*Submitted to the Annals of Applied Statistics*

# DISTRIBUTED SUBSAMPLING AND QUASI DECORRELATED SCORE FOR CLUSTER DATA: AN APPLICATION TO BEIJING MULTI-SITE AIR QUALITY

BY JUNZHUO GAO[1], LEI WANG[1,a] AND JUN SHAO[2]

[1]*School of Statistics and Data Science, Nankai University,* [a]*lwangstat@nankai.edu.cn*

[2]*Department of Statistics, University of Wisconsin-Madison*

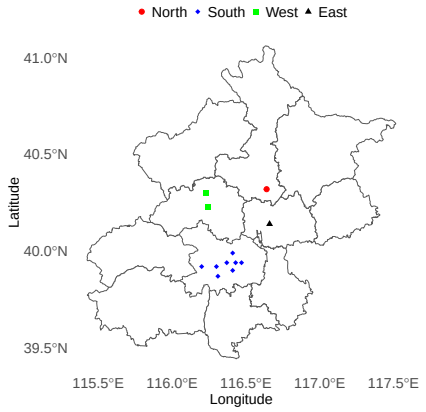
Forecasting and controlling PM2.5 emissions is crucial for environmental protection and public health. To analyze the Beijing multi-site air quality dataset on regional and seasonal effects in PM2.5 emissions, which has large-scale distributed cluster/longitudinal data and high-dimensional covariates, we develop a unified cluster subsampling method for generalized linear models (GLMs) to downsize the data volume and reduce computational burden. To incorporate the within-subject correlation, a weighted generalized estimation equations under an informative working correlation structure is considered and a novel optimal subsampling criterion including both the A- and L-optimality is proposed. For low-dimensional GLMs, the resulting optimal subsample estimators are consistent and asymptotically normal with explicitly derived asymptotic covariance matrices. For the preconceived low-dimensional parameter in high-dimensional GLMs, a quasi decorrelated score function is developed to mitigate the effect from nuisance parameter estimation. Our proposed method is evaluated by simulation. By applying our method to the Beijing multi-site air quality dataset, we reveal that the PM2.5 emissions in the south part of Beijing have a U-shaped seasonal effect in the order of winter, spring, summer, and autumn, and a regional aggregation effect in winter of the southeastern of Beijing.

## 1. Introduction.

1.1. *Beijing multi-site air quality dataset.* Air pollution has serious impacts on environment, climate change, human health, and economy. It is shown that a major air pollutant is from PM2.5, which refers to tiny particles smaller than 2.5 microns in diameter. Exposure to high levels of PM2.5 is linked to increased mortality rates and can exacerbate conditions like asthma and bronchitis. Forecasting and controlling PM2.5 concentration has been regarded as one of the most important issues for protecting public health and ensuring cleaner air quality (Zhang et al., 2012, 2017).

Our study is motivated by the Beijing multi-site air quality dataset (https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data), which comprises emission of PM2.5 and meteorological variables from 12 nationally-controlled monitoring sites in all the urban and rural districts in Beijing from March 1st, 2013 to February 28th, 2017. Although this dataset has been analyzed by some researchers (Yan et al., 2021; Li, Liu and Zhao, 2022; Fan, Lin and Yu, 2024), a thorough statistical analysis of these spatio-seasonal dynamic data with a high-dimensional covariate vector is not available. To enhance the interpretability of the model (Chu, Kadane and Davidson, 2010), temperature, pressure, dew point temperature, wind speed, and their high order terms and interaction terms result in 210 covariates in this dataset (including a constant term for an intercept).

*Keywords and phrases:* Generalized linear models, high-dimensional nuisance parameter, optimal distributed cluster subsampling, quasi score and decorrelated score, within-subject correlation.




(a) Locations of the 12 monitoring sites

| North | | | | | | South | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.00 | 0.94 | 0.90 | 0.87 | 0.84 | 0.80 | 1.00 | 0.95 | 0.91 | 0.86 | 0.83 | 0.79 |
| 0.94 | 1.00 | 0.95 | 0.91 | 0.88 | 0.84 | 0.95 | 1.00 | 0.95 | 0.91 | 0.87 | 0.83 |
| 0.90 | 0.95 | 1.00 | 0.95 | 0.91 | 0.87 | 0.91 | 0.95 | 1.00 | 0.95 | 0.91 | 0.87 |
| 0.87 | 0.91 | 0.95 | 1.00 | 0.95 | 0.91 | 0.86 | 0.91 | 0.95 | 1.00 | 0.95 | 0.91 |
| 0.84 | 0.88 | 0.91 | 0.95 | 1.00 | 0.95 | 0.83 | 0.87 | 0.91 | 0.95 | 1.00 | 0.95 |
| 0.80 | 0.84 | 0.87 | 0.91 | 0.95 | 1.00 | 0.79 | 0.83 | 0.87 | 0.91 | 0.95 | 1.00 |
| West | | | | | | East | | | | | |
| 1.00 | 0.94 | 0.90 | 0.86 | 0.84 | 0.81 | 1.00 | 0.95 | 0.90 | 0.85 | 0.81 | 0.77 |
| 0.94 | 1.00 | 0.94 | 0.90 | 0.87 | 0.83 | 0.95 | 1.00 | 0.95 | 0.90 | 0.86 | 0.82 |
| 0.90 | 0.94 | 1.00 | 0.94 | 0.90 | 0.87 | 0.90 | 0.95 | 1.00 | 0.95 | 0.90 | 0.87 |
| 0.86 | 0.90 | 0.94 | 1.00 | 0.94 | 0.90 | 0.85 | 0.90 | 0.95 | 1.00 | 0.95 | 0.91 |
| 0.84 | 0.87 | 0.90 | 0.94 | 1.00 | 0.94 | 0.81 | 0.86 | 0.90 | 0.95 | 1.00 | 0.95 |
| 0.81 | 0.83 | 0.87 | 0.90 | 0.94 | 1.00 | 0.77 | 0.82 | 0.87 | 0.91 | 0.95 | 1.00 |

(b) Correlation matrices of log PM2.5 emissions

FIG 1. *Beijing multi-site air quality dataset*

According to geographical location and administrative division, we group the 12 monitoring sites in Beijing into 4 regions, north, south, west, and east, as shown in Figure 1(a), and consider seasonal PM2.5 effects of spring, summer, autumn, and winter. This adds 6 additional region and season dummy variables, resulting in a total of 216 covariates in the Beijing multi-site air quality dataset. A methodology to assess effects of main covariates (region and season) not statistically affected by a large number of other covariates needs to be developed.

1.2. *Challenges in analysis.* With the rapid advancement in technology, large-scale data are stored independently in different sites, known as *distributed data*, commonly used by various industries and social organizations to store and manage user records across multiple facilities, ensuring accessibility, security, and scalability for scientific research, treatment planning, and feedback administration. For the Beijing multi-site air quality dataset, a total of 420,768 observations are consolidated into four regions with 35,064, 280,512, 70,128, and 35,064 observations, respectively. Other examples include the observational health data sciences and informatics consortium with over 82 clinical databases around the world (Hripcsak et al., 2015), the communities and crime data stored in different regions designated by the United States Census Bureau, Walmart's data from different locations around the world, and the clinical information of patients stored at different hospitals.

To address the challenge posed by large-scale distributed data with high computational cost, privacy and security concerns, and administrative management, subsampling techniques have been developed, which involves selecting a random subset of the entire dataset and reasonably balancing the need for computational efficiency and the demand for accurate and representative analysis (Ma, Mahoney and Yu, 2015; Wang, Zhu and Ma, 2018; Li and Meng, 2020; Ai et al., 2021a,b; Meng et al., 2021; Wang and Ma, 2021; Zhang, Ning and Ruppert, 2021; Ma et al., 2022; Zhang et al., 2023; Yu, Liu and Wang, 2023; Han et al., 2023; Ye, Yu and Ai, 2025; Wu et al., 2024). More literature can be found in reviews by Yao and Wang (2021), Gao et al. (2022), Li et al. (2024), Yu, Ai and Ye (2024) and references therein. To process large-scale distributed data, distributed subsampling approaches have been proposed, where subsamples are taken from each site and then analyzed collectively (Zhang and Wang, 2021; Zuo et al., 2021; Yu et al., 2022). Although the full dataset of Beijing multi-site air quality is available, distributed subsampling for administrative reasons and/or quick instant investigations may still be needed.

Unfortunately, no work has been done on distributed cluster/longitudinal data in the literature, which presents a significant gap in research and application. Cluster/longitudinal data refer to data collected over a short period of time from the same individual, and have been

ubiquitous in modern society and scientific fields. The Beijing multi-site air quality dataset contains 5,844, 46,752, 11,688, and 5,844 clusters in four regions, respectively, with 6 repeated observations within each cluster. Figure 1(b) indicates that the longitudinal emission of PM2.5 (after taking logarithm) is highly correlated within each cluster. For accurately analyzing cluster/longitudinal data, it is important to consider the correlation within each cluster. When the full dataset from all sites is available and the dimension of covariate vector is fixed and small, the classical parameter estimator is a solution to the quasi score equation in a marginal generalized linear model (GLM) for cluster/longitudinal data with estimated correlation, which is valid regardless of whether the correlation is consistently estimated or not (Liang and Zeger, 1986; Xie and Yang, 2003; Balan and Schiopu-Kratina, 2005). Since the full dataset is often not accessible due to reasons previously given, subsampling for distributed clustered data needs to be developed.

Another challenge is from the rapid increase in covariate dimension with technological development, which may affect the performance of traditional estimators, let alone for the subsample estimator. As described in Section 1.1, the Beijing multi-site air quality dataset contains a total of 216 covariates. Penalization methods on the quasi score equation have been developed to handle high-dimensional covariates (Wang, 2011; Wang, Zhou and Qu, 2012). Recently, there have been significant interests and widespread attentions in studying the relationship between the response variable and a covariate sub-vector with a small dimension, while treating the remaining large number of parameters related with other covariates as nuisances (Zhang and Zhang, 2014; Ning and Liu, 2017; Fang, Ning and Li, 2020; Li, Li and Ma, 2021; Cheng et al., 2022). In the Beijing multi-site air quality dataset, the low dimension sub-vector is formed by region and season dummy variables if their effects are the main focus, while effects of other covariates can be regarded as nuisances. The key to this approach is constructing a decorrelated score (for the parameter of interest) uncorrelated with the score of the high-dimensional nuisance parameter. Under subsampling, Gao, Wang and Lian (2024) proposed a decorrelated score for low-dimensional parameters of interest. However, the existing decorrelated score approach cannot be applied to cluster/longitudinal data, because it is based on likelihood and score, whereas the quasi score equation for cluster/longitudinal data is only based on a marginal GLM model and a within cluster correlation structure. Consequently, there is a need to construct decorrelated score for cluster/longitudinal data, with or without subsampling.

1.3. *Our approaches and results.* In this paper, we aim to address the challenges in the era of large data volume and high covariate dimension. Our contributions in methodology are twofold. First, in Section 2 we develop a *distributed cluster subsampling* for cluster/longitudinal data. To maintain and incorporate the correlation structure within each cluster, we sample clusters instead of individual observations. To pursue efficient distributed cluster subsampling, we propose an optimal distributed cluster subsampling scheme, which incorporates the within-subject correlation for each cluster and includes A- and L-optimality criteria (Wang, Zhu and Ma, 2018; Wang and Ma, 2021; Ai et al., 2021b) as special cases. Second, in Section 3 we propose a quasi decorrelated score, which still enjoys an orthogonality property that a decorrelated score possesses. The resulting estimator of the main low dimensional parameter vector of interest is asymptotically normal and its asymptotic efficiency is not affected by the estimation of nuisance parameter vector with possibly high dimension. When the full dataset is not available, we apply quasi decorrelated score jointly with distributed cluster subsampling developed in Section 2. Optimal distributed cluster subsampling schemes are also constructed when quasi decorrelated score is used. Thus, we extend the optimal decorrelated score subsampling (Gao, Wang and Lian, 2024) to optimal distributed cluster subsampling via quasi decorrelated score.

Asymptotic normality of estimators are established under some regularity conditions in Sections 2 and 3, useful for obtaining optimal distributed cluster subsampling schemes and for asymptotic inference together with derived consistent estimators of asymptotic covariance matrices. Simulation results are presented in Section 4 to supplement the theoretical finding.

In Section 5, we analyze the PM2.5 emissions in Beijing and reveal significant regional and seasonal effects. Specifically, the PM2.5 emissions in south region of Beijing have a U-shaped effect in the order of winter, spring, summer, and autumn, with the highest effect in winter, lowest effect in summer, and in-between effects of spring and autumn with higher effect in spring. Furthermore, there is a regional aggregation effect in winter, i.e., higher PM2.5 emissions in the southeastern part of Beijing, lower PM2.5 emissions in the northwestern part, and gradually decline from the heart of city towards countryside. Explanations to these phenomena can be found in Section 5. Understanding how PM2.5 emissions vary by region and season can help identify which populations are the most at risk during which times of the year. This can lead to more targeted public health interventions and policies to reduce exposure and associated health risks. Insights into regional and seasonal variations in PM2.5 can also help the development of more effective air quality regulations and pollution control strategies tailored to specific areas and times of the year.

The paper is concluded with some discussions in Section 6.

## 2. Distributed cluster subsampling.

2.1. *Review of the quasi score equation.* Let $\boldsymbol{Y}_{ki} = (y_{ki1}, \ldots, y_{kim_{ki}})^{\mathrm{T}} \in \mathbb{R}^{m_{ki}}$ and $\boldsymbol{X}_{ki} = (\boldsymbol{x}_{ki1}, \ldots, \boldsymbol{x}_{kim_{ki}})^{\mathrm{T}} \in \mathbb{R}^{m_{ki}\times p}$ be clustered response and covariate vectors, respectively, from the $i$th subject in the $k$th site, where $\mathbb{R}^d$ is the $d$-dimensional Euclidean space, $\boldsymbol{a}^{\mathrm{T}}$ is the transpose of vector $\boldsymbol{a}$, $m_{ki}$'s are cluster sizes bounded by a fixed constant $m = \max_{i,k} m_{ki}$, $k = 1, \ldots, K$, $i = 1, \ldots, n_k$, and $n_k$ is the number of clusters in site $k$. $(\boldsymbol{X}_{ki}, \boldsymbol{Y}_{ki})$'s are independent, but observations within cluster $(\boldsymbol{X}_{ki}, \boldsymbol{Y}_{ki})$ are correlated with an unknown correlation matrix $\mathrm{corr}(\boldsymbol{Y}_{ki} \mid \boldsymbol{X}_{ki})$.

We assume that the $j$th marginal probability density $f(y \mid \boldsymbol{x})$ of $y_{kij}$ given $\boldsymbol{x}_{kij}$, $j = 1, \ldots, m_{ki}$ (with respect to a dominating measure $\nu$, e.g., the counting measure for discrete $y$ or Lebesgue measure for continuous $y$) follows a generalized linear model (McCullagh and Nelder, 1989) with canonical link, i.e.,

$$f(y \mid \boldsymbol{x}) = h(y) \exp\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}\, y - \psi(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})\}, \tag{1}$$

where $h(\cdot)$ and $\psi(\cdot)$ are known functions, $\psi$ is third-order continuously differentiable, and $\boldsymbol{\beta}$ is an unknown $p$-dimensional parameter vector. An unknown dispersion parameter can be added to model (1) (McCullagh and Nelder, 1989), but we focus on the case of no dispersion parameter for ease of presentation, as the general case can be treated similarly and the practical implement issues are discussed in Section 4.1. Note that the density of $y_{kij}$ given $\boldsymbol{x}_{kij}$ depends on $j$ and site $k$ through the value of $\boldsymbol{x}_{kij}$.

If the full dataset $\mathcal{F} = \{(\boldsymbol{X}_{ki}, \boldsymbol{Y}_{ki}), i = 1, \ldots, n_k, k = 1, \ldots, K\}$ from all sites with a total of $n = \sum_{k=1}^{K} n_k$ clusters is available, and if $p$ is fixed and small, then the customary estimator $\hat{\boldsymbol{\beta}}_{\mathcal{F}}$ of $\boldsymbol{\beta}$ is a solution to the following quasi score equation (Liang and Zeger, 1986),

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} \boldsymbol{X}_{ki}^{\mathrm{T}}\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{b})\hat{\boldsymbol{R}}_{ki}^{-1}\{\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{b})\}^{-1}\{\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_{ki}(\boldsymbol{b})\} = \boldsymbol{0}, \quad \boldsymbol{b} \in \mathbb{R}^p. \tag{2}$$

where $\boldsymbol{0}$ denotes the vector of zeros, $\boldsymbol{\mu}_{ki}(\boldsymbol{b}) = (\dot{\psi}(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_{ki1}), \ldots, \dot{\psi}(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_{kim_{ki}}))^{\mathrm{T}}$, $\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{b})$ is the diagonal matrix of order $m_{ki}$ whose $j$th diagonal entry is $\{\ddot{\psi}(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_{kij})\}^{1/2}$, $\dot{\psi}$ and $\ddot{\psi} > 0$ are the first and second-order derivatives of $\psi$, and $\hat{\boldsymbol{R}}_{ki}$ is an estimator of the correlation

matrix $\operatorname{corr}(\boldsymbol{Y}_{ki} \mid \boldsymbol{X}_{ki})$ based on a working model that may be incorrect. Examples of $\hat{\boldsymbol{R}}_{ki}$ can be found in Section 4.1 and in Liang and Zeger (1986).

Although the full data estimator $\hat{\boldsymbol{\beta}}_{\mathcal{F}}$ is asymptotically valid for estimating $\boldsymbol{\beta}$ with a fixed and low $p$ regardless of whether $\hat{\boldsymbol{R}}_{ki}$'s are consistent or not (Liang and Zeger, 1986), the full dataset $\mathcal{F}$ is often not available as discussed in Section 1, or a quick instant investigation is needed. Thus, we are motivated to consider the following subsampling technique.

2.2. *Optimal subsampling.* In this section we focus on fixed and low covariate dimension $p$. High covariate dimension is considered in Section 3. To maintain and incorporate the correlation structure within each cluster, we sample clusters instead of individual observations; once a cluster is selected, all its observations are extracted together. Specifically, in distributed cluster subsampling, we take a random cluster subsample of size $r_k$ with replacement from $\{(\boldsymbol{X}_{ki}, \boldsymbol{Y}_{ki}), i = 1, \ldots, n_k\}$ for each site $k$, where each $(\boldsymbol{X}_{ki}, \boldsymbol{Y}_{ki})$ is selected with probability $\pi_{ki}$ under the constraint $\sum_{i=1}^{n_k} \pi_{ki} = 1$ and independently across $k = 1, \ldots, K$. Choices of $\pi_{ki}$ and $r_k$ are discussed later.

Let $\mathcal{S} = \{(\boldsymbol{X}^*_{ki}, \boldsymbol{Y}^*_{ki}), i = 1, \ldots, r_k, k = 1, \ldots, K\}$ be the subsample, $\boldsymbol{A}^*_{ki}$ and $\boldsymbol{\mu}^*_{ki}$ be $\boldsymbol{A}_{ki}$ and $\boldsymbol{\mu}_{ki}$, respectively, with $(\boldsymbol{X}_{ki}, \boldsymbol{Y}_{ki})$ replaced by $(\boldsymbol{X}^*_{ki}, \boldsymbol{Y}^*_{ki})$, and $\pi^*_{ki}$ be the associated probability of selecting $(\boldsymbol{X}^*_{ki}, \boldsymbol{Y}^*_{ki})$. The distributed cluster subsampling estimator $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ is a solution to the following weighted quasi score equation

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{r_k} \frac{\boldsymbol{X}^{*\mathrm{T}}_{ki}\boldsymbol{A}^{*1/2}_{ki}(\boldsymbol{b})\hat{\boldsymbol{R}}^{*-1}_{ki}\{\boldsymbol{A}^{*1/2}_{ki}(\boldsymbol{b})\}^{-1}\{\boldsymbol{Y}^*_{ki} - \boldsymbol{\mu}^*_{ki}(\boldsymbol{b})\}}{r_k \pi^*_{ki}} = \boldsymbol{0}, \quad \boldsymbol{b} \in \mathbb{R}^p, \tag{3}$$

where $\hat{\boldsymbol{R}}^*_{ki}$ is an estimator of correlation matrix $\operatorname{corr}(\boldsymbol{Y}_{ki} \mid \boldsymbol{X}_{ki})$ based on $\mathcal{S}$, following the approach of obtaining $\hat{\boldsymbol{R}}_{ki}$ in (2) (see examples in Liang and Zeger (1986) and in Section 4.1). Inverse inclusion probability $r_k\pi^*_{ki}$ weighting in (3) ensures that the left hand side of (3) is unbiased under subsampling for (2) with $\hat{\boldsymbol{R}}_{ki}$ replaced by $\boldsymbol{R}_{ki} =$ limit of $\hat{\boldsymbol{R}}^*_{ki}$.

To establish asymptotic properties of $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$, we list some regularity conditions. All limits without further specification are obtained as $\min_{k\le K} n_k \to \infty$ and $r \to \infty$, where $r = \sum_{k=1}^{K} r_k$ is the total number of sampled clusters, and $\xrightarrow{d}$ denotes convergence in distribution.

(A.1) The covariate values vary in a bounded set, $n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k} E(\|\boldsymbol{Y}_{ki}\|^{6+\delta})$ is bounded for a constant $\delta > 0$, where $\|\boldsymbol{c}\|$ denotes the $L_2$-norm of a vector $\boldsymbol{c}$, and the Fisher information matrix $\boldsymbol{J} = \sum_{k=1}^{K}\sum_{i=1}^{n_k} E\{\boldsymbol{X}^{\mathrm{T}}_{ki}\boldsymbol{A}^{1/2}_{ki}(\boldsymbol{\beta})\boldsymbol{R}^{-1}_{ki}\boldsymbol{A}^{1/2}_{ki}(\boldsymbol{\beta})\boldsymbol{X}_{ki}\}/n$ converges to a positive definite matrix, where $\boldsymbol{R}_{ki}$ is the limit of estimated correlation matrix $\operatorname{corr}(\boldsymbol{Y}_{ki} \mid \boldsymbol{X}_{ki})$.

(A.2) $0 < \max_{i=1,\ldots,n_k,k=1,\ldots,K} \lambda_{\min}(\boldsymbol{R}_{ki}) \le \max_{i=1,\ldots,n_k,k=1,\ldots,K} \lambda_{\max}(\boldsymbol{R}_{ki}) < \infty$, where $\lambda_{\min}(\boldsymbol{R}_{ki})$ and $\lambda_{\max}(\boldsymbol{R}_{ki})$ are the minimum and maximum eigenvalues of $\boldsymbol{R}_{ki}$, respectively.

(A.3) $\max_{i=1,\ldots,n_k,k=1,\ldots,K} r/(nr_k\pi_{ki})$ is bounded in probability.

Assumption (A.1) contains commonly used conditions on the covariates, responses, and model structure (Wang, Zhou and Qu, 2012; Blazère, Loubes and Gamboa, 2014; Ai et al., 2021b; Zhang and Wang, 2021; Zuo et al., 2021; Zhang and Jia, 2022). The boundedness of covariates does not require to know the bounds, and can be achieved by an appropriate transformation that preserves the relationship of covariates and response. In the Beijing multi-site air quality dataset, each meteorological variable is within a range; for example, the temperature ranges between $-20^{\circ}$C and $42^{\circ}$C. For practical implementation, continuous covariates can be centralized and standardized to reduce heterogeneity. Assumption (A.2) is

a restriction on the estimated correlation matrix for longitudinal data (Balan and Schiopu-Kratina, 2005). Assumption (A.3) is typical in survey sampling and can be achieved when we design a subsampling strategy; all our subsampling plans considered later in this section satisfy Assumption (A.3). The proof of following result and all other technical proofs are in the Supplementary Material.

THEOREM 2.1. *Suppose that Assumptions (A.1)-(A.3) hold.*
*(i) With probability tending to* 1*, for any* $\epsilon > 0$*, there exist fixed* $\Delta_\epsilon$ *and* $r_\epsilon$ *such that for all* $r > r_\epsilon$*,* $P\big(\|\hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_{\mathcal{F}}\|^2 \geq \Delta_\epsilon / r \mid \mathcal{F}\big) < \epsilon$*.*
*(ii) If, in addition,* $r/n \to 0$*, then*

$$(\boldsymbol{J}^{-1}\boldsymbol{V}\boldsymbol{J}^{-1})^{-1/2}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}_p), \tag{4}$$

*where* $\boldsymbol{I}_d$ *denotes the identity matrix of order* $d$*,* $\boldsymbol{J}$ *is given in Assumption (A.1),*

$$\boldsymbol{V} = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \frac{\big[\boldsymbol{X}_{ki}^{\mathrm{T}}\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta})\boldsymbol{R}_{ki}^{-1}\{\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta})\}^{-1}\{\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_{ki}(\boldsymbol{\beta})\}\big]^{\otimes 2}}{r_k \pi_{ki} n^2},$$

*and* $\mathbf{c}^{\otimes 2} = \mathbf{c}\mathbf{c}^{\mathrm{T}}$ *for a vector* $\mathbf{c}$*.*

REMARK 2.1. *Theorem 2.1(i) gives a low bound on the* $L_2$*-distance between the subsample estimator* $\hat{\boldsymbol{\beta}}_s$ *and the full data estimator* $\hat{\boldsymbol{\beta}}_{\mathcal{F}}$*. Theorem 2.1(ii) shows that* $\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}$ *is asymptotically normal with mean* $\mathbf{0}$ *and asymptotic covariance matrix* $\boldsymbol{J}^{-1}\boldsymbol{V}\boldsymbol{J}^{-1}$*. Under Assumptions (A.1)-(A.3),* $\boldsymbol{J}^{-1}$ *is asymptotically well defined and* $r\boldsymbol{V}$ *converges in probability to a positive definite matrix and, hence,* $\boldsymbol{J}^{-1}\boldsymbol{V}\boldsymbol{J}^{-1}$ *is of the order* $r^{-1}$*. The condition* $r/n \to 0$ *in Theorem 2.1(ii) is typically satisfied when the cluster subsampling is needed. If* $r/n \to c > 0$*, then* $\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}$ *is still asymptotically normal with mean* $\mathbf{0}$ *and a positive definite covariance matrix. Here, we omit the related discussion since* $r/n \to c > 0$ *does not often occur in subsampling applications.*

REMARK 2.2. *The results in Theorem 2.1 hold for any given set of subsampling probabilities* $\pi_{ki}$*'s and sizes* $r_k$*'s satisfying Assumption (A.3). A simple subsampling plan is uniform within each site* $k$*, i.e.,* $\pi_{ki} = n_k^{-1}$ *for all* $i$*, with sizes* $r_k$*'s proportionally allocated, i.e.,* $r_k/n_k = r/n$ *for all* $k$*. Proportional allocation is frequently applied in sample surveys. For uniform subsampling with proportional allocation, Assumption (A.3) holds as* $r/(nr_k\pi_{ki}) = rn_k/(nr_k) = 1$ *for all* $k$*.*

Although uniform subsampling with proportional allocation is simple and Theorem 2.1 is applicable, it is not the most efficient subsampling strategy. The next theorem shows that we can find an optimal distributed cluster subsampling strategy that minimizes the trace of asymptotic covariance matrix of the transformation $\boldsymbol{L}\hat{\boldsymbol{\beta}}_s$, where $\boldsymbol{L}$ is a known, fixed, and nonsingular $s \times p$ matrix. Examples of $\boldsymbol{L}\boldsymbol{\beta}$ are a component of $\boldsymbol{\beta}$, a linear contrast of components of $\boldsymbol{\beta}$, $\boldsymbol{L} = \boldsymbol{I}_p$ corresponding to A-optimality criterion (Wang, Zhu and Ma, 2018; Ai et al., 2021b), and $\boldsymbol{L} = \boldsymbol{J}$ corresponding to L-optimality criterion (Wang and Ma, 2021; Ai et al., 2021b).

THEOREM 2.2. *Suppose that Assumptions (A.1)-(A.3) hold. For all* $\pi_{ki}$*'s and* $r_k$*'s under the constraint that* $\sum_{i=1}^{n_k}\pi_{ki} = 1$ *and* $\sum_{k=1}^{K} r_k = r$ *with a pre-chosen* $r$ *satisfying* $r/n \to 0$*, if the distributed cluster subsampling probability for selecting cluster* $(k, i)$ *is* $\pi_{ki}^L = \tau_{ki}/\big(\sum_{i=1}^{n_k}\tau_{ki}\big)$ *and subsample size in site* $k$ *is* $r_k^L = r\big(\sum_{i=1}^{n_k}\tau_{ki}\big)/\big(\sum_{l=1}^{K}\sum_{i=1}^{n_l}\tau_{li}\big)$ *with*

$$\tau_{ki} = \|\boldsymbol{L}\boldsymbol{J}^{-1}\boldsymbol{X}_{ki}^{\mathrm{T}}\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta})\boldsymbol{R}_{ki}^{-1}\{\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta})\}^{-1}\{\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_{ki}(\boldsymbol{\beta})\}\|,$$

*for $i = 1, \ldots, n_k, k = 1, \ldots, K$, then* $\operatorname{trace}(\boldsymbol{L}\boldsymbol{J}^{-1}\boldsymbol{V}\boldsymbol{J}^{-1}\boldsymbol{L}^{\mathrm{T}})$*, the trace of asymptotic covariance matrix of $\boldsymbol{L}\hat{\boldsymbol{\beta}}_s$, attains its minimum.*

REMARK 2.3. *The optimality criterion in Theorem 2.2 uses a general $\boldsymbol{L}$ and incorporates the within-subject correlation for each cluster and, thus, is more general than the existing ones in (Wang, Zhu and Ma, 2018; Ai et al., 2021b; Wang and Ma, 2021). The computation complexity of $\tau_{ki}$ for $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$ is $O(np^2s + m^2ns)$. To further calculate $\pi^L_{ki}$ for $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$, we need $O(n)$ memory in total.*

The optimal distributed cluster subsampling scheme in Theorem 2.2 cannot be applied directly, since $\tau_{ki}$ depends on unknown $\boldsymbol{\beta}$ and $\boldsymbol{R}_{ki}$. The situation is similar to the use of optimal stratified sampling in surveys with sites as strata, where we have to estimate the optimal sampling plan based on a pilot study. Therefore, we assume that we first draw an initial subsample $\mathcal{S}_I = \{(\boldsymbol{X}^{*I}_{ki}, \boldsymbol{Y}^{*I}_{ki}), i = 1, \ldots, r^I_k, k = 1, \ldots, K\}$ of size $r^I = \sum_{k=1}^K r^I_k$ with proportionally allocated uniform subsampling as previously described, obtain initial estimators $\hat{\boldsymbol{\beta}}_{s_I}$ and $\hat{\boldsymbol{R}}^{*I}_{ki}$ based on the initial distributed cluster subsample $\mathcal{S}_I$, and then estimate $\tau_{ki}$ by replacing $\boldsymbol{\beta}$ and $\boldsymbol{R}_{ki}$ in $\tau_{ki}$ with $\hat{\boldsymbol{\beta}}_{s_I}$ and $\hat{\boldsymbol{R}}^{*I}_{ki}$ based on initial subsample $\mathcal{S}_I$, respectively.

After $\tau_{ki}$ is estimated, we draw a distributed cluster subsample with estimated optimal probabilities and sizes, independent of the initial subsample, and then calculate $\hat{\boldsymbol{\beta}}_s$ using (3) based on the subsample $\mathcal{S}$. It is shown in the Supplementary Material that $\hat{\boldsymbol{\beta}}_s$ based on this procedure has the asymptotic distribution given in Theorem 2.1(ii) with minimized trace of $\boldsymbol{L}\boldsymbol{J}^{-1}\boldsymbol{V}\boldsymbol{J}^{-1}\boldsymbol{L}^{\mathrm{T}}$. The performance of this optimal distributed cluster subsampling procedure is checked by simulation in Section 4.1. In particular, it is shown in simulation that this optimal procedure produces more efficient estimators than those based on uniform subsampling. Regarding the determination of $r^I$ and $r$, $r^I \le r$ is usually adopted in practice since the second-step subsample is more efficient than the first-step subsample. In addition, $r$ is primarily determined by the desired estimation precision, the cost/time of measuring and collecting the samples. For example, we consider the trace of asymptotic covariance matrix is no greater than a prespecified positive constant $C_0$, i.e., $\operatorname{trace}(\boldsymbol{L}\boldsymbol{J}^{-1}\boldsymbol{V}\boldsymbol{J}^{-1}\boldsymbol{L}^{\mathrm{T}}) \le C_0$.

For $\boldsymbol{Y}_{ki}$ being very close to $\boldsymbol{\mu}_{ki}(\hat{\boldsymbol{\beta}}_{s_I})$, the corresponding estimated $\hat{\tau}_{ki}$ can be too small. To protect the quasi score function from being inflated, we propose to truncate $\hat{\tau}_{ki}$ at a specified threshold such as $10^{-6}$. To conduct asymptotic inference on $\boldsymbol{\beta}$, we can apply Theorem 2.1(ii) with $\boldsymbol{J}^{-1}\boldsymbol{V}\boldsymbol{J}^{-1}$ estimated by the consistent estimator $\hat{\boldsymbol{J}}^{-1}\hat{\boldsymbol{V}}\hat{\boldsymbol{J}}^{-1}$, where

$$\hat{\boldsymbol{J}} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{r_k}\frac{\boldsymbol{X}^{*\mathrm{T}}_{ki}\boldsymbol{A}^{*1/2}_{ki}(\hat{\boldsymbol{\beta}}_s)\hat{\boldsymbol{R}}^{*-1}_{ki}\boldsymbol{A}^{*1/2}_{ki}(\hat{\boldsymbol{\beta}}_s)\boldsymbol{X}^{*}_{ki}}{r_k\pi^{*}_{ki}},$$

$$\hat{\boldsymbol{V}} = \sum_{k=1}^{K}\sum_{i=1}^{r_k}\frac{1}{(nr_k\pi^{*}_{ki})^2}\left[\boldsymbol{X}^{*\mathrm{T}}_{ki}\boldsymbol{A}^{*1/2}_{ki}(\hat{\boldsymbol{\beta}}_s)\hat{\boldsymbol{R}}^{*-1}_{ki}\{\boldsymbol{A}^{*1/2}_{ki}(\hat{\boldsymbol{\beta}}_s)\}^{-1}\{\boldsymbol{Y}^{*}_{ki} - \boldsymbol{\mu}^{*}_{ki}(\hat{\boldsymbol{\beta}}_s)\}\right]^{\otimes 2},$$

and $\pi_{ki}$'s and $r_k$'s are the given or estimated optimal probabilities and sizes. The performance of $\hat{\boldsymbol{J}}^{-1}\hat{\boldsymbol{V}}\hat{\boldsymbol{J}}^{-1}$ is checked by simulation in Section 4.1. The above etimation and inference procedures are summarized in Algorithm S1 in the Supplementary Material.

**3. Quasi decorrelated score function.** As described in Section 1, estimators $\hat{\boldsymbol{\beta}}_{\mathcal{F}}$ and $\hat{\boldsymbol{\beta}}_s$ are not suitable for high dimensional $\boldsymbol{\beta}$. As an alternative, we focus on the situation where we are only interested in the coefficient corresponding to a covariate sub-vector $\boldsymbol{z}$ with a small dimension $q \ll p$. That is, $\boldsymbol{x} = (\boldsymbol{z}^{\mathrm{T}}, \boldsymbol{u}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\beta} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{z} + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{u}$, the $q$-dimensional $\boldsymbol{\theta}$ is what we are interested, $\boldsymbol{u}$ is extraneous although it is related with $y$ and/or $\boldsymbol{z}$, and $\boldsymbol{\gamma}$ is a high dimensional nuisance parameter. In the Beijing multi-site air

quality dataset, $K = 4$, $n_1 = 5,844$, $n_2 = 46,752$, $n_3 = 11,688$, $n_4 = 5,844$, $m_{ki} = 6$, and $p = 216$. Although the sample cluster size $n$ of the full dataset and $r$ of subsample can be larger or much larger than $p$, it is still not good and unnecessary to include all covariates in estimation. We select a $q = 6$ dimensional $\boldsymbol{z}$ with regional and seasonal dummy variables, since our main interest is in analyzing regional and seasonal PM2.5 emissions in Beijing, and we treat the rest 210 covariates as the extraneous covariates in $\boldsymbol{u}$, including an intercept term, temperature, pressure, dew point temperature, wind speed, as well as their high order terms and interaction terms.

We develop a decorrelated score approach in this section. We first assume that the full dataset $\mathcal{F}$ is available.

For non-cluster data ($m_{ki} \equiv 1$) with log likelihood $\ell(\boldsymbol{\beta})$ of $\boldsymbol{\beta} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\theta}$ is a $q$-dimensional parameter vector of interest with a small $q$ and $\boldsymbol{\gamma}$ is a nuisance parameter vector with high dimension $p - q$, the following decorrelated score function of $\boldsymbol{\theta}$ is proposed in Ning and Liu (2017),

$$\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{W}) = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\beta}) - \boldsymbol{W}^{\mathrm{T}} \nabla_{\boldsymbol{\gamma}} \ell(\boldsymbol{\beta}), \tag{5}$$

where $\boldsymbol{W} = [E\{\nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \ell(\boldsymbol{\beta})\}]^{-1} E\{\nabla_{\boldsymbol{\gamma}\boldsymbol{\theta}} \ell(\boldsymbol{\beta})\}$, $\nabla_{\boldsymbol{a}}$ is the partial derivative with respect to $\boldsymbol{a}$, and $\nabla_{\boldsymbol{a}\boldsymbol{b}} = \nabla_{\boldsymbol{b}} \nabla_{\boldsymbol{a}}$. The function $\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{W})$ in (5) is called decorrelated score because it is uncorrelated with the score function for nuisance $\boldsymbol{\gamma}$, which is equivalent to the orthogonality property

$$E\{\nabla_{\boldsymbol{\gamma}} \boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{W})\} = \boldsymbol{0}, \tag{6}$$

useful in separating the effect of estimating $\boldsymbol{\gamma}$ from the estimation of $\boldsymbol{\theta}$.

In our problem with cluster data, however, equation (2) is only a quasi score equation, because the joint distribution of $\boldsymbol{Y}_{ki}$ given $\boldsymbol{X}_{ki}$ is not fully specified. Thus, the idea of decorrelated score cannot be directly applied. In the following, we derive a quasi decorrelated score function. With $\boldsymbol{x} = (\boldsymbol{z}^{\mathrm{T}}, \boldsymbol{u}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, the left hand side of (2) is a $p$-dimensional vector whose first $q$ components form the quasi score for $\boldsymbol{\theta}$ and last $p - q$ components form the quasi score for $\boldsymbol{\gamma}$. With any $(p - q) \times q$ matrix $\boldsymbol{W}$, a $q$-dimensional quasi score for $\boldsymbol{\theta}$ analogous to (5) is

$$\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{R}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\boldsymbol{Z}_{ki} - \boldsymbol{U}_{ki} \boldsymbol{W})^{\mathrm{T}} \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}) \boldsymbol{R}_{ki}^{-1} \{\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta})\}^{-1} \{\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_{ki}(\boldsymbol{\beta})\}, \tag{7}$$

where $\boldsymbol{X}_{ki} = (\boldsymbol{Z}_{ki}, \boldsymbol{U}_{ki})$, $\boldsymbol{Z}_{ki} \in \mathbb{R}^{m_{ki} \times q}$, $\boldsymbol{U}_{ki} \in \mathbb{R}^{m_{ki} \times (p-q)}$, $\boldsymbol{R} = (\boldsymbol{R}_{ki}, i = 1, \ldots, n_k, k = 1, \ldots, K)$, and $\boldsymbol{R}_{ki}$'s are limits of $\hat{\boldsymbol{R}}_{ki}$'s in (2) (not necessary the true correlation matrices). We propose to use a $\boldsymbol{W}$ in (7) satisfying the following analog of the orthogonality property (6),

$$E\{\nabla_{\boldsymbol{\gamma}} \boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{R})\} = \boldsymbol{0}. \tag{8}$$

The function $\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{R})$ in (7) satisfying (8) is called a *quasi decorrelated score*, although orthogonality property (8) does not necessarily produce a $\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{R})$ uncorrelated with the quasi score for nuisance parameter $\boldsymbol{\gamma}$.

After some derivation, it can be shown that the following matrix satisfies (8),

$$\boldsymbol{W} = \arg\min_{\boldsymbol{\omega}} \sum_{k=1}^{K} \sum_{i=1}^{n_k} E \| \boldsymbol{R}_{ki}^{-1/2} \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}) (\boldsymbol{Z}_{ki} - \boldsymbol{U}_{ki} \boldsymbol{\omega}) \|^2, \tag{9}$$

i.e.,

$$\boldsymbol{W} = \Big[ \sum_{k=1}^{K} \sum_{i=1}^{n_k} E\{\boldsymbol{U}_{ki}^{\mathrm{T}} \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}) \boldsymbol{R}_{ki}^{-1} \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}) \boldsymbol{U}_{ki}\} \Big]^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} E\{\boldsymbol{U}_{ki}^{\mathrm{T}} \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}) \boldsymbol{R}_{ki}^{-1} \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}) \boldsymbol{Z}_{ki}\},$$

where $\|\boldsymbol{C}\|$ denotes the Frobenious norm for a matrix $\boldsymbol{C}$.

To use $\boldsymbol{W}$ in (9), we need to estimate $\boldsymbol{\beta}$ and $\boldsymbol{R}$. When the full dataset $\mathcal{F}$ is available, we still should not use (2) to estimate $\boldsymbol{\beta}$ with a high dimension $p$. We consider the following lasso type initial estimator of $\boldsymbol{\beta}$ when full dataset $\mathcal{F}$ is available,

$$\tilde{\boldsymbol{\beta}}_{\mathcal{F}} = \arg\min_{\boldsymbol{b}} \Big[ \sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} \{\psi(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_{kij}) - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_{kij}\, y_{kij}\} \Big] + n\lambda_1 \|\boldsymbol{b}\|_1,$$

where $\lambda_1$ is a penalization parameter for high dimension and $\|\cdot\|_1$ is the $L_1$-norm. With $\tilde{\boldsymbol{\beta}}_{\mathcal{F}}$, we estimate $\boldsymbol{R}$ by $\tilde{\boldsymbol{R}}_{\mathcal{F}} = (\tilde{\boldsymbol{R}}_{ki}, i = 1, \ldots, n_k, k = 1, \ldots, K)$, where $\tilde{\boldsymbol{R}}_{ki}$ is estimated correlation using the same method in Liang and Zeger (1986) when $\mathcal{F}$ is available. To estimate $\boldsymbol{W}$ in (9), since the row dimension of $\boldsymbol{W}$ is $p - q$, we also use a lasso type estimator

$$\tilde{\boldsymbol{W}}_{\mathcal{F}} = \arg\min_{\boldsymbol{\omega}} \Big\{ \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|\tilde{\boldsymbol{R}}_{ki}^{-1/2} \boldsymbol{A}_{ki}^{1/2}(\tilde{\boldsymbol{\beta}}_{\mathcal{F}})(\boldsymbol{Z}_{ki} - \boldsymbol{U}_{ki}\boldsymbol{\omega})\|^2 \Big\} + n\lambda_2 \|\boldsymbol{\omega}\|_{2,1},$$

where $\|\boldsymbol{\omega}\|_{2,1}$ is the sum of $L_2$-norms of rows of $\boldsymbol{\omega}$ and $\lambda_2$ is a penalization parameter. The computation of $\tilde{\boldsymbol{\beta}}_{\mathcal{F}}$, $\tilde{\boldsymbol{R}}_{\mathcal{F}}$, and $\tilde{\boldsymbol{W}}_{\mathcal{F}}$ can be done by R packages `glmnet` (Hastie, Qian and Tay, 2021), `PGEE` (Wang, Zhou and Qu, 2012) and `pqr` (Cheng et al., 2022).

Based on the full dataset $\mathcal{F}$, the quasi decorrelated score estimator $\hat{\boldsymbol{\theta}}_{\mathcal{F}}$ of the main parameter $\boldsymbol{\theta}$ is a solution to

$$\boldsymbol{g}(\boldsymbol{\vartheta}, \tilde{\boldsymbol{\gamma}}_{\mathcal{F}}, \tilde{\boldsymbol{W}}_{\mathcal{F}}, \tilde{\boldsymbol{R}}_{\mathcal{F}}) = \boldsymbol{0}, \qquad \boldsymbol{\vartheta} \in \mathbb{R}^q,$$

where $\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{W}, \boldsymbol{R})$ is $\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{R})$ in (7) with $\boldsymbol{\beta} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ and $\tilde{\boldsymbol{\gamma}}_{\mathcal{F}}$ is the vector of last $p - q$ components of $\tilde{\boldsymbol{\beta}}_{\mathcal{F}}$. Note that we adopt a pseudo-type score estimation in which $(\boldsymbol{\gamma}, \boldsymbol{W}, \boldsymbol{R})$ is substituted by $(\tilde{\boldsymbol{\gamma}}_{\mathcal{F}}, \tilde{\boldsymbol{W}}_{\mathcal{F}}, \tilde{\boldsymbol{R}}_{\mathcal{F}})$ prior to scoring. We do not re-estimate the nuisance parameter $\boldsymbol{\gamma}$ after we obtain $\hat{\boldsymbol{\theta}}_{\mathcal{F}}$. To see the effect of having orthogonality property (8), by Taylor's expansion, it is shown in the Supplementary Material that

$$\begin{aligned} \boldsymbol{0} &= \boldsymbol{g}(\boldsymbol{\beta}, \tilde{\boldsymbol{W}}_{\mathcal{F}}, \tilde{\boldsymbol{R}}_{\mathcal{F}}) + \nabla_{\boldsymbol{\theta}} \boldsymbol{g}(\boldsymbol{\beta}, \tilde{\boldsymbol{W}}_{\mathcal{F}}, \tilde{\boldsymbol{R}}_{\mathcal{F}})(\hat{\boldsymbol{\theta}}_{\mathcal{F}} - \boldsymbol{\theta}) \\ &\quad + \nabla_{\boldsymbol{\gamma}} \boldsymbol{g}(\boldsymbol{\beta}, \tilde{\boldsymbol{W}}_{\mathcal{F}}, \tilde{\boldsymbol{R}}_{\mathcal{F}})(\tilde{\boldsymbol{\gamma}}_{\mathcal{F}} - \boldsymbol{\gamma}) + \text{ a lower order term} \\ &= \boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{R}) + E\{\nabla_{\boldsymbol{\theta}} \boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{R})\}(\hat{\boldsymbol{\theta}}_{\mathcal{F}} - \boldsymbol{\theta}) + \text{a lower order term}, \end{aligned}$$

because of (8). This indicates that the estimation of high dimensional nuisance parameter $\boldsymbol{\gamma}$ does not affect the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{\mathcal{F}}$, an advantage of using quasi decorrelated score; see Figure S1 in the Supplementary Material for illustration.

When the full dataset $\mathcal{F}$ is not available, we apply the quasi decorrelated score approach together with distributed cluster subsampling developed in Section 2. Let $\mathcal{S} = \{(\boldsymbol{X}_{ki}^*, \boldsymbol{Y}_{ki}^*), i = 1, \ldots, r_k, k = 1, \ldots, K\}$ be a distributed cluster subsampe as described in Section 2, $\boldsymbol{A}_{ki}^{*1/2}$ and $\pi_{ki}^*$ be the same as those in (3),

$$\tilde{\boldsymbol{\beta}}_{\mathcal{S}} = \arg\min_{\boldsymbol{b}} \Big\{ \sum_{k=1}^{K} \sum_{i=1}^{r_k} \sum_{j=1}^{m_{ki}} \frac{\psi(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_{kij}^*) - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_{kij}^*\, y_{kij}^*}{r_k \pi_{ki}^*} \Big\} + n\lambda_1 \|\boldsymbol{b}\|_1, \tag{10}$$

$\tilde{\boldsymbol{R}}_{ki}^*$ be the correlation estimator based on $\mathcal{S}$ and $\tilde{\boldsymbol{\beta}}_{\mathcal{S}}$ in (10), using the method in Liang and Zeger (1986) (see Section 4.2), $(\boldsymbol{Z}_{ki}^*, \boldsymbol{U}_{ki}^*) = \boldsymbol{X}_{ki}^*$, and

$$\tilde{\boldsymbol{W}}_{s} = \arg\min_{\boldsymbol{\omega}} \Big\{ \sum_{k=1}^{K} \sum_{i=1}^{r_k} \frac{\|\tilde{\boldsymbol{R}}_{ki}^{*-1/2} \boldsymbol{A}_{ki}^{*1/2}(\tilde{\boldsymbol{\beta}}_{s})(\boldsymbol{Z}_{ki}^* - \boldsymbol{U}_{ki}^*\boldsymbol{\omega})\|^2}{r_k \pi_{ki}^*} \Big\} + n\lambda_2 \|\boldsymbol{\omega}\|_{2,1}.$$

We approximate quasi decorrelated score $\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{R})$ in (7) by the subsample and then estimate the main parameter $\boldsymbol{\theta}$ of interest by $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$, the quasi decorrelated score estimator based on the subsample $\mathcal{S}$, as a solution to

$$\sum_{k=1}^{K}\sum_{i=1}^{r_k}\frac{(\boldsymbol{Z}^*_{ki}-\boldsymbol{U}^*_{ki}\tilde{\boldsymbol{W}}_{\mathcal{S}})^{\mathrm{T}}\boldsymbol{A}^{*1/2}_{ki}(\boldsymbol{\vartheta},\tilde{\boldsymbol{\gamma}}_{\mathcal{S}})\tilde{\boldsymbol{R}}^{*-1}_{ki}\{\boldsymbol{A}^{*1/2}_{ki}(\boldsymbol{\vartheta},\tilde{\boldsymbol{\gamma}}_{\mathcal{S}})\}^{-1}\{\boldsymbol{Y}^*_{ki}-\boldsymbol{\mu}^*_{ki}(\boldsymbol{\vartheta},\tilde{\boldsymbol{\gamma}}_{\mathcal{S}})\}}{r_k\pi^*_{ki}}=\boldsymbol{0},$$

$\boldsymbol{\vartheta}\in\mathbb{R}^q$, where $\tilde{\boldsymbol{\gamma}}_{\mathcal{S}}$ containing the last $p-q$ components of $\tilde{\boldsymbol{\beta}}_{\mathcal{S}}$ is used to substituting for $\boldsymbol{\gamma}$.

To establish asymptotic properties for $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{F}}$, we need the following conditions for high dimensional $\boldsymbol{\beta}$.

(A.1$^\dagger$) The covariate values vary in a bounded set, $\sup_{t\geq 1}t^{-1}[E\{|y_{kij}-\dot{\psi}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_{kij})|^t\}]^{1/t}$ is bounded, and $\boldsymbol{J}_W=\sum_{k=1}^{K}\sum_{i=1}^{n_k}E\{(\boldsymbol{Z}_{ki}-\boldsymbol{U}_{ki}\boldsymbol{W})^{\mathrm{T}}\boldsymbol{A}^{1/2}_{ki}(\boldsymbol{\beta})\boldsymbol{R}^{-1}_{ki}\boldsymbol{A}^{1/2}_{ki}(\boldsymbol{\beta})(\boldsymbol{Z}_{ki}-\boldsymbol{U}_{ki}\boldsymbol{W})\}/n$ converges to a positive definite matrix.

(A.4) Elements of $\boldsymbol{\beta}$ and $\boldsymbol{W}$ vary in a bounded set, $\boldsymbol{\beta}$ and $\boldsymbol{W}$ are sparse in the sense that $r^{-1/2}\max(s_{\beta},s_{W})\log p\log r\to 0$, where $s_{\beta}$ is the number of nonzero components of $\boldsymbol{\beta}$ and $s_{W}$ is the number of nonzero rows of $\boldsymbol{W}$, and the penalization parameters $\lambda_1$ and $\lambda_2$ are chosen to be bounded by $r^{-1/2}\sqrt{\log p}$.

(A.5) For any set $\mathcal{J}\subset\{1,\ldots,p\}$ and any vector $\boldsymbol{v}$ belonging to the cone $\mathcal{C}(\mathcal{J},\alpha)=\{\boldsymbol{v}\in\mathbb{R}^p:\|\boldsymbol{v}_{\mathcal{J}^c}\|_1\leq\alpha\|\boldsymbol{v}_{\mathcal{J}}\|_1\}$, where $\boldsymbol{v}_{\mathcal{J}}$ is the vector containing components of $\boldsymbol{v}$ with indices in $\mathcal{J}$ and $\mathcal{J}^c$ is the complement of $\mathcal{J}$, there exists a constant $C>0$ such that

$$\inf_{\boldsymbol{0}\neq\boldsymbol{v}\in\mathcal{C}(\mathcal{J},\alpha)}\sum_{k=1}^{K}\sum_{i=1}^{r_k}\sum_{j=1}^{m_{ki}}\frac{(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}^*_{kij})^2\ddot{\psi}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}^*_{kij})}{nr_k\pi^*_{ki}\|\boldsymbol{v}\|^2}\geq C.$$

Assumption (A.1$^\dagger$) is a version of Assumption (A.1) for high dimensional $\boldsymbol{\beta}$, which is stronger than Assumption (A.1) but standard for analyzing high-dimensional generalized linear models (Wang, Zhou and Qu, 2012; Fang, Ning and Li, 2020). The sparsity condition for $\boldsymbol{\beta}$ and $\boldsymbol{W}$ in Assumption (A.4) is assumed in Ning and Liu (2017) and Cheng et al. (2022). Assumption (A.5) is for lasso type estimation, which gives a restricted eigenvalue condition (Raskutti, Wainwright and Yu, 2010; Fang, Ning and Li, 2020; Cheng et al., 2022) and the necessary curvature within a cone.

THEOREM 3.1. *Suppose that Assumptions (A.1$^\dagger$) and (A.2)-(A.5) hold.*
*(i) With probability tending to* 1*, for any $\epsilon>0$, there exist fixed $\Delta_\epsilon$ and $r_\epsilon$ such that for all $r>r_\epsilon$, $P\big(\|\hat{\boldsymbol{\theta}}_{\mathcal{S}}-\boldsymbol{\theta}\|^2\geq\Delta_\epsilon/r\big)<\epsilon$, where $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ is the proposed quasi decorrelated score estimators based on the subsample $\mathcal{S}$.*
*(ii) If, in addition, $r/n\to 0$, then*

$$(\boldsymbol{J}_W^{-1}\boldsymbol{V}_W\boldsymbol{J}_W^{-1})^{-1/2}(\hat{\boldsymbol{\theta}}_{\mathcal{S}}-\boldsymbol{\theta})\xrightarrow{d}N(\boldsymbol{0},\boldsymbol{I}_q),$$

*where $\boldsymbol{J}_W$ is given in Assumption (A.1$^\dagger$) and*

$$\boldsymbol{V}_W=\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{\big[(\boldsymbol{Z}_{ki}-\boldsymbol{U}_{ki}\boldsymbol{W})^{\mathrm{T}}\boldsymbol{A}^{1/2}_{ki}(\boldsymbol{\beta})\boldsymbol{R}^{-1}_{ki}\{\boldsymbol{A}^{1/2}_{ki}(\boldsymbol{\beta})\}^{-1}\{\boldsymbol{Y}_{ki}-\boldsymbol{\mu}_{ki}(\boldsymbol{\beta})\}\big]^{\otimes 2}}{r_k\pi_{ki}n^2}.$$

Theorem 3.1(ii) actually holds with $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ replaced by $\hat{\boldsymbol{\theta}}_{\mathcal{F}}$ when the full dataset $\mathcal{F}$ is available, with $r$ changed to $n$ in Assumption (A.4), function related with subsampling changed to its conditional expectation in Assumption (A.5), and $r_k\pi_{ki}$ in $\boldsymbol{V}_W$ changed to 1. Assumptions and proofs based on the full dataset are given in the Supplementary Material.

It is of interest to compare the asymptotic covariance matrices of $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$, i.e., $\boldsymbol{J}^{-1}\boldsymbol{V}\boldsymbol{J}^{-1}$ in Theorem 2.1 and $\boldsymbol{J}_W^{-1}\boldsymbol{V}_W\boldsymbol{J}_W^{-1}$ in Theorem 3.1, when $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ follows (4). To proceed, we block $\boldsymbol{J}$ and $\boldsymbol{V}$ in Theorem 2.1 as

$$\boldsymbol{J} = \begin{pmatrix} \boldsymbol{J}_{ZZ} & \boldsymbol{J}_{ZU} \\ \boldsymbol{J}_{ZU}^{\mathrm{T}} & \boldsymbol{J}_{UU} \end{pmatrix} \quad \text{and} \quad \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{ZZ} & \boldsymbol{V}_{ZU} \\ \boldsymbol{V}_{ZU}^{\mathrm{T}} & \boldsymbol{V}_{UU} \end{pmatrix},$$

where

$$\boldsymbol{J}_{ZU} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} E\{\boldsymbol{Z}_{ki}^{\mathrm{T}}\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta})\boldsymbol{R}_{ki}^{-1}\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta})\boldsymbol{U}_{ki}\},$$

$$\boldsymbol{V}_{ZU} = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \frac{\boldsymbol{Z}_{ki}^{\mathrm{T}}\big[\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta})\boldsymbol{R}_{ki}^{-1}\{\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta})\}^{-1}\{\boldsymbol{Y}_{ki}-\boldsymbol{\mu}_{ki}(\boldsymbol{\beta})\}\big]^{\otimes 2}\boldsymbol{U}_{ki}}{r_k\pi_{ki}n^2},$$

$\boldsymbol{J}_{ZZ}$ and $\boldsymbol{V}_{ZZ}$ are $\boldsymbol{J}_{ZU}$ and $\boldsymbol{V}_{ZU}$, respectively, with $\boldsymbol{U}_{ki}$ replaced by $\boldsymbol{Z}_{ki}$, and $\boldsymbol{J}_{UU}$ and $\boldsymbol{V}_{UU}$ are $\boldsymbol{J}_{ZU}$ and $\boldsymbol{V}_{ZU}$, respectively, with $\boldsymbol{Z}_{ki}$ replaced by $\boldsymbol{U}_{ki}$. Applying the inverse of a block matrix and using the fact that $\boldsymbol{W} = \boldsymbol{J}_{UU}^{-1}\boldsymbol{J}_{ZU}^{\mathrm{T}}$ by (9), we obtain that $\boldsymbol{J}_W$ in Theorem 3.1 is equal to $\boldsymbol{J}_{ZZ} - \boldsymbol{J}_{ZU}\boldsymbol{J}_{UU}^{-1}\boldsymbol{J}_{ZU}^{\mathrm{T}}$ and

$$\boldsymbol{J}^{-1} = \begin{pmatrix} \boldsymbol{J}_W^{-1} & -\boldsymbol{J}_W^{-1}\boldsymbol{J}_{ZU}\boldsymbol{J}_{UU}^{-1} \\ -\boldsymbol{J}_{UU}^{-1}\boldsymbol{J}_{ZU}^{\mathrm{T}}\boldsymbol{J}_W^{-1} & \boldsymbol{J}_{UU}^{-1}\boldsymbol{J}_{ZU}^{\mathrm{T}}\boldsymbol{J}_W^{-1}\boldsymbol{J}_{ZU}\boldsymbol{J}_{UU}^{-1} + \boldsymbol{J}_{UU}^{-1} \end{pmatrix}.$$

Consequently, the first $q \times q$ sub-matrix of $\boldsymbol{J}^{-1}\boldsymbol{V}\boldsymbol{J}^{-1}$, which is the asymptotic covariance matrix of the first $q$ elements of $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ as estimators of elements in $\boldsymbol{\theta}$, is equal to

$$\boldsymbol{J}_W^{-1}(\boldsymbol{V}_{ZZ} - \boldsymbol{V}_{ZU}\boldsymbol{J}_{UU}^{-1}\boldsymbol{J}_{ZU}^{\mathrm{T}} - \boldsymbol{J}_{ZU}\boldsymbol{J}_{UU}^{-1}\boldsymbol{V}_{ZU}^{\mathrm{T}} + \boldsymbol{J}_{ZU}\boldsymbol{J}_{UU}^{-1}\boldsymbol{V}_{UU}\boldsymbol{J}_{UU}^{-1}\boldsymbol{J}_{ZU}^{\mathrm{T}})\boldsymbol{J}_W^{-1} = \boldsymbol{J}_W^{-1}\boldsymbol{V}_W\boldsymbol{J}_W^{-1},$$

by the definition of $\boldsymbol{V}_W$ in Theorem 3.1 and the fact that $\boldsymbol{W} = \boldsymbol{J}_{UU}^{-1}\boldsymbol{J}_{ZU}^{\mathrm{T}}$. In other words, the asymptotic efficiency of $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ is the same as that of the first $q$ elements of $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ when $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ follows (4), not affected by the estimation of $\boldsymbol{\gamma}$. The advantage of quasi decorrelated score appears when $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ does not perform well (i.e., result (4) does not hold) due to high dimensionality of $\boldsymbol{\beta}$, in which case $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ is still asymptotically normal with the same efficiency as that of $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ for low dimensional $\boldsymbol{\beta}$. The same conclusion can be obtained when we compare $\hat{\boldsymbol{\beta}}_{\mathcal{F}}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{F}}$.

Similar to Theorem 2.2 in Section 2, we have the following result for the optimal distributed cluster subsampling when quasi decorrelated score is applied. Let $\boldsymbol{L}$ be a known, fixed, and nonsingular $s \times q$ matrix.

THEOREM 3.2. *Suppose that Assumptions (A.1$^\dagger$) and (A.2)-(A.5) hold. For all $\pi_{ki}$'s and $r_k$'s under the constraint that $\sum_{i=1}^{n_k}\pi_{ki} = 1$ and $\sum_{k=1}^{K} r_k = r$ with a pre-chosen $r$ satisfying $r/n \to 0$, if the distributed cluster subsampling probability for selecting cluster $(k,i)$ is $\pi_{ki}^L = \tau_{ki}/\big(\sum_{i=1}^{n_k}\tau_{ki}\big)$ and subsample size in site $k$ is $r_k^L = r\big(\sum_{i=1}^{n_k}\tau_{ki}\big)/\big(\sum_{l=1}^{K}\sum_{i=1}^{n_l}\tau_{li}\big)$ with*

$$\tau_{ki} = \|\boldsymbol{L}\boldsymbol{J}_W^{-1}(\boldsymbol{Z}_{ki} - \boldsymbol{U}_{ki}\boldsymbol{W})^{\mathrm{T}}\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta})\boldsymbol{R}_{ki}^{-1}\{\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta})\}^{-1}\{\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_{ki}(\boldsymbol{\beta})\}\|,$$

*$i = 1,\ldots,n_k, k = 1,\ldots,K$, then $\mathrm{trace}(\boldsymbol{L}\boldsymbol{J}_W^{-1}\boldsymbol{V}_W\boldsymbol{J}_W^{-1}\boldsymbol{L}^{\mathrm{T}})$, the trace of asymptotic covariance matrix of $\boldsymbol{L}\hat{\boldsymbol{\theta}}_{\mathcal{S}}$, attains its minimum.*

Similarly, the optimal $\tau_{ki}$ in Theorem 3.2 has to be estimated by using an initial distributed cluster subsample as discussed in Section 2. After $\tau_{ki}$ is estimated, we draw a distributed cluster subsample $\mathcal{S}$ with estimated optimal probabilities and sizes, independent of the initial subsample, and then calculate $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ based on $\mathcal{S}$. It is shown in the Supplementary Material that $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ based on this procedure has the asymptotic distribution given in Theorem 3.1(ii) with

minimized trace of $\boldsymbol{L}\boldsymbol{J}_W^{-1}\boldsymbol{V}_W\boldsymbol{J}_W^{-1}\boldsymbol{L}^{\mathrm{T}}$. The performance of this optimal distributed cluster subsampling procedure is checked by simulation in Section 4.2.

To conduct asymptotic inference on $\boldsymbol{\theta}$, we can apply Theorem 3.1(ii) with $\boldsymbol{J}_W^{-1}\boldsymbol{V}_W\boldsymbol{J}_W^{-1}$ estimated by the consistent estimator $\hat{\boldsymbol{J}}_W^{-1}\hat{\boldsymbol{V}}_W\hat{\boldsymbol{J}}_W^{-1}$, where

$$\hat{\boldsymbol{J}}_W = \sum_{k=1}^{K}\sum_{i=1}^{r_k} \frac{(\boldsymbol{Z}_{ki}^* - \boldsymbol{U}_{ki}^*\tilde{\boldsymbol{W}}_s)^{\mathrm{T}}\boldsymbol{A}_{ki}^{*1/2}(\tilde{\boldsymbol{\beta}}_s)\tilde{\boldsymbol{R}}_{ki}^{*-1}\boldsymbol{A}_{ki}^{*1/2}(\tilde{\boldsymbol{\beta}}_s)(\boldsymbol{Z}_{ki}^* - \boldsymbol{U}_{ki}^*\tilde{\boldsymbol{W}}_s)}{nr_k\pi_{ki}^*},$$

$$\hat{\boldsymbol{V}}_W = \sum_{k=1}^{K}\sum_{i=1}^{r_k} \frac{\left[(\boldsymbol{Z}_{ki}^* - \boldsymbol{U}_{ki}^*\tilde{\boldsymbol{W}}_s)^{\mathrm{T}}\boldsymbol{A}_{ki}^{*1/2}(\tilde{\boldsymbol{\beta}}_s)\tilde{\boldsymbol{R}}_{ki}^{*-1}\{\boldsymbol{A}_{ki}^{*1/2}(\tilde{\boldsymbol{\beta}}_s)\}^{-1}\{\boldsymbol{Y}_{ki}^* - \boldsymbol{\mu}_{ki}^*(\tilde{\boldsymbol{\beta}}_s)\}\right]^{\otimes 2}}{(nr_k\pi_{ki}^*)^2},$$

when distributed cluster subsample $\mathcal{S}$ is used, and $r_k\pi_{ki}^*$ in $\hat{\boldsymbol{J}}_W$, $(r_k\pi_{ki}^*)^2$ in $\hat{\boldsymbol{V}}_W$ should be replaced by 1 when $\mathcal{F}$ is used. The performance of $\hat{\boldsymbol{J}}_W^{-1}\hat{\boldsymbol{V}}_W\hat{\boldsymbol{J}}_W^{-1}$ is checked by simulation in Section 4.2. We summarize the aforementioned procedures in Algorithm S2 in the Supplementary Material.

**4. Simulation study.** We assess the performance of our methods using simulations from linear model. We also have similar simulation results under logistic and Poisson models given in the Supplementary Material. We consider $K = 5$, $m_{ki} = 3$ for all $k$ and $i$, $n_k = 5(k+1)\times 10^4$, $k = 1,\ldots,5$, and $n = \sum_k n_k = 10^6$.

4.1. *Distributed cluster subsampling.* We first consider $p = 7$ and quasi decorrelated score is not applied. The simulation studies distributed cluster subsampling in Section 2 with the uniform and the optimal subsampling schemes. We generate correlated data in cluster $(k,i)$ according to

$$\boldsymbol{Y}_{ki} \mid \boldsymbol{X}_{ki} \sim N(\boldsymbol{X}_{ki}\boldsymbol{\beta},\ \boldsymbol{\Sigma}_{y|\boldsymbol{x}}), \tag{11}$$

where $\boldsymbol{\beta} = (1,1,1,1,1,1,1)^{\mathrm{T}}$ is 7-dimensional and $\boldsymbol{\Sigma}_{y|\boldsymbol{x}}$ is $3\times 3$ compound symmetry matrix with diagonal elements equal to 1 and off-diagonal elements equal to 0.8. The first column of $\boldsymbol{X}_{ki}\in\mathbb{R}^{3\times 7}$ has all elements equal to 1, corresponding to an intercept effect. For all $(k,i)$, the three rows of $\boldsymbol{X}_{ki}$ are independent and identically distributed and each row of $\boldsymbol{X}_{ki}$ excluding the constant first element is generated from the 6-dimensional normal distribution $N(\boldsymbol{0},\boldsymbol{\Sigma}_{\boldsymbol{x}}/10)$, where $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ has $(t,s)$th element equal to $0.2^{|t-s|}$, $1\le t,s\le 6$.

We evaluate and compare the performance of $\hat{\boldsymbol{\beta}}_s$ with four different methods to obtain the distributed subsample.

(a) The proposed A-optimal distributed cluster subsampling in Theorem 2.2 ($\boldsymbol{L} = \boldsymbol{I}_p$).
(b) The proposed L-optimal distributed cluster subsampling in Theorem 2.2 ($\boldsymbol{L} = \boldsymbol{J}$).
(c) The uniform distributed cluster subsampling with proportional allocation.
(d) The A-optimal individual subsampling scheme proposed by Zhang and Wang (2021) and Zuo et al. (2021) that ignores the within cluster correlation.

Method (d) is considered to see the advantages of our proposed cluster subsampling. We consider $r = 300, 500, 700$ and 900 as the sizes of distributed cluster subsampling. For A- and L-optimal schemes, we use an independent initial subsample of size 300.

To check the effect of estimated correlation $\hat{\boldsymbol{R}}_{ki}^*$, we consider three working models for correlation structure, the independence model (IND), the first-order autoregressive model (AR), and the compound symmetry model (CS), where the CS model is correct and the IND and AR models are incorrect. Under the IND working model, $\hat{\boldsymbol{R}}_{ki}^* = \boldsymbol{I}_{m_{ki}}$. Under the CS

working model, all diagonal elements of $\hat{\boldsymbol{R}}^*_{ki}$ are equal to 1 and all off-diagonal elements of $\hat{\boldsymbol{R}}^*_{ki}$ are equal to

$$\frac{1}{\hat{\sigma}^2}\sum_{k=1}^{K}\sum_{i=1}^{r_k}\frac{1}{nr_k\pi^*_{ki}}\sum_{j\neq j'}\frac{v^*_{kij}v^*_{kij'}}{m_{ki}(m_{ki}-1)}, \quad v^*_{kij}=\frac{y^*_{kij}-\dot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathcal{S},\mathrm{IND}}\boldsymbol{x}^*_{kij})}{\{\ddot{\psi}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathcal{S},\mathrm{IND}}\boldsymbol{x}^*_{kij})\}^{1/2}},$$

where $\hat{\sigma}^2=\sum_{k=1}^{K}\sum_{i=1}^{r_k}\sum_{j=1}^{m_{ki}}(y^*_{kij}-\hat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathcal{S},\mathrm{IND}}\boldsymbol{x}^*_{kij})^2/(nr_k\pi^*_{ki}m_{ki})$ and $\hat{\boldsymbol{\beta}}_{\mathcal{S},\mathrm{IND}}$ is a solution of (3) under the IND working model with $\hat{\boldsymbol{R}}^*_{ki}=\boldsymbol{I}_{m_{ki}}$. Under the AR working model, the $(j,j')$th off-diagonal element of $\hat{\boldsymbol{R}}^*_{ki}$ is $\hat{\phi}^{|j-j'|}$, where

$$\hat{\phi}=\frac{1}{\hat{\sigma}^2}\sum_{k=1}^{K}\sum_{i=1}^{r_k}\frac{1}{nr_k\pi^*_{ki}}\sum_{|j-j'|=1}\frac{v^*_{kij}v^*_{kij'}}{2(m_{ki}-1)}.$$

We evaluate the performance of methods (a)-(d) in terms of the following quantities: (1) the average of 7 absolute values of biases of estimated $\boldsymbol{\beta}$-components; (2) the summation of 7 estimated variances of $\boldsymbol{\beta}$-components, which is equal to the trace of $\hat{\boldsymbol{J}}^{-1}\hat{\boldsymbol{V}}\hat{\boldsymbol{J}}^{-1}$ in Section 2; (3) the ratio of the average of standard error (se) using $\hat{\boldsymbol{J}}^{-1}\hat{\boldsymbol{V}}\hat{\boldsymbol{J}}^{-1}$ to the average of standard deviation (sd); (4) the average of 7 coverage probabilities of 95% asymptotic confidence intervals for $\boldsymbol{\beta}$-components. Based on 500 simulation replications, Figure 2 and Table 1 show the simulated results for these four quantities. The following are our findings.

(1) All estimators have small biases, i.e., the average of absolute values of biases are smaller than 4%. All estimators are robust against the misspecification of working correlation model.

(2) In terms of estimated variance, method (b) with L-optimal subsampling is worse than method (a), because the average of estimated variances is a measure in favor of the A-optimal scheme. Method (c) based on uniform subsampling is worse than methods (a) and (b), although it is simpler.

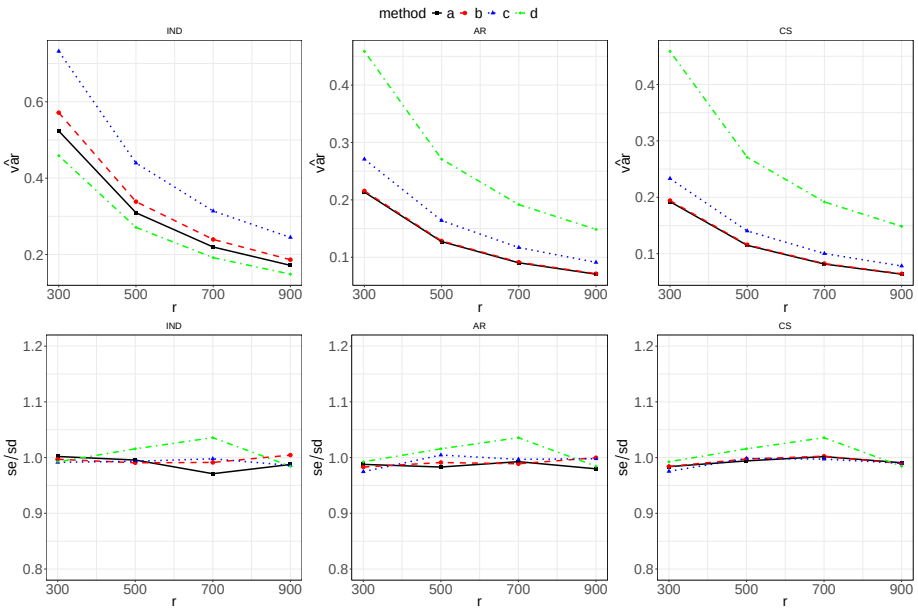



FIG 2. *Summation of estimated variances of $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$-components (*v̂ar*) and ratio of the average of standard error (*se*) to the average of standard deviation (*sd*) based on 500 simulations and low dimensional $\boldsymbol{\beta}$ setting of linear model in Section 4.1*

TABLE 1

*Average of absolute biases of $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$-components ($|\text{bias}|$) and average of coverage probabilities (CP) of $95\%$ asymptotic confidence intervals, based on $500$ simulations and low dimensional $\boldsymbol{\beta}$ setting in Section 4.1*

| | | | Results under linear model (11) | | | |
|---|---|---|---|---|---|---|
| | | | Method | | | |
| Correlation model | $r$ | Quantity | (a) | (b) | (c) | (d) |
| IND | 300 | $\|\text{bias}\|$ | 0.0354 | 0.0361 | 0.0321 | 0.0283 |
| | | CP | 0.9489 | 0.9454 | 0.9483 | 0.9446 |
| | 500 | $\|\text{bias}\|$ | 0.0316 | 0.0328 | 0.0318 | 0.0335 |
| | | CP | 0.9386 | 0.9446 | 0.9471 | 0.9471 |
| | 700 | $\|\text{bias}\|$ | 0.0295 | 0.0299 | 0.0346 | 0.0319 |
| | | CP | 0.9386 | 0.9426 | 0.9466 | 0.9457 |
| | 900 | $\|\text{bias}\|$ | 0.0273 | 0.0304 | 0.0292 | 0.0331 |
| | | CP | 0.9400 | 0.9429 | 0.9434 | 0.9383 |
| AR | 300 | $\|\text{bias}\|$ | 0.0094 | 0.0078 | 0.0111 | 0.0283 |
| | | CP | 0.9457 | 0.9446 | 0.9449 | 0.9446 |
| | 500 | $\|\text{bias}\|$ | 0.0122 | 0.0122 | 0.0156 | 0.0335 |
| | | CP | 0.9511 | 0.9497 | 0.9477 | 0.9471 |
| | 700 | $\|\text{bias}\|$ | 0.0107 | 0.0124 | 0.0133 | 0.0319 |
| | | CP | 0.9426 | 0.9411 | 0.9494 | 0.9457 |
| | 900 | $\|\text{bias}\|$ | 0.0081 | 0.0093 | 0.0088 | 0.0331 |
| | | CP | 0.9471 | 0.9489 | 0.9520 | 0.9383 |
| CS | 300 | $\|\text{bias}\|$ | 0.0156 | 0.0135 | 0.0133 | 0.0283 |
| | | CP | 0.9437 | 0.9463 | 0.9454 | 0.9446 |
| | 500 | $\|\text{bias}\|$ | 0.0117 | 0.0123 | 0.0159 | 0.0335 |
| | | CP | 0.9526 | 0.9520 | 0.9451 | 0.9471 |
| | 700 | $\|\text{bias}\|$ | 0.0137 | 0.0134 | 0.0153 | 0.0319 |
| | | CP | 0.9429 | 0.9417 | 0.9477 | 0.9457 |
| | 900 | $\|\text{bias}\|$ | 0.0131 | 0.0108 | 0.0103 | 0.0331 |
| | | CP | 0.9423 | 0.9511 | 0.9497 | 0.9383 |

(3) Under IND model, our proposed method (a) using A-optimal subsampling scheme can be slightly worse than method (d), which also uses A-optimal subsampling. This is because the cluster subsampling method using incorrect IND working model cannot take advantage of within-cluster correlation information. Under the AR or CS model, method (a) is much better than method (d), although the AR model is still incorrect. Under the AR model even method (c) is better than method (d).

(4) In terms of CP, all methods produce results close to $95\%$. Variance estimator $\hat{\boldsymbol{J}}^{-1}\hat{\boldsymbol{V}}\hat{\boldsymbol{J}}^{-1}$ for methods (a)-(c) generally perform well.

(5) As the CS model is correct, the estimators using CS as working model have smaller variances than those under wrong working models. Although the AR model is incorrect, it still leads to substantially lower variances compared with the IND working model. This indicates that it is important to model the correlation even though the working model may not be correct, rather than to ignore the correlation entirely.

The findings from simulation agree with our theoretical results in Theorems 2.1 and 2.2.

4.2. *Distributed cluster subsampling with quasi decorrelated score.* Our second simulation considers a $q = 4$ dimensional parameter $\boldsymbol{\theta}$ of interest and a $p - q = 900$ dimensional nuisance parameter $\boldsymbol{\gamma}$. Thus, distributed cluster subsampling is applied with the quasi decorrelated score approach in Section 3. We generate correlated data in cluster $(k, i)$ according to

$$\boldsymbol{Y}_{ki} \mid \boldsymbol{X}_{ki} \sim N\Big(\boldsymbol{Z}_{ki}\boldsymbol{\theta} + \boldsymbol{U}_{ki}\boldsymbol{\gamma},\ \boldsymbol{\Sigma}_{y|\boldsymbol{x}}\Big), \tag{12}$$

where $\boldsymbol{\theta} = (1, 1, 1, 1)^{\mathrm{T}}$, $\boldsymbol{\gamma}$ has first three elements equal to 1 and rest equal to 0, and $\boldsymbol{\Sigma}_{y|\boldsymbol{x}}$ is the same as that in Section 4.1. The first 4 columns of $\boldsymbol{X}_{ki} \in \mathbb{R}^{3\times 904}$ are columns of $\boldsymbol{Z}_{ki} \in \mathbb{R}^{3\times 4}$ and the rest $p - q = 900$ columns of $\boldsymbol{X}_{ki}$ are columns of $\boldsymbol{U}_{ki} \in \mathbb{R}^{3\times 900}$, where the first column of $\boldsymbol{U}_{ki}$ have all elements equal to 1, corresponding to an intercept effect. Each row of $\boldsymbol{X}_{ki}$ excluding the constant fifth element is generated from the 903-dimensional normal distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}}/10)$, where $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ has $(t, s)$th element equal to $0.2^{|t-s|}$, $1 \le t, s \le 903$.

We evaluate and compare the performance of $\hat{\boldsymbol{\theta}}_s$ based on the quasi decorrelated score approach and methods (a)-(c) in Section 4.1 to obtain the distributed cluster subsample, with $\boldsymbol{L} = \boldsymbol{I}_q$ in the A-optimal distributed cluster subsampling and $\boldsymbol{L} = \boldsymbol{J}_W$ in the L-optimal distributed cluster subsampling. In addition, we include an oracle method knowing that the last 897 components of $\boldsymbol{\gamma}$ are zeros so that estimation focuses on 7 coefficients including components of $\boldsymbol{\theta}$ without decorrelating, using A-optimal distributed cluster subsampling. This is called the oracle method and is added to see the effect of decorrelation in our proposed estimators.

We consider $r = 300, 500, 700$ and 900 as the sizes of distributed cluster subsampling. For A- and L-optimal schemes, we use an independent initial subsample of size 300. Estimation of the correlation matrix is carried out the same as in Section 4.1 under 3 working models, IND, AR, and CS, with $\hat{\boldsymbol{\beta}}_{s,\mathrm{IND}}$ replaced by lasso estimator $\tilde{\boldsymbol{\beta}}_s$ in (10) that is also based on IND working model. We evaluate the performance of methods (a)-(c) and the oracle method in terms of the following quantities: (1) the average of 4 absolute values of biases of estimated $\boldsymbol{\theta}$-components; (2) the summation of 4 estimated variances of $\boldsymbol{\theta}$-components, which is equal to the trace of $\hat{\boldsymbol{J}}_W^{-1}\hat{\boldsymbol{V}}_W\hat{\boldsymbol{J}}_W^{-1}$ in Section 3; (3) the ratio of the average of standard error using $\hat{\boldsymbol{J}}_W^{-1}\hat{\boldsymbol{V}}_W\hat{\boldsymbol{J}}_W^{-1}$ to the average of standard deviation; (4) the average of 4 coverage probabilities of 95% asymptotic confidence intervals for $\boldsymbol{\theta}$-components.

Based on 500 simulation replications, Figure 3 and Table 2 show the simulated results for these four quantities and we have the following findings.

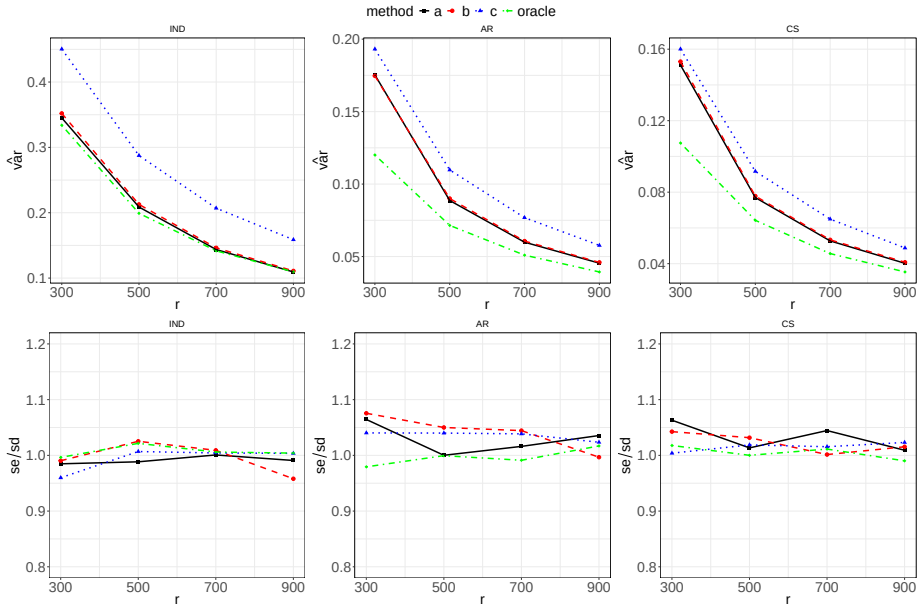



FIG 3. *Summation of estimated variances of $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$-components* (v̂ar) *and ratio of the average of standard error* (se) *to the average of standard deviation* (sd) *based on* 500 *simulations and high dimensional $\boldsymbol{\beta}$ setting of linear model in Section 4.2*

TABLE 2

*Average of absolute biases of $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$-components ($|\text{bias}|$) and average of coverage probabilities (CP) of 95% asymptotic confidence intervals based on 500 simulations and high dimensional $\boldsymbol{\beta}$ setting in Section 4.2*

| | | | Results under linear model (12) | | | |
|---|---|---|---|---|---|---|
| | | | Method | | | |
| Correlation model | $r$ | Quantity | (a) | (b) | (c) | oracle |
| IND | 300 | $\|\text{bias}\|$ | 0.0351 | 0.0364 | 0.0471 | 0.0076 |
| | | CP | 0.9460 | 0.9500 | 0.9330 | 0.9256 |
| | 500 | $\|\text{bias}\|$ | 0.0203 | 0.0354 | 0.0376 | 0.0075 |
| | | CP | 0.9470 | 0.9570 | 0.9510 | 0.9476 |
| | 700 | $\|\text{bias}\|$ | 0.0227 | 0.0311 | 0.0267 | 0.0026 |
| | | CP | 0.9420 | 0.9490 | 0.9510 | 0.9516 |
| | 900 | $\|\text{bias}\|$ | 0.0221 | 0.0237 | 0.0179 | 0.0054 |
| | | CP | 0.9455 | 0.9315 | 0.9540 | 0.9416 |
| AR | 300 | $\|\text{bias}\|$ | 0.0340 | 0.0326 | 0.0251 | 0.0040 |
| | | CP | 0.9505 | 0.9585 | 0.9520 | 0.9415 |
| | 500 | $\|\text{bias}\|$ | 0.0180 | 0.0251 | 0.0210 | 0.0025 |
| | | CP | 0.9505 | 0.9555 | 0.9535 | 0.9516 |
| | 700 | $\|\text{bias}\|$ | 0.0149 | 0.0195 | 0.0233 | 0.0023 |
| | | CP | 0.9455 | 0.9510 | 0.9505 | 0.9436 |
| | 900 | $\|\text{bias}\|$ | 0.0160 | 0.0202 | 0.0141 | 0.0034 |
| | | CP | 0.9540 | 0.9485 | 0.9470 | 0.9275 |
| CS | 300 | $\|\text{bias}\|$ | 0.0282 | 0.0307 | 0.0279 | 0.0070 |
| | | CP | 0.9525 | 0.9485 | 0.9480 | 0.9476 |
| | 500 | $\|\text{bias}\|$ | 0.0190 | 0.0237 | 0.0227 | 0.0020 |
| | | CP | 0.9450 | 0.9535 | 0.9510 | 0.9436 |
| | 700 | $\|\text{bias}\|$ | 0.0156 | 0.0225 | 0.0221 | 0.0056 |
| | | CP | 0.9540 | 0.9405 | 0.9465 | 0.9435 |
| | 900 | $\|\text{bias}\|$ | 0.0167 | 0.0204 | 0.0150 | 0.0026 |
| | | CP | 0.9505 | 0.9470 | 0.9480 | 0.9355 |

(1) All estimators have small biases. All estimators are robust against the misspecification of correlation model as the biases are all small.
(2) In terms of estimated variance, our proposed method (a) using decorrelation and A-optimal subsampling scheme can be slightly worse or better than the oracle method, which also uses A-optimal subsampling. This is because the oracle method aims at minimizing the asymptotic variance of $\boldsymbol{\beta}$ rather than $\boldsymbol{\theta}$, although it uses the information that $\boldsymbol{\beta}$ is 7-dimensional and doesn't estimate the high dimensional nuisance parameter. Method (b) with decorrelation and L-optimal subsampling is still worse than method (a). Method (c) based on decorrelation and uniform subsampling is the worst, although sampling is simple.
(3) In terms of CP, all methods produce results close to 95% except for a few cases where CP values are below 93%, even for the oracle method. Variance estimator $\hat{\boldsymbol{J}}_W^{-1}\hat{\boldsymbol{V}}_W\hat{\boldsymbol{J}}_W^{-1}$ for methods (a)-(c) generally perform well.
(4) The performance of estimators relative to the working models IND, AR, and CS are similar to that in Section 4.1.

The average computational time for the subsampling methods and full dataset are placed in the Supplementary Material.

**5. Analysis of Beijing multi-site air quality dataset.** We illustrate our proposed methods and analyze the effects of region and season variables $\boldsymbol{Z}_{ki}$ on PM2.5 emissions. Following the notation in Section 3, we are interested in estimation of a 6-dimensional $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)^{\mathrm{T}}$, where $\theta_1$, $\theta_2$, and $\theta_3$ are the effects for regions north, west, and east

versus baseline south, respectively, and $\theta_4$, $\theta_5$, and $\theta_6$ are the effects for seasons spring, summer, and autumn versus baseline winter, respectively. We consider the linear regression in (12) with $\boldsymbol{Y}_{ki}$ being the logarithm of PM2.5 emissions, $\boldsymbol{X}_{ki}\boldsymbol{\beta} = \boldsymbol{Z}_{ki}\boldsymbol{\theta} + \boldsymbol{U}_{ki}\boldsymbol{\gamma}$, and $\boldsymbol{U}_{ki}$ being an intercept term and 209 centralized and standardized extraneous covariates. In this application, the full dataset is available. In addition to the quasi decorrelated score with the full dataset, we also apply the method with distributed cluster subsampling as described in Section 3. This gives us an opportunity to check how quasi decorrelated score with subsampling preforms, compared with the full dataset analysis.

We consider an initial subsample of size $r^{\prime} = 100$ and four choices of $r = 100, 500, 1,000$, and $1,500$, which give $r/n = 0.14\%$, $0.71\%$, $1.43\%$, and $2.14\%$, respectively. Since data within each cluster is correlated as shown in Figure 1(b), we adopt AR and CS correlation models. Before data analysis, about $2\%$ of missing values are imputed by the before-and-after method in Fan, Lin and Yu (2024). In subsamples, the proportion of sampled imputed values is still around $2\%$ for any $r$ used. The results from statistical analysis depend on the imputation method, regardless of whether full data or subsamples are used. However, the effect is not serious since the proportion of missing values is small.

Table 3 presents the point estimates and standard errors (se's) of $\boldsymbol{\theta}$ using the full dataset and optimal subsample methods (a) and (b) described in Section 4.2 based on four choices of $r$. Estimates and se's based on uniform subsample (method (c)) are not shown in Table 3 but given in the Supplementary Material, since the simulation results in Section 4 show that uniform subsample is less efficient than methods (a) and (b). The estimates based on subsample with $r = 100$ can be very different from those based on $r =$ 500, 1,000, and 1,500, indicating that $r$ as small as 100 (only $0.14\%$ of the full dataset size and even smaller than the number of covariates $p = 216$) is not appropriate. For subsample with $r =$ 500, 1,000, 1,500, conclusions about significance of regional and seasonal covariates effects are the same as those of full dataset analysis, although magnitudes of estimates may not be always close. The following are our findings and conclusions from analysis of this dataset with any $r \geq 500$ or the full dataset.

The effects $\theta_1$, $\theta_2$, and $\theta_3$ are negative at significance level 0.05, although $\theta_3$ is only marginally significant from 0. From a regional perspective, it can be seen that PM2.5 emissions in region south during winter are more serious than those in regions north and west, but comparable to those in region east, which indicates an aggregation effect in south gradually decline from the heart of city towards countryside. An explanation to this phenomenon is that region south (e.g., Dongsi and Gucheng) has high population density, dense industries, and heavy vehicle emissions, while two mountains, Yanshan and Taihang, are around the north and west of Beijing such that the air pollutants emitted from region south are not easily diffused to regions north and west.

The effects $\theta_4$, $\theta_5$ and $\theta_6$ are significantly negative at level 0.05. Moreover, it can be seen that $|\hat{\theta}_5|$ is the largest, $|\hat{\theta}_6|$ is larger than $|\hat{\theta}_4|$, and their se's are comparable. From a seasonal perspective, the emission of PM2.5 in region south during winter is significantly higher compared to spring, summer and autumn. Our result also indicates that the emission of PM2.5 in region south is the highest in winter, lowest in summer, and much higher in spring compared with autumn, which shows a U-shaped variation. Some explanations are: Beijing citizens consume enormous amounts of coal for heating in cold and dry winters; the rainy summer season is beneficial to the diffusion and deposition of air pollutants; in spring, the air pollutants are generated by fireworks during the spring festival and construction activities; and in autumn, coal/biomass combustion are increased.

Our analysis confirms significant regional and seasonal effects in PM2.5 emissions. The significant increase in PM2.5 during winter in region south suggests stricter controls on coal/biomass combustion and industrial activities should be taken to reduce the emissions.

TABLE 3
*Point estimates and standard errors (se's) using full dataset and subsampling methods (a) and (b) in Section 5*

| Correlation model | | Method | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Full dataset | (a) | | | | (b) | | | |
| | | $n = 70,128$ | $r = 100$ | 500 | $1,000$ | $1,500$ | $r = 100$ | 500 | $1,000$ | $1,500$ |
| AR | $\hat{\theta}_1$ | -0.1912 | -0.4579 | -0.2650 | -0.2792 | -0.2663 | -0.5362 | -0.2893 | -0.2724 | -0.2913 |
| | se | 0.0118 | 0.2118 | 0.1084 | 0.0784 | 0.0671 | 0.2290 | 0.1257 | 0.0790 | 0.0734 |
| | $\hat{\theta}_2$ | -0.2219 | -0.4245 | -0.2672 | -0.2455 | -0.2384 | -0.4786 | -0.3190 | -0.3651 | -0.2799 |
| | se | 0.0087 | 0.1912 | 0.0991 | 0.0726 | 0.0626 | 0.1767 | 0.0931 | 0.0669 | 0.0656 |
| | $\hat{\theta}_3$ | -0.0570 | -0.2219 | -0.1801 | -0.1552 | -0.0698 | -0.2551 | -0.2274 | -0.1034 | -0.0813 |
| | se | 0.0114 | 0.2084 | 0.1184 | 0.0857 | 0.0707 | 0.2270 | 0.1243 | 0.1047 | 0.0698 |
| | $\hat{\theta}_4$ | -0.3420 | -0.3342 | -0.4466 | -0.6247 | -0.5499 | -0.4781 | -0.5657 | -0.6229 | -0.5181 |
| | se | 0.0081 | 0.1441 | 0.0822 | 0.0658 | 0.0555 | 0.1389 | 0.0817 | 0.0634 | 0.0521 |
| | $\hat{\theta}_5$ | -1.1883 | -1.1554 | -1.2916 | -1.3433 | -1.3100 | -1.0299 | -1.4079 | -1.4254 | -1.3591 |
| | se | 0.0105 | 0.1802 | 0.1061 | 0.0837 | 0.0660 | 0.1955 | 0.1114 | 0.0797 | 0.0677 |
| | $\hat{\theta}_6$ | -0.7198 | -0.5798 | -0.7157 | -0.8551 | -0.8229 | -0.6865 | -0.8050 | -0.8821 | -0.8819 |
| | se | 0.0091 | 0.1550 | 0.0837 | 0.0641 | 0.0562 | 0.1304 | 0.0753 | 0.0586 | 0.0502 |
| CS | $\hat{\theta}_1$ | -0.1458 | -0.4520 | -0.2697 | -0.2776 | -0.1850 | -0.6169 | -0.3104 | -0.2434 | -0.2474 |
| | se | 0.0118 | 0.2189 | 0.1088 | 0.0800 | 0.0688 | 0.2423 | 0.1193 | 0.0957 | 0.0773 |
| | $\hat{\theta}_2$ | -0.1806 | -0.4441 | -0.2627 | -0.2995 | -0.1847 | -0.5446 | -0.2896 | -0.2718 | -0.2521 |
| | se | 0.0087 | 0.1935 | 0.0981 | 0.0763 | 0.0634 | 0.1775 | 0.0936 | 0.0672 | 0.0593 |
| | $\hat{\theta}_3$ | -0.0197 | -0.2814 | -0.1785 | -0.0812 | -0.0520 | -0.2956 | -0.2232 | -0.0507 | -0.0526 |
| | se | 0.0115 | 0.2124 | 0.1110 | 0.0835 | 0.0689 | 0.2205 | 0.1274 | 0.0920 | 0.0837 |
| | $\hat{\theta}_4$ | -0.3282 | -0.3060 | -0.4376 | -0.4949 | -0.4168 | -0.4351 | -0.5021 | -0.5153 | -0.3978 |
| | se | 0.0085 | 0.1433 | 0.0818 | 0.0676 | 0.0621 | 0.1407 | 0.0801 | 0.0685 | 0.0616 |
| | $\hat{\theta}_5$ | -1.3755 | -1.2535 | -1.3933 | -1.4552 | -1.3894 | -1.0760 | -1.4424 | -1.5231 | -1.3670 |
| | se | 0.0112 | 0.1954 | 0.1082 | 0.0879 | 0.0772 | 0.1930 | 0.1137 | 0.0865 | 0.0822 |
| | $\hat{\theta}_6$ | -0.7542 | -0.5336 | -0.7021 | -0.7970 | -0.7454 | -0.7015 | -0.7354 | -0.8272 | -0.7790 |
| | se | 0.0094 | 0.1502 | 0.0832 | 0.0655 | 0.0617 | 0.1325 | 0.0759 | 0.0651 | 0.0608 |

To reduce the PM2.5 emissions in region south during winter, energy-saving policies and end-of-pipe control measures can be applied compared with the other regions. Our results help the development of more effective air quality regulations and pollution control strategies tailored to specific areas and times of the year and lead to more targeted public health interventions and policies to reduce exposure and associated health risks.

Some of our conclusions are similar to the reports in Fan, Lin and Yu (2024) and Li, Liu and Zhao (2022). But our results are more statistically reliable since decorrelated quasi score is applied to handle high dimensionality of covariate. Furthermore, our analysis based on subsamples reveals that subsampling with an appropriate $r$ (e.g., $r \geq 500$) can be used to achieve almost the same purpose as the full dataset analysis but saves amount of computation, although computational saving is not the only reason for distributed subsampling (other reasons are privacy, security, and administrative management as discussed in Section 1.2). The computational time for the full dataset is about 510 seconds, and for subsamples (including calculation of optimal subsampling probabilities and sampling) are about 8, 14, 20, and 28 seconds with $r = 100$, 500, 1,000, and 1,500, respectively, using R (version 4.1.0) based on Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz with memory 125 GB.

**6. Discussion.** To investigate regional and seasonal PM2.5 effects using the Beijing multi-site air quality dataset, we choose the baseline level as region south and winter. Similar analyses can be carried out by choosing other baselines and/or other covariates of interest. Other statistical methods can also be applied and studied; for example, the quadratic inference function (Qu, Lindsay and Li, 2000) instead of the quasi score equation, the longitudinal

quantile regression or generalized partial linear models instead of the linear GLM for mean effects of covariates, and Poisson subsampling (Yu et al., 2022) as a better way to deal with the memory constraint problem in distributed cluster subsampling. When there is a non-negligible proportion of missing values and outliers, methods of handling missing values and outliers in distributed subsampling and quasi decorrelaed score deserve further research.

**Acknowledgments.** The authors would like to thank the Editor, an Associate Editor and two anonymous referees for their constructive comments that improved the paper's quality. Lei Wang was supported by the National Natural Science Foundation of China (Grant No. 12271272).

## SUPPLEMENTARY MATERIAL

**Supplement to "Distributed subsampling and quasi decorrelated score for cluster data: an application to Beijing multi-site air quality"**. The Supplementary Material provides additional details, simulations, technical lemmas, and proofs for theorems.

**Code**. R code for the simulation study and real data analysis is provided.

## REFERENCES

AI, M., WANG, F., YU, J. and ZHANG, H. (2021a). Optimal subsampling for large-scale quantile regression. *Journal of Complexity* **62** 101512.

AI, M., YU, J., ZHANG, H. and WANG, H. (2021b). Optimal subsampling algorithms for big data regressions. *Statistica Sinica* **31** 749–772.

BALAN, R. M. and SCHIOPU-KRATINA, I. (2005). Asymptotic results with generalized estimating equations for longitudinal data. *The Annals of Statistics* **33** 522–541.

BLAZÈRE, M., LOUBES, J.-M. and GAMBOA, F. (2014). Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension. *IEEE Transactions on Information Theory* **60** 2303–2318.

CHENG, C., FENG, X., HUANG, J. and LIU, X. (2022). Regularized projection score estimation of treatment effects in high-dimensional quantile regression. *Statistica Sinica* **32** 23–41.

CHU, N., KADANE, J. B. and DAVIDSON, C. I. (2010). Using statistical regressions to identify factors influencing PM2.5 concentrations: the Pittsburgh supersite as a case study. *Aerosol Science and Technology* **44** 766–774.

FAN, Y., LIN, N. and YU, L. (2024). Distributed quantile regression for longitudinal big data. *Computational Statistics* **39** 751–779.

FANG, E. X., NING, Y. and LI, R. (2020). Test of significance for high-dimensional longitudinal data. *The Annals of Statistics* **48** 2622–2645.

GAO, J., WANG, L. and LIAN, H. (2024). Optimal decorrelated score subsampling for generalized linear models with massive data. *Science China Mathematics* **67** 405–430.

GAO, Y., LIU, W., WANG, H., WANG, X., YAN, Y. and ZHANG, R. (2022). A review of distributed statistical inference. *Statistical Theory and Related Fields* **6** 89–99.

HAN, Y., YU, J., ZHANG, N., MENG, C., MA, P., ZHONG, W. and ZOU, C. (2023). Leverage classifier: Another look at support vector machine. *Statistica Sinica* to appear.

HASTIE, T., QIAN, J. and TAY, K. (2021). An Introduction to glmnet. *CRAN R Repositary*.

HRIPCSAK, G., DUKE, J. D., SHAH, N. H., REICH, C. G., HUSER, V., SCHUEMIE, M. J., SUCHARD, M. A., PARK, R. W., WONG, I. C. K., RIJNBEEK, P. R. et al. (2015). Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Studies in Health Technology and Informatics* **216** 574–578.

LI, M., LI, R. and MA, Y. (2021). Inference in high dimensional linear measurement error models. *Journal of Multivariate Analysis* **184** 104759.

LI, D., LIU, J. and ZHAO, Y. (2022). Forecasting of PM2. 5 concentration in Beijing using hybrid deep learning framework based on attention mechanism. *Applied Sciences* **12** 11155.

LI, T. and MENG, C. (2020). Modern subsampling methods for large-scale least squares regression. *International Journal of Cyber-Physical Systems* **2** 1–28.

LI, X., GAO, Y., CHANG, H., HUANG, D., MA, Y., PAN, R., QI, H., WANG, F., WU, S., XU, K. et al. (2024). A selective review on statistical methods for massive data computation: distributed computing, subsampling, and minibatch techniques. *Statistical Theory and Related Fields* **8** 163–185.

LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

MA, P., MAHONEY, M. W. and YU, B. (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research* **16** 861–911.

MA, P., CHEN, Y., ZHANG, X., XING, X., MA, J. and MAHONEY, M. (2022). Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. *Journal of Machine Learning Research* **23** 1–45.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

MENG, C., XIE, R., MANDAL, A., ZHANG, X., ZHONG, W. and MA, P. (2021). LowCon: a design-based subsampling approach in a misspecified linear model. *Journal of Computational and Graphical Statistics* **30** 694–708.

NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* **45** 158–195.

QU, A., LINDSAY, B. G. and LI, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87** 823–836.

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research* **11** 2241–2259.

WANG, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics* **39** 389–417.

WANG, H. and MA, Y. (2021). Optimal subsampling for quantile regression in big data. *Biometrika* **108** 99–112.

WANG, L., ZHOU, J. and QU, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68** 353–360.

WANG, H., ZHU, R. and MA, P. (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association* **113** 829–844.

WU, X., HUO, Y., REN, H. and ZOU, C. (2024). Optimal subsampling via predictive inference. *Journal of the American Statistical Association* **119** 2844–2856.

XIE, M. and YANG, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *The Annals of Statistics* **31** 310–347.

YAN, R., LIAO, J., YANG, J., SUN, W., NONG, M. and LI, F. (2021). Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. *Expert Systems with Applications* **169** 114513.

YAO, Y. and WANG, H. (2021). A review on optimal subsampling methods for massive datasets. *Journal of Data Science* **19** 151–172.

YE, Z., YU, J. and AI, M. (2025). Optimal subsampling for multinomial logistic models with big data. *Statistica Sinica* **35** 1–25.

YU, J., AI, M. and YE, Z. (2024). A review on design inspired subsampling for big data. *Statistical Papers* **65** 467–510.

YU, J., LIU, J. and WANG, H. (2023). Information-based optimal subdata selection for non-linear models. *Statistical Papers* **64** 1069–1093.

YU, J., WANG, H., AI, M. and ZHANG, H. (2022). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association* **117** 265–276.

ZHANG, H. and JIA, J. (2022). Elastic-net regularized high-dimensional negative binomial regression: consistency and weak signal detection. *Statistica Sinica* **32** 181–207.

ZHANG, T., NING, Y. and RUPPERT, D. (2021). Optimal sampling for generalized linear models under measurement constraints. *Journal of Computational and Graphical Statistics* **30** 106–114.

ZHANG, H. and WANG, H. (2021). Distributed subdata selection for big data via sampling-based approach. *Computational Statistics and Data Analysis* **153** 107072.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 217–242.

ZHANG, X. Y., WANG, Y., NIU, T., ZHANG, X., GONG, S., ZHANG, Y. and SUN, J. (2012). Atmospheric aerosol compositions in China: spatial/temporal variability, chemical signature, regional haze distribution and comparisons with global aerosols. *Atmospheric Chemistry and Physics* **12** 779–799.

ZHANG, S., GUO, B., DONG, A., HE, J., XU, Z. and CHEN, S. X. (2017). Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **473** 20170457.

ZHANG, J., MENG, C., YU, J., ZHANG, M., ZHONG, W. and MA, P. (2023). An optimal transport approach for selecting a representative subsample with application in efficient kernel density estimation. *Journal of Computational and Graphical Statistics* **32** 329–339.

ZUO, L., ZHANG, H., WANG, H. and SUN, L. (2021). Optimal subsample selection for massive logistic regression with distributed data. *Computational Statistics* **36** 2535–2562.