Statistica Sinica: Supplement

OPTIMAL SUBSAMPLING FOR MULTINOMIAL LOGISTIC MODELS WITH BIG DATA

Zhiqiang Ye[1], Jun Yu[2], Mingyao Ai[1]

LMAM, School of Mathematical Sciences and Center for Statistical Science, Peking University [1]

School of Mathematics and Statistics, Beijing Institute of Technology[2] -

Supplementary Material

In this supplementary material, we provide the explicit forms of categorical probabilities and their derivatives, prove the theorems in the paper, and present some additional simulation results.

S1. Explicit Forms of Categorical Probabilities and their Derivatives

This section is dedicated to presenting the explicit forms of $\pi_{ij}(\boldsymbol{\beta})$'s and their derivatives, which are important parts in searching the maximum likelihood estimator and in the theoretical proofs. The categorical probability $\pi_{ij}(\boldsymbol{\beta})$ for Models (2.1)-(2.4) can be calculated directly, and the first derivative of $\pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ can be gotten through

$$\frac{\partial \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \pi_{ij}(\boldsymbol{\beta}) \frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \tag{S1.1}$$

as long as $\partial \log \pi_{ij}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ is obtained. Denote the cumulative categorical probability by $\xi_{ij}(\boldsymbol{\beta})$, i.e., $\xi_{i0}(\boldsymbol{\beta}) = 0$, $\xi_{ij}(\boldsymbol{\beta}) = \pi_{i1}(\boldsymbol{\beta}) + \cdots + \pi_{ij}(\boldsymbol{\beta}), j = 1, \ldots, J$. By direct calculations, the explicit forms of $\pi_{ij}(\boldsymbol{\beta})$ and $\partial \log \pi_{ij}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ are presented below for Models (2.1)-(2.4), respectively.

(i) Model (2.1).

The categorical probability is

$$\pi_{ij}(\boldsymbol{\beta}) = \begin{cases} \frac{\exp\{\boldsymbol{x}_{i(0)}^T \boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T \boldsymbol{\beta}_j\}}{1+\sum_{k=1}^{J-1} \exp\{\boldsymbol{x}_{i(0)}^T \boldsymbol{\beta}_0 + \boldsymbol{x}_{i(k)}^T \boldsymbol{\beta}_k\}} & j = 1, \ldots, J-1, \\ \frac{1}{1+\sum_{k=1}^{J-1} \exp\{\boldsymbol{x}_{i(0)}^T \boldsymbol{\beta}_0 + \boldsymbol{x}_{i(k)}^T \boldsymbol{\beta}_k\}} & j = J. \end{cases}$$

Recall $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{J-1}^T)^T$, the first derivative of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_0$ is

$$\frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_0} = \begin{cases} \pi_{iJ}(\boldsymbol{\beta}) \boldsymbol{x}_{i(0)} & j \neq J, \\ (\pi_{iJ}(\boldsymbol{\beta}) - 1) \boldsymbol{x}_{i(0)} & j = J, \end{cases}$$

and the first derivative of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_k$, $k = 1, \ldots, J-1$, is

$$\frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} = \begin{cases} (1 - \pi_{ik}(\boldsymbol{\beta})) \boldsymbol{x}_{i(k)} & k = j, \\ -\pi_{ik}(\boldsymbol{\beta}) \boldsymbol{x}_{i(k)} & k \neq j. \end{cases}$$

(ii) Model (2.2).

The categorical probability is

$$\pi_{ij}(\boldsymbol{\beta}) = \begin{cases} \frac{\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(1)}^T\boldsymbol{\beta}_1\}}{1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(1)}^T\boldsymbol{\beta}_1\}} & j=1, \\ \frac{\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j\}}{1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j\}} - \frac{\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}\}}{1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}\}} & j=2,\ldots,J-1, \\ \frac{1}{1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(J-1)}^T\boldsymbol{\beta}_{J-1}\}} & j=J. \end{cases}$$

The first derivative of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_0$ is

$$\frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_0} = (1 - \xi_{ij}(\boldsymbol{\beta}) - \xi_{i,j-1}(\boldsymbol{\beta}))\boldsymbol{x}_{i(0)},$$

and the first derivative of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_k$, $k = 1, \ldots, J-1$, is

$$\frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} = \begin{cases} -\frac{\xi_{ik}(\boldsymbol{\beta})(1-\xi_{ik}(\boldsymbol{\beta}))}{\pi_{i,k+1}(\boldsymbol{\beta})}\boldsymbol{x}_{i(k)} & k=j-1, \\ \frac{\xi_{ik}(\boldsymbol{\beta})(1-\xi_{ik}(\boldsymbol{\beta}))}{\pi_{ik}(\boldsymbol{\beta})}\boldsymbol{x}_{i(k)} & k=j. \end{cases}$$

(iii) Model (2.3).

The categorical probability is

$$\pi_{ij}(\boldsymbol{\beta}) = \begin{cases} \frac{\prod_{k=j}^{J-1}\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(k)}^T\boldsymbol{\beta}_k\}}{1+\sum_{l=1}^{J-1}\left(\prod_{k=l}^{J-1}\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(k)}^T\boldsymbol{\beta}_k\}\right)} & j=1,\ldots,J-1, \\ \frac{1}{1+\sum_{l=1}^{J-1}\left(\prod_{k=l}^{J-1}\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(k)}^T\boldsymbol{\beta}_k\}\right)} & j=J. \end{cases}$$

The first derivative of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_0$ is

$$\frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_0} = \left(J - j - \sum_{l=1}^{J-1}\xi_{il}(\boldsymbol{\beta})\right)\boldsymbol{x}_{i(0)},$$

and the first derivative of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_k$, $k = 1, \ldots, J-$

1, is

$$\frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} = \begin{cases} -\xi_{ik}(\boldsymbol{\beta})\boldsymbol{x}_{i(k)} & k < j, \\ (1-\xi_{ik}(\boldsymbol{\beta}))\boldsymbol{x}_{i(k)} & k \geq j. \end{cases}$$

(iv) Model (2.4).

The categorical probability is

$$\pi_{ij}(\boldsymbol{\beta}) = \begin{cases} \frac{\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j\}}{\prod_{k=1}^{j}\left(1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(k)}^T\boldsymbol{\beta}_k\}\right)} & j = 1, \ldots, J-1, \\ \frac{1}{\prod_{k=1}^{J-1}\left(1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(k)}^T\boldsymbol{\beta}_k\}\right)} & j = J. \end{cases}$$

The first derivative of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_0$ is

$$\frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_0} = \left(1 - \sum_{l=1}^{j} \frac{\pi_{il}(\boldsymbol{\beta})}{1-\xi_{i,l-1}(\boldsymbol{\beta})}\right)\boldsymbol{x}_{i(0)},$$

and the first derivative of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_k$, $k = 1, \ldots, J-1$, is

$$\frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} = \begin{cases} -\frac{\pi_{ik}(\boldsymbol{\beta})}{1-\xi_{i,k-1}(\boldsymbol{\beta})}\boldsymbol{x}_{i(k)} & k < j, \\ \frac{1-\xi_{ik}(\boldsymbol{\beta})}{1-\xi_{i,k-1}(\boldsymbol{\beta})}\boldsymbol{x}_{i(k)} & k = j, \\ 0 & k > j. \end{cases}$$

Note that all the first order derivatives contain the categorical probabilities and the predictors only. Recall that the first order derivatives of $\pi_{ij}(\boldsymbol{\beta})$ can be obtained through Equation (S1.1). Thus, one can easily calculate the corresponding second order derivatives. For Model (2.1), we have

$$\frac{\partial^2 \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0^T} = \pi_{iJ}(\boldsymbol{\beta})(\pi_{iJ}(\boldsymbol{\beta})-1)\boldsymbol{x}_{i(0)}\boldsymbol{x}_{i(0)}^T,$$

$$\frac{\partial^2 \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_k^T} = -\pi_{iJ}(\boldsymbol{\beta})\pi_{ik}(\boldsymbol{\beta})\boldsymbol{x}_{i(0)}\boldsymbol{x}_{i(k)}^T,$$

$$\frac{\partial^2 \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_l^T} = \begin{cases} -\pi_{ik}(\boldsymbol{\beta})(1-\pi_{ik}(\boldsymbol{\beta}))\boldsymbol{x}_{i(k)}\boldsymbol{x}_{i(k)}^T & l = k, \\ \pi_{ik}(\boldsymbol{\beta})\pi_{il}(\boldsymbol{\beta})\boldsymbol{x}_{i(k)}\boldsymbol{x}_{i(l)}^T & l \neq k, \end{cases}$$

where $j = 1, \ldots, J, k = 1, \ldots, J-1, l = 1, \ldots, J-1$. The second order derivatives of $\log \pi_{ij}(\boldsymbol{\beta})$ in Models (2.2)-(2.4) are similar to calculate and thus we omit it for simplicity.

S2. Theoretical Proofs

To proof Theorem 1, we start with introducing the following two lemmas.

Lemma S1. *For Models* (2.1)-(2.4)*, there exist two constant C_1, C_2 such that*

$$\max_{1 \leq j \leq J} |\log \pi_{ij}(\boldsymbol{\beta})| \leq C_1 \|\boldsymbol{x}_i\| \|\boldsymbol{\beta}\| + C_2,$$

for any $i = 1, \ldots, N$.

Proof of Lemma S1. We prove this lemma for the four models separately. Utilizing the facts that $\|\boldsymbol{x}_{i(k)}\| \leq \|\boldsymbol{x}_i\|, \|\boldsymbol{\beta}_k\| \leq \|\boldsymbol{\beta}\|$, for $i = 1, \ldots, N, k = 0, \ldots, J-1$, and $\log(1+x) < 1 + \log x$, for all $x \geq 1$, we have the following conclusions.

(i) Model (2.1).

Directly calculation yields

$$\begin{aligned}
&|\log \pi_{ij}(\boldsymbol{\beta})| \\
\leq &\max_{1\leq j\leq J-1}\{|\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j|\} + \log\left(1+\sum_{j=1}^{J-1}\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j\}\right) \\
\leq &2\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\| + \log\left(1+(J-1)\exp\{2\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|\}\right) \\
\leq &2\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\| + 1 + \log(J-1) + 2\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\| \\
= &4\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\| + 1 + \log(J-1).
\end{aligned}$$

(ii) Model (2.2).

The cases $j = 1$ and $j = J$ are similar to the cases in Model (2.1). When $j = 2, \ldots, J-1$, by the mean-value theorem, there exists $\xi \in (\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}, \boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j)$ such that

$$\pi_{ij}(\boldsymbol{\beta}) = \frac{e^\xi}{(1+e^\xi)^2}\left(\boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j - \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}\right).$$

Since

$$\begin{aligned}
&\left|\log\frac{\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(k)}^T\boldsymbol{\beta}_k\}}{\left(1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(k)}^T\boldsymbol{\beta}_k\}\right)^2}\right| \\
\leq &|\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(k)}^T\boldsymbol{\beta}_k| + 2\log\left(1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(k)}^T\boldsymbol{\beta}_k\}\right) \\
\leq &2\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\| + 2\log(1+\exp\{2\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|\}) \\
\leq &2\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\| + 2(1+2\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|)
\end{aligned}$$

$$=6\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|+2,$$

for $k=j-1, j$. Since the maximum point of the function $|\log(e^t/(1+e^t)^2)|$ with t belonging to a closed interval is the endpoint, and $\xi \in (\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}, \boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j)$, then

$$\left|\log\frac{e^\xi}{(1+e^\xi)^2}\right| \leq 6\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|+2.$$

Combining Assumption 2, we have

$$|\log\pi_{ij}(\boldsymbol{\beta})| \leq 6\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|+2+|\log c_0|.$$

(iii) Model (2.3).

Simple calculation yields

$$\begin{aligned}
&|\log\pi_{ij}(\boldsymbol{\beta})|\\
\leq&\left|\sum_{k=j}^{J-1}\left(\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(k)}^T\boldsymbol{\beta}_k\right)\right|+\log\left[1+\sum_{j=1}^{J-1}\left(\prod_{k=j}^{J-1}\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0+\boldsymbol{x}_{i(k)}^T\boldsymbol{\beta}_k\}\right)\right]\\
\leq&2(J-1)\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|+\log\left(1+(J-1)\exp\{2(J-1)\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|\}\right)\\
\leq&2(J-1)\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|+1+\log(J-1)+2(J-1)\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|\\
=&4(J-1)\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|+1+\log(J-1).
\end{aligned}$$

(iv) Model (2.4).

We have

$$\begin{aligned}
&|\log \pi_{ij}(\boldsymbol{\beta})| \\
\leq & \max_{1\leq j\leq J-1} \left|\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j\right| + \log\left[\prod_{k=1}^{J-1}\left(1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(k)}^T\boldsymbol{\beta}_k\}\right)\right] \\
\leq & 2\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\| + (J-1)\log(1+\exp\{2\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|\}) \\
\leq & 2\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\| + (J-1)\left(1+2\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\|\right) \\
= & 2J\|\boldsymbol{x}_i\|\|\boldsymbol{\beta}\| + J - 1.
\end{aligned}$$

Combining the four cases, this lemma has been proved. □

Lemma S2. *For Models* (2.1)-(2.4)*, there exists a constant C, such that*

$$\begin{aligned}
\left\|\frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\| &\leq C\|\boldsymbol{x}_i\|, \\
\left\|\frac{\partial^2 \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\right\|_F &\leq C\|\boldsymbol{x}_i\|^2, \\
\left\|\frac{\partial^3 \log \pi_{ij}(\boldsymbol{\beta})}{\partial \bar{\beta}_l\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\right\|_F &\leq C\|\boldsymbol{x}_i\|^3,
\end{aligned}$$

for $i = 1, \ldots, N, j = 1, \ldots, J$, $l = 1, \ldots, d$, and $\bar{\beta}_l$ is the lth item of $\boldsymbol{\beta}$, where $\|\cdot\|_F$ denotes the Frobenius norm of the corresponding matrix.

Proof of Lemma S2. As illustrated in Section S1, it is easily to calculate the second and third order derivatives of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ through the explicit forms of $\pi_{ij}(\boldsymbol{\beta})$ and the first order derivatives of $\log \pi_{ij}(\boldsymbol{\beta})$, with Equation (S1.1). Now we prove this lemma for the four models separately.

(i) Model (2.1).

By the explicit forms of the first order and second order derivatives in Section S1, we know that

$$\left\| \frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right\| \le \sum_{k=0}^{J-1} \|\boldsymbol{x}_{i(k)}\| \le J\|\boldsymbol{x}_i\|,$$

and

$$\left\| \frac{\partial^2 \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\|_F \le \sum_{k=0}^{J-1} \sum_{l=0}^{J-1} \|\boldsymbol{x}_{i(k)}\| \|\boldsymbol{x}_{i(l)}\| \le J^2 \|\boldsymbol{x}_i\|.$$

Note that the first order derivative of $\pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_0$ is

$$\frac{\partial \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_0} = \begin{cases} \pi_{ij}(\boldsymbol{\beta})\pi_{iJ}(\boldsymbol{\beta})\boldsymbol{x}_{i(0)} & j \neq J, \\ \pi_{iJ}(\boldsymbol{\beta})(\pi_{iJ}(\boldsymbol{\beta}) - 1)\boldsymbol{x}_{i(0)} & j = J, \end{cases}$$

and the first order derivative of $\pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_k$, $k = 1, \ldots, J-1$, is

$$\frac{\partial \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} = \begin{cases} \pi_{ij}(\boldsymbol{\beta})(1 - \pi_{ik}(\boldsymbol{\beta}))\boldsymbol{x}_{i(k)} & k = j, \\ -\pi_{ij}(\boldsymbol{\beta})\pi_{ik}(\boldsymbol{\beta})\boldsymbol{x}_{i(k)} & k \neq j. \end{cases}$$

Thus, we have

$$\left\| \frac{\partial \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} \right\| \le \|\boldsymbol{x}_{i(k)}\|,$$

for $k = 0, 1, \ldots, J-1$. Combining with the explicit form of $\partial^2 \log \pi_{ij}(\boldsymbol{\beta})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T$, it follows that

$$\left\| \frac{\partial^3 \log \pi_{ij}(\boldsymbol{\beta})}{\partial \bar{\beta}_l \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\|_F \le J^2 \|\boldsymbol{x}_i\|^3,$$

for any $l = 1, \ldots, d$.

(ii) Model (2.2).

The cases $j = 1$ and $j = J$ are similar to the case in Model (2.1) and we omit it for simplicity. For $j = 2, \ldots, J-1$, we have

$$\begin{aligned}
&\log \pi_{ij}(\boldsymbol{\beta})\\
=&\log\left(\frac{\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j\}}{1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j\}} - \frac{\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}\}}{1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}\}}\right)\\
=&(\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}) + \log\left(\exp\{\boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j - \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}\} - 1\right)\\
&-\log\left(1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j\}\right) - \log\left(1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}\}\right).
\end{aligned}$$

The first order derivative of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_0$ is

$$\left(1 - \frac{\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j\}}{1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j\}} - \frac{\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}\}}{1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}\}}\right)\boldsymbol{x}_{i(0)},$$

which implies $\|\partial \log \pi_{ij}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}_0\| \le 2\|\boldsymbol{x}_i\|$. Since the first and second order derivatives of $e^t/(1+e^t)$ with respect to t are bounded by 1, thus we have $\|\partial^2 \log \pi_{ij}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_k^T\|_F \le 2\|\boldsymbol{x}_i\|^2$, $\|\partial^3 \log \pi_{ij}(\boldsymbol{\beta})/\partial \bar{\beta}_l \partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_k^T\|_F \le 2\|\boldsymbol{x}_i\|^3$, for $k = 0, \ldots, J-1, l = 1, \ldots, p$.

Now we turns to consider the first order derivative of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_j$ (the case $\boldsymbol{\beta}_{j-1}$ is similar), i.e.,

$$\left(\frac{\exp\{\boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j - \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}\}}{\exp\{\boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j - \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1}\} - 1} - \frac{\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j\}}{1+\exp\{\boldsymbol{x}_{i(0)}^T\boldsymbol{\beta}_0 + \boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j\}}\right)\boldsymbol{x}_{i(j)}.$$

By Assumption 2, $\boldsymbol{x}_{i(j)}^T\boldsymbol{\beta}_j - \boldsymbol{x}_{i(j-1)}^T\boldsymbol{\beta}_{j-1} > c_0$. Since the function

$e^t/(e^t-1), e^t/(1+e^t)$, and their first, second order derivatives are all bounded with $t \in (c_0, +\infty)$, this lemma holds.

(iii) Model (2.3).

The first order derivative of $\pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_0$ is

$$\frac{\partial \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_0} = \pi_{ij}(\boldsymbol{\beta}) \left(J - j + \sum_{l=1}^{J-1} \xi_{il}(\boldsymbol{\beta}) \right) \boldsymbol{x}_{i(0)},$$

and the first order derivative of $\pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_k$, $k = 1, \ldots, J-1$, is

$$\frac{\partial \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} = \begin{cases} -\pi_{ij}(\boldsymbol{\beta}) \xi_{ik}(\boldsymbol{\beta}) \boldsymbol{x}_{i(k)} & k < j, \\ \pi_{ij}(\boldsymbol{\beta})(1 - \xi_{ik}(\boldsymbol{\beta})) \boldsymbol{x}_{i(k)} & k \geq j, \end{cases}$$

which implies that

$$\left\| \frac{\partial \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} \right\| \leq J \|\boldsymbol{x}_i\|,$$

for $k = 0, 1, \ldots, J-1$. By using the same method in case (1), this lemma holds.

(iv) Model (2.4).

The first order derivative of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_0$ is

$$\frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_0} = \begin{cases} \left(1 - \sum_{l=1}^{j} \frac{\exp\{\boldsymbol{x}_{i(0)}^T \boldsymbol{\beta}_0 + \boldsymbol{x}_{i(l)}^T \boldsymbol{\beta}_l\}}{1 + \exp\{\boldsymbol{x}_{\boldsymbol{i(0)}}^{\boldsymbol{T}} \boldsymbol{\beta}_0 + \boldsymbol{x}_{i(l)}^T \boldsymbol{\beta}_l\}}\right) \boldsymbol{x}_{i(0)} & j < J, \\ -\left(\sum_{l=1}^{J-1} \frac{\exp\{\boldsymbol{x}_{\boldsymbol{i(0)}}^{\boldsymbol{T}} \boldsymbol{\beta}_0 + \boldsymbol{x}_{\boldsymbol{i(l)}}^{\boldsymbol{T}} \boldsymbol{\beta}_l\}}{1 + \exp\{\boldsymbol{x}_{\boldsymbol{i(0)}}^{\boldsymbol{T}} \boldsymbol{\beta}_0 + \boldsymbol{x}_{i(l)}^T \boldsymbol{\beta}_l\}}\right) \boldsymbol{x}_{i(0)} & j = J, \end{cases}$$

and the first order derivative of $\log \pi_{ij}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_k$, $k = 1, \ldots, J-1$, is

$$\frac{\partial \log \pi_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} = \begin{cases} -\frac{\exp\{\boldsymbol{x}_{i(0)}^T \boldsymbol{\beta}_0 + \boldsymbol{x}_{\boldsymbol{i(k)}}^{\boldsymbol{T}} \boldsymbol{\beta}_k\}}{1+\exp\{\boldsymbol{x}_{\boldsymbol{i(0)}}^{\boldsymbol{T}} \boldsymbol{\beta}_0 + \boldsymbol{x}_{i(k)}^T \boldsymbol{\beta}_k\}} \boldsymbol{x}_{i(k)} & k < j, \\ \frac{1}{1+\exp\{\boldsymbol{x}_{i(0)}^T \boldsymbol{\beta}_0 + \boldsymbol{x}_{i(k)}^T \boldsymbol{\beta}_k\}} \boldsymbol{x}_{i(k)} & k = j, \\ 0 & k > j. \end{cases}$$

Since the first order and second order derivatives of functions $e^t/(1+e^t)$ with respect to t are bounded, this lemma holds.

Combining the four cases above, we have proved this lemma. □

Now we turn to prove Theorem 1.

Proofs of Theorem 1. Recall that $\ell(\boldsymbol{\beta})$ is the log-likelihood function on the full dataset. For the subsample, the weighted log-likelihood function (2.6) can be written as

$$\ell^*(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{R_i}{p_i} \boldsymbol{\delta}_i^T \log \boldsymbol{\pi}_i(\boldsymbol{\beta}),$$

where R_i is the Bernoulli variable with probability p_i, $\boldsymbol{\delta}_i = (\mathbb{I}(y_i = 1), \ldots, \mathbb{I}(y_i = J))^T$, and $\boldsymbol{\pi}_i(\boldsymbol{\beta}) = (\boldsymbol{\pi}_{i1}(\boldsymbol{\beta}), \ldots, \boldsymbol{\pi}_{iJ}(\boldsymbol{\beta}))^T$.

Direct calculation shows that,

$$\begin{aligned} E\left(\frac{1}{N}\ell^*(\boldsymbol{\beta})\middle| \mathcal{F}_N\right) &= \frac{1}{N}\ell(\boldsymbol{\beta}), \\ \operatorname{Var}\left(\frac{1}{N}\ell^*(\boldsymbol{\beta})\middle| \mathcal{F}_N\right) &= \frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{1}{p_i} - 1\right)\left(\boldsymbol{\delta}_i^T \log \boldsymbol{\pi}_i(\boldsymbol{\beta})\right)^2 \end{aligned}$$

$$\le \frac{1}{N^2}\sum_{i=1}^{N}\frac{\left(\boldsymbol{\delta}_i^T \log \boldsymbol{\pi}_i(\boldsymbol{\beta})\right)^2}{p_i}$$

$$\le \left(\max_{i=1,\dots,N}\frac{1}{Np_i}\right)\sum_{i=1}^{N}\frac{\left(\boldsymbol{\delta}_i^T \log \boldsymbol{\pi}_i(\boldsymbol{\beta})\right)^2}{N}.$$

By Lemma S1 and Assumptions 4, 5, we have

$$\mathrm{Var}\left(\frac{1}{N}\ell^*(\boldsymbol{\beta})\middle|\mathcal{F}_N\right) = O_P(n^{-1}).$$

Thus, as $n \to \infty$, $\ell^*(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}) \to 0$ in conditional probability given $\mathcal{F}_N$ for all $\boldsymbol{\beta}$. Note that the parameter space is compact, $\hat{\boldsymbol{\beta}}_{sub}$ and $\hat{\boldsymbol{\beta}}_{full}$ are the unique global maximums of the continuous concave functions $\ell^*(\boldsymbol{\beta})$ and $\ell(\boldsymbol{\beta})$, respectively. Thus, from Theorem 5.9 and its remark in Van der Vaart (1998), we obtain that

$$\|\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full}\| = o_{P|\mathcal{F}_N}(1). \tag{S2.2}$$

Using Taylor's Theorem (Ferguson, 1996),

$$0 = \dot{\ell}_k^*(\hat{\boldsymbol{\beta}}_{sub}) = \dot{\ell}_k^*(\hat{\boldsymbol{\beta}}_{full}) + \frac{\partial \dot{\ell}_k^*(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full}) + R_k,$$

where $\dot{\ell}_k^*(\cdot)$ is the partial derivative of $\ell^*(\cdot)$ with respect to the kth item of the parameter vector $\boldsymbol{\beta}$, and

$$R_k = \left(\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full}\right)^T \int_0^1\int_0^1 \frac{\partial^2 \dot{\ell}_k^*(\hat{\boldsymbol{\beta}}_{full} + uv(\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full}))}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} v du dv \left(\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full}\right).$$

By Lemma S2, we have

$$\left\|\frac{1}{N}\frac{\partial \dot{\ell}_k^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\right\| \le \frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{p_i}C\|\boldsymbol{x}_i\|^3,$$

where C is a constant. By Markov inequality,

$$P\left(\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{p_i}C\|\boldsymbol{x}_i\|^3 \geq \tau \middle| \mathcal{F}_N\right) \leq \frac{C}{N\tau}\sum_{i=1}^{N}E\left(\frac{R_i\|\boldsymbol{x}_i\|^3}{p_i}\middle| \mathcal{F}_N\right)$$
$$= \frac{C}{N\tau}\sum_{i=1}^{N}\|\boldsymbol{x}_i\|^3$$
$$\to 0,$$

as $\tau \to \infty$. Thus

$$\frac{1}{N}\sup_{u,v}\left\|\frac{\partial^2 \dot{\ell}_k^*(\hat{\boldsymbol{\beta}}_{full} + uv(\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full}))}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right\| = O_{P|\mathcal{F}_N}(1),$$

and then

$$\frac{1}{N}R_k = O_{P|\mathcal{F}_N}\left(\|\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full}\|^2\right).$$

Denote

$$M_N^*(\hat{\boldsymbol{\beta}}_{full}) = \frac{1}{N}\frac{\partial^2 \ell^*(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T},$$

and hence,

$$\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full} = -M_N^{*-1}(\hat{\boldsymbol{\beta}}_{full})\left(\frac{1}{N}\frac{\partial \ell^*(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}} + O_{P|\mathcal{F}_N}\left(\|\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full}\|^2\right)\right). \tag{S2.3}$$

Direct calculation yields

$$E(M_N^*(\hat{\boldsymbol{\beta}}_{full})|\mathcal{F}_N) = M_N(\hat{\boldsymbol{\beta}}_{full}).$$

For any component $M_N^{*(j_1j_2)}(\hat{\boldsymbol{\beta}}_{full})$ of $M_N^*(\hat{\boldsymbol{\beta}}_{full})$, where $1 \leq j_1, j_2 \leq d$,

$$\text{Var}\left(M_N^{*(j_1j_2)}(\hat{\boldsymbol{\beta}}_{full})|\mathcal{F}_N\right)$$

$$
\begin{aligned}
&=\sum_{i=1}^{N}\frac{p_i(1-p_i)}{p_i^2}\left[\frac{1}{N}\sum_{j=1}^{J}\mathbb{I}(y_i=j)\left(\frac{\partial^2\log\pi_{ij}(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{(j_1j_2)}\right]^2\\
&\leq\sum_{i=1}^{N}\frac{1}{p_i}\left[\frac{1}{N}\sum_{j=1}^{J}\mathbb{I}(y_i=j)\left(\frac{\partial^2\log\pi_{ij}(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{(j_1j_2)}\right]^2\\
&\leq\left(\max_{i=1,\ldots,N}\frac{1}{Np_i}\right)\frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{J}\mathbb{I}(y_i=j)\left(\frac{\partial^2\log\pi_{ij}(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{(j_1j_2)}\right]^2\\
&\leq\left(\max_{i=1,\ldots,N}\frac{1}{Np_i}\right)\frac{1}{N}\sum_{i=1}^{N}C^2\|\boldsymbol{x}_i\|^4.
\end{aligned}
$$

Combined with Markov's inequality and Assumption 4, 5, we have that

$$
M_N^*(\hat{\boldsymbol{\beta}}_{full})-M_N(\hat{\boldsymbol{\beta}}_{full})=O_{P|\mathcal{F}_N}(n^{-1/2}). \tag{S2.4}
$$

Note that

$$
E\left(\frac{1}{N}\frac{\partial\ell^*(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}}\middle|\mathcal{F}_N\right)=\frac{1}{N}\frac{\partial\ell(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}}=0, \tag{S2.5}
$$

and

$$
\begin{aligned}
&\operatorname{Var}\left(\frac{1}{N}\frac{\partial\ell^*(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}}\middle|\mathcal{F}_N\right)\\
=&\frac{1}{N^2}\sum_{i=1}^{N}\frac{1-p_i}{p_i}\left(\frac{\partial\log\boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}^T}\right)^T\boldsymbol{\delta}_i\boldsymbol{\delta}_i^T\left(\frac{\partial\log\boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}^T}\right),\\
\leq&\left(\max_{i=1,\ldots,N}\frac{1}{Np_i}\right)\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial\log\boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}^T}\right)^T\boldsymbol{\delta}_i\boldsymbol{\delta}_i^T\left(\frac{\partial\log\boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_{full})}{\partial\boldsymbol{\beta}^T}\right),
\end{aligned} \tag{S2.6}
$$

whose elements are bounded by

$$
\left(\max_{i=1,\ldots,N}\frac{1}{Np_i}\right)\frac{1}{N}\sum_{i=1}^{N}C^2\|\boldsymbol{x}_i\|^2=O_P(n^{-1}).
$$

By Markov's inequality, we know that

$$\frac{1}{N}\frac{\partial \ell^*(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}} = O_{P|\mathcal{F}_N}(n^{-1/2}). \tag{S2.7}$$

Note that Equation (S2.4) indicates that $M_N^{*-1}(\hat{\boldsymbol{\beta}}_{full}) = O_{P|\mathcal{F}_N}(1)$. Combining this with Equations (S2.2), (S2.3) and (S2.7),we have

$$\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full} = O_{P|\mathcal{F}_N}(n^{-1/2}) + o_{P|\mathcal{F}_N}(\|\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full}\|) = O_{P|\mathcal{F}_N}(n^{-1/2}). \tag{S2.8}$$

Denote

$$\boldsymbol{\eta}_i = \frac{R_i}{Np_i}\left(\frac{\partial \log \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}^T}\right)^T \boldsymbol{\delta}_i, i = 1, \ldots, N.$$

Then

$$\frac{1}{N}\frac{\partial \ell^*(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N \boldsymbol{\eta}_i.$$

For any $\varepsilon > 0$,

$$\begin{aligned}
&\sum_{i=1}^N E\left(\|\boldsymbol{\eta}_i\|^2 \mathbb{I}(\|\boldsymbol{\eta}_i\| > \varepsilon)\middle| \mathcal{F}_N\right)\\
\leq &\sum_{i=1}^N \frac{1}{\varepsilon} E\left(\|\boldsymbol{\eta}_i\|^3 \mathbb{I}(\|\boldsymbol{\eta}_i\| > \varepsilon)\middle| \mathcal{F}_N\right)\\
\leq &\sum_{i=1}^N \frac{1}{\varepsilon} E\left(\|\boldsymbol{\eta}_i\|^3\middle| \mathcal{F}_N\right)\\
= &\sum_{i=1}^N \frac{1}{\varepsilon}\frac{1}{N^3 p_i^2}\left\|\left(\frac{\partial \log \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}^T}\right)^T \boldsymbol{\delta}_i\right\|^3\\
\leq &\frac{1}{\varepsilon}\left(\max_{i=1,\ldots,N}\frac{1}{(Np_i)^2}\right)\frac{C^3}{N}\sum_{i=1}^N \|\boldsymbol{x}_i\|^3
\end{aligned}$$

$$=o_P(1),$$

by Assumption 4 and Lemma S2. By Equations (S2.5), (S2.6), the Lindeberg-Feller central limit theorem (see Van der Vaart, 1998, Proposition 2.27), conditional on $\mathcal{F}_N$ in probability,

$$\frac{1}{N} V_{Nc}^{-1/2}(\hat{\boldsymbol{\beta}}_{full}) \frac{\partial \ell^*(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}} \to N(\mathbf{0}, I_d),$$

where

$$V_{Nc}(\hat{\boldsymbol{\beta}}_{full}) = \frac{1}{N^2} \sum_{i=1}^{N} \frac{1-p_i}{p_i} \left(\frac{\partial \log \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}^T} \right)^T \boldsymbol{\delta}_i \boldsymbol{\delta}_i^T \left(\frac{\partial \log \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}^T} \right).$$

Equations (S2.3), (S2.8) imply

$$\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full} = -\frac{1}{N} M_N^{*-1}(\hat{\boldsymbol{\beta}}_{full}) \dot{\ell}^*(\hat{\boldsymbol{\beta}}_{full}) + O_{P|\mathcal{F}_N}(n^{-1}). \tag{S2.9}$$

From Equation (S2.4),

$$\begin{aligned} & M_N^{*-1}(\hat{\boldsymbol{\beta}}_{full}) - M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) \\ =& -M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})(M_N^*(\hat{\boldsymbol{\beta}}_{full}) - M_N(\hat{\boldsymbol{\beta}}_{full}))M_N^{*-1}(\hat{\boldsymbol{\beta}}_{full}) \\ =& O_{P|\mathcal{F}_N}(n^{-1/2}). \end{aligned} \tag{S2.10}$$

Based on Assumptions 3-5,

$$V = M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) V_{Nc}(\hat{\boldsymbol{\beta}}_{full}) M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) = O_P(n^{-1}).$$

Combining with Equations (S2.9), (S2.10), it follows that

$$V^{-1/2}(\hat{\boldsymbol{\beta}}_{sub} - \hat{\boldsymbol{\beta}}_{full})$$

$$
\begin{aligned}
&= - V^{-1/2} M_N^{*-1}(\hat{\boldsymbol{\beta}}_{full}) \frac{1}{N} \frac{\partial \ell^*(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}} + O_{P|\mathcal{F}_N}(n^{-1/2}) \\
&= - V^{-1/2} M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) \frac{1}{N} \frac{\partial \ell^*(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}} - V^{-1/2}(M_N^{*-1}(\hat{\boldsymbol{\beta}}_{full}) \\
&\quad - M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})) \frac{1}{N} \frac{\partial \ell^*(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}} + O_{P|\mathcal{F}_N}(n^{-1/2}) \\
&= - V^{-1/2} M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) V_{Nc}^{1/2}(\hat{\boldsymbol{\beta}}_{full}) V_{Nc}^{-1/2}(\hat{\boldsymbol{\beta}}_{full}) \frac{1}{N} \frac{\partial \ell^*(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}} + O_{P|\mathcal{F}_N}(n^{-1/2}).
\end{aligned}
$$

Theorem 1 holds by Slutsky's Theorem (Ferguson, 1996) and the fact that

$$
V^{-1/2} M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) V_{Nc}^{1/2}(\hat{\boldsymbol{\beta}}_{full}) (V^{-1/2} M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) V_{Nc}^{1/2}(\hat{\boldsymbol{\beta}}_{full}))^T = I_d.
$$

□

Proof of Theorem 2. In order to minimize the AMSE, i.e., tr(V), it is sufficient to solve the following optimization problem:

$$
\begin{aligned}
\min \quad & \sum_{i=1}^{N} \frac{1}{p_i} \left\| M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full}) \right\|^2, \\
\text{s.t.} \quad & \sum_{i=1}^{N} p_i = n, 0 < p_i \leq 1, i = 1, \ldots, N.
\end{aligned}
$$

This problem is essentially the same with the optimization problem (A.29) in Ai et al. (2021), therefore we omit the rest proofs for simplicity. □

The following lemma is needed to proof Theorem 3.

Lemma S3. *(Achlioptas, 2003, Theorem 1.1) Let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ be an arbitrary set of points, where $\boldsymbol{z}_i \in \mathcal{R}^d$, $T_3 \in \mathbb{R}^{r_3 \times d}$ be a JLT, $\nu_0, \beta > 0$. If r_3 satisfies*

$$
r_3 \geq \frac{4 + 2\beta}{\nu_0^2/2 - \nu_0^3/3} \log(N+1),
$$

with probability at least $1-(N+1)^{-\beta}$,

$$(1-\nu_0)\|\boldsymbol{z}_i\|^2 \le \|T_3\boldsymbol{z}_i\|^2 \le (1+\nu_0)\|\boldsymbol{z}_i\|^2.$$

Proof of Theorem 3. By the definition of T_2 and $\widehat{M_N}(\hat{\boldsymbol{\beta}}_{full})$, we know that

$$\widehat{M_N}(\hat{\boldsymbol{\beta}}_{full}) = \frac{1}{r_2}\sum_{i=1}^{N}\left(\tilde{R}_i\sum_{j=1}^{J}\mathbb{I}(y_i=j)\frac{\partial \log \pi_{ij}(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T}\right),$$

where $\tilde{R}_i$ is the number of times that the ith row of $\sqrt{N/r_2}I_N$ is selected, $(\tilde{R}_1,\ldots,\tilde{R}_N) = \text{Multinomial}(r_2; 1/N,\ldots,1/N)$. Directly calculation yields that

$$E\left(\widehat{M_N}(\hat{\boldsymbol{\beta}}_{full})|\mathcal{F}_N\right) = M_N(\hat{\boldsymbol{\beta}}_{full}).$$

For any component $\widehat{M_N}^{(j_1j_2)}(\hat{\boldsymbol{\beta}}_{full})$ of $\widehat{M_N}(\hat{\boldsymbol{\beta}}_{full})$, where $1 \le j_1, j_2 \le d$,

$$\begin{aligned}
&\text{Var}\left(\widehat{M_N}^{(j_1j_2)}(\hat{\boldsymbol{\beta}}_{full})|\mathcal{F}_N\right)\\
=&\frac{1}{r_2N}\sum_{i=1}^{N}\left[\sum_{j=1}^{J}\mathbb{I}(y_i=j)\left(\frac{\partial^2 \log \pi_{ij}(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{(j_1j_2)} - M_N^{(j_1j_2)}(\hat{\boldsymbol{\beta}}_{full})\right]^2\\
=&\frac{1}{r_2N}\sum_{i=1}^{N}\left[\sum_{j=1}^{J}\mathbb{I}(y_i=j)\left(\frac{\partial^2 \log \pi_{ij}(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{(j_1j_2)}\right]^2 - \frac{1}{r_2}\left(M_N^{(j_1j_2)}(\hat{\boldsymbol{\beta}}_{full})\right)^2\\
\le&\frac{1}{r_2N}\sum_{i=1}^{N}\left[\sum_{j=1}^{J}\mathbb{I}(y_i=j)\left(\frac{\partial^2 \log \pi_{ij}(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{(j_1j_2)}\right]^2.
\end{aligned}$$

Thus, by Lemma S2,

$$E\left(\|\widehat{M_N}(\hat{\boldsymbol{\beta}}_{full}) - M_N(\hat{\boldsymbol{\beta}}_{full})\|_F^2|\mathcal{F}_N\right) = \sum_{j_1=1}^{d}\sum_{j_2=1}^{d}\text{Var}\left(\widehat{M_N}^{(j_1j_2)}(\hat{\boldsymbol{\beta}}_{full})|\mathcal{F}_N\right)$$

$$\leq \frac{C^2}{r_2 N} \sum_{i=1}^{N} \|\boldsymbol{x}_i\|^4.$$

Combining with Markov's inequality, we have that

$$P\left(\left| \|\widehat{M_N}(\hat{\boldsymbol{\beta}}_{full}) - M_N(\hat{\boldsymbol{\beta}}_{full})\|_F > \sqrt{\frac{C^2}{r_2 \nu_2} \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{x}_i\|^4} \right| \mathcal{F}_N \right) \leq \nu_2. \tag{S2.11}$$

It follows that $\|\widehat{M_N}(\hat{\boldsymbol{\beta}}_{full}) - M_N(\hat{\boldsymbol{\beta}}_{full})\|_F = O_{P|\mathcal{F}_N}(r_2^{-1/2})$. There exists $\gamma \in (0, 1]$ such that $\lambda_{min}\left(\widehat{M_N}(\hat{\boldsymbol{\beta}}_{full})\right) \geq \gamma \lambda_{min}\left(M_N(\hat{\boldsymbol{\beta}}_{full})\right)$, where $\lambda_{min}(\cdot)$ denotes the minimal eigenvalue of the corresponding matrix.

Let $\beta = -\log \nu_1 / \log(N+1)$, $\nu_0 = \sqrt{(12\log(N+1) - 6\log \nu_1)/r_3}$, then $\nu_0 \leq 1/2$ and r_3 satisfies the condition in Lemma S3, with probability at least $1 - \nu_1$,

$$\left\| T_3 \widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full}) \right\| \leq \sqrt{1 + \nu_0} \left\| \widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full}) \right\|. \tag{S2.12}$$

Combine (S2.11) and (S2.12), with probability at least $(1 - \nu_1)(1 - \nu_2)$, conditional on $\mathcal{F}_N$,

$$\begin{aligned}
&\left| \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| - \|T_3 \widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| \right| \\
\leq &\left| \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| - \|\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| \right| \\
&\quad + \left| \|\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| - \|T_3 \widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| \right| \\
\leq &\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full}) - \widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| +
\end{aligned}$$

$$
\begin{aligned}
&\max\left\{\sqrt{1+\nu_0}-1, 1+\sqrt{1+\nu_0}\right\}\|\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full})\|_F\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| \\
\leq&\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\|_F\|\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full})\|_F\|M_N(\hat{\boldsymbol{\beta}}_{full})-\widehat{M_N}(\hat{\boldsymbol{\beta}}_{full})\|_F\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|+ \\
&\nu_0\|\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full})\|_F\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| \\
\leq&\lambda_{max}(M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}))\lambda_{max}(\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full}))dC\sqrt{\frac{1}{r_2\nu_2}\frac{1}{N}\sum_{i=1}^N\|\boldsymbol{x}_i\|^4}\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| \\
&+\nu_0\sqrt{d}\lambda_{max}\left(\widehat{M_N}^{-1}(\hat{\boldsymbol{\beta}}_{full})\right)\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| \\
=&\lambda_{min}^{-1}(M_N(\hat{\boldsymbol{\beta}}_{full}))\lambda_{min}^{-1}(\widehat{M_N}(\hat{\boldsymbol{\beta}}_{full}))dC\sqrt{\frac{1}{r_2\nu_2}\frac{1}{N}\sum_{i=1}^N\|\boldsymbol{x}_i\|^4}\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| \\
&+\sqrt{\frac{12\log(N+1)-6\log\nu_1}{r_3}}d\lambda_{min}(\widehat{M_N}(\hat{\boldsymbol{\beta}}_{full}))^{-1}\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| \\
=&\frac{\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|}{\gamma\lambda_{min}^2(M_N(\hat{\boldsymbol{\beta}}_{full}))}\sqrt{\frac{d^2C^2}{r_2\nu_2}\frac{1}{N}\sum_{i=1}^N\|\boldsymbol{x}_i\|^4} \\
&+\frac{\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|}{\gamma\lambda_{min}(M_N(\hat{\boldsymbol{\beta}}_{full}))}\sqrt{\frac{12\log(N+1)-6\log\nu_1}{r_3}}d.
\end{aligned}
$$

Since using a JLT has no much benefit when $r_3 \geq d$, the second term of the previous formula is omitted, which completes the proof. □

Proof of Theorem 4. Note that $n_0 n^{-1/2} \to 0$, we focus on the subsamples drawn in the second step only since the contribution of the first step subsamples to the likelihood function is $o_{P|\mathcal{F}_N}(n^{-1/2})$.

We reuse the notation of $\ell^*(\boldsymbol{\beta})$ to represent the subsampling likelihood function, i.e.,

$$\ell^*(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{R_i}{\breve{p}_i \wedge 1} \boldsymbol{\delta}_i^T \log \boldsymbol{\pi}_i(\boldsymbol{\beta}),$$

where $\breve{p}_i$ is defined in (4.10), and $R_i = 1$ if and only if $(\boldsymbol{x}_i, y_i)$ is in the subsample.

Since $\breve{p}_i$ satisfies $\breve{p}_i \geq \rho n/N$, $\max_{i=1,\dots,N}(N(\breve{p}_i \wedge 1))^{-1} = O_P(n^{-1})$, the subsampling probabilities satisfy Assumption 5. Using the same method in the proof of Theorem 1, we have

$$\frac{1}{N} \breve{V}_{Nc}^{-1/2}(\hat{\boldsymbol{\beta}}_{full}) \frac{\partial \ell^*(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}} \to N(\mathbf{0}, I_d),$$

in distribution, conditional on $\hat{\boldsymbol{\beta}}_{pilot}$ and $\mathcal{F}_N$, where

$$\breve{V}_{Nc}(\hat{\boldsymbol{\beta}}_{full}) = \frac{1}{N^2} \sum_{i=1}^{N} \frac{1 - (\breve{p}_i \wedge 1)}{\breve{p}_i \wedge 1} \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i^T(\hat{\boldsymbol{\beta}}_{full}).$$

The distance between $\breve{V}_{Nc}$ and $\grave{V}_{Nc}$ can be quantified as

$$\begin{aligned}
\|\breve{V}_{Nc}(\hat{\boldsymbol{\beta}}_{full}) - \grave{V}_{Nc}(\hat{\boldsymbol{\beta}}_{full})\|_F =& \left\| \frac{1}{N^2} \sum_{i=1}^{N} \left(\frac{1}{\breve{p}_i \wedge 1} - \frac{1}{\grave{p}_i \wedge 1} \right) \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full}) \boldsymbol{u}_i^T(\hat{\boldsymbol{\beta}}_{full}) \right\|_F \\
\leq& \frac{1}{N^2} \sum_{i=1}^{N} \left(\left| \frac{1}{\breve{p}_i \wedge 1} - \frac{1}{\grave{p}_i \wedge 1} \right| \|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|^2 \right) \\
\leq& \left(\max_{i=1,\dots,N} \frac{1}{N \grave{p}_i} \right) \frac{1}{N} \sum_{i=1}^{N} \left(\left| \frac{\grave{p}_i \wedge 1}{\breve{p}_i \wedge 1} - 1 \right| \|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|^2 \right) \\
\leq& \frac{C^2}{\rho n} \frac{1}{N} \sum_{i=1}^{N} \left(\left| \frac{\grave{p}_i \wedge 1}{\breve{p}_i \wedge 1} - 1 \right| \|\boldsymbol{x}_i\|^2 \right),
\end{aligned}$$

where the last inequality holds by Lemma S2.

Let $\breve{h}_i = \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\|$, $\grave{h}_i = \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|$, it follows that

$$\begin{aligned}
\left|\frac{\grave{p}_i \wedge 1}{\breve{p}_i \wedge 1} - 1\right| &\leq \frac{|\grave{p}_i - \breve{p}_i|}{\breve{p}_i} \\
&\leq \frac{N}{\rho n}|\grave{p}_i - \breve{p}_i| \\
&= \frac{N}{\rho n}\left|\frac{n\grave{h}_i}{\sum_{i=1}^N \grave{h}_i} - \frac{n\breve{h}_i}{\sum_{i=1}^N \breve{h}_i}\right| \\
&\leq \frac{N}{\rho}\left(\left|\frac{\grave{h}_i}{\sum_{i=1}^N \grave{h}_i} - \frac{\breve{h}_i}{\sum_{i=1}^N \grave{h}_i}\right| + \left|\frac{\breve{h}_i}{\sum_{i=1}^N \grave{h}_i} - \frac{\breve{h}_i}{\sum_{i=1}^N \breve{h}_i}\right|\right) \\
&= \frac{1}{\rho}\frac{1}{N^{-1}\sum_{i=1}^N \grave{h}_i}\left(|\grave{h}_i - \breve{h}_i| + \left|1 - \frac{\sum_{i=1}^N \grave{h}_i}{\sum_{i=1}^N \breve{h}_i}\right|\breve{h}_i\right).
\end{aligned}$$

Let $\boldsymbol{\beta}_t$ denote the true value of parameter vector. By Assumption 4 and Lemma S2, it follows that

$$\begin{aligned}
\|M_N(\hat{\boldsymbol{\beta}}_{full}) - M_N(\boldsymbol{\beta}_t)\|_F \leq& \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^J \mathbb{I}(y_i = j)\left\|\frac{\partial \log \pi_{ij}(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T} - \frac{\partial \log \pi_{ij}(\boldsymbol{\beta}_t)}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T}\right\|_F \\
\leq& \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^J \mathbb{I}(y_i = j)C\|\hat{\boldsymbol{\beta}}_{full} - \boldsymbol{\beta}_t\|\|\boldsymbol{x}_i\| \\
=& o_P(1),
\end{aligned}$$

where the last equality is because $\|\hat{\boldsymbol{\beta}}_{full} - \boldsymbol{\beta}_t\| = o_{P|\mathcal{F}_N}(1) = o_P(1)$ (see Xiong and Li, 2008, Theorem 3.3). Combining this with Assumption 3, $M_N(\boldsymbol{\beta}_t)$ is positive definite.

By the law of large number, we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\grave{h}_i &= \frac{1}{N}\sum_{i=1}^{N}\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| \\
&= \frac{1}{N}\sum_{i=1}^{N}\|M_N^{-1}(\boldsymbol{\beta}_t)\boldsymbol{u}_i(\boldsymbol{\beta}_t)\| + o_P(1) \\
&\geq \lambda_{\min}(M_N^{-1}(\boldsymbol{\beta}_t))\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{u}_i(\boldsymbol{\beta}_t)\| + o_P(1) \\
&= \lambda_{\min}(M_N^{-1}(\boldsymbol{\beta}_t))E\|\boldsymbol{u}_1(\boldsymbol{\beta}_t)\| + o_P(1),
\end{aligned}$$

thus,

$$\left(\frac{1}{N}\sum_{i=1}^{N}\grave{h}_i\right)^{-1} = O_P(1). \tag{S2.13}$$

In the same way, we have

$$\frac{1}{N}\sum_{i=1}^{N}\grave{h}_i^2 = O_P(1). \tag{S2.14}$$

By the triangle inequality, it follows that

$$\begin{aligned}
|\grave{h}_i - \breve{h}_i| &= \Big|\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| - \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\|\Big| \\
&\leq \Big|\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\| - \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\|\Big| \\
&\quad + \Big|\|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\| - \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\|\Big| \\
&\leq \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\|_F\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full}) - \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\| \\
&\quad + \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) - M_N^{-1}(\hat{\boldsymbol{\beta}}_{pilot})\|_F\|\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{pilot})\| \\
&= o_P(1),
\end{aligned}$$

where the last equality holds by noting Lemma S2, and

$$M_N^{-1}(\hat{\boldsymbol{\beta}}_{full}) - M_N^{-1}(\hat{\boldsymbol{\beta}}_{pilot}) = M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})(M_N(\hat{\boldsymbol{\beta}}_{full}) - M_N(\hat{\boldsymbol{\beta}}_{pilot}))M_N^{-1}(\hat{\boldsymbol{\beta}}_{pilot}).$$

Thus, we have

$$\frac{1}{N}\sum_{i=1}^{N}(\breve{h}_i - \grave{h}_i)^2 = o_P(1), \tag{S2.15}$$

$$\left|1 - \frac{\sum_{i=1}^{N}\grave{h}_i}{\sum_{i=1}^{N}\breve{h}_i}\right| = o_P(1). \tag{S2.16}$$

Combining Equations (S2.13), (S2.14), (S2.15), (S2.16), and Assumption 4, it can be seen that

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left|\frac{\grave{p}_i \wedge 1}{\breve{p}_i \wedge 1} - 1\right| \|\boldsymbol{x}_i\|^2 \\
\leq& \frac{1}{\rho N^{-1}\sum_{i=1}^{N}\grave{h}_i}\left(\frac{1}{N}\sum_{i=1}^{N}|\grave{h}_i - \breve{h}_i|\|\boldsymbol{x}_i\|^2 + \left|1 - \frac{\sum_{i=1}^{N}\grave{h}_i}{\sum_{i=1}^{N}\breve{h}_i}\right|\frac{1}{N}\sum_{i=1}^{N}\breve{h}_i\|\boldsymbol{x}_i\|^2\right) \\
\leq& \frac{1}{\rho N^{-1}\sum_{i=1}^{N}\grave{h}_i}\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}|\grave{h}_i - \breve{h}_i|^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_i\|^4}\right) \\
&+ \frac{1}{\rho N^{-1}\sum_{i=1}^{N}\grave{h}_i}\left(\left|1 - \frac{\sum_{i=1}^{N}\grave{h}_i}{\sum_{i=1}^{N}\breve{h}_i}\right|\sqrt{\frac{1}{N}\sum_{i=1}^{N}\breve{h}_i^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_i\|^4}\right) \\
=& o_P(1).
\end{aligned}$$

That is $\|\breve{V}_{Nc}(\hat{\boldsymbol{\beta}}_{full}) - \grave{V}_{Nc}(\hat{\boldsymbol{\beta}}_{full})\|_F = o_P(n^{-1})$.

Therefore, the desired results follow by Slutsky's theorem (Ferguson, 1996) and

$$\grave{V}^{-1/2}M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\breve{V}_{Nc}^{1/2}(\hat{\boldsymbol{\beta}}_{full})(\grave{V}^{-1/2}M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\breve{V}_{Nc}^{1/2}(\hat{\boldsymbol{\beta}}_{full}))^T$$

$$
\begin{aligned}
&=\grave{V}^{-1/2}M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\breve{V}_{Nc}(\hat{\boldsymbol{\beta}}_{full})M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\grave{V}^{-1/2}\\
&=\grave{V}^{-1/2}M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\grave{V}_{Nc}(\hat{\boldsymbol{\beta}}_{full})M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\grave{V}^{-1/2}+o_{P|\mathcal{F}_N}(n^{-1/2})\\
&=I_d+o_{P|\mathcal{F}_N}(n^{-1/2}).
\end{aligned}
$$

□

Proof of Theorem 5. Note that the subsampling probabilities in Algorithm 2 naturally satisfied Assumption 5. Thus Theorem 1 implies that $\sqrt{n}\|\hat{\boldsymbol{\beta}}_{ts}-\hat{\boldsymbol{\beta}}_{full}\|=O_{P|\mathcal{F}_N}(1)=O_P(1)$, where the last equality comes from Xiong and Li (2008). By the natural of the MLE, one can show that $\sqrt{N}\|\hat{\boldsymbol{\beta}}_{full}-\boldsymbol{\beta}_{true}\|=O_P(1)$.

Since $n\leq N$ by the virtual of Poisson sampling, it follows that $\|\hat{\boldsymbol{\beta}}_{ts}-\boldsymbol{\beta}_{true}\|\leq\|\hat{\boldsymbol{\beta}}_{ts}-\hat{\boldsymbol{\beta}}_{full}\|+\|\hat{\boldsymbol{\beta}}_{full}-\boldsymbol{\beta}_{true}\|=O_P(n^{-1/2})$. The result follows.

□

Proof of Remark 6. Note that the subsampling probabilities in Algorithm 2 naturally satisfied Assumption 5. This is a direct result from Theorem 1 by the definition of $\hat{\boldsymbol{\beta}}_{full}$.

□

Proof of Lemma 1. Without loss of generality, we only show the case $i=1$ here.

Let $\ell_{-1}(\boldsymbol{\beta}) = \sum_{i=2}^{N} \sum_{j=1}^{J} \mathbb{I}(y_i = j) \log \pi_{ij}(\boldsymbol{\beta})$. Using Taylor's Theorem (Ferguson, 1996), it follows that

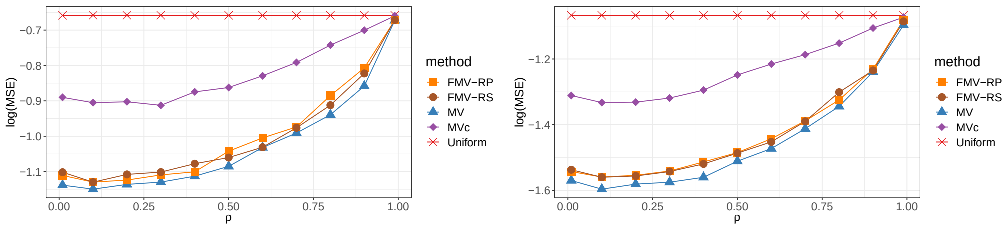
$$\mathbf{0} = \dot{\ell}_{-1}(\hat{\boldsymbol{\beta}}_{-1}) = \dot{\ell}_{-1}(\hat{\boldsymbol{\beta}}_{full}) + \frac{\partial \dot{\ell}_{-1}(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}}_{-1} - \hat{\boldsymbol{\beta}}_{full}) + O_P(\|\hat{\boldsymbol{\beta}}_{-1} - \hat{\boldsymbol{\beta}}_{full}\|^2),$$

where $\dot{\ell}_{-1}(\cdot)$ is the partial derivative of $\ell_{-1}(\cdot)$ with respect to the $\boldsymbol{\beta}$, and the remainder term comes from the similar techniques as Theorem 1 under Assumptions 1–4. Note that $\mathbf{0} = \dot{\ell}(\hat{\boldsymbol{\beta}}_{full})$. One can conclude that $\dot{\ell}_{-1}(\hat{\boldsymbol{\beta}}_{full}) = -\boldsymbol{u}_1(\hat{\boldsymbol{\beta}}_{full})$. Since the outliers are finite, the contribution of the likelihood can be ignored as $N \to \infty$. Thus both $\|\hat{\boldsymbol{\beta}}_{-1} - \boldsymbol{\beta}_{true}\|^2$ and $\|\hat{\boldsymbol{\beta}}_{full} - \boldsymbol{\beta}_{true}\|^2$ are $O_P(N^{-1})$. Let $M_{N-1}(\hat{\boldsymbol{\beta}}_{full}) = -N^{-1}\frac{\partial \dot{\ell}_{-1}(\hat{\boldsymbol{\beta}}_{full})}{\partial \boldsymbol{\beta}^T}$. Simple calculation yields $M_N^{-1}(\boldsymbol{\beta}_{true}) - M_{N-1}^{-1}(\boldsymbol{\beta}_{true}) = o_P(1)$. Thus the desired result holds by direct calculation. □

S3. Additional Simulation Results

In this section, we explore the effect of different ρ with fixed n_0 and n. We choose data generated in Case 1 as an example. The results are given in Figure S1 with $n_0 = 400$, $r_1 = r_2 = 5000$, $r_3 = 10$, for the cases $n = 1000, 1600$, respectively.

From Figure S1, one can see that $\rho = 0.1$ and 0.2 are slightly better than the case $\rho = 0$ which echoes the discussion in Section 4 that a proper specified ρ can lead a more stable and robust estimator. When ρ is close to



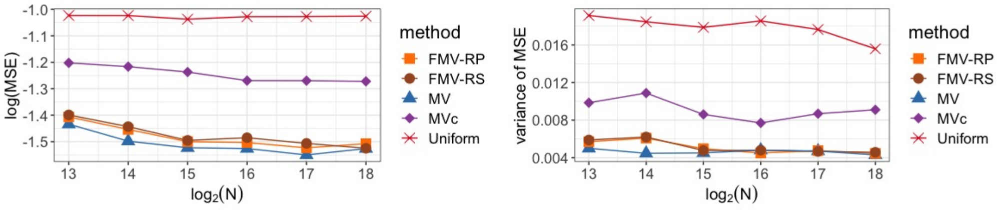
(a) $n = 1000$

(b) $n = 1600$

Figure S1: The log of MSE for Model (6.12) using data generated from Case 1 with different ρ based on MV, FMV-RP, FMV-RS, MVc, and Uniform methods, where $n_0 = 400$, $r_1 = r_2 = 5000$, $r_3 = 10$.

one, the performances of all the methods are similar since the probabilities are close to the uniform subsampling probabilities. This reflects that the optimal subsampling indeed improve the statistical efficiency within the same computing budget n.

In the following, we will illustrate the impact of the full sample size N. To ease the presentation, we take Case 1 with N varying from 2^{13} to 2^{18} as an example. The empirical mean squared error (MSE) of the resultant estimator $K^{-1}\sum_{k=1}^{K}\|\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}^{(k)} - \boldsymbol{\beta}_{true}\|^2$ are reported in Figure S2. Here we opt to report the distance between $\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}^{(k)}$ and $\boldsymbol{\beta}_{true}$, since $\hat{\boldsymbol{\beta}}_{full}$ changes according to different full samples.



(a) Log MSE

(b) Variance of MSE

Figure S2: The log of MSE and variance of MSE for Model (6.12) with different N varying from 2^{13} to 2^{18} based on MV, MVc, FMV-RP, FMV-RS, and Uniform subsampling, where $n_0 = 400, n = 1600$, $\rho = 0.2$, $r_1 = r_2 = 5000$, $r_3 = 10$. The data are generated under Case 1 at the beginning of Section 6.1.

Compared with the increase of n, the relative change is small as N increases. This echoes the result in Theorem 1 that the convergence rate is $O(n^{1/2})$, which does not depend on N. One can see that when N changes from 2^{13} to 2^{18}, there is a little change for the MSE. This is because the M in the optimal subsampling probabilities derived in Theorem 2 can not be simply specified as ∞ when n/N is not that small. According to the discussion in Remark 3, one can expect that our method assigns the inclusion probabilities to be one for some informative data points. Let $\mathcal{S}_1$ be the set consisting of such informative points. One can expect that our method spares the excess probability, i.e., $\sum_{\mathcal{S}_1}(n\hbar_i^{MV}/\sum_{j=1}^{N}\hbar_j^{MV}) - |\mathcal{S}_1|$, to the less informative data points, where $\hbar_i^{MV} = \|M_N^{-1}(\hat{\boldsymbol{\beta}}_{full})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{full})\|$ given in Theorem 2, and $|A|$ denotes the cardinality of set A. However, as N increases, there are more informative points and M can be specified as $+\infty$, which implies the inclusion probabilities are all strictly less than one and the excess probability is zero. Thus, we do not need to spare the excess probability to the less informative data points.

To evaluate the computing time of the fast approximation algorithm introduced in Section 3, we use the data generated in Case 1 as an example. All the computations are carried out on a MacBook Pro with a 3.1GHz Intel Core i5 processor and 8GB memory. Results on the average computing time (in seconds) for using the different methods with different expected subsample sizes and using the full data are reported in Table S1.

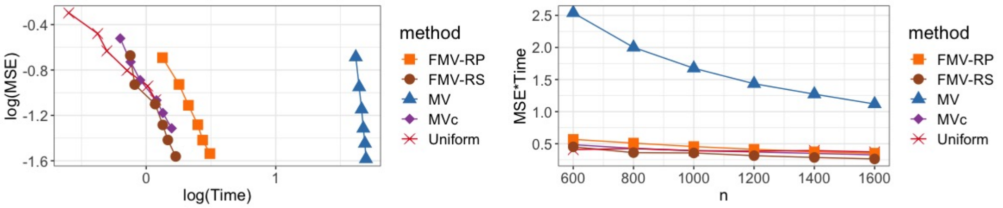
Compared with estimating parameters on the full data, all the subsampling methods use less time. As expected, both the FMV-RP and the FMV-RS methods are faster than the MV method. It is clear to see that FMV-RS takes less time than FMV-RP since FMV-RS does not perform a subsampled randomized Hadamard transform. The FMV-RS and MVc methods have similar performance in terms of computing time. From Table S1, one can clearly see that fitting the multinomial logistic regression

Table S1: Average Computing time (in seconds) with different n varying from 600 to 1600. All the settings are the same as the main text.

n	600	800	1000	1200	1400	1600
MV	5.043	5.176	5.276	5.345	5.412	5.455
FMV-RP	1.134	1.288	1.385	1.489	1.546	1.635
FMV-RS	0.884	0.915	1.071	1.136	1.181	1.255
MVc	0.818	0.886	0.954	1.082	1.137	1.217
Uniform	0.549	0.687	0.738	0.862	1.012	1.075
Full data computing time seconds: 35.122						

model takes most of the computing time. Thus, the sampling cost especially for FMV-RP, FMV-RS, and MVc sampling methods can be ignored.

To take a close look at the trade-off between computational cost and estimation efficiency, we also illustrate the results for MSE against computation times under Case 1. As suggested by a reviewer, another indicator MSE times computing time is also reported in Figure S3.



(a) MSE v.s. Time.

(b) MSE × Time.

Figure S3: Graphs showing the trade-off between computational time (in seconds) and MSE for the five approaches under Case 1 as in the main text.

From Figure S3, we can see that the uniform subsampling has its own advantage when the computational budget is very limited. Both MVc and FMV-RS produce a smaller MSE compared with the uniform subsampling method in the same computing time when n is more than 1000. From Figure S3(a), one can see a cross between the uniform subsampling and FMV-RS, which implies the advantage becomes more evident with the increase in the used CPU time. Compared with the MV approach, both FMV-RP and FMV-RS methods yield merely the same MSE with less computing time, which echoes the discussions in Section 3. From Figure S3(b), we can see that the FMV-RP, FMV-RS, and MVc methods have similar performance compared with the uniform subsampling approach. In addition, one may expect that as the subsample size n increases, but is still much smaller than N, the performance of the proposed methods will be better than the uniform subsampling since the difference between the uniform subsampling and MV methods becomes small. It is worth mentioning that the FMV-RP, FMV-RS, and MVc methods still have their own advantage. For example, if the available memory only allows the analysis of a subsample of size n while the computational time is relatively cheap, then our approaches may be preferable as it often results in the same statistical accuracy with fewer subdata points. This also holds for the MV method.

To see the computational cost under different full data sizes N, we also present the results on the average computing time (in seconds) for using the five methods with $n = 1600$. The results are reported in Table S2. For reference, we also report the computing time of the full data approach.

From Table S2, it is clear that the sampling costs of the proposed methods increase as N increases, since the computational costs of FMV-RP, FMV-RS, and MVc are $O(Nd)$ and the computational cost of MV is $O(Nd^2)$. This echoes the discussions in Section 3. As in Table S1, the advantages of the FMV-RS and FMV-RP compared with the uniform subsampling are still obvious, since in most cases, the FMV-RS and FMV-RP reduce 40% MSE while they do not require 40% additional time of the uniform subsampling.

Table S2: Average Computing time (in seconds) with different N varying from 2^{13} to 2^{18}. Here we fix $n = 1600$, $n_0 = 400$, $\rho = 0.2$, $r_1 = r_2 = 5000$, $r_3 = 10$.

N	2^{13}	2^{14}	2^{15}	2^{16}	2^{17}	2^{18}
MV	1.588	2.116	3.220	5.455	10.125	18.792
FMV-RP	1.178	1.272	1.433	1.635	2.162	3.241
FMV-RS	1.096	1.142	1.211	1.255	1.498	1.920
MVc	1.065	1.072	1.118	1.217	1.380	1.615
Uniform	1.061	1.064	1.068	1.075	1.137	1.167
Full data	5.541	11.171	20.517	35.122	76.785	137.582

It is worth mentioning that Theorem 4 enables us to draw inference on $\boldsymbol{\beta}$. Under the big data setups, the expected subsample size n is much less than the full data size N. If $n = o(N)$, then the full data MLE $\hat{\boldsymbol{\beta}}_{full}$ in Theorem 4 can be replaced by the true parameter. We take β_{01} as an example and construct a 95% confidence interval for it. The MVc and uniform subsampling methods are also implemented for comparison. Since the methods in four cases have similar performance, we only report the first two cases for brevity and the results are reported in Table S3 with $n_0 = 400$,

$r_1 = r_2 = 5000, r_3 = 10$, and $n = 800, 1200, 1600$.

Clearly, the subsampling methods based on MV, FMV-RP, FMV-RS, MVc have similar performances and are uniformly better than the uniform subsampling method. Moreover, the lengths of confidence intervals decrease when the expected subsample size n increases.

For nominal categorical data, Yao and Wang (2019) used subsampling with replacement in the softmax regression under non-proportional odds assumption. To compare our methods with the method proposed by Yao and Wang (2019), we take Model (2.1) with $J = 3$ and non-proportional odds assumption as an example. The MVc method mentioned in Section 6 and the uniform subsampling method are also considered for comparison. For reference, we list the model used in this section as follows.

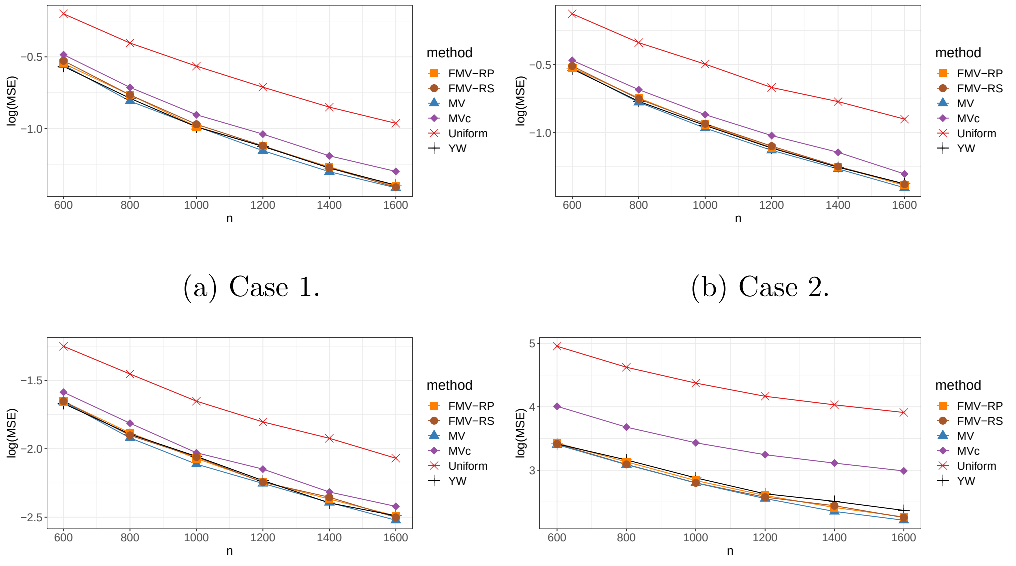
$$\begin{aligned} \log\left(\frac{\pi_{i1}}{\pi_{i3}}\right) &= \boldsymbol{x}_i^T \boldsymbol{\beta}_1, \\ \log\left(\frac{\pi_{i2}}{\pi_{i3}}\right) &= \boldsymbol{x}_i^T \boldsymbol{\beta}_2. \end{aligned} \tag{S3.17}$$

Here we set $\boldsymbol{\beta}_1 = 0.5 \times \mathbf{1}_{15}$, $\boldsymbol{\beta}_2 = \mathbf{1}_{15}$, where $\mathbf{1}_{15}$ is a 15 dimensional all-ones vector. The corresponding covariate $\boldsymbol{x}_i \in \mathbb{R}^{30}$ with $N = 2^{16}$ is generated in the same scenarios in Section 6.1. We set $n_0 = 400$, $\rho = 0.2$, $r_1 = r_2 = 5000$, $r_3 = 10$, and the expected subsample size n to be 600, 800, 1000, 1200, 1400, and 1600. We report the results in Figure S4, where "YW" represents the method proposed by Yao and Wang (2019).

Table S3: Empirical coverage probabilities and average lengths of confidence intervals for Cases 1 and 2.

			MV	FMV-RP	FMV-RS	MVc	Uniform
Case 1	800	Coverage	0.9349	0.9389	0.9369	0.9329	0.9399
		Length	0.3293	0.3325	0.3327	0.3297	0.3505
	1200	Coverage	0.9469	0.9409	0.9449	0.9449	0.9449
		Length	0.2812	0.2839	0.2842	0.2816	0.3009
	1600	Coverage	0.9499	0.9329	0.9339	0.9419	0.9239
		Length	0.2488	0.2516	0.2518	0.2499	0.2675
Case 2	800	Coverage	0.9389	0.9539	0.9599	0.9501	0.9449
		Length	0.3554	0.3579	0.3573	0.3579	0.3661
	1200	Coverage	0.9409	0.9499	0.9519	0.9419	0.9489
		Length	0.3035	0.3072	0.3085	0.3081	0.3147
	1600	Coverage	0.9509	0.9409	0.9459	0.9449	0.9459
		Length	0.2701	0.2731	0.2743	0.2747	0.2793

Empirical coverage probabilities and average lengths of confidence intervals for β_{01} in Model (6.12) with different n based on MV, FMV-RP, FMV-RS, MVc, and Uniform methods, where $n_0 = 400, \rho = 0.2$, $r_1 = r_2 = 5000$, $r_3 = 10$.



(a) Case 1.

(b) Case 2.

(c) Case 3.

(d) Case 4.

Figure S4: The log of MSE for Model (S3.17) with different n based on MV, MVc, FMV-RP, FMV-RS, Uniform, and the method proposed in Yao and Wang (2019), where $n_0 = 400$, $\rho = 0.2$, $r_1 = r_2 = 5000$, $r_3 = 10$. The different distributions of covariates are listed in the beginning of Section 6.1.

As shown in Figure S4, the empirical MSEs for YW are close to that for our methods and are uniformly smaller than the MVc and uniform subsampling methods. Nevertheless, YW needs to calculate all the inclusion probabilities at once, which takes a large memory to implement and may be infeasible in the big data setting. The Poisson subsampling, compared with subsampling with replacement, also has a high estimation efficiency with nonuniform subsampling probabilities.

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences 66*(4), 671–687.

Ai, M., F. Wang, J. Yu, and H. Zhang (2021). Optimal subsampling for large-scale quantile regression. *Journal of Complexity 62*, 101512.

Ferguson, T. S. (1996). *A Course in Large Sample Theory.* Chapman & Hall.

Van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Xiong, S. and G. Li (2008). Some results on the convergence of conditional distributions. *Statistics & Probability letters 78*(18), 3249–3253.

Yao, Y. and H. Wang (2019). Optimal subsampling for softmax regression. *Statistical Papers 60*(2), 585–599.