hat{\mathbf{V}}^{-1} + \hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}^2}{\partial\rho^2}\mathbf{A}\right),$$

$$\ddot{\mathbf{B}}_{22} = \frac{\partial}{\partial\rho}\dot{\mathbf{B}}_2 = -\left(2\dot{\mathbf{B}}_2\frac{\partial\hat{\mathbf{V}}}{\partial\rho}\hat{\mathbf{V}}^{-1} + 3\mathbf{B}\frac{\partial^2\hat{\mathbf{V}}}{\partial\rho^2}\hat{\mathbf{V}}^{-1} + 2\mathbf{B}\frac{\partial\hat{\mathbf{V}}}{\partial\rho}\mathbf{B} + \hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}^3}{\partial\rho^3}\hat{\mathbf{V}}^{-1} + \hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}^2}{\partial\rho^2}\mathbf{B}\right).$$

Note, given the structure of the covariance matrix $\mathbf{V}$ as compound symmetric, $\mathbf{A}$, $\dot{\mathbf{A}}_1$, $\dot{\mathbf{A}}_2$, $\ddot{\mathbf{A}}_{11}$, $\ddot{\mathbf{A}}_{12}$, $\ddot{\mathbf{A}}_{21}$, $\ddot{\mathbf{A}}_{22}$, and $\mathbf{B}$, $\dot{\mathbf{B}}_1$, $\dot{\mathbf{B}}_2$, $\ddot{\mathbf{B}}_{11}$, $\ddot{\mathbf{B}}_{12}$, $\ddot{\mathbf{B}}_{21}$, $\ddot{\mathbf{B}}_{22}$ are all finite.

Then, for $1 \leq j \leq d$,

$$\begin{aligned}\frac{\partial^2\dot{\mathrm{M}}_r^{*j}(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} &= \frac{1}{r}\sum_{i=1}^{r}\frac{1}{\eta_i^*}\begin{bmatrix}\frac{\partial^2\dot{\mathrm{p}}_i^j}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \frac{\dot{\mathrm{p}}_i^j}{\partial\boldsymbol{\beta}\partial\boldsymbol{\xi}'}\\ \frac{\dot{\mathrm{p}}_i^j}{\partial\boldsymbol{\xi}\partial\boldsymbol{\beta}'} & \frac{\dot{\mathrm{p}}_i^j}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}'}\end{bmatrix}\\ &= \frac{1}{r}\sum_{i=1}^{r}\frac{1}{\eta_i^*}\begin{bmatrix}\mathbf{0} & \mathbf{X}_i^{*\prime}\mathbf{A}\mathbf{x}_{i,j}^* & \mathbf{X}_i^{*\prime}\mathbf{A}\mathbf{x}_{i,j}^*\\ \mathbf{x}_{i,j}^{*\prime}\mathbf{A}\mathbf{X}_i^* & \mathbf{x}_{i,j}^{*\prime}\dot{\mathbf{A}}_1\boldsymbol{\varepsilon}_i^* & \mathbf{x}_{i,j}^{*\prime}\dot{\mathbf{A}}_2\boldsymbol{\varepsilon}_i^*\\ \mathbf{x}_{i,j}^{*\prime}\mathbf{B}\mathbf{X}_i^* & \mathbf{x}_{i,j}^{*\prime}\dot{\mathbf{B}}_1\boldsymbol{\varepsilon}_i^* & \mathbf{x}_{i,j}^{*\prime}\dot{\mathbf{B}}_2\boldsymbol{\varepsilon}_i^*\end{bmatrix}.\end{aligned}$$

For any matrix $\mathbf{Q}$ whose components are finite,

$$
\begin{aligned}
&P\left(\left\|\int_0^1\int_0^1 \frac{1}{mr}\sum_{i=1}^{r}\frac{\mathbf{X}_i^{*\prime}\mathbf{Q}\mathbf{x}_{i,j}^{*}}{\eta_i^{*}}vdudv\right\| \geq \alpha\middle|\mathcal{D}_m\right)\\
&= P\left(\left\|\frac{1}{2mr}\sum_{i=1}^{r}\frac{\mathbf{X}_i^{*\prime}\mathbf{Q}\mathbf{x}_{i,j}^{*}}{\eta_i^{*}}\right\| \geq \alpha\middle|\mathcal{D}_m\right)\\
&\leq \frac{1}{2mr\alpha}\sum_{i=1}^{r}\mathbb{E}\left[\frac{\mathbf{X}_i^{*\prime}\mathbf{Q}\mathbf{x}_{i,j}^{*}}{\eta_i^{*}}\right] = \frac{1}{2m\alpha}\sum_{i=1}^{m}\mathbf{X}_i^{\prime}\mathbf{Q}\mathbf{x}_{i,j} = \mathcal{O}_p(1)\,,
\end{aligned}
$$

where the last equality is from Assumption 1. We then have

$$
R^j = \left\|\int_0^1\int_0^1 \frac{\partial^2\dot{\mathrm{M}}_r^{*j}(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}vdudv\right\| = \mathcal{O}_{p|\mathcal{D}_m}(\|\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}\|^2) = o_{P|\mathcal{D}_m}(1)\,.
$$

For $j = d+1$

$$
\begin{aligned}
\frac{\partial^2\dot{\mathrm{M}}_r^{*j}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} &= \frac{1}{r}\sum_{i=1}^{r}\frac{1}{\eta_i^{*}}\begin{bmatrix}\frac{\partial^2\dot{\mathbf{p}}^j}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \frac{\dot{\mathbf{p}}^j}{\partial\boldsymbol{\beta}\partial\boldsymbol{\xi}'}\\ \frac{\dot{\mathbf{p}}^j}{\partial\boldsymbol{\xi}\partial\boldsymbol{\beta}'} & \frac{\dot{\mathbf{p}}^j}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}'}\end{bmatrix}\\
&= \frac{1}{r}\sum_{i=1}^{r}\frac{1}{\eta_i^{*}}\begin{bmatrix}\mathbf{X}_i^{*\prime}\mathbf{A}\mathbf{X}_i^{*} & \mathbf{X}_i^{*\prime}\dot{\mathbf{A}}_1\boldsymbol{e}_i^{*} & \mathbf{X}_i^{*\prime}\dot{\mathbf{A}}_2\boldsymbol{e}_i^{*}\\ \boldsymbol{e}_i^{*\prime}\dot{\mathbf{A}}_1\mathbf{X}_i^{*} & \mathrm{K}_{r2,2}^{*d+1} & \mathrm{K}_{r2,3}^{*d+1}\\ \boldsymbol{e}_i^{*\prime}\dot{\mathbf{A}}_2\mathbf{X}_i^{*} & \mathrm{K}_{r2,2}^{*d+1} & \mathrm{K}_{r3,3}^{*d+1}\end{bmatrix}\,,
\end{aligned}
$$

where

$$
\mathrm{K}_{r2,2}^{*d+1} = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{e}_i^{*\prime}\ddot{\mathbf{A}}_{11}\boldsymbol{e}_i^{*} - \dot{\mathbf{A}}_1\frac{\partial\hat{\mathbf{V}}}{\partial\sigma^2} - 2\mathbf{A}\frac{\partial^2\hat{\mathbf{V}}}{\partial\sigma^4} - \hat{\mathbf{V}}^{-1}\frac{\partial^3\hat{\mathbf{V}}}{\partial\sigma^6}\right),
$$

$$
\mathrm{K}_{r2,3}^{*d+1} = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{e}_i^{*\prime}\ddot{\mathbf{A}}_{12}\boldsymbol{e}_i^{*} - \dot{\mathbf{A}}_2\frac{\partial\hat{\mathbf{V}}}{\partial\sigma^2} - \mathbf{A}\frac{\partial^2\hat{\mathbf{V}}}{\partial\sigma^2\partial\rho} - \mathbf{B}\frac{\partial^2\hat{\mathbf{V}}}{\partial\sigma^4} - \hat{\mathbf{V}}^{-1}\frac{\partial^3\hat{\mathbf{V}}}{\partial\sigma^4\partial\rho}\right),
$$

and

$$
\mathrm{K}_{r3,3}^{*d+1} = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{e}_i^{*\prime}\ddot{\mathbf{A}}_{22}\boldsymbol{e}_i^{*} - \dot{\mathbf{B}}_2\frac{\partial\hat{\mathbf{V}}}{\partial\sigma^2} - 2\mathbf{B}\frac{\partial^2\hat{\mathbf{V}}}{\partial\sigma^2\partial\rho} - \hat{\mathbf{V}}^{-1}\frac{\partial^3\hat{\mathbf{V}}}{\partial\sigma^4\partial\rho}\right).
$$

For any matrix $\mathbf{Q}$ whose components are finite,

$$
\begin{aligned}
&P\left(\left\|\int_0^1\int_0^1 \frac{1}{mr}\sum_{i=1}^{r}\frac{\mathbf{X}_i^{*\prime}\mathbf{Q}\boldsymbol{e}_i^{*}}{\eta_i^{*}}vdudv\right\| \geq \alpha\middle|\mathcal{D}_m\right)\\
&= P\left(\left\|\frac{1}{2mr}\sum_{i=1}^{r}\frac{\mathbf{X}_i^{*\prime}\mathbf{Q}\boldsymbol{e}_i^{*}}{\eta_i^{*}}\right\| ge\alpha\middle|\mathcal{D}_m\right)
\end{aligned}
$$

$$\leq \frac{1}{2mr\alpha}\sum_{i=1}^{r}\mathbb{E}\left[\frac{\mathbf{X}_i^{*\prime}\mathbf{Q}\boldsymbol{e}_i^*}{\eta_i^*}\right]$$
$$= \frac{1}{2m\alpha}\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{Q}\boldsymbol{e}_i = \mathcal{O}_p(1)\,,$$

where the last equality is from Assumptions 1 and 2 and by using the Cauchy-Schwarz inequality. Also,

$$P\left(\left\|\int_0^1\int_0^1\frac{1}{mr}\sum_{i=1}^{r}\frac{\boldsymbol{e}_i^{*\prime}\mathbf{Q}\boldsymbol{e}_i^*}{\eta_i^*}vdudv\right\|\geq\alpha\middle|\mathcal{D}_m\right)$$
$$= P\left(\left\|\frac{1}{2mr}\sum_{i=1}^{r}\frac{\boldsymbol{e}_i^{*\prime}\mathbf{Q}\boldsymbol{e}_i^*}{\eta_i^*}\right\|\geq\alpha\middle|\mathcal{D}_m\right)$$
$$\leq \frac{1}{2mr\alpha}\sum_{i=1}^{r}\mathbb{E}\left[\frac{\boldsymbol{e}_i^{*\prime}\mathbf{Q}\boldsymbol{e}_i^*}{\eta_i^*}\right]$$
$$= \frac{1}{2m\alpha}\sum_{i=1}^{m}\boldsymbol{e}_i'\mathbf{Q}\boldsymbol{e}_i = \mathcal{O}_{p|\mathcal{D}_m}(1)\,,$$

where the last equality is from Assumption 2, and we have

$$R^j = \left\|\int_0^1\int_0^1\frac{\partial^2\dot{\mathrm{M}}_r^{*j}(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}vdudv\right\| = \mathcal{O}_{p|\mathcal{D}_m}(\|\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}\|^2) = o_{P|\mathcal{D}_m}(1).$$

For $j = d+2$

$$\frac{\partial^2\dot{\mathrm{M}}_r^{*j}(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{\eta_i^*}\begin{bmatrix}\frac{\partial^2\dot{\mathbf{p}}^j}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \frac{\dot{\mathbf{p}}^j}{\partial\boldsymbol{\beta}\partial\boldsymbol{\xi}'}\\ \frac{\dot{\mathbf{p}}^j}{\partial\boldsymbol{\xi}\partial\boldsymbol{\beta}'} & \frac{\dot{\mathbf{p}}^j}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}'}\end{bmatrix}$$
$$= \frac{1}{r}\sum_{i=1}^{r}\frac{1}{\eta_i^*}\begin{bmatrix}\mathbf{X}_i^{*\prime}\mathbf{B}\mathbf{X}_i^* & \mathbf{X}_i^{*\prime}\dot{\mathbf{B}}_1\boldsymbol{e}_i^* & \mathbf{X}_i^{*\prime}\dot{\mathbf{B}}_2\boldsymbol{e}_i^*\\ \boldsymbol{e}_i^{*\prime}\dot{\mathbf{B}}_1\mathbf{X}_i^* & \mathrm{K}_{r2,2}^{*d+2} & \mathrm{K}_{r2,3}^{*d+2}\\ \boldsymbol{e}_i^{*\prime}\dot{\mathbf{B}}_2\mathbf{X}_i^* & \mathrm{K}_{r2,3}^{*d+2} & \mathrm{K}_{r3,3}^{*d+2}\end{bmatrix},$$

where

$$\mathrm{K}_{r2,2}^{*d+2} = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{e}_i^{*\prime}\ddot{\mathbf{B}}_{11}\boldsymbol{e}_i^* - \dot{\mathbf{A}}_1\frac{\partial\hat{\mathbf{V}}}{\partial\rho} - 2\mathbf{A}\frac{\partial^2\hat{\mathbf{V}}}{\partial\rho\partial\sigma^2} - \hat{\mathbf{V}}^{-1}\frac{\partial^3\hat{\mathbf{V}}}{\partial\rho\partial\sigma^4}\right),$$

$$\mathrm{K}_{r2,3}^{*d+2} = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{e}_i^{*\prime}\ddot{\mathbf{B}}_{12}\boldsymbol{e}_i^* - \dot{\mathbf{A}}_2\frac{\partial\hat{\mathbf{V}}}{\partial\rho} - \mathbf{A}\frac{\partial^2\hat{\mathbf{V}}}{\partial\rho^2} - \mathbf{B}\frac{\partial^2\hat{\mathbf{V}}}{\partial\rho\partial\sigma^2} - \hat{\mathbf{V}}^{-1}\frac{\partial^3\hat{\mathbf{V}}}{\partial\rho\partial\sigma^2\partial\rho}\right),$$

and

$$\mathrm{K}_{r3,3}^{*d+2} = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{e}_i^{*\prime}\ddot{\mathbf{B}}_{22}\boldsymbol{e}_i^* - \dot{\mathbf{B}}_2\frac{\partial\hat{\mathbf{V}}}{\partial\rho} - 2\mathbf{B}\frac{\partial^2\hat{\mathbf{V}}}{\partial\rho^2} - \hat{\mathbf{V}}^{-1}\frac{\partial^3\hat{\mathbf{V}}}{\partial\rho^3}\right).$$

In this case, we similarly have

$$R^j = \left\| \int_0^1 \int_0^1 \frac{\partial^2 \dot{\mathrm{M}}_r^{*j}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} v du dv \right\| = \mathcal{O}_{p|\mathcal{D}_m}(\|\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\|^2) = o_{P|\mathcal{D}_m}(1)\,.$$

Thus,

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{\xi}} - \hat{\boldsymbol{\xi}} \end{bmatrix} = \left[-\ddot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) \right]^{-1} \dot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) + o_P(1)\,.$$

Letting

$$\boldsymbol{\tau}_i = \frac{1}{m} \begin{bmatrix} \frac{\mathbf{x}_i^{*\prime} \hat{\mathbf{V}}^{-1} (\mathbf{y}_i^* - \mathbf{x}_i^* \hat{\boldsymbol{\beta}})}{\eta_i^*} \\ \frac{\mathrm{tr}\left[\hat{\mathbf{V}}^{-1} \left((\mathbf{y}_i^* - \mathbf{x}_i^* \hat{\boldsymbol{\beta}})(\mathbf{y}_i^* - \mathbf{x}_i^* \hat{\boldsymbol{\beta}})' - \hat{\mathbf{V}} \right) \hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}}{\partial \sigma^2} \right]}{2\eta_i^*} \\ \frac{\mathrm{tr}\left[\hat{\mathbf{V}}^{-1} \left((\mathbf{y}_i^* - \mathbf{x}_i^* \hat{\boldsymbol{\beta}})(\mathbf{y}_i^* - \mathbf{x}_i^* \hat{\boldsymbol{\beta}})' - \hat{\mathbf{V}} \right) \hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}}{\partial \rho} \right]}{2\eta_i^*} \end{bmatrix},$$

we have

$$\dot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) = \frac{1}{r} \sum_{i=1}^r \boldsymbol{\tau}_i\,.$$

Given the full data $\mathcal{D}_m$, $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_r$ are i.i.d with

$$\mathbb{E}[\boldsymbol{\tau}_i] = \frac{1}{m} \sum_{i=1}^m \begin{bmatrix} \mathbf{X}_i{}' \hat{\mathbf{V}}^{-1} (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}) \\ \frac{\mathrm{tr}\left[\hat{\mathbf{V}}^{-1} \left((\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}})(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}})' - \hat{\mathbf{V}} \right) \hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}}{\partial \sigma^2} \right]}{2} \\ \frac{\mathrm{tr}\left[\hat{\mathbf{V}}^{-1} \left((\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}})(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}})' - \hat{\mathbf{V}} \right) \hat{\mathbf{V}}^{-1} \frac{\partial \hat{\mathbf{V}}}{\partial \rho} \right]}{2} \end{bmatrix} = \mathbf{0}\,.$$

By Assumptions 1 and 2, we have

$$\mathbb{V}(\boldsymbol{\tau}_i \mid \mathcal{D}_m) = \frac{1}{m^2} \sum_{i=1}^m \frac{\dot{\mathbf{p}}_i(\mathbf{X}_i, \mathbf{y}_i;\ \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) \dot{\mathbf{p}}_i'(\mathbf{X}_i, \mathbf{y}_i;\ \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}))}{\eta_i} = r \boldsymbol{\Sigma}_c$$

$$= \frac{1}{m^2} \sum_{i=1}^m \frac{1}{\eta_i} \dot{\mathbf{p}}_i \dot{\mathbf{p}}_i' = \mathcal{O}_p(1)\,.$$

Meanwhile, for every $\alpha > 0$ and some $\delta > 0$,

$$\sum_{i=1}^r \mathbb{E}\left[\|r^{-1/2} \boldsymbol{\tau}_i\|^2 \mathbf{1}\{\|\boldsymbol{\tau}_i\| > r^{1/2}\alpha\} \middle| \mathcal{D}_m \right]$$

$$= \frac{1}{r} \sum_{i=1}^r \mathbb{E}\left[\frac{\|\boldsymbol{\tau}_i\|^{(2+\delta)}}{\|\boldsymbol{\tau}_i\|^\delta} \mathbf{1}\{\|\boldsymbol{\tau}_i\| > r^{1/2}\alpha\} \middle| \mathcal{D}_m \right]$$

$$\leq \frac{1}{r^{1+\delta/2}\alpha^\delta} \sum_{i=1}^r \mathbb{E}\left[\|\boldsymbol{\tau}_i\|^{2+\delta} \mathbf{1}\{\|\boldsymbol{\tau}_i\| > r^{1/2}\alpha\} \mid \mathcal{D}_m \right]$$

$$\leq \frac{1}{r^{1+\delta/2}\alpha^\delta} \sum_{i=1}^r \mathbb{E}\left[\|\boldsymbol{\tau}_i\|^{2+\delta} \mid \mathcal{D}_m \right] \leq \frac{1}{r^{\delta/2}\alpha^\delta m^{2+\delta}} \sum_{i=1}^m \frac{\|\dot{\mathbf{p}}_i\|^{2+\delta}}{\eta_i^{1+\delta}}$$

$$= o_p(r^{-\delta/2}) = o_p(1)\,,$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, and the second last equality is from Assumptions 3, 4 and equation (A1).

This shows that Lindeberg's condition is satisfied in probability. Based on the above result, by the Lindeberg-Feller central limit theorem, conditionally on $\mathcal{D}_m$,

$$\boldsymbol{\Sigma}_c^{-1/2}\dot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) = \frac{1}{r^{1/2}}\{\mathbb{V}(\boldsymbol{\tau}_i \mid \mathcal{D}_m)\}^{-1/2}\sum_{i=1}^{r}\boldsymbol{\tau}_i \to \mathcal{N}(\mathbf{0}, \mathbf{I}_{(d+2)\times(d+2)})\,.$$

Previously, we had

$$\begin{bmatrix}\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\xi}}-\hat{\boldsymbol{\xi}}\end{bmatrix} = \left[-\ddot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\right]^{-1}\dot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) + o_p(1).$$

Multiplying both sides by $\boldsymbol{\Sigma}^{-1/2}$ and using lemma 1 we have,

$$\begin{aligned}
&\boldsymbol{\Sigma}^{-1/2}\begin{bmatrix}\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\xi}}-\hat{\boldsymbol{\xi}}\end{bmatrix} = \boldsymbol{\Sigma}^{-1/2}\left[-\ddot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\right]^{-1}\dot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) + o_p(1)\\
&= -\boldsymbol{\Sigma}^{-1/2}\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\\
&\quad - \boldsymbol{\Sigma}^{-1/2}\left(\ddot{\mathbf{M}}_r^{*-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) - \ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\right)\dot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) + o_p(1)\\
&= -\boldsymbol{\Sigma}^{-1/2}\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\boldsymbol{\Sigma}_c^{1/2}\boldsymbol{\Sigma}_c^{-1/2}\dot{\mathbf{M}}_r^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}}) + o_p(1)\,.
\end{aligned}$$

Since

$$\boldsymbol{\Sigma}^{-1/2}\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\boldsymbol{\Sigma}_c^{1/2}\left(\boldsymbol{\Sigma}^{-1/2}\ddot{\mathbf{M}}_m^{-1}\boldsymbol{\Sigma}_c^{1/2}\right)' = \mathbf{I}\,,$$

applying Slutsky's Theorem and we have

$$\boldsymbol{\Sigma}^{-1/2}\begin{pmatrix}\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\xi}}-\hat{\boldsymbol{\xi}}\end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0},\mathbf{I})\,,$$

Then, using the A-optimality, to minimize the MSE of $\left((\tilde{\boldsymbol{\beta}}',\tilde{\boldsymbol{\xi}}')' - (\hat{\boldsymbol{\beta}}',\hat{\boldsymbol{\xi}}')\right)$ is equivalent to minimizing the trace of $\boldsymbol{\Sigma}$, so that,

$$\mathrm{tr}(\boldsymbol{\Sigma}) = \mathrm{tr}\left(\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\boldsymbol{\Sigma}_c\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\right) = \frac{1}{r}\mathrm{tr}\left(\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\mathbf{S}\mathbf{S}'\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\right),$$

where

$$\mathbf{S} = \begin{bmatrix}
\frac{\mathbf{x}_1'\hat{\mathbf{V}}^{-1}(\mathbf{y}_1-\mathbf{x}_1\hat{\boldsymbol{\beta}})}{m\sqrt{\eta_1}} & \cdots & \frac{\mathbf{x}_m'\hat{\mathbf{V}}^{-1}(\mathbf{y}_m-\mathbf{x}_m\hat{\boldsymbol{\beta}})}{m\sqrt{\eta_m}}\\
\frac{\mathrm{tr}\left[\hat{\mathbf{V}}^{-1}((\mathbf{y}_1-\mathbf{x}_1\hat{\boldsymbol{\beta}})(\mathbf{y}_1-\mathbf{x}_1\hat{\boldsymbol{\beta}})'-\hat{\mathbf{V}})\hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}}{\partial\sigma^2}\right]}{2m\sqrt{\eta_1}} & \cdots & \frac{\mathrm{tr}\left[\hat{\mathbf{V}}^{-1}((\mathbf{y}_m-\mathbf{x}_m\hat{\boldsymbol{\beta}})(\mathbf{y}_m-\mathbf{x}_m\hat{\boldsymbol{\beta}})'-\hat{\mathbf{V}})\hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}}{\partial\sigma^2}\right]}{2m\sqrt{\eta_m}}\\
\frac{\mathrm{tr}\left[\hat{\mathbf{V}}^{-1}((\mathbf{y}_1-\mathbf{x}_1\hat{\boldsymbol{\beta}})(\mathbf{y}_1-\mathbf{x}_1\hat{\boldsymbol{\beta}})'-\hat{\mathbf{V}})\hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}}{\partial\rho}\right]}{2m\sqrt{\eta_1}} & \cdots & \frac{\mathrm{tr}\left[\hat{\mathbf{V}}^{-1}((\mathbf{y}_m-\mathbf{x}_m\hat{\boldsymbol{\beta}})(\mathbf{y}_m-\mathbf{x}_m\hat{\boldsymbol{\beta}})'-\hat{\mathbf{V}})\hat{\mathbf{V}}^{-1}\frac{\partial\hat{\mathbf{V}}}{\partial\rho}\right]}{2m\sqrt{\eta_m}}
\end{bmatrix}$$

$$= \frac{1}{m}\left[\frac{\dot{\mathbf{p}}_i}{\sqrt{\eta_1}} \cdots \frac{\dot{\mathbf{p}}_m}{\sqrt{\eta_m}}\right],$$

and

$$\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\mathbf{S} = \frac{1}{m}\left[\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\frac{\dot{\mathbf{p}}_i}{\sqrt{\eta_1}} \cdots \ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\frac{\dot{\mathbf{p}}_m}{\sqrt{\eta_m}}\right].$$

Appendix C Proof of Theorem 2

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{\Sigma}) &= \mathrm{tr}\left(\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\boldsymbol{\Sigma}_c\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\right) = \frac{\mathrm{tr}\left(\sum_{i=1}^m \frac{1}{\eta_i}\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\dot{\mathbf{p}}_i'\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\right)}{rm^2} \\
&= \frac{1}{rm^2}\sum_{i=1}^m \mathrm{tr}\left(\frac{\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\dot{\mathbf{p}}_i'\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})}{\eta_i}\right) = \frac{1}{rm^2}\sum_{i=1}^m \frac{\|\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\|^2}{\eta_i} \\
&= \frac{1}{rm^2}\sum_{i=1}^m \eta_i \sum_{i=1}^m \frac{\|\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\|^2}{\eta_i} \\
&\geq \frac{1}{rm^2}\left(\sum_{i=1}^m \|\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\|\right)^2,
\end{aligned}$$

where the last step is from the Cauchy-Schwarz inequality. Equality holds if and only if $\eta_i \propto \|\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\|$.

Appendix D Proof of Theorem 3

$$\begin{aligned}
\mathrm{tr}(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}') &= \mathrm{tr}\left(\mathbf{T}\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\boldsymbol{\Sigma}_c\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\mathbf{T}'\right) \\
&= \frac{\mathrm{tr}\left(\sum_{i=1}^m \frac{1}{\eta_i}\mathbf{T}\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\dot{\mathbf{p}}_i'\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\mathbf{T}'\right)}{rm^2} \\
&= \frac{1}{rm^2}\sum_{i=1}^m \mathrm{tr}\left(\frac{\mathbf{T}\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\dot{\mathbf{p}}_i'\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\mathbf{T}'}{\eta_i}\right) \\
&= \frac{1}{rm^2}\sum_{i=1}^m \frac{\|\mathbf{T}\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\|^2}{\eta_i} \\
&= \frac{1}{rm^2}\sum_{i=1}^m \eta_i \sum_{i=1}^m \frac{\|\mathbf{T}\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\|^2}{\eta_i} \\
&\geq \frac{1}{rm^2}\left(\sum_{i=1}^m \|\mathbf{T}\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\|\right)^2,
\end{aligned}$$

where the last step is from the Cauchy-Schwarz inequality. Equality holds if and only if $\eta_i \propto \|\mathbf{T}\ddot{\mathbf{M}}_m^{-1}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})\dot{\mathbf{p}}_i\|$.

Appendix E Proof of Remark 2

To show the subsampling probabilities $\tilde{\eta}_{\alpha,i}^{\text{opt}}$ in (11) satisfy Assumption 1 under condition given in equation (12), plugging $\tilde{\eta}_{\alpha,i}^{\text{opt}}$ in we have,

for $1 \leq j \leq d$,

$$\frac{1}{m^2}\sum_{i=1}^{m}\frac{\|\mathbf{x}_{i,j}\|^4}{\tilde{\eta}_{\alpha,i}^{\text{opt}}} \leq \frac{1}{m^2}\sum_{i=1}^{m}\frac{\|\mathbf{x}_{i,j}\|^4}{\frac{\alpha}{m}} = \frac{1}{\alpha}\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}_{i,j}\|^4 = \mathcal{O}_p(1).$$

To show the subsampling probabilities $\tilde{\eta}_{\alpha,i}^{\text{opt}}$ in (11) satisfy Assumption 2 under condition given in equation (12), plugging $\tilde{\eta}_{\alpha,i}^{\text{opt}}$ in we have

$$\begin{aligned}
&\frac{1}{m^2}\sum_{i=1}^{m}\frac{\|\boldsymbol{e}_i\|^4}{\tilde{\eta}_{\alpha,i}^{\text{opt}}} \leq \frac{1}{m^2}\sum_{i=1}^{m}\frac{\|\boldsymbol{e}_i\|^4}{\frac{\alpha}{m}} = \frac{1}{\alpha}\frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{e}_i\|^4 \\
&= \frac{1}{\alpha}\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{X}_i\boldsymbol{\beta}+\boldsymbol{\varepsilon}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}}\|^4 \leq \frac{1}{\alpha}\frac{1}{m}\sum_{i=1}^{m}\Big(\|\mathbf{X}_i\boldsymbol{\beta}-\mathbf{X}_i\hat{\boldsymbol{\beta}}\|+\|\boldsymbol{\varepsilon}_i\|\Big)^4 \\
&= \mathcal{O}_p(1).
\end{aligned}$$

To show the subsampling probabilities $\tilde{\eta}_{\alpha,i}^{\text{opt}}$ in (11) satisfy Assumption 4 under condition given in equation (12), plugging $\tilde{\eta}_{\alpha,i}^{\text{opt}}$ in we have,

for some $\delta > 0$,

$$\begin{aligned}
&\frac{1}{m^2}\sum_{i=1}^{m}\frac{\|\boldsymbol{e}_i\|^{4+2\delta}}{(\tilde{\eta}_{\alpha,i}^{\text{opt}})^{1+\delta}} \leq \frac{1}{m^2}\sum_{i=1}^{m}\frac{\|\boldsymbol{e}_i\|^{4+2\delta}}{(\frac{\alpha}{m})^{1+\delta}} = \frac{m^\delta}{\alpha^{1+\delta}}\frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{e}_i\|^{4+2\delta} \\
&= \frac{m^\delta}{\alpha^{1+\delta}}\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{X}_i\boldsymbol{\beta}+\boldsymbol{\varepsilon}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}}\|^{4+2\delta} \\
&\leq \frac{m^\delta}{\alpha^{1+\delta}}\frac{1}{m}\sum_{i=1}^{m}\Big(\|\mathbf{X}_i\boldsymbol{\beta}-\mathbf{X}_i\hat{\boldsymbol{\beta}}\|+\|\boldsymbol{\varepsilon}_i\|\Big)^{4+2\delta} \\
&= \mathcal{O}_p(1).
\end{aligned}$$

Using the above and equation (12), we can also show

$$\begin{aligned}
&\frac{1}{m^2}\sum_{i=1}^{m}\frac{|\mathbf{x}'_{i,j}\boldsymbol{e}_i|^{2+\delta}}{(\tilde{\eta}_{\alpha,i}^{\text{opt}})^{1+\delta}} \leq \frac{1}{m^2}\sum_{i=1}^{m}\frac{|\mathbf{x}'_{i,j}\boldsymbol{e}_i|^{2+\delta}}{(\frac{\alpha}{m})^{1+\delta}} = \frac{m^\delta}{\alpha^{1+\delta}}\frac{1}{m}\sum_{i=1}^{m}|\mathbf{x}'_{i,j}\boldsymbol{e}_i|^{2+\delta} \\
&\leq \frac{m^\delta}{\alpha^{1+\delta}}\sqrt{\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}_{i,j}\|^{4+2\delta}\frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{e}_i\|^{4+\delta}} \\
&= \mathcal{O}_p(1).
\end{aligned}$$

This shows that the subsampling probabilities $\tilde{\eta}_{\alpha,i}^{\text{opt}}$ in (11) also satisfies Assumption 3 under the condition given in equation (12).

Using similar procedures as above, we can show the subsampling probabilities $\tilde{\eta}_{\alpha,i}^{\text{Lopt}}$ in (11) satisfy Assumption 1-4 under condition given in equation (12).

References

Ai M, Wang F, Yu J, et al (2021a) Optimal subsampling for large-scale quantile regression. Journal of Complexity 62:101,512

Ai M, Yu J, Zhang H, et al (2021b) Optimal subsampling algorithms for big data regressions. Statistica Sinica 31:749–772

Avron H, Maymounkov PB, Toledo S (2010) Blendenpik: Supercharging lapack's leastsquares solver. SIAM Journal on Scientific Computing 32:1–24

Diggle PJ, Heagerty P, Liang KY, et al (2013) Analysis of Longitudinal Data, 2nd edn. Oxford University Press

Drineas P, Mahoney MW, Muthukrishnan S (2006) Sampling algorithms for l_2 regression and applications. In: Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms, pp 1127–1136

Drineas P, Mahoney MW, Muthukrishnan S, et al (2010) Faster least squares approximation. Numerische Mathematik 117(2):219–249

Drineas P, Magdon-Ismail M, Mahoney MW, et al (2012) Fast approximation of matrix coherence and statistical leverage. Journal of Machine Learning Research 13:3475–3506

Fithian W, Hastie T (2014) Local case-control sampling: Efficient subsampling in imbalanced data sets. Annals of Statistics 42(5):1693–1724

Hong D, Shyr Y (2007) Quantitative Medical Data Analysis Using Mathematical Tools and Statistical Techniques. World Scientific, https://doi.org/10.1142/6345

Jennrich RI, Schluchter MD (1986) Unbalanced repeated-measures models with structured covariance matrices. Biometrics 42(4):805–820

Kaplan D, Kaplan D, Sage Publications i (2004) The SAGE Handbook of Quantitative Methodology for the Social Sciences. The Sage handbook of, SAGE Publications, URL https://books.google.com/books?id=k1M34kAj4VwC

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38(4):963–974

Li T, Meng C (2021) Modern subsampling methods for large-scale least squares regression. arXiv preprint arXiv:210501552

Ma P, Mahoney MW, Yu B (2015) A statistical perspective on algorithmic leveraging. Journal of Machine Learning Research 16:861–991

Meng C, Xie R, Mandal A, et al (2020) Lowcon: A design-based subsampling approach in a misspecified linear model. Journal of Computational and Graphical Statistics in press

Pusponegoro NH, Rachmawati RN, Notodiputro KA, et al (2017) Linear mixed model for analyzing longitudinal data: A simulation study of children growth differences. Procedia Computer Science 116:284–291

van der Vaart A (1998) Asympotic Statistics. Cambridge University Press

Wang H, Ma Y (2021) Optimal subsampling for quantile regression in big data. Biometrika 108:99–112

Wang H, Zhu R, Ma P (2018) Optimal subsampling for large sample logistic regression. Journal of the American Statistical Association 13(522):829–844

Wang H, Yang M, Stufken J (2019) Information-based optimal subdata selection for big data linear regression. Journal of the American Statistical Association 114(525):393–405

Yang T, Zhang L, Jin R, et al (2015) An explicit sampling dependent spectral error bound for column subset selection. ArXiv preprint

Yao Y, Wang H (2019) Optimal subsampling for softmax regression. Statistical Papers 60(2):235–249

Yu J, Wang H, Ai M, et al (2021) Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. Journal of the American Statistical Association in press

Zhao J, Wang C, Totton SC, et al (2019) Reporting and analysis of repeated measurements in preclinical animals experiments. PLoS ONE 14(8):e0220,879

Zhu R (2018) Gradient-based sampling: An adaptive importance sampling for least-squares. Proceedings of the 30th International Conference on Neural Information Processing Systems 29:406–414