SCIENCE CHINA
Mathematics

• ARTICLES •

February 2024 Vol. 67 No. 2: 405–430
https://doi.org/10.1007/s11425-022-2057-8

# Optimal decorrelated score subsampling for generalized linear models with massive data

Junzhuo Gao[1], Lei Wang[1,*] & Heng Lian[2]

[1]*School of Statistics and Data Science & LPMC, Nankai University, Tianjin* 300071*, China;*
[2]*Department of Mathematics, City University of Hong Kong, Hong Kong, China*

*Email: junzhuogao*1012@163*.com, lwangstat@nankai.edu.cn, henglian@cityu.edu.hk*

Received April 19, 2022; accepted November 1, 2022; published online June 29, 2023

**Abstract** In this paper, we consider the unified optimal subsampling estimation and inference on the low-dimensional parameter of main interest in the presence of the nuisance parameter for low/high-dimensional generalized linear models (GLMs) with massive data. We first present a general subsampling decorrelated score function to reduce the influence of the less accurate nuisance parameter estimation with the slow convergence rate. The consistency and asymptotic normality of the resultant subsample estimator from a general decorrelated score subsampling algorithm are established, and two optimal subsampling probabilities are derived under the A- and L-optimality criteria to downsize the data volume and reduce the computational burden. The proposed optimal subsampling probabilities provably improve the asymptotic efficiency of the subsampling schemes in the low-dimensional GLMs and perform better than the uniform subsampling scheme in the high-dimensional GLMs. A two-step algorithm is further proposed to implement, and the asymptotic properties of the corresponding estimators are also given. Simulations show satisfactory performance of the proposed estimators, and two applications to census income and Fashion-MNIST datasets also demonstrate its practical applicability.

**Keywords** A-optimality, decorrelated score subsampling, high-dimensional inference, L-optimality, massive data
**MSC(2020)** 62R07, 62H12

**Citation:** Gao J Z, Wang L, Lian H. Optimal decorrelated score subsampling for generalized linear models with massive data. Sci China Math, 2024, 67: 405–430, https://doi.org/10.1007/s11425-022-2057-8

## 1 Introduction

With the rapid growth in modern science and technology, massive data have become ubiquitous in, such as, sociology, biology, and physics. At the same time, the extraordinary amount of data also challenges researchers in conducting data analysis since the traditional statistical methods are no longer applicable. For example, the US census with large-sample data provides fundamental information for people to study socio-economic issues (see [28]), but extracting useful information efficiently and quickly is notoriously difficult. Let $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$ be an independent and identically distributed (i.i.d.) sample from $(y, \boldsymbol{x})$, where $y \in \mathbb{R}$ is a univariate response and $\boldsymbol{x} \in \mathbb{R}^p$ is the covariate. For generalized linear models (GLMs) with a canonical link, it is assumed that the conditional distribution of $y$ given $\boldsymbol{x}$ is

$$f(y \mid \boldsymbol{\beta}, \boldsymbol{x}) = h(y) \exp\{y(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}) - \psi(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})\},$$

* Corresponding author

 

where $h(t)$ and $\psi(t)$ are specific functions, $\boldsymbol{\beta}$ is the unknown parameter and assumed to be in a compact set. Several approaches have been investigated to extract useful information from large-scale data for the GLMs (see, e.g., [3, 14, 23, 38]). Among these methods, the subsampling technique is an effective way of taking random subsamples of small size from the massive data as a surrogate to downsize the data volume. The key idea to subsampling is to find acceptable probabilities for each sample and draw observations according to the chosen sampling scheme. Since the small subsample estimator will not be as accurate as the full data estimator, it is crucial to design a good subsampling strategy. More specifically, Ma et al. [18] proposed an efficient subsampling method for linear regression models based on normalized statistical leverage scores. Wang et al. [28] developed an optimal subsampling procedure for logistic regression based on A- and L-optimality criteria inspired by the optimal experimental design. Ma et al. [19] conducted statistical inference of randomized numerical linear algebra algorithms to identify optimal sampling probabilities. Ai et al. [3] and Wang and Ma [27] extended the idea of A- and L-optimality criteria to GLMs and quantile regression, respectively. Zhang et al. [37] proposed optimal sampling under measurement constraints for GLMs. Poisson subsampling (see, e.g., [1, 33]) and distributed subsampling (see [36]) frameworks have been gradually investigated. More literature can be found in [32] and the references therein.

Recently, massive data with high-dimensional covariates are now routinely encountered in many applications. For example, in a Fashion-MNIST dataset (see [30]), there are 12,000 grayscale images of fashion products in two classes, i.e., sneakers and ankle boots, with $10 \times 10$ pixels represented by a covariate vector. One important goal is to distinguish them by training a classifier. However, most of these existing subsampling studies are based on the low-dimensional GLM assumption, and they cannot be applied in the high-dimensional case, i.e., 100-dimensional pixels, due to the singularity of the subsampling design matrix. On the other hand, it is well known that these two different fashion products can be classified by some important low-dimensional pixels, and the rest of them are extraneous pixels of secondary importance. Realizing that not all of the effects of the covariates are our concerned parameters, many authors are interested in the problem of statistical inference on the low-dimensional parameters for both the low-dimensional and high-dimensional regression models (see, e.g., [5, 6, 20]). In this paper, we consider the covariate effects containing two components, i.e., $\boldsymbol{x} \in \mathbb{R}^p$ contains low-dimensional covariates of main interest $\boldsymbol{z} \in \mathbb{R}^d$ and probably extraneous covariates $\boldsymbol{u} \in \mathbb{R}^q$, and $\boldsymbol{\beta}$ can be decomposed into two parts $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, corresponding to $\boldsymbol{z}$ and $\boldsymbol{u}$, respectively. Thus, the conditional distribution of $y \mid \boldsymbol{x}$ becomes

$$f(y \mid \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{z}, \boldsymbol{u}) = h(y) \exp\{y(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{z} + \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{u}) - \psi(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{z} + \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{u})\}. \tag{1.1}$$

Our main interest is to estimate and make inference on the preconceived low-dimensional parameter $\boldsymbol{\theta}$ in the presence of some nuisance parameter $\boldsymbol{\gamma}$. Two scenarios that differ in whether $q$ is small or large (comparable to or even much larger than the subsample size) are considered. On the one hand, when $q$ is smaller than the subsample size, although the popular optimal subsampling criteria for $\boldsymbol{\beta} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ can still be implemented, it is worthwhile to point out that these existing subsampling probabilities obtained by minimizing the total asymptotic variance of the targeted subsample estimator of $\boldsymbol{\beta}$ are no longer the optimal subsampling probabilities for our concerned parameter $\boldsymbol{\theta}$, which can be verified from our theory in Section 4 and simulation results in Section 5. On the other hand, when $q$ is large or even larger than the subsample size, the existing subsampling schemes fail. Comparatively, attention to the subsampling strategies for the estimation and inference on the low-dimensional parameter for both the low-dimensional and high-dimensional GLMs is limited.

To alleviate the adverse influence caused by nuisance parameters, Zhang and Zhang [34], van de Geer et al. [25], Javanmard and Montanari [13] and Ning and Liu [20] proposed the decorrelated score function, which is uncorrelated with the score function of the nuisance parameter $\boldsymbol{\gamma}$, to conduct hypothesis tests or construct confidence intervals for the preconceived low-dimensional parameter $\boldsymbol{\theta}$. Lately, Fang et al. [8] applied this method to longitudinal data, Li et al. [17] studied the high-dimensional linear models with the measurement error and Cheng et al. [5] extended the regularized projection score estimation method to the high-dimensional quantile regression model. However, all these decorrelated score functions are valid only when the underlying population is infinite. Hence, it is impossible to directly apply the existing

decorrelated score method to the subsampling schemes within the finite population. Compared with Ning and Liu [20], three difficulties described in Subsection 2.2 make both parameter estimation and statistical inference for the subsampling schemes complicated and bring some technical challenges.

To our knowledge, this problem has not been previously investigated and is considerably more complicated than the existing subsampling studies. Our main contributions are in two aspects.

(1) With the nuisance parameter $\boldsymbol{\gamma}$, the first objective of our study is to propose a general subsampling decorrelated score function to reduce the influence of the less accurate nuisance parameter estimation with the slow convergence rate. To be specific, given general subsampling probabilities, we first propose a new projection matrix in order to retain the asymptotic efficiency of our concerned parameter $\boldsymbol{\theta}$ and then construct the general subsampling decorrelated score function naturally. We show that our proposed subsample estimator enjoys consistency and asymptotic normality and also achieves statistical efficiency as the weighted maximum likelihood estimator (MLE) in [3]. In addition, the proposed estimator is less sensitive to small perturbations of the nuisance parameter.

(2) To pursue more efficient subsampling procedures, two optimal subsampling probabilities, i.e., A- and L-optimality criteria, are proposed by minimizing the asymptotic mean squared error of the resulting subsample estimator and the trace of the asymptotic covariance matrix for a linearly transformed subsample estimator. Compared with [3], it can be seen that our proposed subsample estimators achieve smaller asymptotic variance for our concerned parameter $\boldsymbol{\theta}$ in the low-dimensional case. Furthermore, our method is also suitable for the high-dimensional case, which significantly promotes the study of the high-dimensional subsampling schemes. A two-step algorithm is proposed to approximate the optimal subsampling probabilities in practice, and the asymptotic properties of the resultant estimators are also constructed.

The rest of this paper is organized as follows. In Section 2, we first review the optimal subsampling procedures for the low-dimensional GLMs proposed by Ai et al. [3] and then propose a general subsampling decorrelated score function for both the low-dimensional and high-dimensional cases. In Section 3, we establish theoretical results for the proposed subsample estimators. In Section 4, we derive two optimal decorrelated score subsampling strategies based on the A- and L-optimality criteria and further give a two-step algorithm in practice. In Section 5, we present numerical studies to illustrate our method. In Section 6, we show two real data applications to validate our proposed method further. We conclude this paper in Section 7.

## 2 Methodology

To facilitate the presentation, denote the full data matrix by $\mathcal{F}_n = (\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$ is the covariate matrix and $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ is the vector of responses. Denote by $\dot{\psi}(t)$ and $\ddot{\psi}(t)$ the first and second derivatives of $\psi(t)$, respectively, by $\nabla_{\mathcal{S}} f(\boldsymbol{x})$ the gradient of $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}_{\mathcal{S}}$ for $\mathcal{S} \subset \{1, \ldots, p\}$ and by $\mathcal{S}^c$ the complement of $\mathcal{S}$. For a square matrix $\boldsymbol{S}$, denote by $\mathrm{tr}(\boldsymbol{S})$ the trace of $\boldsymbol{S}$. For a vector $\boldsymbol{v}$, denote by $\|\boldsymbol{v}\|$ the Euclidean norm and $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$. Given $a, b \in \mathbb{R}$, the maximum and minimum of $a$ and $b$ are denoted by $a \vee b$ and $a \wedge b$, respectively. We first review the subsampling algorithm for the GLMs in Subsection 2.1 and then propose a general subsampling decorrelated score function based on a new projection matrix in Subsection 2.2.

### 2.1 Review of the subsampling algorithm for GLMs

Take a random subsample of size $r$ using sampling with replacement from the full data $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$ according to the probabilities $\pi_i$ satisfying $\sum_{i=1}^n \pi_i = 1$. Here, $\pi_i$ may depend on the full data. Denote the subsample by $\{(y_i^*, \boldsymbol{x}_i^*), i = 1, \ldots, r\}$ with associated subsampling probabilities $\pi_i^*$. Wang et al. [28] and Ai et al. [3] obtained the subsample estimator for $\boldsymbol{\beta}$ via minimizing the following weighted loss function:

$$L^*(\boldsymbol{\beta}) = \frac{1}{nr} \sum_{i=1}^{r} \frac{1}{\pi_i^*} [-y_i^* (\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i^*) + \psi(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i^*)]. \tag{2.1}$$

Equivalently, they solved the following weighted score function:

$$\nabla_{\boldsymbol{\beta}} L^*(\boldsymbol{\beta}) = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}[-y_i^* + \dot{\psi}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i^*)]\boldsymbol{x}_i^* = \mathbf{0}. \tag{2.2}$$

Denote the MLE of $\boldsymbol{\beta}$ with the full data by

$$\boldsymbol{\beta}_{\mathrm{F}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n}[-y_i(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i) + \psi(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)], \tag{2.3}$$

which is well defined for both low-dimensional and high-dimensional cases since $n \gg p$. Under some conditions, [3, Theorem 2] concludes that the asymptotic covariance matrix of the resulting estimator of $\boldsymbol{\beta}$ given $\mathcal{F}_n$ is

$$\boldsymbol{\Sigma} = \boldsymbol{\mathcal{J}}^{-1}\boldsymbol{\Sigma}_c\boldsymbol{\mathcal{J}}^{-1},$$

where

$$\boldsymbol{\mathcal{J}} = \frac{1}{n}\sum_{i=1}^{n}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)\boldsymbol{x}_i^{\otimes 2}, \quad \boldsymbol{\Sigma}_c = \frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}[y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2\boldsymbol{x}_i^{\otimes 2}.$$

Noticing that $\boldsymbol{\Sigma}_c$ depends on $\pi_i$, they proposed to obtain two optimal subsampling probabilities by minimizing $\mathrm{tr}(\boldsymbol{\Sigma})$ or $\mathrm{tr}(\boldsymbol{\Sigma}_c)$ and put forward a two-step optimal subsampling algorithm. It is worth pointing out that their optimal subsampling schemes can only be established when the dimension of $\boldsymbol{\beta}$ is relatively smaller than the subsample size $r$. However, it remains challenging to develop subsampling procedures for the high-dimensional $\boldsymbol{\beta}$, where the dimension of $\boldsymbol{\beta}$ can be much larger than $r$. In this case, the subsample estimator in (2.2) is not well defined such that a sparsity assumption imposed on the unknown parameter $\boldsymbol{\beta}$ is necessary, i.e., most components of $\boldsymbol{\beta}$ are zero. Unfortunately, it is notoriously difficult to derive a tractable limiting distribution and an asymptotic covariance matrix for the regularized estimator due to the existence of nuisance parameter $\boldsymbol{\gamma}$ (see [7]).

Alternatively, we are interested in the problem of estimation and statistical inference on some preconceived low-dimensional parameters in both the low-dimensional and high-dimensional regression models. To be specific, we consider that $\boldsymbol{\beta} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ corresponding to $\boldsymbol{x} = (\boldsymbol{z}^{\mathrm{T}}, \boldsymbol{u}^{\mathrm{T}})^{\mathrm{T}}$ in (1.1) and then the weighted loss function (2.1) can be rewritten as

$$L^*(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}[-y_i^*(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{z}_i^* + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{u}_i^*) + \psi(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{z}_i^* + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{u}_i^*)].$$

Two scenarios are considered where $q$ is small or large (comparable to or even much larger than $r$). To address these problems, we next construct a new type of subsampling score function for $\boldsymbol{\theta}$ and show that the resulting subsample estimator of $\boldsymbol{\theta}$ is asymptotically normal in both scenarios. The key strategy of our proposed procedure is a subsampling decorrelated score function to handle the impact of both the low-dimensional and high-dimensional nuisance parameters.

### 2.2 General subsampling decorrelated score function

In this subsection, we propose a novel general subsampling decorrelated score function. Here, general subsampling means that we do not impose distributions or specify values of the subsampling probability. Inspired by the decorrelated score method (see [20]), we are motivated to find a projection matrix $\boldsymbol{W}_{\mathrm{F}} \in \mathbb{R}^{d\times q}$ and then construct the subsampling decorrelated score function for $\boldsymbol{\theta}$ such that the estimator solved from this score function enjoys consistency and asymptotic normality. However, it is not easy to apply the existing decorrelated score method to subsampling schemes directly based on the following three reasons.

(i) The population in the subsampling problem is finite, and all the theoretical results should be derived based on $\mathcal{F}_n$, which is different from the traditional infinite population results of Ning and Liu [20].

(ii) A well-defined projection matrix for the finite population depends on the full data sample $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$ and model structure. Therefore, it should be pointed out that directly applying the popular definition of the projection matrix

$$\boldsymbol{W}'_{\mathrm{F}} = \arg\min_{\boldsymbol{W}} \mathrm{E}(\|\nabla_{\boldsymbol{\theta}} L^*(\boldsymbol{\beta}_{\mathrm{F}}) - \boldsymbol{W}\nabla_{\boldsymbol{\gamma}} L^*(\boldsymbol{\beta}_{\mathrm{F}})\|^2 \mid \mathcal{F}_n)$$

is unreasonable, since this type of projection matrix depends on the subsampling probability $\pi_i$, i.e.,

$$\begin{aligned}\boldsymbol{W}'_{\mathrm{F}} &= \mathrm{E}(\nabla_{\boldsymbol{\theta}} L^*(\boldsymbol{\beta}_{\mathrm{F}})\nabla_{\boldsymbol{\gamma}} L^*(\boldsymbol{\beta}_{\mathrm{F}})^{\mathrm{T}} \mid \mathcal{F}_n)\mathrm{E}(\nabla_{\boldsymbol{\gamma}} L^*(\boldsymbol{\beta}_{\mathrm{F}})\nabla_{\boldsymbol{\gamma}} L^*(\boldsymbol{\beta}_{\mathrm{F}})^{\mathrm{T}} \mid \mathcal{F}_n)^{-1} \\ &= \left\{\frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}[-y_i + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2\boldsymbol{z}_i\boldsymbol{u}_i^{\mathrm{T}}\right\}\left\{\frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}[-y_i + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2\boldsymbol{u}_i^{\otimes 2}\right\}^{-1}.\end{aligned}$$

In this case, different subsampling probabilities $\pi_i$ will result in different projection matrices for the finite population. Thus $\boldsymbol{W}'_{\mathrm{F}}$ cannot be used in the subsampling schemes.

(iii) There is no explicit quantification of the benefit of the decorrelated score in a subsampling setting, and many important inference-related questions remain unanswered.

To proceed, we construct the following weighted subsampling decorrelated score function based on a novel projection matrix:

$$S^*(\boldsymbol{\theta}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}}) = \nabla_{\boldsymbol{\theta}} L^*(\boldsymbol{\theta}, \boldsymbol{\gamma}_{\mathrm{F}}) - \boldsymbol{W}_{\mathrm{F}}\nabla_{\boldsymbol{\gamma}} L^*(\boldsymbol{\theta}, \boldsymbol{\gamma}_{\mathrm{F}}) = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}[-y_i^* + \dot{\psi}(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{z}_i^* + \boldsymbol{\gamma}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{u}_i^*)](\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i^*),$$

where

$$\begin{aligned}\boldsymbol{W}_{\mathrm{F}} &= \arg\min_{\boldsymbol{W}} \; E\left(\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)\|\boldsymbol{z}_i^* - \boldsymbol{W}\boldsymbol{u}_i^*\|^2 \,\middle|\, \mathcal{F}_n\right) \\ &= [\mathrm{E}(\nabla^2_{\boldsymbol{\theta}\boldsymbol{\gamma}} L^*(\boldsymbol{\beta}_{\mathrm{F}}) \mid \mathcal{F}_n)][\mathrm{E}(\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}} L^*(\boldsymbol{\beta}_{\mathrm{F}}) \mid \mathcal{F}_n)]^{-1} \\ &= \left[\frac{1}{n}\sum_{i=1}^{n}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)\boldsymbol{z}_i\boldsymbol{u}_i^{\mathrm{T}}\right]\left[\frac{1}{n}\sum_{i=1}^{n}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)\boldsymbol{u}_i^{\otimes 2}\right]^{-1}. \end{aligned} \tag{2.4}$$

It is clear that $\boldsymbol{W}_{\mathrm{F}}$ does not depend on $\pi_i$ and is different from $\boldsymbol{W}'_{\mathrm{F}}$. The main reason is that in the subsampling schemes $\mathrm{E}(\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} L^*(\boldsymbol{\beta}_{\mathrm{F}}) \mid \mathcal{F}_n)$ may not equal $\mathrm{E}(\nabla_{\boldsymbol{\beta}} L^*(\boldsymbol{\beta}_{\mathrm{F}})\nabla_{\boldsymbol{\beta}} L^*(\boldsymbol{\beta}_{\mathrm{F}})^{\mathrm{T}} \mid \mathcal{F}_n)$ anymore. Furthermore, it should be pointed out that our proposed $\boldsymbol{W}_{\mathrm{F}}$ and $S^*(\boldsymbol{\theta}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}})$ are suitable for both the low-dimensional and high-dimensional models in the massive data ($n \gg p$).

**Remark 2.1.** An important feature of $S^*(\boldsymbol{\theta}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}})$ is that it is essentially a weighted score function and the corresponding weights are inverses of the subsampling probabilities, which coincides with classic sampling techniques (see [11]).

**Remark 2.2.** It can be seen that $\boldsymbol{W}_{\mathrm{F}}$ is the weighted least square regression coefficient between $\boldsymbol{z}$ and $\boldsymbol{u}$. Once we have $\boldsymbol{W}_{\mathrm{F}}$, it can be verified that $\boldsymbol{W}_{\mathrm{F}}$ satisfies the following orthogonality property:

$$\mathrm{E}\left(\frac{\partial}{\partial\boldsymbol{\gamma}}\left\{\sum_{i=1}^{r}\frac{1}{\pi_i^*}[-y_i^* + \dot{\psi}(\boldsymbol{\theta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{z}_i^* + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{u}_i^*)](\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i^*)\right\}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_{\mathrm{F}}} \,\middle|\, \mathcal{F}_n\right) = \boldsymbol{0},$$

which enables the convergence rate of the subsample estimator of $\boldsymbol{\theta}$ derived from the forthcoming equation (2.8), not be affected by the initial estimator of $\boldsymbol{\gamma}$. Compared with the projection matrix $\boldsymbol{W}_0 = [\mathrm{E}(\ddot{\psi}(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{x})\boldsymbol{z}\boldsymbol{u}^{\mathrm{T}})][\mathrm{E}(\ddot{\psi}(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{x})\boldsymbol{u}^{\otimes 2})]^{-1}$ based on the infinite population in [20], where $\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} \mathrm{E}[-y(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}) + \psi(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})]$, it can be shown that $\boldsymbol{W}_{\mathrm{F}}$ is $\sqrt{n}$-consistent to $\boldsymbol{W}_0$ (see [15]). More interpretations can also be referred to Li et al. [17], Han et al. [10] and Cheng et al. [5].

Although $\boldsymbol{\beta}_{\mathrm{F}}$ and $\boldsymbol{W}_{\mathrm{F}}$ in (2.3) and (2.4) are well defined, it is impractical to obtain $\boldsymbol{\beta}_{\mathrm{F}}$ and $\boldsymbol{W}_{\mathrm{F}}$ using the entire data due to limited storage. Thus, we propose the following two kinds of subsample estimators

$\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{W}}$ which differ in whether $q$ is small or large. When $q$ is small, $\hat{\boldsymbol{\beta}}$ is solved from (2.2) and $\hat{\boldsymbol{W}}$ is defined as

$$\hat{\boldsymbol{W}} = \left[\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i^*)\boldsymbol{z}_i^*(\boldsymbol{u}_i^*)^{\mathrm{T}}\right]\left[\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i^*)(\boldsymbol{u}_i^*)^{\otimes 2}\right]^{-1}. \tag{2.5}$$

When $q$ is large, however, $\boldsymbol{\beta}$ and $\boldsymbol{W}$ cannot be estimated as above due to the singularity of the design matrix. In order to build sparse models and identify relevant predictors to the response variable, we need to modify the initial estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{W}}$. To be specific,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}[-y_i^*(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i^*) + \psi(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i^*)] + \lambda_1\|\boldsymbol{\beta}\|_1, \tag{2.6}$$

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i^*)\|\boldsymbol{z}_i^* - \boldsymbol{W}\boldsymbol{u}_i^*\|^2 + \lambda_2\sum_{j=1}^{q}\|\boldsymbol{w}_j\|, \tag{2.7}$$

where $\boldsymbol{w}_j$ is the $j$-th column of $\boldsymbol{W}$ and $\lambda_1$ and $\lambda_2$ are regularized parameters. It should be pointed out that (2.6) is a weighted loss function with the LASSO (least absolute shrinkage and selection operator) penalty (see [24]) such that a consistent initial estimate of $\boldsymbol{\beta}_{\mathrm{F}}$ conditional on $\mathcal{F}_n$ can be obtained based on the subsample, while (2.7) is a weighted group LASSO for the multi-response regression (see, e.g., [21, 29]). Finally, we propose to solve the decorrelated score subsample estimator of $\boldsymbol{\theta}$, namely $\tilde{\boldsymbol{\theta}}$, via

$$\hat{S}^*(\boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{W}}) = \nabla_{\boldsymbol{\theta}} L^*(\boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}) - \hat{\boldsymbol{W}}\nabla_{\boldsymbol{\gamma}} L^*(\boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}) = \boldsymbol{0}. \tag{2.8}$$

We summarize the general decorrelated score subsampling procedure in Algorithm 1.

**Algorithm 1** General decorrelated score subsampling algorithm

**Step 1** (Sampling)**.** Assign subsampling probabilities $\pi_i$ $(i = 1, \ldots, n)$ for all the data points and draw a random subsample of size $r$ $(\ll n)$ with replacement. Denote the subsample by $(y_i^*, \boldsymbol{x}_i^*)$ with $\pi_i^*$, respectively, for $i = 1, \ldots, r$.

**Step 2** (Estimation)**.** Solve the subsampling decorrelated score function (2.8) to get the estimate $\tilde{\boldsymbol{\theta}}$ based on the subsample, where $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{W}}$ are solved by (2.2) and (2.5) for the low-dimensional case or by (2.6) and (2.7) for the high-dimensional case, respectively.

## 3 Asymptotic properties

In this section, we establish the asymptotic properties of $\tilde{\boldsymbol{\theta}}$ obtained by Algorithm 1. When $q$ is fixed and small, we need some regularity assumptions listed as follows.

**Assumption 3.1.** *The covariate $\boldsymbol{x}$ is bounded by a constant almost surely, i.e., there exists a constant $L > 0$ such that $\|\boldsymbol{x}\| \leqslant L$ almost surely.*

**Assumption 3.2.** *For all $\boldsymbol{x} \in [-L, L]^p$, $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}$ is an interior point of the parameter space*

$$\boldsymbol{\Theta} = \left\{\theta \in \mathbb{R} : \int h(t)\exp(\theta t)\mu(dt) < \infty\right\}$$

*with $\mu$ being the dominating measure.*

**Assumption 3.3.** *Assume $\boldsymbol{\beta}$ lies in a compact domain*

$$\Lambda_B = \{\boldsymbol{\beta} \in \mathbb{R}^p : \forall\, \boldsymbol{x} \in [-L, L]^p, \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} \in \boldsymbol{\Theta},\ \|\boldsymbol{\beta}\| \leqslant B\}$$

*for some large constant $B$.*

**Assumption 3.4.** *As $n \to \infty$, the symmetric matrix*

$$\boldsymbol{J} = n^{-1}\sum_{i=1}^{n}\ddot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_i)(\boldsymbol{z}_i - \boldsymbol{W}_F\boldsymbol{u}_i)\boldsymbol{z}_i^{\mathrm{T}}$$

*goes to a positive-definite matrix in probability.*

**Assumption 3.5.** *It holds that* $\max_{1\leqslant i\leqslant n}(n\pi_i)^{-1} = O_{\mathrm{P}}(1)$.

Assumptions 3.1–3.3 are commonly used in the literature (see, e.g., [4, 35]). Assumption 3.4 is used to derive the asymptotic covariance matrix. Assumption 3.5 imposes a sufficient condition for the subsampling probabilities. The following theorem presents the consistency of $\tilde{\boldsymbol{\theta}}$ to the full data MLE $\boldsymbol{\theta}_{\mathrm{F}}$.

**Theorem 3.6.** *Under Assumptions* 3.1–3.5*, as* $n \to \infty$ *and* $r \to \infty$*,* $\tilde{\boldsymbol{\theta}}$ *is consistent with the full data MLE* $\boldsymbol{\theta}_F$ *in the conditional probability given* $\mathcal{F}_n$*. Moreover, the rate of convergence is* $r^{-1/2}$*, i.e., with probability approaching one, for any* $\epsilon > 0$*, there exist finite* $\Delta_\epsilon$ *and* $r_\epsilon$ *such that* $\mathrm{P}(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_F\| \geqslant r^{-1/2}\Delta_\epsilon \mid \mathcal{F}_n) < \epsilon$ *for all* $r > r_\epsilon$.

**Remark 3.7.** Although Theorem 3.6 presents the result that $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{F}} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2})$, it implies that $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{F}} = O_{\mathrm{P}}(r^{-1/2})$ as well (see, e.g., [3,31]). Hence, Theorem 3.6 indicates that the proposed subsample estimator is $\sqrt{r}$-consistent to the full data MLE under the unconditional distribution.

Besides consistency, we derive the asymptotic distribution of $\tilde{\boldsymbol{\theta}}$ conditional on $\mathcal{F}_n$.

**Theorem 3.8.** *Under the assumptions of Theorem* 3.6*, as* $n \to \infty$ *and* $r \to \infty$*, conditional on* $\mathcal{F}_n$ *in probability,*

$$\boldsymbol{V}^{-1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_F) \to N(\boldsymbol{0}, \boldsymbol{I})$$

*in distribution, where* $\boldsymbol{V} = \boldsymbol{J}^{-1}\boldsymbol{V}_c\boldsymbol{J}^{-1}$ *and*

$$\boldsymbol{J} = \frac{1}{n}\sum_{i=1}^{n} \ddot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_i)(\boldsymbol{z}_i - \boldsymbol{W}_F\boldsymbol{u}_i)\boldsymbol{z}_i^{\mathrm{T}}, \boldsymbol{V}_c = \frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}[y_i - \dot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_i)]^2(\boldsymbol{z}_i - \boldsymbol{W}_F\boldsymbol{u}_i)^{\otimes 2}. \tag{3.1}$$

**Remark 3.9.** Theorem 3.8 establishes asymptotic normality of the proposed estimator $\tilde{\boldsymbol{\theta}}$ given $\mathcal{F}_n$. The definition of $\boldsymbol{W}_{\mathrm{F}}$ leads to a symmetric matrix $\boldsymbol{J}$, and it is also known as the Schur complement in linear algebra. It is worthwhile to point out that the asymptotic covariance matrix $\boldsymbol{V}$ is exactly the submatrix of $\boldsymbol{\Sigma}$ corresponding to the subsample estimator of $\boldsymbol{\theta}$ in [3]. In other words, in the low-dimensional settings and given the same general subsampling probabilities, our proposed estimator $\tilde{\boldsymbol{\theta}}$ solved from (2.8) follows the same asymptotic distribution as the general subsample estimator in [3] without the decorrelated score, which indicates that the decorrelated score subsampling method is in favor of constructing any sub-estimator of $\boldsymbol{\beta}$ and deriving the explicit asymptotic covariance matrix immediately. Otherwise, to seek the asymptotic covariance matrix of a sub-estimator, one may have to calculate the whole $\boldsymbol{\Sigma} = \boldsymbol{\mathcal{J}}^{-1}\boldsymbol{\Sigma}_c\boldsymbol{\mathcal{J}}^{-1}$ and then find the corresponding submatrix, which is computationally expensive and may not be clearly expressed in theory.

Next, we pursue the asymptotic properties of $\tilde{\boldsymbol{\theta}}$ obtained by Algorithm 1 when $q$ is large and some further assumptions are needed.

**Assumption 3.10.** *Assume that*

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} \mathrm{E}[-y(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}) + \psi(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})]$$

*is sparse with* $s_l = |\mathcal{S}_l|$*, where* $\mathcal{S}_l = \{j : \boldsymbol{\beta}_{0j} \neq 0, j = 1, \ldots, p\}$*,*

$$\boldsymbol{W}_0 = [\mathrm{E}(\ddot{\psi}(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{x})\boldsymbol{z}\boldsymbol{u}^{\mathrm{T}})][\mathrm{E}(\ddot{\psi}(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{x})\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}})]^{-1}$$

*is sparse with* $s_h = |\mathcal{S}_h|$*, where* $\mathcal{S}_h = \{j : \boldsymbol{w}_{0j} \neq \boldsymbol{0}, j = 1, \ldots, q\}$*, and* $(s_l \vee s_h)\log p/\sqrt{r} = o(1)$.

**Assumption 3.11.** *For any set* $\mathcal{S} \subset \{1, \ldots, p\}$ *and any vector* $\boldsymbol{v}$ *belonging to the cone*

$$\mathcal{C}(\mathcal{S}, \alpha) = \{\boldsymbol{v} \in \mathbb{R}^p : \|\boldsymbol{v}_{\mathcal{S}^c}\|_1 \leqslant \alpha\|\boldsymbol{v}_{\mathcal{S}}\|_1\},$$

*it holds that*

$$L^*(\boldsymbol{\beta}_0 + \boldsymbol{v}) - L^*(\boldsymbol{\beta}_0) - \nabla_{\boldsymbol{\beta}}L^*(\boldsymbol{\beta}_0)^{\mathrm{T}}\boldsymbol{v} \geqslant \gamma\|\boldsymbol{v}\|^2,$$

*where* $\gamma > 0$ *is a constant.*

**Assumption 3.12.** *For any set* $\mathcal{S}' \subset \{1,\dots,q\}$ *and any vector* $\boldsymbol{v}'$ *belonging to the cone*

$$\mathcal{C}'(\mathcal{S}',\alpha') = \{\boldsymbol{v}' \in \mathbb{R}^q : \|\boldsymbol{v}'_{\mathcal{S}'^c}\|_1 \leqslant \alpha' \|\boldsymbol{v}'_{\mathcal{S}'}\|_1\},$$

*it holds that*

$$\inf_{\mathbf{0}\neq \boldsymbol{v}' \in \mathcal{C}'(\mathcal{S}',\alpha')} \frac{(\boldsymbol{v}')^{\mathrm{T}} \nabla^2_{\gamma\gamma} L^*(\boldsymbol{\beta}_0) \boldsymbol{v}'}{\|\boldsymbol{v}'\|^2} \geqslant \kappa > 0.$$

Assumption 3.10 emphasizes the sparsity for both $\boldsymbol{\beta}_0$ and $\boldsymbol{W}_0$, which is widely used in the literature (see, e.g., [5, 20]). Assumption 3.11 is the restricted strong convexity condition (see [12, pp. 310–311]), which requires the weighted loss function to be a strongly convex function when restricted to the cone $\mathcal{C}(\mathcal{S},\alpha)$. Assumption 3.12 is the restricted eigenvalue condition (see, e.g., [5, 8, 22]) for the submatrix $\nabla^2_{\gamma\gamma} L^*(\boldsymbol{\beta}_0)$ corresponding to the nuisance parameter and provides the necessary curvature within a cone.

**Theorem 3.13.** *Under Assumptions* 3.1–3.5 *and* 3.10–3.12*, as* $n \to \infty$ *and* $r \to \infty$*,* $\tilde{\boldsymbol{\theta}}$ *is consistent to the full data MLE* $\boldsymbol{\theta}_F$ *in the conditional probability given* $\mathcal{F}_n$*. Moreover, the rate of convergence is* $r^{-1/2}$*, i.e., with probability approaching one, for any* $\epsilon > 0$*, there exist finite* $\Delta_\epsilon$ *and* $r_\epsilon$ *such that*

$$\mathrm{P}(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_F\| \geqslant r^{-1/2}\Delta_\epsilon \mid \mathcal{F}_n) < \epsilon$$

*for all* $r > r_\epsilon$*.*

**Theorem 3.14.** *Under the assumptions of Theorem* 3.13*, as* $n \to \infty$ *and* $r \to \infty$*, conditional on* $\mathcal{F}_n$ *in probability,*

$$\boldsymbol{V}^{-1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_F) \to N(\mathbf{0}, \boldsymbol{I})$$

*in distribution, where* $\boldsymbol{V}$ *is defined in Theorem* 3.8*.*

**Remark 3.15.** Theorems 3.13 and 3.14 show that in the high-dimensional case $\tilde{\boldsymbol{\theta}}$ still has the same consistency and asymptotic normality results as those in the low-dimensional case. Even the intermediate estimators $\hat{\boldsymbol{\beta}}$ in (2.6) and $\hat{\boldsymbol{W}}$ in (2.7) have lower convergence rates than (2.2) and (2.5), respectively. However, it should be pointed out that the subsampling method of Ai et al. [3] fails in high-dimensional settings.

## 4 Optimal decorrelated score subsampling strategies

In this section, we consider optimal subsampling strategies. In Subsection 4.1, we present two subsampling probabilities based on A- and L-optimality criteria. In Subsection 4.2, we discuss the implementation issue.

### 4.1 Optimal subsampling probabilities

Considering that the asymptotic covariance matrix in (3.1) depends on the subsampling probability, in this subsection, we propose some efficient subsampling procedures by choosing the optimal subsampling probability. Notice that the asymptotic mean squared error of $\tilde{\boldsymbol{\theta}}$ equals $\mathrm{tr}(\boldsymbol{V})$, and then the A-optimality criterion is proposed to pursue the smallest value of $\mathrm{tr}(\boldsymbol{V})$. Besides, the L-optimality criterion is considered further to minimize $\mathrm{tr}(\boldsymbol{V}_c)$, which reduces the computing time without sacrificing much estimation efficiency. Theorems 4.1 and 4.2 establish two optimal subsampling probabilities based on the A- and L-optimality criteria, respectively.

**Theorem 4.1** (A-optimality)**.** *If the subsampling probability is chosen such that*

$$\pi_i^{\mathrm{dmV}} = \frac{|y_i - \dot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_i)| \|\boldsymbol{J}^{-1}(\boldsymbol{z}_i - \boldsymbol{W}_F\boldsymbol{u}_i)\|}{\sum_{j=1}^n |y_j - \dot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_j)| \|\boldsymbol{J}^{-1}(\boldsymbol{z}_j - \boldsymbol{W}_F\boldsymbol{u}_j)\|}, \quad i = 1,\dots,n,$$

*then* $\mathrm{tr}(\boldsymbol{V})$ *attains its minimum.*

**Theorem 4.2** (L-optimality). *If the subsampling probability is chosen such that*

$$\pi_i^{\mathrm{dmVc}} = \frac{|y_i - \dot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_i)|\|\boldsymbol{z}_i - \boldsymbol{W}_F\boldsymbol{u}_i\|}{\sum_{j=1}^n |y_j - \dot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_j)|\|\boldsymbol{z}_j - \boldsymbol{W}_F\boldsymbol{u}_j\|}, \quad i = 1, \ldots, n,$$

*then* $\mathrm{tr}(\boldsymbol{V}_c)$ *attains its minimum.*

**Remark 4.3.** It is worthwhile to point out that the optimal probabilities proposed in [3] are based on minimizing the asymptotic covariance matrix for the subsample estimator of $\boldsymbol{\beta}$ in the low-dimensional case. However, minimizing $\mathrm{tr}(\boldsymbol{\Sigma})$ or $\mathrm{tr}(\boldsymbol{\Sigma}_c)$ with respect to $\boldsymbol{\beta}$ is not equivalent to minimizing $\mathrm{tr}(\boldsymbol{V})$ or $\mathrm{tr}(\boldsymbol{V}_c)$ with respect to $\boldsymbol{\theta}$. To clarify this explicitly, we take $\mathrm{tr}(\boldsymbol{V})$ as an example. It can be shown that

$$\begin{aligned}\mathrm{tr}(\boldsymbol{V}) &= \frac{1}{rn^2}\sum_{j=1}^n \pi_j \sum_{i=1}^n \frac{1}{\pi_i}[y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2\|\boldsymbol{J}^{-1}(\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)\|^2 \\ &\geqslant \frac{1}{rn^2}\left\{\sum_{i=1}^n |y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)|\|\boldsymbol{J}^{-1}(\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)\|\right\}^2,\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Notice that the equality holds if and only if $\pi_i \propto |y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)|\|\boldsymbol{J}^{-1}(\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)\|$ and thus our proposed subsampling probabilities achieve the smallest asymptotic variance for the subsample estimator of $\boldsymbol{\theta}$. Moreover, the optimal probabilities are suitable for the low-dimensional case and the high-dimensional case, which can be seen from our simulation results in Section 5.

**Remark 4.4.** Compared with $\pi_i^{\mathrm{dmV}}$, the L-optimality subsampling probabilities $\pi_i^{\mathrm{dmVc}}$ do not depend on $\boldsymbol{J}^{-1}$ and thus are easier to calculate in practice.

### 4.2 Two-step practical algorithm

The optimal subsampling probabilities $\pi_i^{\mathrm{dmV}}$ and $\pi_i^{\mathrm{dmVc}}$ cannot be directly implemented since they depend on the MLE $\boldsymbol{\beta}_{\mathrm{F}}$ and $\boldsymbol{W}_{\mathrm{F}}$ based on the entire data. Thus, we propose a two-step algorithm. In the first step, a subsample of size $r_1$ is taken to get pilot estimates of $\boldsymbol{\beta}_{\mathrm{F}}$ and $\boldsymbol{W}_{\mathrm{F}}$, which will be used to approximate the optimal subsampling probability for drawing a more informative subsample of size $r_2$ in the second step. We still denote these two intermediate estimators by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{W}}$ but the final estimator by $\breve{\boldsymbol{\theta}}$. In the high-dimensional case, (2.6) and (2.7) can be solved by R functions "glmnet" and "mvr" (see [5]) respectively in the R programming language. We present the following two-step algorithm in Algorithm 2 and Theorems 4.5 and 4.6 show the asymptotic properties of $\breve{\boldsymbol{\theta}}$ obtained from Algorithm 2.

**Algorithm 2** Optimal decorrelated score subsampling algorithm

**Step 1.** Run Algorithm 1 with the subsample size $r_1$ to obtain the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{W}}$ using the uniform subsampling probabilities. Replace $\boldsymbol{\beta}_{\mathrm{F}}$ and $\boldsymbol{W}_{\mathrm{F}}$ with $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{W}}$, respectively, and then get the approximate optimal subsampling probabilities corresponding to a chosen optimality criterion.

**Step 2.** Draw a subsample of size $r_2$ using the approximate optimal subsampling probabilities calculated in Step 1 with replacement. Obtain the estimate $\breve{\boldsymbol{\theta}}$ based on the subsample of size $r_2$ according to Algorithm 1.

**Theorem 4.5.** *Under Assumptions* 3.1–3.5 *and* $\sqrt{r_2}/r_1 \to 0$ *in the low-dimensional case or Assumptions* 3.1–3.5, 3.10–3.12 *and* $\sqrt{r_2}(s_l \vee s_h)\log p/r_1 \to 0$ *in the high-dimensional case, for the estimator* $\breve{\boldsymbol{\theta}}$ *obtained from Algorithm* 2, *as* $r_1 \to \infty$, $r_2 \to \infty$ *and* $n \to \infty$, *with probability approaching one, there exist finite* $\Delta_\epsilon$ *and* $r_\epsilon$ *such that* $\mathrm{P}(\|\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}_F\| \geqslant r_2^{-1/2}\Delta_\epsilon \mid \mathcal{F}_n) < \epsilon$ *for any* $\epsilon > 0$ *and all* $r_2 > r_\epsilon$.

**Theorem 4.6.** *Under the assumptions of Theorem* 4.5, *as* $r_1 \to \infty$, $r_2 \to \infty$ *and* $n \to \infty$, *conditional on* $\mathcal{F}_n$,

$$\boldsymbol{V}_{\mathrm{opt}}^{-1/2}(\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}_F) \to N(\boldsymbol{0}, \boldsymbol{I})$$

*in distribution, where* $\boldsymbol{V}_{\mathrm{opt}} = \boldsymbol{J}^{-1}\boldsymbol{V}_{c,\mathrm{opt}}\boldsymbol{J}^{-1}$,

$$\boldsymbol{V}_{c,\mathrm{opt}} = \frac{1}{nr_2}\sum_{i=1}^n \frac{[y_i - \dot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_i)]^2(\boldsymbol{z}_i - \boldsymbol{W}_F\boldsymbol{u}_i)^{\otimes 2}}{|y_i - \dot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_i)|\|\boldsymbol{J}^{-1}(\boldsymbol{z}_i - \boldsymbol{W}_F\boldsymbol{u}_i)\|} \cdot \frac{1}{n}\sum_{j=1}^n |y_j - \dot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_j)|\|\boldsymbol{J}^{-1}(\boldsymbol{z}_j - \boldsymbol{W}_F\boldsymbol{u}_j)\|$$

*for the estimator obtained from Algorithm* 2 *based on the estimated subsampling probabilities for* $\pi_i^{\mathrm{dmV}}$, *and*

$$\boldsymbol{V}_{c,\mathrm{opt}} = \frac{1}{nr_2}\sum_{i=1}^{n}\frac{[y_i-\dot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_i)]^2(\boldsymbol{z}_i-\boldsymbol{W}_F\boldsymbol{u}_i)^{\otimes 2}}{|y_i-\dot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_i)|\|\boldsymbol{z}_i-\boldsymbol{W}_F\boldsymbol{u}_i\|}\cdot\frac{1}{n}\sum_{j=1}^{n}|y_j-\dot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_j)|\|\boldsymbol{z}_j-\boldsymbol{W}_F\boldsymbol{u}_j\|$$

*for the estimator obtained from Algorithm* 2 *based on the estimated subsampling probabilities for* $\pi_i^{\mathrm{dmVc}}$.

In order to conduct statistical inference for the corresponding estimator, we adopt the method of moments to estimate the covariance matrix of $\breve{\boldsymbol{\theta}}$ using $\breve{\boldsymbol{V}} = \breve{\boldsymbol{J}}^{-1}\breve{\boldsymbol{V}}_c\breve{\boldsymbol{J}}^{-\mathrm{T}}$, where

$$\breve{\boldsymbol{J}} = \frac{1}{nr_2}\sum_{i=1}^{r_2}\frac{1}{\hat{\pi}_i^*}\ddot{\psi}(\breve{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{z}_i^* + \hat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{u}_i^*)(\boldsymbol{z}_i^* - \hat{\boldsymbol{W}}\boldsymbol{u}_i^*)\boldsymbol{z}_i^{*\mathrm{T}}$$

and

$$\breve{\boldsymbol{V}}_c = \frac{1}{n^2r_2^2}\sum_{i=1}^{r_2}\frac{1}{(\hat{\pi}_i^*)^2}[y_i^* - \dot{\psi}(\breve{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{z}_i^* + \hat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{u}_i^*)]^2(\boldsymbol{z}_i^* - \hat{\boldsymbol{W}}\boldsymbol{u}_i^*)^{\otimes 2}.$$

Here, we refer to $\hat{\pi}_i^*$ as the estimator of $\pi_i^{\mathrm{dmV}}$ or $\pi_i^{\mathrm{dmVc}}$ for the selected subsample $\{(y_i^*, \boldsymbol{x}_i^*), i = 1, \ldots, r_2\}$ in the second subsampling step.

## 5 Numerical studies

In this section, we conduct simulation studies to assess the finite sample performance of the proposed estimators. In Subsections 5.1 and 5.2, we present simulation results for linear regression and logistic regression models, respectively, based on both the low-dimensional and high-dimensional settings.

### 5.1 Linear regression

We generate data of size $n = 10^5$ from the following linear regression model:

$$y_i = \alpha + \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{z}_i + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{u}_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{z}_i \in \mathbb{R}^d$, $\boldsymbol{u}_i \in \mathbb{R}^{q-1}$ and $(\boldsymbol{z}_i^{\mathrm{T}}, \boldsymbol{u}_i^{\mathrm{T}})^{\mathrm{T}}$ is generated from the multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma}_x$ whose $(j,k)$-th element $\sigma_{jk} = 0.5^{|j-k|}$ for $1 \leqslant j, k \leqslant d+q-1$.

In the low-dimensional case, we set $d = 2$, $q = 8$ and all the elements of $(\alpha, \boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ to 1. The random error $\epsilon_i$ is generated independently from one of the following four distributions respectively: (i) the normal error: $\epsilon_i \sim N(0, 2^2)$; (ii) the $t$ error: $\epsilon_i \sim t(3)$; (iii) the heteroscedastic normal error: $\epsilon_i \sim |z_{i1}|N(0, 3^2)$; (iv) the exponential error: $\epsilon_i \sim \mathrm{Exp}(0.2)$. In the high-dimensional case, we set $q = 700$ and the additional elements of $\boldsymbol{\gamma}$ to 0. The random error $\epsilon_i$ is generated independently from one of the following four distributions respectively: (i) the normal error: $\epsilon_i \sim N(0, 1)$; (ii) the $t$ error: $\epsilon_i \sim t(3)$; (iii) the heteroscedastic normal error: $\epsilon_i \sim |z_{i1}|N(0, 1)$; (iv) the exponential error: $\epsilon_i \sim \mathrm{Exp}(1)$. Set $r_1$ equaling 400 and $r_2$ equaling 400, 600, 800 and 1,000. We compare the following different subsampling methods:

(a) Our proposed dmV and dmVc methods based on the decorrelated score subsampling probabilities $\pi_i^{\mathrm{dmV}}$ and $\pi_i^{\mathrm{dmVc}}$ defined in Subsection 4.2.

(b) The mV and mVc subsampling methods proposed by Ai et al. [3] with the optimal probabilities through minimizing $\mathrm{tr}(\boldsymbol{\Sigma})$ or $\mathrm{tr}(\boldsymbol{\Sigma}_c)$, which are treated as benchmarks and only can be used in the low-dimensional case.

(c) The uniform subsampling method (unif) and uniform decorrelated score subsampling method (dunif), which solves the decorrelated score subsample estimator via

$$\hat{S}^*(\boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{W}}) = \nabla_{\boldsymbol{\theta}} L^*(\boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}) - \hat{\boldsymbol{W}}\nabla_{\boldsymbol{\gamma}} L^*(\boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}) = \boldsymbol{0}$$

from (2.8) with the uniform subsampling probabilities.

To be specific, we evaluate the mean squared error (MSE) based on the true parameter, the average of the empirical coverage probabilities ACP $= d^{-1}\sum_{j=1}^{d}\mathrm{CP}(\theta_j)$ and the average length (AL) of the 95% confidence intervals for each element $\theta_j$ of $\boldsymbol{\theta}$. As in [27], in our simulations, we do not combine the two-step subsamples to perform estimation since if we are willing to handle estimation under size $r_1+r_2$, then we could have chosen a better subsample by setting the second step subsample size to $r_1+r_2$ directly. When $q=700$, we compare our proposed dmV and dmVc methods with the unif (using the LASSO penalty) and dunif methods. Since there is no asymptotic distribution expression for the unif method, we use the percentile bootstrap to obtain the ACP and AL based on 200 replications.

The simulation results are listed in Tables 1 and 2 based on 500 replications. In the low-dimensional case,

(1) **MSEs**: it can be seen that (i) the proposed dmV subsampling strategy always results in the smallest MSEs while the uniform subsampling strategy always results in the largest MSEs among all the scenarios; (ii) compared with the mV and mVc subsampling strategies proposed by Ai et al. [3], our proposed dmV and dmVc strategies always yield smaller MSEs, respectively. This finding agrees with our theoretical results that we aim to minimize the asymptotic MSEs for the subsample estimator of $\boldsymbol{\theta}$ or the linearly transformed subsample estimator of $\boldsymbol{\theta}$ rather than the whole subsample estimator of $\boldsymbol{\beta}$; (iii) the proposed dmV subsampling strategy performs better than the dmVc method in most cases since the dmV method aims to minimize the asymptotic MSE for the subsample estimator of $\boldsymbol{\theta}$; (iv) when the subsample size increases, the MSEs of all the subsampling methods decrease.

**Table 1** MSE (×10), ACP and AL for low-dimensional linear regression

| $r_2$ | | dmV | dmVc | mV | mVc | unif | dmV | dmVc | mV | mVc | unif |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | (i) | | | | | (ii) | | |
| 400 | MSE | 0.169 | 0.196 | 0.212 | 0.234 | 0.306 | 0.083 | 0.099 | 0.114 | 0.127 | 0.235 |
| | ACP | 0.947 | 0.940 | 0.948 | 0.944 | 0.944 | 0.953 | 0.947 | 0.944 | 0.936 | 0.944 |
| | AL | 0.180 | 0.191 | 0.207 | 0.212 | 0.244 | 0.128 | 0.135 | 0.147 | 0.151 | 0.206 |
| 600 | MSE | 0.122 | 0.129 | 0.137 | 0.144 | 0.214 | 0.059 | 0.067 | 0.077 | 0.085 | 0.154 |
| | ACP | 0.947 | 0.947 | 0.953 | 0.950 | 0.939 | 0.939 | 0.939 | 0.942 | 0.930 | 0.950 |
| | AL | 0.147 | 0.155 | 0.167 | 0.171 | 0.199 | 0.104 | 0.110 | 0.119 | 0.122 | 0.168 |
| 800 | MSE | 0.099 | 0.105 | 0.107 | 0.107 | 0.159 | 0.047 | 0.053 | 0.055 | 0.061 | 0.123 |
| | ACP | 0.928 | 0.936 | 0.950 | 0.961 | 0.945 | 0.940 | 0.933 | 0.946 | 0.948 | 0.950 |
| | AL | 0.127 | 0.134 | 0.144 | 0.148 | 0.173 | 0.090 | 0.095 | 0.102 | 0.105 | 0.148 |
| 1,000 | MSE | 0.081 | 0.082 | 0.084 | 0.085 | 0.132 | 0.038 | 0.043 | 0.044 | 0.048 | 0.104 |
| | ACP | 0.923 | 0.941 | 0.947 | 0.953 | 0.943 | 0.935 | 0.929 | 0.953 | 0.939 | 0.940 |
| | AL | 0.114 | 0.120 | 0.129 | 0.132 | 0.154 | 0.080 | 0.085 | 0.091 | 0.093 | 0.133 |
| | | | | (iii) | | | | | (iv) | | |
| 400 | MSE | 0.370 | 0.424 | 0.447 | 0.479 | 1.138 | 1.030 | 1.114 | 1.385 | 1.442 | 1.938 |
| | ACP | 0.949 | 0.947 | 0.947 | 0.935 | 0.940 | 0.942 | 0.945 | 0.922 | 0.927 | 0.943 |
| | AL | 0.265 | 0.282 | 0.295 | 0.303 | 0.458 | 0.417 | 0.441 | 0.479 | 0.489 | 0.608 |
| 600 | MSE | 0.251 | 0.291 | 0.302 | 0.303 | 0.775 | 0.694 | 0.752 | 0.832 | 0.894 | 1.294 |
| | ACP | 0.944 | 0.944 | 0.933 | 0.949 | 0.940 | 0.927 | 0.934 | 0.934 | 0.944 | 0.951 |
| | AL | 0.215 | 0.229 | 0.237 | 0.243 | 0.376 | 0.339 | 0.358 | 0.386 | 0.396 | 0.496 |
| 800 | MSE | 0.188 | 0.220 | 0.228 | 0.234 | 0.585 | 0.521 | 0.560 | 0.600 | 0.638 | 0.999 |
| | ACP | 0.943 | 0.935 | 0.942 | 0.944 | 0.944 | 0.933 | 0.936 | 0.944 | 0.944 | 0.938 |
| | AL | 0.186 | 0.197 | 0.204 | 0.209 | 0.327 | 0.293 | 0.310 | 0.333 | 0.341 | 0.429 |
| 1,000 | MSE | 0.158 | 0.178 | 0.170 | 0.191 | 0.460 | 0.415 | 0.463 | 0.461 | 0.493 | 0.771 |
| | ACP | 0.934 | 0.936 | 0.950 | 0.938 | 0.944 | 0.924 | 0.925 | 0.944 | 0.944 | 0.939 |
| | AL | 0.166 | 0.176 | 0.181 | 0.186 | 0.293 | 0.262 | 0.277 | 0.297 | 0.304 | 0.384 |

**Table 2** MSE (×10), ACP and AL for high-dimensional linear regression

| $r_2$ | | dmV | dmVc | dunif | unif | dmV | dmVc | dunif | unif |
|---|---|---|---|---|---|---|---|---|---|
| | | (i) | | | | (ii) | | | |
| 400 | MSE | 0.044 | 0.050 | 0.080 | 0.151 | 0.104 | 0.110 | 0.227 | 0.431 |
| | ACP | 0.930 | 0.936 | 0.932 | 0.763 | 0.927 | 0.938 | 0.954 | 0.753 |
| | AL | 0.089 | 0.094 | 0.122 | 0.236 | 0.130 | 0.137 | 0.207 | 0.397 |
| 600 | MSE | 0.032 | 0.034 | 0.056 | 0.098 | 0.069 | 0.083 | 0.155 | 0.292 |
| | ACP | 0.921 | 0.938 | 0.937 | 0.772 | 0.923 | 0.920 | 0.949 | 0.761 |
| | AL | 0.073 | 0.077 | 0.099 | 0.192 | 0.106 | 0.112 | 0.169 | 0.322 |
| 800 | MSE | 0.026 | 0.026 | 0.044 | 0.072 | 0.057 | 0.063 | 0.113 | 0.213 |
| | ACP | 0.910 | 0.934 | 0.932 | 0.783 | 0.913 | 0.916 | 0.956 | 0.767 |
| | AL | 0.063 | 0.067 | 0.086 | 0.166 | 0.092 | 0.097 | 0.146 | 0.279 |
| 1,000 | MSE | 0.022 | 0.022 | 0.037 | 0.057 | 0.050 | 0.052 | 0.094 | 0.170 |
| | ACP | 0.898 | 0.923 | 0.922 | 0.798 | 0.897 | 0.914 | 0.953 | 0.775 |
| | AL | 0.056 | 0.059 | 0.077 | 0.148 | 0.082 | 0.087 | 0.132 | 0.250 |
| | | (iii) | | | | (iv) | | | |
| 400 | MSE | 0.046 | 0.051 | 0.128 | 0.197 | 0.041 | 0.044 | 0.077 | 0.155 |
| | ACP | 0.936 | 0.939 | 0.938 | 0.814 | 0.927 | 0.941 | 0.950 | 0.735 |
| | AL | 0.089 | 0.095 | 0.153 | 0.285 | 0.083 | 0.088 | 0.122 | 0.235 |
| 600 | MSE | 0.031 | 0.034 | 0.093 | 0.135 | 0.028 | 0.031 | 0.051 | 0.099 |
| | ACP | 0.922 | 0.928 | 0.933 | 0.843 | 0.920 | 0.935 | 0.948 | 0.765 |
| | AL | 0.073 | 0.077 | 0.125 | 0.235 | 0.068 | 0.072 | 0.099 | 0.191 |
| 800 | MSE | 0.025 | 0.028 | 0.068 | 0.097 | 0.023 | 0.024 | 0.041 | 0.074 |
| | ACP | 0.906 | 0.912 | 0.935 | 0.847 | 0.907 | 0.929 | 0.933 | 0.782 |
| | AL | 0.063 | 0.067 | 0.109 | 0.205 | 0.059 | 0.062 | 0.086 | 0.165 |
| 1,000 | MSE | 0.020 | 0.023 | 0.057 | 0.078 | 0.019 | 0.020 | 0.035 | 0.059 |
| | ACP | 0.909 | 0.901 | 0.926 | 0.860 | 0.898 | 0.923 | 0.919 | 0.783 |
| | AL | 0.056 | 0.060 | 0.097 | 0.184 | 0.053 | 0.055 | 0.077 | 0.147 |

(2) **ACPs**: all the subsampling methods enjoy results close to 0.95, which coincides with the asymptotic normal property and illustrates the reasonableness of the covariance matrix formula.

(3) **ALs**: we find that (i) the proposed dmV and dmVc methods enjoy much smaller lengths, compared with other methods; (ii) when the subsample size increases, the ALs of all the subsampling methods decrease.

In the high-dimensional case, our proposed dmV and dmVc methods still result in much smaller MSEs, compared with the unif and dunif methods. Moreover, the comparison of the dunif and unif methods also shows that the decorrelated score subsampling method has obvious advantages in estimation and statistical inference. In terms of ALs and ACPs, our methods also perform well. However, it can be seen that the unif method has low ACPs and large ALs due to the large biases.

### 5.2 Logistic regression

We generate data of size $n = 10^5$ from the following logistic regression model:

$$y_i \sim \text{Bernoulli}(p_i), \quad \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{z}_i + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{u}_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{z}_i \in \mathbb{R}^d$ and $\boldsymbol{u}_i \in \mathbb{R}^{q-1}$. In the low-dimensional case, we set $d = 2$, $q = 4$ and all the elements of $(\alpha, \boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ to 0.5. The covariate $(\boldsymbol{z}_i^{\mathrm{T}}, \boldsymbol{u}_i^{\mathrm{T}})^{\mathrm{T}}$ is generated independently from one of the following four distributions respectively: (i) $N(\boldsymbol{0}, \boldsymbol{\Sigma}_x/2)$ with $\sigma_{jk} = 0.5^{|j-k|}$; (ii) $N(\boldsymbol{0}, \boldsymbol{\Sigma}_x/2)$ with $\sigma_{jk} = 0.5^{I(j\neq k)}$; (iii) $t_3(\boldsymbol{0}, \boldsymbol{\Sigma}_x)/10$ with $\sigma_{jk} = 0.5^{|j-k|}$; (iv) a mixture of $0.2N(\boldsymbol{1}, \boldsymbol{\Sigma}_x/2)$ and $0.2N(-\boldsymbol{1}, \boldsymbol{\Sigma}_x/2)$ with $\sigma_{jk} = 0.5^{|j-k|}$ for $1 \leqslant j, k \leqslant d+q-1$. In the high-dimensional case, we set $q = 700$ and the additional elements of $\boldsymbol{\gamma}$ to 0.

The covariate is generated independently from one of the following four distributions respectively: (i) $N(\mathbf{0}, \mathbf{\Sigma}_x/2)$ with $\sigma_{jk} = 0.5^{|j-k|}$; (ii) $N(\mathbf{0}, \mathbf{\Sigma}_x/2)$ with $\sigma_{jk} = \xi^{|j-k|}$, $\xi \sim U[0.1, 0.3]$; (iii) $t_5(\mathbf{0}, \mathbf{\Sigma}_x)/2$ with $\sigma_{jk} = 0.5^{|j-k|}$; (iv) a mixture of $N(\mathbf{0.5}, \mathbf{\Sigma}_x)$ and $N(-\mathbf{0.5}, \mathbf{\Sigma}_x)$ with $\sigma_{jk} = 0.5^{|j-k|}$ for $1 \leqslant j, k \leqslant d+q-1$. Tables 3 and 4 show the results for the logistic regression. It can be seen that we have similar conclusions to those in the linear regression model.

## 6 Two real data applications

In this section, we evaluate the performance of our proposed methods using two real datasets in the low-dimensional and high-dimensional cases, respectively.

### 6.1 Census income dataset

We use the census income dataset (see [16]) from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Census+Income) to illustrate our methods in the low-dimensional case. There are $n = 32{,}561$ observations, and the response variable is whether a person makes over 50,000 dollars per year. We conduct logistic regression to investigate the effects on income based on five covariates: the age ($x_1$), capital loss ($x_2$), final weight, highest level of education in the numerical form, and hours worked per week. The first two covariates are treated as the variables of interest. The subsample size is set to $r_1$ equaling 400 and $r_2$ equaling 400, 600, 800 and 1,000. Table 5 presents the estimation and inference results for the full data estimate and five candidates described in Section 5.

**Table 3** MSE (×10), ACP and AL for low-dimensional logistic regression

| $r_2$ | | dmV | dmVc | mV | mVc | unif | dmV | dmVc | mV | mVc | unif |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | (i) | | | | | (ii) | | |
| 400 | MSE | 0.497 | 0.542 | 0.606 | 0.686 | 0.857 | 0.582 | 0.638 | 0.716 | 0.759 | 0.961 |
| | ACP | 0.961 | 0.960 | 0.951 | 0.943 | 0.950 | 0.948 | 0.947 | 0.941 | 0.946 | 0.957 |
| | AL | 0.320 | 0.333 | 0.346 | 0.357 | 0.410 | 0.343 | 0.344 | 0.367 | 0.377 | 0.445 |
| 600 | MSE | 0.351 | 0.385 | 0.418 | 0.453 | 0.556 | 0.401 | 0.422 | 0.468 | 0.495 | 0.650 |
| | ACP | 0.951 | 0.952 | 0.946 | 0.946 | 0.962 | 0.950 | 0.948 | 0.951 | 0.943 | 0.954 |
| | AL | 0.260 | 0.271 | 0.280 | 0.289 | 0.334 | 0.278 | 0.280 | 0.297 | 0.305 | 0.361 |
| 800 | MSE | 0.266 | 0.306 | 0.313 | 0.334 | 0.407 | 0.300 | 0.322 | 0.346 | 0.367 | 0.481 |
| | ACP | 0.948 | 0.938 | 0.950 | 0.949 | 0.956 | 0.947 | 0.942 | 0.944 | 0.943 | 0.947 |
| | AL | 0.225 | 0.235 | 0.241 | 0.249 | 0.288 | 0.241 | 0.242 | 0.256 | 0.263 | 0.312 |
| 1,000 | MSE | 0.223 | 0.249 | 0.252 | 0.266 | 0.338 | 0.244 | 0.263 | 0.277 | 0.297 | 0.399 |
| | ACP | 0.943 | 0.934 | 0.946 | 0.942 | 0.950 | 0.953 | 0.927 | 0.953 | 0.943 | 0.941 |
| | AL | 0.201 | 0.210 | 0.216 | 0.223 | 0.258 | 0.215 | 0.216 | 0.229 | 0.235 | 0.279 |
| | | | | (iii) | | | | | (iv) | | |
| 400 | MSE | 8.199 | 9.111 | 9.639 | 13.200 | 14.960 | 6.274 | 6.719 | 7.136 | 7.900 | 8.445 |
| | ACP | 0.960 | 0.954 | 0.956 | 0.952 | 0.953 | 0.945 | 0.962 | 0.947 | 0.951 | 0.938 |
| | AL | 1.240 | 1.301 | 1.330 | 1.580 | 1.662 | 1.080 | 1.163 | 1.157 | 1.224 | 1.290 |
| 600 | MSE | 5.568 | 5.982 | 6.346 | 8.817 | 9.569 | 4.065 | 4.549 | 4.975 | 5.535 | 5.627 |
| | ACP | 0.943 | 0.954 | 0.941 | 0.943 | 0.943 | 0.945 | 0.959 | 0.929 | 0.938 | 0.947 |
| | AL | 1.002 | 1.047 | 1.063 | 1.265 | 1.335 | 0.877 | 0.946 | 0.938 | 0.993 | 1.049 |
| 800 | MSE | 4.488 | 4.676 | 4.820 | 6.174 | 6.561 | 3.175 | 3.621 | 3.661 | 3.936 | 4.327 |
| | ACP | 0.935 | 0.946 | 0.939 | 0.948 | 0.957 | 0.936 | 0.938 | 0.931 | 0.944 | 0.953 |
| | AL | 0.866 | 0.903 | 0.912 | 1.082 | 1.144 | 0.758 | 0.817 | 0.809 | 0.857 | 0.905 |
| 1,000 | MSE | 3.718 | 3.875 | 3.928 | 4.854 | 5.204 | 2.520 | 2.837 | 2.997 | 3.106 | 3.421 |
| | ACP | 0.933 | 0.943 | 0.929 | 0.950 | 0.951 | 0.939 | 0.939 | 0.932 | 0.945 | 0.951 |
| | AL | 0.773 | 0.806 | 0.812 | 0.964 | 1.018 | 0.676 | 0.729 | 0.722 | 0.765 | 0.809 |

**Table 4** MSE (×10), ACP and AL for high-dimensional logistic regression

| $r_2$ | | dmV | dmVc | dunif | unif | dmV | dmVc | dunif | unif |
|---|---|---|---|---|---|---|---|---|---|
| | | (i) | | | | (ii) | | | |
| 400 | MSE | 0.503 | 0.522 | 0.798 | 1.563 | 0.338 | 0.342 | 0.563 | 2.101 |
| | ACP | 0.941 | 0.945 | 0.920 | 0.677 | 0.955 | 0.958 | 0.921 | 0.398 |
| | AL | 0.300 | 0.315 | 0.363 | 0.553 | 0.267 | 0.269 | 0.308 | 0.432 |
| 600 | MSE | 0.340 | 0.345 | 0.558 | 1.150 | 0.214 | 0.249 | 0.354 | 1.472 |
| | ACP | 0.933 | 0.947 | 0.929 | 0.692 | 0.951 | 0.944 | 0.922 | 0.446 |
| | AL | 0.245 | 0.257 | 0.300 | 0.516 | 0.218 | 0.220 | 0.254 | 0.424 |
| 800 | MSE | 0.266 | 0.282 | 0.453 | 0.926 | 0.182 | 0.199 | 0.297 | 1.120 |
| | ACP | 0.928 | 0.933 | 0.919 | 0.672 | 0.945 | 0.943 | 0.927 | 0.469 |
| | AL | 0.212 | 0.223 | 0.261 | 0.473 | 0.189 | 0.191 | 0.221 | 0.396 |
| 1,000 | MSE | 0.221 | 0.233 | 0.360 | 0.741 | 0.149 | 0.163 | 0.243 | 0.912 |
| | ACP | 0.928 | 0.932 | 0.918 | 0.693 | 0.949 | 0.941 | 0.923 | 0.469 |
| | AL | 0.190 | 0.199 | 0.234 | 0.440 | 0.169 | 0.170 | 0.199 | 0.366 |
| | | (iii) | | | | (iv) | | | |
| 400 | MSE | 0.629 | 0.753 | 1.063 | 2.003 | 0.274 | 0.320 | 0.546 | 1.157 |
| | ACP | 0.929 | 0.917 | 0.918 | 0.628 | 0.924 | 0.915 | 0.923 | 0.610 |
| | AL | 0.327 | 0.343 | 0.406 | 0.555 | 0.210 | 0.222 | 0.282 | 0.483 |
| 600 | MSE | 0.460 | 0.536 | 0.743 | 1.548 | 0.195 | 0.234 | 0.364 | 0.814 |
| | ACP | 0.913 | 0.912 | 0.912 | 0.652 | 0.911 | 0.900 | 0.919 | 0.642 |
| | AL | 0.267 | 0.279 | 0.336 | 0.532 | 0.171 | 0.181 | 0.233 | 0.431 |
| 800 | MSE | 0.367 | 0.430 | 0.531 | 1.178 | 0.165 | 0.187 | 0.279 | 0.634 |
| | ACP | 0.903 | 0.899 | 0.923 | 0.684 | 0.895 | 0.891 | 0.914 | 0.636 |
| | AL | 0.231 | 0.242 | 0.292 | 0.512 | 0.148 | 0.157 | 0.203 | 0.389 |
| 1,000 | MSE | 0.300 | 0.338 | 0.432 | 0.965 | 0.142 | 0.159 | 0.239 | 0.525 |
| | ACP | 0.905 | 0.903 | 0.928 | 0.689 | 0.880 | 0.886 | 0.899 | 0.634 |
| | AL | 0.206 | 0.216 | 0.263 | 0.483 | 0.133 | 0.140 | 0.182 | 0.353 |

In view of point estimates, it can be seen that our proposed estimates are closer to the full data estimate in general compared with other subsampling estimates. When the subsample size increases, the standard errors (SEs) of all the subsampling estimates decrease. For any fixed subsample size, our proposed methods produce smaller SEs than other subsampling methods, showing that the proposed asymptotic covariance matrix formula works well in practice. Moreover, the dmV subsampling estimates perform better than the dmVc subsampling estimates in terms of the SEs, which coincides with our theoretical results. All the estimates show that the effects of $x_1$ and $x_2$ are significantly positive at level 0.05 according to the 95% confidence intervals (CIs). The positive effect of $x_1$ can be explained by the fact that elder people may have rich experience and tend to have higher income. It is interesting that capital loss has a positive effect on income. The possible reason is that people with high income are willing and able to invest their money. The investigations on this real data example support our theoretical results in the low-dimensional case.

### 6.2 Fashion-MNIST dataset

We further apply our methods to the high-dimensional Fashion-MNIST dataset (see [30]) for illustration. There are $n = 12{,}000$ grayscale images of fashion products belonging to sneakers and ankle boots. The response variable is whether an image belongs to the sneakers class, and the middle $10 \times 10 = 100$ pixels of the images are treated as a feature vector in $[0,1]^{100}$. In this classification problem, we are interested in pixel267 ($x_1$) and pixel268 ($x_2$) since they have essential effects on distinguishing between sneakers and ankle boots. Therefore, the effects of pixel267 ($x_1$) and pixel268 ($x_2$) are viewed as parameters of interest, and the remaining ones are nuisance parameters. We conduct logistic regression

**Table 5** Estimation and inference results for the census income dataset

| $r_2$ | | full | dmV | dmVc | mV | mVc | unif |
|---|---|---|---|---|---|---|---|
| 400 | $x_1$ | 0.637 | 0.558 | 0.494 | 0.419 | 0.353 | 0.797 |
| | SE | 0.016 | 0.096 | 0.105 | 0.114 | 0.130 | 0.149 |
| | CI | [0.606, 0.669] | [0.370, 0.745] | [0.289, 0.700] | [0.195, 0.643] | [0.097, 0.609] | [0.504, 1.090] |
| | $x_2$ | 0.234 | 0.361 | 0.252 | 0.394 | 0.451 | 0.438 |
| | SE | 0.013 | 0.067 | 0.084 | 0.120 | 0.111 | 0.137 |
| | CI | [0.209, 0.260] | [0.229, 0.493] | [0.088, 0.416] | [0.159, 0.628] | [0.234, 0.668] | [0.170, 0.706] |
| 600 | $x_1$ | 0.637 | 0.540 | 0.535 | 0.495 | 0.411 | 0.765 |
| | SE | 0.016 | 0.080 | 0.091 | 0.097 | 0.103 | 0.118 |
| | CI | [0.606, 0.669] | [0.384, 0.696] | [0.357, 0.713] | [0.305, 0.684] | [0.210, 0.613] | [0.534, 0.996] |
| | $x_2$ | 0.234 | 0.364 | 0.229 | 0.301 | 0.387 | 0.368 |
| | SE | 0.013 | 0.054 | 0.066 | 0.095 | 0.087 | 0.115 |
| | CI | [0.209, 0.260] | [0.259, 0.469] | [0.099, 0.359] | [0.115, 0.488] | [0.216, 0.558] | [0.142, 0.595] |
| 800 | $x_1$ | 0.637 | 0.545 | 0.566 | 0.474 | 0.442 | 0.781 |
| | SE | 0.016 | 0.065 | 0.069 | 0.084 | 0.089 | 0.102 |
| | CI | [0.606, 0.669] | [0.418, 0.671] | [0.431, 0.700] | [0.310, 0.639] | [0.267, 0.617] | [0.582, 0.980] |
| | $x_2$ | 0.234 | 0.342 | 0.230 | 0.299 | 0.357 | 0.358 |
| | SE | 0.013 | 0.045 | 0.053 | 0.081 | 0.079 | 0.110 |
| | CI | [0.209, 0.260] | [0.253, 0.431] | [0.125, 0.335] | [0.141, 0.457] | [0.202, 0.512] | [0.143, 0.574] |
| 1,000 | $x_1$ | 0.637 | 0.568 | 0.561 | 0.419 | 0.388 | 0.697 |
| | SE | 0.016 | 0.057 | 0.061 | 0.077 | 0.087 | 0.087 |
| | CI | [0.606, 0.669] | [0.457, 0.679] | [0.442, 0.681] | [0.268, 0.570] | [0.218, 0.558] | [0.527, 0.868] |
| | $x_2$ | 0.234 | 0.320 | 0.245 | 0.348 | 0.355 | 0.229 |
| | SE | 0.013 | 0.040 | 0.046 | 0.071 | 0.072 | 0.086 |
| | CI | [0.209, 0.260] | [0.241, 0.399] | [0.154, 0.336] | [0.209, 0.487] | [0.214, 0.497] | [0.060, 0.397] |

and the subsample size is set to $r_1$ equaling 200 and $r_2$ equaling 400, 600, 800 and 1,000. Table 6 presents the estimation and inference results for the full data estimate and three candidates described in Section 5. The SE of the unif method is constructed using the bootstrap with replication size 200. For parameter point estimates, we find that our proposed estimates are closer to the full data approach in most cases compared with the unif and dunif methods. When the subsample size increases, the total SEs decrease for all the subsampling methods. For any fixed subsample size, our proposed two estimates have smaller SEs than the unif and dunif estimates. Moreover, the dmV estimates yield smaller SEs than the dmVc estimates since the dmV subsampling strategy aims to minimize the asymptotic MSE. Our proposed two estimates show that the effects of $x_1$ and $x_2$ are significantly negative at level 0.05 according to the 95% CIs, consistent with those from the full data approach. However, the dunif method indicates that $x_1$ is insignificant when $r_2 = 600$ because of the relatively large SE caused by the uniform subsampling probabilities. The experiments on this real data example support our theoretical findings in the high-dimensional case.

## 7 Conclusions

In this paper, we investigate the nonuniform decorrelated score subsampling methods of GLMs to overcome the computation bottleneck and influence of the potential low convergence rate. The asymptotic properties for the general decorrelated score subsample estimator are derived, and then two optimal subsampling probabilities are proposed according to the A- and L-optimality criteria. Furthermore, we develop a two-step algorithm to approximate the optimal decorrelated score subsampling strategies. We fix the first step subsampling size $r_1$ and vary the second step subsampling size $r_2$ in our simulation studies and real data applications. One may adopt an increasing $r_1$ with respect to $r_2$ to obtain

**Table 6** Estimation and inference results for the Fashion-MNIST dataset

| $r_2$ | | full | dmV | dmVc | dunif | unif |
|---|---|---|---|---|---|---|
| 400 | $x_1$ | $-1.494$ | $-1.444$ | $-1.699$ | $-1.927$ | $-1.277$ |
| | SE | 0.220 | 0.410 | 0.557 | 0.878 | 0.741 |
| | CI | $[-1.925, -1.063]$ | $[-2.248, -0.640]$ | $[-2.790, -0.608]$ | $[-3.647, -0.207]$ | $[-2.850, 0.000]$ |
| | $x_2$ | $-1.290$ | $-1.638$ | $-1.536$ | $-1.564$ | $-1.472$ |
| | SE | 0.188 | 0.310 | 0.315 | 0.555 | 0.697 |
| | CI | $[-1.659, -0.922]$ | $[-2.246, -1.030]$ | $[-2.154, -0.919]$ | $[-2.653, -0.475]$ | $[-2.513, 0.000]$ |
| 600 | $x_1$ | $-1.494$ | $-1.363$ | $-1.760$ | $-1.473$ | $-1.606$ |
| | SE | 0.220 | 0.352 | 0.469 | 0.847 | 0.755 |
| | CI | $[-1.925, -1.063]$ | $[-2.054, -0.672]$ | $[-2.679, -0.842]$ | $[-3.134, 0.187]$ | $[-2.898, 0.000]$ |
| | $x_2$ | $-1.290$ | $-1.685$ | $-1.765$ | $-1.406$ | $-1.108$ |
| | SE | 0.188 | 0.260 | 0.276 | 0.471 | 0.683 |
| | CI | $[-1.659, -0.922]$ | $[-2.195, -1.175]$ | $[-2.306, -1.223]$ | $[-2.328, -0.483]$ | $[-2.351, 0.000]$ |
| 800 | $x_1$ | $-1.494$ | $-1.502$ | $-1.642$ | $-1.487$ | $-1.772$ |
| | SE | 0.220 | 0.308 | 0.387 | 0.745 | 0.751 |
| | CI | $[-1.925, -1.063]$ | $[-2.106, -0.898]$ | $[-2.400, -0.884]$ | $[-2.947, -0.028]$ | $[-3.346, -0.451]$ |
| | $x_2$ | $-1.290$ | $-1.582$ | $-1.732$ | $-1.338$ | $-1.283$ |
| | SE | 0.188 | 0.213 | 0.225 | 0.451 | 0.576 |
| | CI | $[-1.659, -0.922]$ | $[-1.999, -1.166]$ | $[-2.174, -1.290]$ | $[-2.221, -0.455]$ | $[-2.125, 0.000]$ |
| 1,000 | $x_1$ | $-1.494$ | $-1.315$ | $-1.403$ | $-1.484$ | $-1.827$ |
| | SE | 0.220 | 0.267 | 0.323 | 0.673 | 0.649 |
| | CI | $[-1.925, -1.063]$ | $[-1.838, -0.792]$ | $[-2.037, -0.770]$ | $[-2.803, -0.165]$ | $[-2.975, -0.463]$ |
| | $x_2$ | $-1.290$ | $-1.567$ | $-1.772$ | $-1.543$ | $-1.367$ |
| | SE | 0.188 | 0.193 | 0.208 | 0.402 | 0.553 |
| | CI | $[-1.659, -0.922]$ | $[-1.945, -1.189]$ | $[-2.180, -1.365]$ | $[-2.331, -0.754]$ | $[-2.309, -0.240]$ |

better MSEs and CPs. However, from the perspective of balancing statistical results and implementation costs, we still suggest a relatively small and fixed $r_1$ to obtain fair intermediate estimates in the first step.

Some interesting issues still merit further research. First, the proposed methods focus on the subsampling with replacement and require that all the optimal probabilities be calculated at once. However, due to the memory constraint, it is infeasible to implement if $n$ is extremely large. To solve this problem, we can apply the Poisson subsampling framework (see, e.g., [1,33]) to select the data points one by one or block by block. Second, due to the storage or transmission burden, large-scale data are usually scattered at multiple locations. In this case, a distributed subsampling method is more useful. Third, it should be pointed out that the formulas in Theorems 4.1 and 4.2 are based on the i.i.d. random errors. However, when the random errors are not i.i.d., the optimal probabilities are different. Fourth, the proposed method has the potential to be extended to quantile regression and some other models.

**Acknowledgements** This work was supported by the Fundamental Research Funds for the Central Universities, National Natural Science Foundation of China (Grant No. 12271272) and the Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin. The authors are grateful to the referees for their insightful comments and suggestions on this article, which have led to significant improvements.

## References

1 Ai M Y, Wang F, Yu J, et al. Optimal subsampling for large-scale quantile regression. J Complexity, 2021, 62: 101512
2 Ai M Y, Yu J, Zhang H M, et al. Optimal subsampling algorithms for big data generalized linear models. arXiv: 1806.06761v1, 2018
3 Ai M Y, Yu J, Zhang H M, et al. Optimal subsampling algorithms for big data regressions. Statist Sinica, 2021, 31: 749–772

4 Blazère M, Loubes J-M, Gamboa F. Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension. IEEE Trans Inform Theory, 2014, 60: 2303–2318
5 Cheng C, Feng X D, Huang J, et al. Regularized projection score estimation of treatment effects in high-dimensional quantile regression. Statist Sinica, 2022, 32: 23–41
6 Duan R, Ning Y, Chen Y. Heterogeneity-aware and communication-efficient distributed statistical inference. Biometrika, 2022, 109: 67–83
7 Fan J Q, Li R Z. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc, 2001, 96: 1348–1360
8 Fang E X, Ning Y, Li R Z. Test of significance for high-dimensional longitudinal data. Ann Statist, 2020, 48: 2622–2645
9 Ferguson T S. A Course in Large Sample Theory. London: Chapman and Hall, 1996
10 Han D X, Huang J, Lin Y Y, et al. Robust post-selection inference of high-dimensional mean regression with heavy-tailed asymmetric or heteroskedastic errors. J Econometrics, 2022, 230: 416–431
11 Hansen M H, Hurwitz W N. On the theory of sampling from finite populations. Ann Math Statist, 1943, 14: 333–362
12 Hastie T, Tibshirani R, Wainwright M. Statistical Learning with Sparsity. The Lasso and Generalizations. Monographs on Statistics and Applied Probability, vol. 143. Boca Raton: CRC Press, 2015
13 Javanmard A, Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res, 2014, 15: 2869–2909
14 Jordan M I, Lee J D, Yang Y. Communication-efficient distributed statistical inference. J Amer Statist Assoc, 2019, 114: 668–681
15 Koenker R, Portnoy S. *M* estimation of multivariate regressions. J Amer Statist Assoc, 1990, 85: 1060–1068
16 Kohavi R. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Menol Park: AAAI Press, 1996, 202–207
17 Li M Y, Li R Z, Ma Y Y. Inference in high dimensional linear measurement error models. J Multivariate Anal, 2021, 184: 104759
18 Ma P, Mahoney M W, Yu B. A statistical perspective on algorithmic leveraging. J Mach Learn Res, 2015, 16: 861–911
19 Ma P, Zhang X L, Xing X, et al. Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. In: Proceedings of Machine Learning Research, vol. 108. Boston: Addison-Wesley, 2020, 1026–1034
20 Ning Y, Liu H. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. Ann Statist, 2017, 45: 158–195
21 Obozinski G, Wainwright M J, Jordan M I. Support union recovery in high-dimensional multivariate regression. Ann Statist, 2011, 39: 1–47
22 Raskutti G, Wainwright M J, Yu B. Restricted eigenvalue properties for correlated Gaussian designs. J Mach Learn Res, 2010, 11: 2241–2259
23 Schifano E D, Wu J, Wang C, et al. Online updating of statistical inference in the big data setting. Technometrics, 2016, 58: 393–403
24 Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Stat Methodol, 1996, 58: 267–288
25 van de Geer S, Bühlmann P, Ritov Y, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann Statist, 2014, 42: 1166–1202
26 van der Vaart A W. Asymptotic Statistics. Cambridge: Cambridge University Press, 1998
27 Wang H Y, Ma Y Y. Optimal subsampling for quantile regression in big data. Biometrika, 2021, 108: 99–112
28 Wang H Y, Zhu R, Ma P. Optimal subsampling for large sample logistic regression. J Amer Statist Assoc, 2018, 113: 829–844
29 Wang W G, Liang Y B, Xing E P. Block regularized Lasso for multivariate multi-response linear regression. J Mach Learn Res, 2013, 14: 608–617
30 Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747, 2017
31 Xiong S F, Li G Y. Some results on the convergence of conditional distributions. Statist Probab Lett, 2008, 78: 3249–3253
32 Yao Y Q, Wang H Y. A review on optimal subsampling methods for massive datasets. J Data Sci, 2021, 19: 151–172
33 Yu J, Wang H Y, Ai M Y, et al. Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. J Amer Statist Assoc, 2022, 117: 265–276
34 Zhang C-H, Zhang S S. Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat Soc Ser B Stat Methodol, 2014, 76: 217–242
35 Zhang H M, Jia J Z. Elastic-net regularized high-dimensional negative binomial regression: Consistency and weak signal detection. Statist Sinica, 2022, 32: 181–207
36 Zhang H X, Wang H Y. Distributed subdata selection for big data via sampling-based approach. Comput Statist Data Anal, 2021, 153: 107072
37 Zhang T, Ning Y, Ruppert D. Optimal sampling for generalized linear models under measurement constraints. J

Comput Graph Stat, 2021, 30: 106–114
38 Zhang Y C, Duchi J C, Wainwright M J. Communication-efficient algorithms for statistical optimization. J Mach Learn Res, 2013, 14: 3321–3363

## Appendix A

Let $\|\boldsymbol{W}\|_{2,1}=\sum_{j=1}^{p}\|\boldsymbol{w}_j\|$ for any matrix $\boldsymbol{W}\in\mathbb{R}^{d\times q}$, where $\boldsymbol{w}_j$ is the $j$-th column of $\boldsymbol{W}$, and $\|\boldsymbol{v}\|$ is the standard $\ell_2$ norm for any vector $\boldsymbol{v}\in\mathbb{R}^q$. For an index set $\mathcal{S}\in\{1,\ldots,q\}$ and a matrix $\boldsymbol{W}\in\mathbb{R}^{d\times q}$, $\boldsymbol{W}_S$ denotes the submatrix of $\boldsymbol{W}$ containing columns of $\boldsymbol{W}$ with indices in $\mathcal{S}$. For a vector $\boldsymbol{v}$, $\boldsymbol{v}_S$ denotes the subvector of $\boldsymbol{v}$ containing elements of $\boldsymbol{v}$ with indices in $\mathcal{S}$. The notation $p_n\lesssim q_n$ means that there exists some constant $C>0$ such that $p_n\leqslant Cq_n$ holds for sufficiently large $n$. For notational simplicity, we use $C$ to denote a generic constant, whose value may change from line to line.

**Lemma A.1.** *Under Assumptions* 3.1–3.5*, as $n\to\infty$ and $r\to\infty$, conditional on $\mathcal{F}_n$ in probability,*

$$\hat{\boldsymbol{W}}-\boldsymbol{W}_F=O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}),\tag{A.1}$$

$$\hat{\boldsymbol{J}}-\boldsymbol{J}=O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}),\tag{A.2}$$

$$\hat{\boldsymbol{K}}=O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}),\tag{A.3}$$

$$\hat{S}^*(\boldsymbol{\theta}_F,\boldsymbol{\gamma}_F,\hat{\boldsymbol{W}})=O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}),\tag{A.4}$$

*where*

$$\hat{\boldsymbol{J}}=\frac{\partial\hat{S}^*(\boldsymbol{\theta}_F,\boldsymbol{\gamma}_F,\hat{\boldsymbol{W}})}{\partial\boldsymbol{\theta}^{\mathrm{T}}}=\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_i^*)(\boldsymbol{z}_i^*-\hat{\boldsymbol{W}}\boldsymbol{u}_i^*)(\boldsymbol{z}_i^*)^{\mathrm{T}},$$

$$\hat{\boldsymbol{K}}=\frac{\partial\hat{S}^*(\boldsymbol{\theta}_F,\boldsymbol{\gamma}_F,\hat{\boldsymbol{W}})}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}=\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}}\boldsymbol{x}_i^*)(\boldsymbol{z}_i^*-\hat{\boldsymbol{W}}\boldsymbol{u}_i^*)(\boldsymbol{u}_i^*)^{\mathrm{T}}.$$

*Proof of Lemma* A.1. (i) To show $\hat{\boldsymbol{W}}-\boldsymbol{W}_{\mathrm{F}}=O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2})$, define

$$\boldsymbol{W}_{\mathrm{F}_1}=\frac{1}{n}\sum_{i=1}^{n}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)\boldsymbol{z}_i\boldsymbol{u}_i^{\mathrm{T}},\quad \boldsymbol{W}_{\mathrm{F}_2}=\frac{1}{n}\sum_{i=1}^{n}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)\boldsymbol{u}_i^{\otimes 2},$$

$$\hat{\boldsymbol{W}}_1=\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i^*)\boldsymbol{z}_i^*(\boldsymbol{u}_i^*)^{\mathrm{T}},\quad \hat{\boldsymbol{W}}_2=\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i^*)(\boldsymbol{u}_i^*)^{\otimes 2},$$

$$\tilde{\boldsymbol{W}}_1=\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)\boldsymbol{z}_i^*(\boldsymbol{u}_i^*)^{\mathrm{T}},\quad \tilde{\boldsymbol{W}}_2=\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)(\boldsymbol{u}_i^*)^{\otimes 2}.$$

It can be seen that

$$\mathrm{E}(\tilde{\boldsymbol{W}}_1\mid\mathcal{F}_n)=\frac{1}{n}\sum_{i=1}^{n}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)\boldsymbol{z}_i\boldsymbol{u}_i^{\mathrm{T}}=\boldsymbol{W}_{\mathrm{F}_1}.$$

For any component $\tilde{W}_1^{j_1j_2}$ of $\tilde{W}_1$ where $1\leqslant j_1,j_2\leqslant p$,

$$\begin{aligned}
&\mathrm{E}(\tilde{W}_1^{j_1j_2}-W_{\mathrm{F}_1}^{j_1j_2}\mid\mathcal{F}_n)^2\\
&=\mathrm{E}\bigg(\frac{1}{r}\sum_{i=1}^{r}\bigg[\frac{1}{n\pi_i^*}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)z_{ij_1}^*u_{ij_2}^*-W_{\mathrm{F}_1}^{j_1j_2}\bigg]\bigg|\,\mathcal{F}_n\bigg)^2\\
&=\frac{1}{r}\sum_{i=1}^{n}\pi_i\bigg[\frac{1}{n\pi_i}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)z_{ij_1}u_{ij_2}-W_{\mathrm{F}_1}^{j_1j_2}\bigg]^2\\
&=\frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}[\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)z_{ij_1}u_{ij_2}]^2-\frac{1}{r}(W_{\mathrm{F}_1}^{j_1j_2})^2\\
&\leqslant\frac{1}{r}\bigg(\max_{1\leqslant i\leqslant n}\frac{\|\boldsymbol{x}_i\|^2}{n\pi_i}\bigg)\sum_{i=1}^{n}\frac{1}{n}(\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i))^2-\frac{1}{r}(W_{\mathrm{F}_1}^{j_1j_2})^2=O_{\mathrm{P}}(r^{-1}).
\end{aligned}$$

By Chebyshev's inequality,

$$\tilde{\boldsymbol{W}}_1 - \boldsymbol{W}_{\mathrm{F}_1} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}).$$

It can be calculated that

$$\hat{\boldsymbol{W}}_1 - \tilde{\boldsymbol{W}}_1 = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}[\ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i^*) - \ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)]\boldsymbol{z}_i^*(\boldsymbol{u}_i^*)^{\mathrm{T}} = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\dddot{\psi}(\bar{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathrm{F}})^{\mathrm{T}}\boldsymbol{x}_i^*\boldsymbol{z}_i^*(\boldsymbol{u}_i^*)^{\mathrm{T}},$$

where $\dddot{\psi}(t)$ is the third-order derivative of $\psi(t)$ and $\bar{\boldsymbol{\beta}}$ lies between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_{\mathrm{F}}$. Since $\dddot{\psi}(t)$ is bound by a constant and $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathrm{F}} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2})$ implied by [2, Theorem 1], we have

$$\hat{\boldsymbol{W}}_1 - \tilde{\boldsymbol{W}}_1 = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2})$$

by the Cauchy-Schwarz inequality and Chebyshev's inequality. Thus,

$$\hat{\boldsymbol{W}}_1 - \boldsymbol{W}_{\mathrm{F}_1} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}).$$

Similarly, it can be proved that

$$\hat{\boldsymbol{W}}_2 - \boldsymbol{W}_{\mathrm{F}_2} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}).$$

Since

$$\hat{\boldsymbol{W}}_2^{-1} - \boldsymbol{W}_{\mathrm{F}_2}^{-1} = -\boldsymbol{W}_{\mathrm{F}_2}^{-1}(\hat{\boldsymbol{W}}_2 - \boldsymbol{W}_{\mathrm{F}_2})\hat{\boldsymbol{W}}_2^{-1} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}),$$

we have

$$\begin{aligned}\hat{\boldsymbol{W}} - \boldsymbol{W}_{\mathrm{F}} &= \hat{\boldsymbol{W}}_1\hat{\boldsymbol{W}}_2^{-1} - \boldsymbol{W}_{\mathrm{F}_1}\boldsymbol{W}_{\mathrm{F}_2}^{-1}\\ &= (\hat{\boldsymbol{W}}_1 - \boldsymbol{W}_{\mathrm{F}_1})(\hat{\boldsymbol{W}}_2^{-1} - \boldsymbol{W}_{\mathrm{F}_2}^{-1}) + \boldsymbol{W}_{\mathrm{F}_1}(\hat{\boldsymbol{W}}_2^{-1} - \boldsymbol{W}_{\mathrm{F}_2}^{-1}) + (\hat{\boldsymbol{W}}_1 - \boldsymbol{W}_{\mathrm{F}_1})\boldsymbol{W}_{\mathrm{F}_2}^{-1}\\ &= O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}).\end{aligned}$$

This completes the proof of (A.1).

(ii) To show $\hat{\boldsymbol{J}} - \boldsymbol{J} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2})$ and $\hat{\boldsymbol{K}} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2})$, define

$$\tilde{\boldsymbol{J}} = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)(\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i^*)(\boldsymbol{z}_i^*)^{\mathrm{T}}.$$

By calculation,

$$\mathrm{E}(\tilde{\boldsymbol{J}} \mid \mathcal{F}_n) = \frac{1}{n}\sum_{i=1}^{n}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)(\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)\boldsymbol{z}_i^{\mathrm{T}} = \boldsymbol{J}.$$

For any component $\tilde{J}^{j_1j_2}$ of $\tilde{\boldsymbol{J}}$ where $1 \leqslant j_1, j_2 \leqslant p$,

$$\begin{aligned}\mathrm{E}(\tilde{J}^{j_1j_2} - J^{j_1j_2} \mid \mathcal{F}_n)^2 &= \mathrm{E}\left(\frac{1}{r}\sum_{i=1}^{r}\left[\frac{1}{n\pi_i^*}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)(\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i^*)_{j_1}z_{ij_2}^* - J^{j_1j_2}\right]\,\middle|\,\mathcal{F}_n\right)^2\\ &= \frac{1}{r}\sum_{i=1}^{n}\pi_i\left[\frac{1}{n\pi_i}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)(\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)_{j_1}z_{ij_2} - J^{j_1j_2}\right]^2\\ &= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}[\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{z}_i)(\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)_{j_1}z_{ij_2}]^2 - \frac{1}{r}(J^{j_1j_2})^2 = O_{\mathrm{P}}(r^{-1}).\end{aligned}$$

By Chebyshev's inequality, we have

$$\tilde{\boldsymbol{J}} - \boldsymbol{J} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}).$$

Since

$$\hat{\boldsymbol{J}} - \tilde{\boldsymbol{J}} = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)(\boldsymbol{W}_{\mathrm{F}} - \hat{\boldsymbol{W}})\boldsymbol{u}_i^*(\boldsymbol{z}_i^*)^{\mathrm{T}} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}),$$

we have

$$\hat{\boldsymbol{J}} - \boldsymbol{J} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}).$$

Define

$$\tilde{\boldsymbol{K}} = \frac{\partial \hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)(\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i^*)(\boldsymbol{u}_i^*)^{\mathrm{T}}.$$

Similar arguments lead to

$$\hat{\boldsymbol{K}} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}).$$

This completes the proof of (A.2) and (A.3).

(iii) To show $\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2})$, we first note that

$$\begin{aligned}
&\mathrm{E}\left(\frac{1}{n}\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}})\,\middle|\,\mathcal{F}_n\right) = \frac{1}{n}\sum_{i=1}^{n}[-y_i + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)](\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i) = \boldsymbol{0},\\
&\mathrm{var}\left(\frac{1}{n}\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}})\,\middle|\,\mathcal{F}_n\right) = \frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}[y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2(\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)^{\otimes 2} = O_{\mathrm{P}}(r^{-1}),
\end{aligned}$$

where the convergence rate in the second equality is implied by [2, Lemma 2]. By Chebyshev's inequality, we have

$$\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}}) = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}).$$

Noting that

$$\begin{aligned}
&\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) - \hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}}) = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}[-y_i^* + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)](\boldsymbol{W}_{\mathrm{F}} - \hat{\boldsymbol{W}})\boldsymbol{u}_i^*,\\
&\mathrm{E}\left(\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}[-y_i^* + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)]\boldsymbol{u}_i^*\,\middle|\,\mathcal{F}_n\right) = \frac{1}{n}\sum_{i=1}^{n}[-y_i + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]\boldsymbol{u}_i = \boldsymbol{0},\\
&\mathrm{var}\left(\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}[-y_i^* + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)]\boldsymbol{u}_i^*\,\middle|\,\mathcal{F}_n\right) = \frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}[y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2\boldsymbol{u}_i^{\otimes 2} = O_{\mathrm{P}}(r^{-1}),
\end{aligned}$$

by Chebyshev's inequality, we have

$$\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}[-y_i^* + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)]\boldsymbol{u}_i^* = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}),$$

and thus

$$\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) - \hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}}) = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1}).$$

Hence,

$$\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}).$$

This completes the proof of (A.4), and the proof of Lemma A.1 is completed. □

*Proof of Theorem* 3.6. By the consistency of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{W}}$, it can be shown that

$$\hat{S}^*(\boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{W}}) - \nabla_{\boldsymbol{\theta}} L^*(\boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}) = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}[y_i - \dot{\psi}(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{z}_i^* + \hat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{u}_i^*)]\hat{\boldsymbol{W}}\boldsymbol{u}_i^* = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}).$$

Applying [26, Theorem 5.9], we obtain $\|\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\| = o_{P\,|\,\mathcal{F}_n}(1)$. By the proof of [2, Theorem 1], we have $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{F}}\| = o_{P\,|\,\mathcal{F}_n}(1)$. Thus, it leads to $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{F}}\| = o_{P\,|\,\mathcal{F}_n}(1)$. Using Taylor's theorem for random variables (see [9]), we have

$$0 = \hat{S}_j^*(\tilde{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{W}}) = \hat{S}_j^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) + \frac{\partial \hat{S}_j^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{F}}) + \frac{\partial \hat{S}_j^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{\mathrm{F}}) + R_j,$$

where the subscript $j$ denotes the $j$-th element of a vector, the $j$-th Lagrange remainder

$$R_j = (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathrm{F}})^{\mathrm{T}} \int_0^1 \int_0^1 \frac{\partial^2 \hat{S}_j^*(\boldsymbol{\beta}_{\mathrm{F}} + uv(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathrm{F}}), \hat{\boldsymbol{W}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} v du dv (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathrm{F}})$$

and $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\theta}}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$. By the consistency of $\hat{\boldsymbol{W}}$ and boundedness of $\dddot{\psi}(t)$, we have

$$\left\| \frac{\partial^2 \hat{S}_j^*(\boldsymbol{\beta}, \hat{\boldsymbol{W}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \right\| = \frac{1}{nr} \left\| \sum_{i=1}^r \frac{\dddot{\psi}(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i^*)}{\pi_i^*} (\boldsymbol{x}_i - \hat{\boldsymbol{W}} \boldsymbol{u}_i)_j \boldsymbol{x}_i^* (\boldsymbol{x}_i^*)^{\mathrm{T}} \right\| = O_{\mathrm{P} \,|\, \mathcal{F}_n}(1)$$

for all $\boldsymbol{\beta}$. Thus

$$\left\| \int_0^1 \int_0^1 \frac{\partial^2 \hat{S}_j^*(\boldsymbol{\beta}_{\mathrm{F}} + uv(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathrm{F}}), \hat{\boldsymbol{W}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} v du dv \right\| = O_{\mathrm{P} \,|\, \mathcal{F}_n}(1).$$

Combining the above equations with the Taylor's expansion, we have

$$\begin{aligned}
\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{F}} &= -\left\{ \frac{\partial \hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \right\}^{-1} \left\{ \hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) + \frac{\partial \hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{\mathrm{F}}) + \boldsymbol{R} \right\} \\
&= -\hat{\boldsymbol{J}}^{-1} \{ \hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) + \hat{\boldsymbol{K}} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{\mathrm{F}}) + O_{\mathrm{P} \,|\, \mathcal{F}_n}(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathrm{F}}\|^2) \} \\
&= O_{\mathrm{P} \,|\, \mathcal{F}_n}(r^{-1/2}) + o_{P \,|\, \mathcal{F}_n}(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{F}}\|),
\end{aligned}$$

which implies that

$$\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{F}} = O_{\mathrm{P} \,|\, \mathcal{F}_n}(r^{-1/2}).$$

This completes the proof of Theorem 3.6. □

*Proof of Theorem* 3.8. Note that

$$\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}}) = \frac{1}{r} \sum_{i=1}^r \frac{1}{n\pi_i^*} [-y_i^* + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}} \boldsymbol{x}_i^*)](\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}} \boldsymbol{u}_i^*) =: \frac{1}{r} \sum_{i=1}^r \boldsymbol{\eta}_i.$$

It can be seen that given $\mathcal{F}_n$, $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_r$ are i.i.d. random variables with mean $\mathbf{0}$ and variance

$$\mathrm{var}(\boldsymbol{\eta}_1 \mid \mathcal{F}_n) = \frac{1}{n^2} \sum_{i=1}^n \frac{1}{\pi_i} [y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}} \boldsymbol{x}_i)]^2 (\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}} \boldsymbol{u}_i)^{\otimes 2} = O_{\mathrm{P} \,|\, \mathcal{F}_n}(1)$$

from [2, Lemma 2]. For some $\delta$ and every $\epsilon > 0$,

$$\begin{aligned}
\sum_{i=1}^r \mathrm{E}(\|r^{-1/2} \boldsymbol{\eta}_i\|^2 I(\|\boldsymbol{\eta}_i\| > r^{1/2} \epsilon) \,|\, \mathcal{F}_n) &\leqslant \frac{1}{r^{1+\delta/2} \epsilon^\delta} \sum_{i=1}^r \mathrm{E}(\|\boldsymbol{\eta}_i\|^{2+\delta} I(\|\boldsymbol{\eta}_i\| > r^{1/2} \epsilon) \,|\, \mathcal{F}_n) \\
&\leqslant \frac{1}{r^{1+\delta/2} \epsilon^\delta} \sum_{i=1}^r \mathrm{E}(\|\boldsymbol{\eta}_i\|^{2+\delta} \,|\, \mathcal{F}_n) \\
&\leqslant \frac{1}{r^{\delta/2} n^{2+\delta} \epsilon^\delta} \sum_{i=1}^n \frac{1}{\pi_i^{1+\delta}} [y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}} \boldsymbol{x}_i)]^{2+\delta} \|\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}} \boldsymbol{u}_i\|^{2+\delta} \\
&\leqslant \frac{1}{r^{\delta/2} \epsilon^\delta} \left( \max_{1 \leqslant i \leqslant n} \frac{\|\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}} \boldsymbol{u}_i\|^{2+\delta}}{(n\pi_i)^{1+\delta}} \right) \sum_{i=1}^n \frac{[y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}} \boldsymbol{x}_i)]^{2+\delta}}{n}.
\end{aligned}$$

Then we obtain

$$\sum_{i=1}^r \mathrm{E}(\|r^{-1/2} \boldsymbol{\eta}_i\|^2 I(\|\boldsymbol{\eta}_i\| > r^{1/2} \epsilon) \,|\, \mathcal{F}_n) \leqslant \frac{1}{r^{\delta/2} \epsilon^\delta} O_{\mathrm{P}}(1) O_{\mathrm{P}}(1) = o_P(1).$$

By the Lindeberg-Feller central limit theorem,

$$\boldsymbol{V}_c^{-1/2} \hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}}) = r^{-1} \boldsymbol{V}_c^{-1/2} \sum_{i=1}^r \boldsymbol{\eta}_i \to N(\mathbf{0}, \boldsymbol{I})$$

in distribution. Noticing that $\boldsymbol{V}_c = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1})$ and $\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) - \hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}}) = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1})$, we have

$$\boldsymbol{V}_c^{-1/2}\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) - \boldsymbol{V}_c^{-1/2}\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \boldsymbol{W}_{\mathrm{F}}) = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}).$$

Applying Slutsky's theorem, conditional on $\mathcal{F}_n$,

$$\boldsymbol{V}_c^{-1/2}\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) \to N(\boldsymbol{0}, \boldsymbol{I})$$

in distribution. It can be seen that

$$\hat{\boldsymbol{J}}^{-1} - \boldsymbol{J}^{-1} = -\boldsymbol{J}^{-1}(\hat{\boldsymbol{J}} - \boldsymbol{J})\hat{\boldsymbol{J}}^{-1} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}), \boldsymbol{V} = \boldsymbol{J}^{-1}\boldsymbol{V}_c\boldsymbol{J}^{-1} = \frac{1}{r}\boldsymbol{J}^{-1}(r\boldsymbol{V}_c)\boldsymbol{J}^{-1} = O_{\mathrm{P}}(r^{-1}).$$

Thus, we have

$$\begin{aligned}\boldsymbol{V}^{-1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{F}}) &= -\boldsymbol{V}^{-1/2}\hat{\boldsymbol{J}}^{-1}\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) + O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2})\\ &= -\boldsymbol{V}^{-1/2}\boldsymbol{J}^{-1}\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) - \boldsymbol{V}^{-1/2}(\hat{\boldsymbol{J}}^{-1} - \boldsymbol{J}^{-1})\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) + O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2})\\ &= -\boldsymbol{V}^{-1/2}\boldsymbol{J}^{-1}\boldsymbol{V}_c^{1/2}\boldsymbol{V}_c^{-1/2}\hat{S}^*(\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\gamma}_{\mathrm{F}}, \hat{\boldsymbol{W}}) + O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}).\end{aligned}$$

Using the fact that

$$(\boldsymbol{V}^{-1/2}\boldsymbol{J}^{-1}\boldsymbol{V}_c^{1/2})(\boldsymbol{V}^{-1/2}\boldsymbol{J}^{-1}\boldsymbol{V}_c^{1/2})^{\mathrm{T}} = \boldsymbol{I}$$

and applying Slutsky's theorem, we obtain

$$\boldsymbol{V}^{-1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{F}}) \to N(\boldsymbol{0}, \boldsymbol{I})$$

in distribution. To further illustrate Remark 3.9, rewrite

$$\begin{aligned}\boldsymbol{\mathcal{J}} &= \frac{1}{n}\sum_{i=1}^{n}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)\begin{pmatrix}\boldsymbol{z}_i^{\otimes 2} & \boldsymbol{z}_i\boldsymbol{u}_i^{\mathrm{T}}\\ \boldsymbol{u}_i\boldsymbol{z}_i^{\mathrm{T}} & \boldsymbol{u}_i^{\otimes 2}\end{pmatrix} := \begin{pmatrix}\boldsymbol{\mathcal{J}}_{11} & \boldsymbol{\mathcal{J}}_{12}\\ \boldsymbol{\mathcal{J}}_{21} & \boldsymbol{\mathcal{J}}_{22}\end{pmatrix},\\ \boldsymbol{\Sigma}_c &= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}[y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2\begin{pmatrix}\boldsymbol{z}_i^{\otimes 2} & \boldsymbol{z}_i\boldsymbol{u}_i^{\mathrm{T}}\\ \boldsymbol{u}_i\boldsymbol{z}_i^{\mathrm{T}} & \boldsymbol{u}_i^{\otimes 2}\end{pmatrix} := \begin{pmatrix}\boldsymbol{\Sigma}_{c11} & \boldsymbol{\Sigma}_{c12}\\ \boldsymbol{\Sigma}_{c21} & \boldsymbol{\Sigma}_{c22}\end{pmatrix}.\end{aligned}$$

Applying the inverse of the block matrix, we obtain

$$\boldsymbol{\mathcal{J}}^{-1} = \begin{pmatrix}\boldsymbol{I} & \boldsymbol{O}\\ -\boldsymbol{\mathcal{J}}_{22}^{-1}\boldsymbol{\mathcal{J}}_{21} & \boldsymbol{I}\end{pmatrix}\begin{pmatrix}(\boldsymbol{\mathcal{J}}_{11} - \boldsymbol{\mathcal{J}}_{12}\boldsymbol{\mathcal{J}}_{22}^{-1}\boldsymbol{\mathcal{J}}_{21})^{-1} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{\mathcal{J}}_{22}^{-1}\end{pmatrix}\begin{pmatrix}\boldsymbol{I} & -\boldsymbol{\mathcal{J}}_{12}\boldsymbol{\mathcal{J}}_{22}^{-1}\\ \boldsymbol{O} & \boldsymbol{I}\end{pmatrix}.$$

Then simple calculation gives the submatrix of $\boldsymbol{\mathcal{J}}^{-1}\boldsymbol{\Sigma}_c\boldsymbol{\mathcal{J}}^{-1}$ corresponding to the estiamted $\boldsymbol{\theta}$, denoted by $\boldsymbol{J}^{-1}\boldsymbol{V}_c\boldsymbol{J}^{-1}$, where

$$\begin{aligned}\boldsymbol{J} &= \boldsymbol{\mathcal{J}}_{11} - \boldsymbol{\mathcal{J}}_{12}\boldsymbol{\mathcal{J}}_{22}^{-1}\boldsymbol{\mathcal{J}}_{21},\\ \boldsymbol{V}_c &= \boldsymbol{\Sigma}_{c11} - \boldsymbol{\Sigma}_{c12}\boldsymbol{\mathcal{J}}_{22}^{-1}\boldsymbol{\mathcal{J}}_{21} - \boldsymbol{\mathcal{J}}_{12}\boldsymbol{\mathcal{J}}_{22}^{-1}\boldsymbol{\Sigma}_{c21} + \boldsymbol{\mathcal{J}}_{12}\boldsymbol{\mathcal{J}}_{22}^{-1}\boldsymbol{\Sigma}_{c22}\boldsymbol{\mathcal{J}}_{22}^{-1}\boldsymbol{\mathcal{J}}_{21}.\end{aligned}$$

Take $\boldsymbol{W}_{\mathrm{F}} = \boldsymbol{\mathcal{J}}_{12}\boldsymbol{\mathcal{J}}_{22}^{-1}$ and the proof is completed. □

**Lemma A.2.** *Under Assumptions* 3.1–3.5 *and* 3.10–3.11, *as* $n \to \infty$ *and* $r \to \infty$, *conditional on* $\mathcal{F}_n$ *in probability,*

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_F = O_{\mathrm{P}\,|\,\mathcal{F}_n}(\sqrt{s_l \log p/r}).$$

*Proof.* Define $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathrm{F}}$. By the definitions of $L^*(\boldsymbol{\beta})$ in (2.1) and $\hat{\boldsymbol{\beta}}$ in (2.6), we have

$$L^*(\hat{\boldsymbol{\beta}}) + \lambda_1\|\hat{\boldsymbol{\beta}}\|_1 \leqslant L^*(\boldsymbol{\beta}_{\mathrm{F}}) + \lambda_1\|\boldsymbol{\beta}_{\mathrm{F}}\|_1.$$

By the convexity of $L^*(\boldsymbol{\beta})$, we obtain

$$L^*(\hat{\boldsymbol{\beta}}) \geqslant L^*(\boldsymbol{\beta}_{\mathrm{F}}) + \nabla_{\boldsymbol{\beta}}L^*(\boldsymbol{\beta}_{\mathrm{F}})^{\mathrm{T}}\hat{\boldsymbol{\Gamma}}.$$

In the event $\{\|\nabla_{\boldsymbol{\beta}} L^*(\boldsymbol{\beta}_{\mathrm{F}})\|_\infty \leqslant \lambda_1/2\}$, we have

$$|\nabla_{\boldsymbol{\beta}} L^*(\boldsymbol{\beta}_{\mathrm{F}})^{\mathrm{T}} \hat{\boldsymbol{\Gamma}}| \leqslant \frac{1}{2}\lambda_1 \|\hat{\boldsymbol{\Gamma}}\|_1.$$

Hence,

$$-\frac{1}{2}\|\hat{\boldsymbol{\Gamma}}\|_1 + \|\hat{\boldsymbol{\beta}}\|_1 \leqslant \|\boldsymbol{\beta}_{\mathrm{F}}\|_1.$$

Let

$$\mathcal{S}_l = \{j : \boldsymbol{\beta}_{0j} \neq 0,\, j = 1, \ldots, p\}.$$

Using the fact that $\boldsymbol{\beta}_{\mathrm{F}} - \boldsymbol{\beta}_0 = O_{\mathrm{P}}(1/\sqrt{n})$ (see [26]) and

$$\|\boldsymbol{\beta}_{\mathrm{F}} + \hat{\boldsymbol{\Gamma}}\|_1 = \|(\boldsymbol{\beta}_{\mathrm{F}} + \hat{\boldsymbol{\Gamma}})_{\mathcal{S}_l}\|_1 + \|(\boldsymbol{\beta}_{\mathrm{F}} + \hat{\boldsymbol{\Gamma}})_{\mathcal{S}_l^c}\|_1 \geqslant \|\boldsymbol{\beta}_{\mathrm{F}}\|_1 - \|\hat{\boldsymbol{\Gamma}}_{\mathcal{S}_l}\|_1 + \|\hat{\boldsymbol{\Gamma}}_{\mathcal{S}_l^c}\|_1 + O_{\mathrm{P}}(p/\sqrt{n}),$$

we obtain

$$\|\hat{\boldsymbol{\Gamma}}_{\mathcal{S}_l^c}\|_1 \leqslant 3\|\hat{\boldsymbol{\Gamma}}_{\mathcal{S}_l}\|_1 + O_{\mathrm{P}}(p/\sqrt{n}),$$

which implies that there exists an $\alpha$ such that $\hat{\boldsymbol{\Gamma}} \in \mathcal{C}(\mathcal{S}_l, \alpha)$. Applying the restricted strong convexity condition, we have

$$\begin{aligned}
\gamma\|\hat{\boldsymbol{\Gamma}}\|^2 &\leqslant L^*(\boldsymbol{\beta}_0 + \hat{\boldsymbol{\Gamma}}) - L^*(\boldsymbol{\beta}_0) - \nabla_{\boldsymbol{\beta}} L^*(\boldsymbol{\beta}_0)^{\mathrm{T}} \hat{\boldsymbol{\Gamma}} \leqslant L^*(\hat{\boldsymbol{\beta}}) - L^*(\boldsymbol{\beta}_{\mathrm{F}}) - \nabla_{\boldsymbol{\beta}} L^*(\boldsymbol{\beta}_{\mathrm{F}})^{\mathrm{T}} \hat{\boldsymbol{\Gamma}} + O_{\mathrm{P}}(1/\sqrt{n}) \\
&\leqslant \lambda_1 \|\boldsymbol{\beta}_{\mathrm{F}}\|_1 - \lambda_1 \|\hat{\boldsymbol{\beta}}\|_1 + \frac{1}{2}\lambda_1 \|\hat{\boldsymbol{\Gamma}}\|_1 + O_{\mathrm{P}}(1/\sqrt{n}) \leqslant \frac{3}{2}\lambda_1 \|\hat{\boldsymbol{\Gamma}}_{\mathcal{S}_l}\|_1 - \frac{1}{2}\lambda_1 \|\hat{\boldsymbol{\Gamma}}_{\mathcal{S}_l^c}\|_1 + O_{\mathrm{P}}(p/\sqrt{n}) \\
&\leqslant \frac{3}{2}\lambda_1 \sqrt{s_l} \|\hat{\boldsymbol{\Gamma}}\| + O_{\mathrm{P}}(p/\sqrt{n}).
\end{aligned}$$

This concludes that

$$\|\hat{\boldsymbol{\Gamma}}\| \leqslant \frac{3}{2\gamma}\sqrt{s_l}\lambda_1$$

with probability approaching one. It remains to calculate the probability of the event

$$\left\{\|\nabla_{\boldsymbol{\beta}} L^*(\boldsymbol{\beta}_{\mathrm{F}})\|_\infty \leqslant \frac{\lambda_1}{2}\right\}.$$

By [2, Lemma 2], we have

$$\operatorname{var}\left(\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}[-y_i^* + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)]\boldsymbol{x}_i^* \,\middle|\, \mathcal{F}_n\right) = \frac{1}{nr}\sum_{i=1}^{n}\frac{1}{n\pi_i}[-y_i + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2 \boldsymbol{x}_i^{\otimes 2} = O_{\mathrm{P}}(r^{-1}).$$

By the union bound and Bernstein's inequality, we have

$$\begin{aligned}
&\mathrm{P}\left(\left\|\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}[-y_i^* + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)]\boldsymbol{x}_i^*\right\|_\infty > \frac{1}{2}\lambda_1 \,\middle|\, \mathcal{F}_n\right) \\
&\quad\leqslant \sum_{j=1}^{p}\mathrm{P}\left(\left|\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}[-y_i^* + \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)]x_{ij}^*\right| > \frac{1}{2}\lambda_1 \,\middle|\, \mathcal{F}_n\right) \leqslant 2p\exp\left\{\frac{-Cr\lambda_1^2}{8}\right\},
\end{aligned}$$

where $C$ is a large constant. Thus,

$$\mathrm{P}\left(\|\nabla_{\boldsymbol{\beta}} L^*(\boldsymbol{\beta}_{\mathrm{F}})\|_\infty \leqslant \frac{\lambda_1}{2}\right) \geqslant 1 - 2p\exp\left\{\frac{-Cr\lambda_1^2}{8}\right\}.$$

Taking $\lambda_1 = \sqrt{\log p/r}$, we have the rate of convergence

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(\sqrt{s_l \log p/r}).$$

This completes the proof. □

**Lemma A.3.** *Under Assumptions* 3.1–3.5 *and* 3.10–3.12, *as* $n \to \infty$ *and* $r \to \infty$, *conditional on* $\mathcal{F}_n$ *in probability,*

$$\hat{\boldsymbol{W}} - \boldsymbol{W}_F = O_{\mathrm{P}\,|\,\mathcal{F}_n}(\sqrt{s_h \log q/r}), \tag{A.5}$$

$$\hat{\boldsymbol{J}} - \boldsymbol{J} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(\sqrt{s_h \log q/r}), \tag{A.6}$$

$$\hat{\boldsymbol{K}} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(\sqrt{s_h \log q/r}), \tag{A.7}$$

$$\hat{S}^*(\boldsymbol{\theta}_F, \boldsymbol{\gamma}_F, \hat{\boldsymbol{W}}) = O_{\mathrm{P}\,|\,\mathcal{F}_n}(r^{-1/2}), \tag{A.8}$$

*where*

$$\hat{\boldsymbol{J}} = \frac{\partial \hat{S}^*(\boldsymbol{\theta}_F, \boldsymbol{\gamma}_F, \hat{\boldsymbol{W}})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} = \frac{1}{nr}\sum_{i=1}^{r} \frac{1}{\pi_i^*} \ddot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}} \boldsymbol{x}_i^*)(\boldsymbol{z}_i^* - \hat{\boldsymbol{W}} \boldsymbol{u}_i^*)(\boldsymbol{z}_i^*)^{\mathrm{T}},$$

$$\hat{\boldsymbol{K}} = \frac{\partial \hat{S}^*(\boldsymbol{\theta}_F, \boldsymbol{\gamma}_F, \hat{\boldsymbol{W}})}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} = \frac{1}{nr}\sum_{i=1}^{r} \frac{1}{\pi_i^*} \ddot{\psi}(\boldsymbol{\beta}_F^{\mathrm{T}} \boldsymbol{x}_i^*)(\boldsymbol{z}_i^* - \hat{\boldsymbol{W}} \boldsymbol{u}_i^*)(\boldsymbol{u}_i^*)^{\mathrm{T}}.$$

*Proof.* To show $\hat{\boldsymbol{W}} - \boldsymbol{W}_{\mathrm{F}} = O_{\mathrm{P}\,|\,\mathcal{F}_n}(\sqrt{s_h \log p/r})$, we observe that

$$\frac{1}{nr}\sum_{i=1}^{r} \frac{1}{\pi_i^*} \ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i) \|\boldsymbol{z}_i^* - \hat{\boldsymbol{W}} \boldsymbol{u}_i^*\|^2 + \lambda_2 \sum_{j=1}^{q} \|\boldsymbol{w}_j\| \leqslant \frac{1}{nr}\sum_{i=1}^{r} \frac{1}{\pi_i^*} \ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i) \|\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}} \boldsymbol{u}_i^*\|^2 + \lambda_2 \sum_{j=1}^{q} \|\boldsymbol{w}_{\mathrm{F}j}\|.$$

Let $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{W}} - \boldsymbol{W}_{\mathrm{F}}$. The above inequality can be written as

$$\begin{aligned}
&\frac{1}{nr}\sum_{i=1}^{r} \frac{1}{\pi_i^*} \ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i^*)(\hat{\boldsymbol{\Delta}} \boldsymbol{u}_i^*)^{\mathrm{T}} \hat{\boldsymbol{\Delta}} \boldsymbol{u}_i^* \\
&\quad \leqslant \frac{2}{nr}\sum_{i=1}^{r} \frac{1}{\pi_i^*} \ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i^*)(\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}} \boldsymbol{u}_i^*)^{\mathrm{T}} \hat{\boldsymbol{\Delta}} \boldsymbol{u}_i^* + \lambda_2 \|\boldsymbol{W}_{\mathrm{F}}\|_{2,1} - \lambda_2 \|\hat{\boldsymbol{W}}\|_{2,1} \\
&\quad \leqslant \sum_{j=1}^{q} \left\{ \|\hat{\boldsymbol{\delta}}_j\| \Big\| \frac{2}{nr}\sum_{i=1}^{r} \frac{1}{\pi_i^*} \ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i^*)(\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}} \boldsymbol{u}_i^*) u_{ij}^* \Big\| \right\} + \lambda_2 \|\boldsymbol{W}_{\mathrm{F}}\|_{2,1} - \lambda_2 \|\hat{\boldsymbol{W}}\|_{2,1},
\end{aligned}$$

where $\hat{\boldsymbol{\delta}}_j = \hat{\boldsymbol{w}}_j - \boldsymbol{w}_{\mathrm{F}j}$ is the $j$-th column of $\hat{\boldsymbol{\Delta}}$. Similar to the proof of (A.2), we have

$$\mathrm{var}\left( \frac{2}{nr}\sum_{i=1}^{r} \frac{1}{\pi_i^*} \ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}} \boldsymbol{x}_i^*)(\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}} \boldsymbol{u}_i^*) u_{ij}^* \,\Big|\, \mathcal{F}_n \right) = O_{\mathrm{P}}(r^{-1}).$$

Applying Bernstein's inequality, we have

$$\begin{aligned}
&\mathrm{P}\left( \max_{1 \leqslant j \leqslant q} \left\| \frac{2}{nr}\sum_{i=1}^{r} \frac{1}{\pi_i^*} \ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}} \boldsymbol{x}_i^*)(\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}} \boldsymbol{u}_i^*) u_{ij}^* \right\| > t \,\Big|\, \mathcal{F}_n \right) \\
&\quad \leqslant dq \max_{1 \leqslant k \leqslant d} \max_{1 \leqslant j \leqslant q} \mathrm{P}\left( \left| \frac{2}{nr}\sum_{i=1}^{r} \frac{1}{\pi_i^*} \ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}} \boldsymbol{x}_i^*)(\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}} \boldsymbol{u}_i^*)_k u_{ij}^* \right| > \frac{t}{\sqrt{d}} \,\Big|\, \mathcal{F}_n \right) \leqslant 2dq \exp\left( -\frac{Crt^2}{2d} \right),
\end{aligned}$$

where $C > 0$ is a large constant. Thus, taking $t = \sqrt{\log q/r}$, we have

$$\max_{1 \leqslant j \leqslant q} \left\| \frac{2}{nr}\sum_{i=1}^{r} \frac{1}{\pi_i^*} \ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i^*)(\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}} \boldsymbol{u}_i^*) u_{ij}^* \right\| \lesssim \max_{1 \leqslant j \leqslant q} \left\| \frac{2}{nr}\sum_{i=1}^{r} \frac{1}{\pi_i^*} \ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}} \boldsymbol{x}_i^*)(\boldsymbol{z}_i^* - \boldsymbol{W}_{\mathrm{F}} \boldsymbol{u}_i^*) u_{ij}^* \right\| \leqslant \sqrt{\log q/r}$$

with probability approaching one. Notice that

$$\boldsymbol{W}_{\mathrm{F}} = \arg\min_{\boldsymbol{W}} \frac{1}{n}\sum_{i=1}^{n} \ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}} \boldsymbol{x}_i) \|\boldsymbol{z}_i - \boldsymbol{W} \boldsymbol{u}_i\|^2 \quad \text{and} \quad \boldsymbol{W}_0 = \arg\min_{\boldsymbol{W}} \mathrm{E}(\ddot{\psi}(\boldsymbol{\beta}_0^{\mathrm{T}} \boldsymbol{x}) \|\boldsymbol{z} - \boldsymbol{W} \boldsymbol{u}\|^2).$$

Based on the results of Koenker and Portnoy [15], we have $\boldsymbol{W}_{\mathrm{F}} = \boldsymbol{W}_0 + O_{\mathrm{P}}(1/\sqrt{n})$. Let

$$\mathcal{S}_h = \{j : \boldsymbol{w}_{0j} \neq \boldsymbol{0},\, j = 1, \ldots, q\}.$$

We have

$$\|\hat{\boldsymbol{W}}\|_{2,1} \geqslant \|(\boldsymbol{W}_{\mathrm{F}})_{\mathcal{S}_h}\|_{2,1} - \|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_h}\|_{2,1} + \|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_h^c}\|_{2,1} + O_{\mathrm{P}}(q/\sqrt{n}).$$

Taking $\lambda_2 = O(\sqrt{\log q/r})$, we obtain

$$\begin{aligned}\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i^*)(\hat{\boldsymbol{\Delta}}\boldsymbol{u}_i^*)^{\mathrm{T}}\hat{\boldsymbol{\Delta}}\boldsymbol{u}_i^* &\leqslant \frac{1}{2}\lambda_2\|\hat{\boldsymbol{\Delta}}\|_{2,1} + \lambda_2\|\boldsymbol{W}_{\mathrm{F}}\|_{2,1} - \lambda_2\|\hat{\boldsymbol{W}}\|_{2,1}\\ &\leqslant \frac{3}{2}\lambda_2\|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_h}\|_{2,1} - \frac{1}{2}\lambda_2\|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_h^c}\|_{2,1} + O_{\mathrm{P}}(q/\sqrt{n}),\end{aligned}$$

which implies that there exists an $\alpha'$ such that $\|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_h^c}\|_{2,1} \leqslant \alpha'\|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_h}\|_{2,1}$ and $\hat{\boldsymbol{\Delta}} \in \mathcal{C}'(\mathcal{S}_h, \alpha')$. Noticing that $\|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_h}\|_{2,1} \leqslant \sqrt{s_h}\|\hat{\boldsymbol{\Delta}}\|$, we have

$$\begin{aligned}\kappa\|\hat{\boldsymbol{\Delta}}\|^2 &\leqslant \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{x}_i^*)(\hat{\boldsymbol{\Delta}}\boldsymbol{u}_i^*)^{\mathrm{T}}\hat{\boldsymbol{\Delta}}\boldsymbol{u}_i^* \lesssim \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i^*)(\hat{\boldsymbol{\Delta}}\boldsymbol{u}_i^*)^{\mathrm{T}}\hat{\boldsymbol{\Delta}}\boldsymbol{u}_i^*\\ &\leqslant \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i^*)(\hat{\boldsymbol{\Delta}}\boldsymbol{u}_i^*)^{\mathrm{T}}\hat{\boldsymbol{\Delta}}\boldsymbol{u}_i^* \lesssim \frac{3}{2}\lambda_2\|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_h}\|_{2,1} - \frac{1}{2}\lambda_2\|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_h^c}\|_{2,1} \leqslant \frac{3}{2}\lambda_2\sqrt{s_h}\|\hat{\boldsymbol{\Delta}}\|.\end{aligned}$$

This concludes that

$$\|\hat{\boldsymbol{\Delta}}\| \lesssim \sqrt{s_h}\lambda_2,$$

and furthermore,

$$\|\hat{\boldsymbol{\Delta}}\|_{2,1} \leqslant 4\|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_h}\|_{2,1} \lesssim s_h\lambda_2.$$

The proof of (A.5) is completed. Furthermore, the proof of (A.6)–(A.8) is similar to that of (A.2)–(A.4) and thus we omit them to save space. □

*Proof of Theorems* 3.13 *and* 3.14. The proof of Theorems 3.13 and 3.14 is similar to the proof of Theorems 3.6 and 3.8, which differs in the convergence rate of $\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{J}}$ and $\hat{\boldsymbol{K}}$. Thus, the proof of Theorems 3.13 and 3.14 is direct by combining Lemmas A.2 and A.3 and Theorems 3.6 and 3.8. □

*Proof of Theorem* 4.1. It holds that

$$\begin{aligned}\mathrm{tr}(\boldsymbol{V}) &= \mathrm{tr}(\boldsymbol{J}^{-1}\boldsymbol{V}_c\boldsymbol{J}^{-1})\\ &= \frac{1}{rn^2}\sum_{i=1}^{n}\mathrm{tr}\left\{\frac{1}{\pi_i}[y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2\boldsymbol{J}^{-1}(\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)^{\otimes 2}\boldsymbol{J}^{-1}\right\}\\ &= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}[y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2\|\boldsymbol{J}^{-1}(\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)\|^2\\ &= \frac{1}{rn^2}\sum_{j=1}^{n}\pi_j\sum_{i=1}^{n}\frac{1}{\pi_i}[y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2\|\boldsymbol{J}^{-1}(\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)\|^2\\ &\geqslant \frac{1}{rn^2}\left\{\sum_{i=1}^{n}|y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)|\|\boldsymbol{J}^{-1}(\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)\|\right\}^2,\end{aligned}$$

where the last step is from the Cauchy-Schwarz inequality and the equality in it holds if and only if $\pi_i \propto |y_i - \dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)|\|\boldsymbol{J}^{-1}(\boldsymbol{z}_i - \boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)\|$. □

*Proof of Theorem* 4.2. It holds that

$$\begin{aligned}\operatorname{tr}(\boldsymbol{V}_c) &= \frac{1}{rn^2}\sum_{i=1}^{n}\operatorname{tr}\left\{\frac{1}{\pi_i}[y_i-\dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2(\boldsymbol{z}_i-\boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i)^{\otimes 2}\right\}\\ &= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}[y_i-\dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2\|\boldsymbol{z}_i-\boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i\|^2\\ &= \frac{1}{rn^2}\sum_{j=1}^{n}\pi_j\sum_{i=1}^{n}\frac{1}{\pi_i}[y_i-\dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)]^2\|\boldsymbol{z}_i-\boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i\|^2\\ &\geqslant \frac{1}{rn^2}\left\{\sum_{i=1}^{n}|y_i-\dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)|\|\boldsymbol{z}_i-\boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i\|\right\}^2,\end{aligned}$$

where the last step is from the Cauchy-Schwarz inequality and the equality in it holds if and only if $\pi_i \propto |y_i-\dot{\psi}(\boldsymbol{\beta}_{\mathrm{F}}^{\mathrm{T}}\boldsymbol{x}_i)|\|\boldsymbol{z}_i-\boldsymbol{W}_{\mathrm{F}}\boldsymbol{u}_i\|$. □

*Proof of Theorems* 4.5 *and* 4.6. For convenience, we only prove the result for the A-optimality criterion in the low-dimensional case. It is clear that Assumption 3.5 is satisfied for uniform subsampling. Together with Assumptions 3.1–3.5, we can obtain the consistent results for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{W}}$, which further indicates that the estimated $\hat{\pi}_i^{\mathrm{dmV}}$ satisfies Assumption 3.5. Using the additional condition $\sqrt{r_2}/r_1 \to 0$, we complete the proof of Theorem 4.5 by applying the proof of Theorem 3.6 carefully. Notice that $\hat{\pi}_i^{\mathrm{dmV}}$ has the same expression as $\pi_i^{\mathrm{dmV}}$ except that $\boldsymbol{\beta}_{\mathrm{F}}$ and $\boldsymbol{W}_{\mathrm{F}}$ are replaced by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{W}}$, and the proof of Theorem 4.6 is completed through the proof of Theorem 3.8 and the continuous mapping theory. □