IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 11, 2020, accepted January 23, 2020, date of publication January 31, 2020, date of current version March 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2970741

Large-Scale Expectile Regression With Covariates Missing at Random

YINGLI PAN, ZHAN LIU, AND WEN CAI
Hubei Key Laboratory of Applied Mathematics, School of Mathematics and Statistics, Hubei University, Wuhan 430062, China
Corresponding author: Zhan Liu (eleen_20040109@163.com)

This work was supported in part by the National Science Foundation of China under Grant 11901175, and in part by the National Social Science Found of China under Grant 18BTJ022.

ABSTRACT Analysis of large volumes of data is very complex due to not only a high level of skewness and heteroscedasticity of variance but also the phenomenon of missing data. Expectile regression is a popular alternative method of analyzing heterogeneous data. In this paper, we consider fitting a linear expectile regression model for estimating conditional expectiles based on a large quantity of data with covariates missing at random. We construct a communication-efficient surrogate loss (CSL) function to estimate model parameters. The asymptotic normality of the proposed estimator is established. A proximal alternating direction method of multipliers (ADMM) algorithm is developed for distributed statistical optimization on a large quantity of data. Simulation studies are performed to assess the finite-sample performance of the proposed method. Survey data from the Behavioral Risk Factor Surveillance System (BRFSS) is used to demonstrate the utility of the proposed method in practice.

INDEX TERMS CSL function, expectile regression, large-scale data, missing at random, proximal ADMM algorithm.

I. INTRODUCTION

Large-scale data, which arise in many fields such as online surveys, genomics and economics, are characterized by a high level of skewness, heteroscedasticity of variance and the phenomenon of missing information. When the size of the dataset becomes extremely large, it may be infeasible to store all of the data on a single machine. The process of statistical analysis of large-scale data must involve data transfers between different storage devices, which becomes a direct cause of a slowdown of computation. When information on covariates is collected, missing data may arise due to a lack of responses. These features pose great challenges for statistical analysis of large-scale data.

The impact of statistical procedures on missing data's processing and data transfers should be considered apart from the usual statistical inference criteria. In recent years, numerous studies on distributed approaches to large-scale statistical optimization problems have been performed [2], [4], [11], [22]. However, these families of distributed methods are characterized by communication complexity. For example, as different machines should be synchronized, communication can impose a significant overhead if the amount of data is very large. It makes sense to study distributed inference that only needs limited synchronization and communication while still enjoying the statistical power guaranteed by having a large-scale data set.

Most of the existing studies of large-scale data analysis have focused on the classic least-squares regression. Least-squares estimates are optimal if the errors are independent, identically distributed and normal. However, uncontrolled inhomogeneity of variance among random errors and genuinely long-tailed error distributions have indistinguishable effects, and may reduce the efficiency of least-squares estimates. Thus, robust alternatives to the least-squares method are definitely needed. Expectile regression, first proposed by Newey and Powell [13] to analyze heterogeneous data, is defined by the asymmetric quadratic loss based on the l_2 norm as follows:

$$\rho_\tau(u) = |\tau - I(u \leq 0)|\, u^2, \quad \text{for } \tau \in (0, 1) \qquad (1.1)$$

where $I(\cdot)$ is the indicator function. Following the landmark paper of Newey and Powell [13], numerous and extensive studies of expectile regression have been

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Asaduzzaman.

performed [16], [25]. Specifically, two expectile regression packages, expectreg [17] and ER-Boost [21], have been made available. Zhao and Zhang [23] further indicate that sample expectiles provide a class of smooth curves as functions of level τ and are robust for heavy-tailed distributions.

In this paper, we study a distributed optimization approach to analyzing large-scale data based on expectile regression with covariates missing at random. The simplest method of handling missing data is complete-case analysis, which deletes all incomplete data. However, Little and Rubin [12] pointed out that doing so may cause biased estimation if the data are not missing at random. The other method of dealing with missing data is imputation [3], [20], [24]. For example, Robins *et al.* [15] studied a new class of consistent weighted semiparametric estimators in the case of data missing at random. Wang *et al.* [18] assessed the statistical performance of a Horvitz and Thompson [8]-type weighted estimator by applying different estimates of selection probabilities.

Applying the inverse probability-weighted method, we define a weighted global loss function and several weighted local loss functions. Based on these functions, we construct a communication-efficient surrogate loss (CSL) function of the weighted global loss function inspired by the idea of Jordan *et al.* [10]. The CSL function can be regarded as a communication-efficient surrogate for the weighted global loss function, and can effectively solve the problems caused by large-scale data stored randomly on multiple machines. During the calculation, our proposed method can be applied on the master machine in each round, and other worker machines only need to compute the gradient based on local data. As to communication, our proposed method based on CSL matches the estimation error bound of the oracle method within a certain number of communications compared to the weighted estimator obtained by the oracle method of using all data simultaneously.

We develop an inference procedure for regression parameters of the linear expectile regression model based on the CSL function. To establish the asymptotic properties of the proposed estimator, we apply the distributed optimization theory [10] and the Lindeberg-Feller central limit theorem. Another challenge arises from the numerical calculation of the proposed estimator. We use an alternating direction method of multipliers (ADMM) algorithm [7], [9]. Boyd [1] showed that an ADMM algorithm is appropriate for large-scale statistical inference and distributed convex optimization problems. Combining the advantages of the CSL function with the ADMM algorithm, we explore a proximal ADMM algorithm for the calculation of our proposed estimator.

The rest of this paper is organized as follows. In Section 2, we fit a linear expectile regression model to distributed data with covariates missing at random. In Section 3, we develop asymptotic properties of the proposed estimation. In Section 4, we propose a proximal ADMM algorithm for the implementation of the proposed estimation. We perform simulation studies to evaluate the finite-sample performance of the proposed method in Section 5. We analyze a dataset from the Behavioral Risk Factor Surveillance System (BRFSS) in Section 6 and provide some concluding remarks in Section 7. The proofs of asymptotic properties are given in the Appendix.

II. DESIGN AND ESTIMATION

Let $\{(X_j, Y_j)_{j=1}^N\} \triangleq \{(X_{ki}, Y_{ki}) : k = 1, 2, \cdots, K, i = 1, 2, \cdots, n\}$ denote $N = nK$ independent identically distributed observations. Suppose that data storage is distributed so that each machine stores a subsample of n observations. Let $\{(X_{ki}, Y_{ki}) : i = 1, 2, \cdots, n\}$ denote the subsample stored on the kth machine $\mathcal{M}_k$ for $k = 1, 2, \cdots, K$. We consider the following linear expectile model:

$$Y_{ki} = X_{ki}^{\mathrm{T}}\beta(\tau) + \epsilon_{ki}(\tau), \quad k = 1, 2, \cdots, K, \ i = 1, 2, \cdots, n, \tag{2.1}$$

where $\beta(\tau)$ is $(p+q)$-dimensional vector of regression parameters of interest, and $\epsilon_{ki}(\tau)$ are random errors with the τth expectile given X_{ki} equal to zero for $\tau \in (0, 1)$. We drop τ in the parameter and error terms for notation simplicity in the following.

If covariate information can be observed for each individual, the following global loss function is widely used for the inference of parameter β [13]:

$$L_N(\beta) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_\tau(Y_{ki} - X_{ki}^{\mathrm{T}}\beta), \tag{2.2}$$

where $\rho_\tau(u)$ is defined by (1.1). Regression parameter β could be estimated by solving the estimation equation $U_N(\beta) = 0$, where

$$U_N(\beta) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}X_{ki}\psi_\tau(Y_{ki} - X_{ki}^{\mathrm{T}}\beta) \tag{2.3}$$

and $\psi_\tau(u) = 2|\tau - I(u \le 0)|u$ is the gradient function of $\rho_\tau(u)$. Note that if we denote the local loss function by $L_k(\beta) = n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_{ki} - X_{ki}^{\mathrm{T}}\beta)$ for $k = 1, 2, \cdots, K$, we have $L_N(\beta) = K^{-1}\sum_{k=1}^{K}L_k(\beta)$.

In many practical applications, the modeling process in statistical inference often encounters missing covariates. For $k = 1, 2, \cdots, K, i = 1, 2, \cdots, n$, we observe a response variable Y_{ki}, a p-dimensional vector W_{ki} of covariates that is always observed, and a vector T_{ki} of covariates of dimension q that may contain some missing components. Let $X_{ki} = (W_{ki}^{\mathrm{T}}, T_{ki}^{\mathrm{T}})^{\mathrm{T}}$ be the $(p+q)$-dimensional vector of covariates. Let R_{ki} denote the indicator variable: $R_{ki} = 1$ if T_{ki} is fully observed, and $R_{ki} = 0$ otherwise. We assume that values of T_{ki} are missing at random, i.e. $P(R_{ki} = 1|Y_{ki}, X_{ki}) = P(R_{ki} = 1|Y_{ki}, W_{ki})$. Additionally, we assume that for an unknown γ and $V_{ki} = (Y_{ki}, W_{ki}^{\mathrm{T}}) \in \mathbb{R}^{(p+1)}$, we establish the logistic regression $P(R_{ki} = 1|Y_{ki}, X_{ki}) = e^{V_{ki}^{\mathrm{T}}\gamma}/(1 + e^{V_{ki}^{\mathrm{T}}\gamma}) \triangleq \pi(V_{ki}, \gamma)$.

To account for the missing data problem, we consider the following inverse probability weights

$$w_{ki} = \frac{1}{\pi(V_{ki}, \widehat{\gamma})}, \quad k = 1, 2, \cdots, K, \ i = 1, 2, \cdots, n, \quad (2.4)$$

where $\widehat{\gamma}$ is the estimator of γ based on the logistic regression model. Using the weights proposed in (2.4), we obtain the following weighted estimation equation:

$$U_w(\beta) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ki} X_{ki} \psi_\tau(Y_{ki} - X_{ki}^{\mathrm{T}} \beta). \quad (2.5)$$

Inspired by the relationship between (2.2) and (2.3), we construct the following weighted global loss function $\widetilde{L}_N(\beta)$ and weighted local loss functions $\widetilde{L}_k(\beta)$:

$$\widetilde{L}_N(\beta) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ki} \rho_\tau(Y_{ki} - X_{ki}^{\mathrm{T}} \beta), \quad (2.6)$$

$$\widetilde{L}_k(\beta) = \frac{1}{n} \sum_{i=1}^{n} w_{ki} \rho_\tau(Y_{ki} - X_{ki}^{\mathrm{T}} \beta), \quad k = 1, 2, \cdots, K. \quad (2.7)$$

The weighted expectile regression estimator is defined as

$$\widetilde{\beta} = \arg\min_{\beta \in \mathcal{B}} \widetilde{L}_N(\beta), \quad (2.8)$$

where $\mathcal{B}$ is the parameter space. The optimization process (2.8) achieves the optimal statistical error. However, large-scale data is usually stored on multiple machines, so a statistical analysis involving the entirety of data must entail data transfers between different machines. Optimizing $\widetilde{L}_N(\beta)$ directly is infeasible under the distributed storage framework.

In this paper, inspired by the idea of communication-efficient surrogate likelihood method [10], we define a CSL function that approximates the weighted global loss function $\widetilde{L}_N(\beta)$ as follows:

$$\widetilde{L}(\beta) = \widetilde{L}_1(\beta) + \left\langle \nabla \widetilde{L}_N(\overline{\beta}) - \nabla \widetilde{L}_1(\overline{\beta}), \beta \right\rangle. \quad (2.9)$$

Here, $\overline{\beta}$ is any initial estimator of β; $\langle \cdot, \cdot \rangle$ denotes the inner product, and ∇ denotes the partial derivative with respect to β. In the distributed learning setting, the proposed estimator using the CSL function as the objective optimization function is defined by

$$\widehat{\beta} = \arg\min_{\beta \in \mathcal{B}} \widetilde{L}(\beta). \quad (2.10)$$

III. ASYMPTOTIC PROPERTIES

In this section, we derive asymptotic properties of the proposed estimator $\widehat{\beta}$ in (2.10). To present the asymptotic results, we introduce some notation. Let β_0 denote the true value of β, and define

$$\epsilon_i = Y_i - X_i^{\mathrm{T}} \beta_0, \quad \overline{\epsilon}_i = Y_i - X_i^{\mathrm{T}} \overline{\beta},$$

$$\Sigma_0 = \mathrm{E}\left[X_i X_i^{\mathrm{T}} \frac{1}{\pi(V_i, \gamma)} \psi_\tau^2(\epsilon_i) \right],$$

$$\overline{\Sigma}_0 = \mathrm{E}\left[X_i X_i^{\mathrm{T}} \frac{1}{\pi(V_i, \gamma)} \psi_\tau^2(\overline{\epsilon}_i) \right],$$

$$\overline{\overline{\Sigma}}_0 = \mathrm{E}\left[X_i X_i^{\mathrm{T}} \frac{1}{\pi(V_i, \gamma)} \psi_\tau(\epsilon_i) \psi_\tau(\overline{\epsilon}_i) \right],$$

$$\Sigma_2 = \mathrm{E}\left[(1 - \pi(V_i, \gamma)) V_i X_i^{\mathrm{T}} \psi_\tau(\epsilon_i) \right],$$

$$\overline{\Sigma}_2 = \mathrm{E}\left[(1 - \pi(V_i, \gamma)) V_i X_i^{\mathrm{T}} \psi_\tau(\overline{\epsilon}_i) \right],$$

$$I(\gamma) = \mathrm{E}\left[V_i V_i^{\mathrm{T}} \pi(V_i, \gamma) (1 - \pi(V_i, \gamma)) \right].$$

We impose the following regularity conditions throughout this paper.

(C1) Parameter space $\mathcal{B}$ is compact. For any i, there exists a compact set $\mathcal{X}$ such that $X_i \in \mathcal{X} \subset \mathbb{R}^{p+q}$.

(C2) Regression errors $\{\epsilon_{ki}\}$ are independent and identically distributed with cumulative distribution function $F(\cdot)$. Furthermore, the τ-expectiles of $\{\epsilon_{ki}\}$ are zero and $\mathrm{E}[\epsilon_{ki}^2 | X_{ki}] < \infty$.

(C3) The MLE $\widehat{\gamma}$ of γ satisfies the regularity conditions of asymptotic normality of MLEs for exponential family models.

(C4) There exists $\alpha > 0$ such that $\pi(V_i, \gamma) > \alpha$ uniformly in i.

(C5) The matrixes Σ_0, $\overline{\Sigma}_0$ and $I(\gamma)$ are positive definite. There exists a positive definite matrix Σ_1 such that $\lim_{N \longrightarrow \infty} \left[N^{-1} \sum_{i=1}^{N} X_i X_i^{\mathrm{T}} \right] = \Sigma_1$.

The asymptotic properties of the proposed estimator $\widehat{\beta}$ are summarized in the following theorem with the proof given in the Appendix.

Theorem 1: If Conditions (C1)-(C5) are satisfied, then

$$\sqrt{N}(\widehat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \frac{1}{4g^2(\tau)} \Sigma_1^{-1} \Sigma \Sigma_1^{-1}), \quad N \to \infty,$$

where $g(\tau) = (1 - \tau)F(0) + \tau(1 - F(0))$ *and*

$$\Sigma = K(\Sigma_0 - \Sigma_2^T I^{-1}(\gamma) \Sigma_2) + (K - 1)(\overline{\Sigma}_0 - \overline{\Sigma}_2^T I^{-1}(\gamma) \overline{\Sigma}_2) - (2K - 2)(\overline{\overline{\Sigma}}_0 - \Sigma_2^T I^{-1}(\gamma) \overline{\Sigma}_2).$$

Remark 1: If we choose the initial value $\overline{\beta}$ *that satisfies* $\overline{\beta} = \beta_0 + o_p(n^{-\frac{1}{2}})$, *e.g.,* $\overline{\beta} = \arg\min_{\beta \in \mathcal{B}} \widetilde{L}_1(\beta)$, *the proof of Theorem 1 shows that*

$$N^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \frac{1}{4g^2(\tau)} \Sigma_1^{-1} (\Sigma_0 - \Sigma_2^T I^{-1}(\gamma) \Sigma_2) \Sigma_1^{-1}),$$

as $N \to \infty$. *Furthermore, if we use true weights* $\pi(V_i, \gamma)$ *instead of estimated weights* $\pi(V_i, \widehat{\gamma})$ *in obtaining the weighted expectile estimator, and denote the estimator by* $\widehat{\beta}_T$, *then*

$$N^{1/2}(\widehat{\beta}_T - \beta_0) \xrightarrow{d} \mathcal{N}(0, \frac{1}{4g^2(\tau)} \Sigma_1^{-1} \Sigma_0 \Sigma_1^{-1}), \quad N \to \infty.$$

IV. PROXIMAL ADMM ALGORITHM

We have established asymptotic properties of the proposed estimator $\widehat{\beta}$. The implementation of the optimization problem (2.10) is difficult and complicated in practice. In this section, we develop a proximal ADMM algorithm for the calculation of the proposed estimator. We set $\overline{\beta}$ to be the current tth

iteration's value $\beta^{(t)}$ to construct the surrogate loss function; then, (2.9) can be rewritten as

$$\widetilde{L}^{(t)}(\beta) = \widetilde{L}_1(\beta) + \left\langle \nabla \widetilde{L}_N(\beta^{(t)}) - \nabla \widetilde{L}_1(\beta^{(t)}), \beta \right\rangle. \quad (4.1)$$

Based on (4.1), we construct the following optimization problem:

$$\beta^{(t)} = \arg\min_{\beta \in \mathcal{B}} \widetilde{L}^{(t)}(\beta). \quad (4.2)$$

We denote by $\{X, Y\}$ the dataset stored on the first machine, where $X = (X_1, X_2, \cdots, X_n)^{\mathrm{T}}$ is a matrix with dimensions $n \times (p+q)$. Y is an $n \times 1$-dimensional vector. We denote $Z = (Z_1, Z_2, \cdots, Z_n)^{\mathrm{T}} \triangleq Y - X\beta$ and $G_\tau(Z) = (1/n)\sum_{i=1}^{n} w_i \rho_\tau(Z_i)$. By convexity, problem (4.2) is equivalent to

$$\begin{cases} \min\limits_{\beta \in \mathcal{B}, Z \in \mathbb{R}^n} & G_\tau(Z) + \left\langle \nabla \widetilde{L}_N(\beta^{(t)}) - \nabla \widetilde{L}_1(\beta^{(t)}), \beta \right\rangle \\ s.t. & X\beta + Z = Y. \end{cases} \quad (4.3)$$

Fix $\rho > 0$; the augmented Lagrangian function of (4.3) is

$$F(\beta, Z, \theta) = G_\tau(Z) + \left\langle \nabla \widetilde{L}_N(\beta^{(t)}) - \nabla \widetilde{L}_1(\beta^{(t)}), \beta \right\rangle - \langle \theta, X\beta + Z - Y \rangle + \frac{\rho}{2}\|X\beta + Z - Y\|_2^2, \quad (4.4)$$

where $\theta \in \mathbb{R}^n$ is the Lagrangian multiplier, and $\|\cdot\|_2$ denotes the l_2 norm in the Euclidean space.

To optimize the Lagrangian function (4.4), the iterations of the classical ADMM algorithm are given by

$$\begin{cases} \beta^{(t+1)} = \arg\min\limits_{\beta \in \mathcal{B}} F(\beta, Z^{(t)}, \theta^{(t)}), \\ Z^{(t+1)} = \arg\min\limits_{Z \in \mathbb{R}^n} F(\beta^{(t+1)}, Z, \theta^{(t)}), \\ \theta^{(t+1)} = \theta^{(t)} - \rho(X\beta^{(t+1)} + Z^{(t+1)} - Y), \end{cases} \quad (4.5)$$

where $(Z^{(t)}, \theta^{(t)})$ denotes the tth iteration of the algorithm for $t \geq 0$. Discarding the constant terms that are independent of the corresponding parameters to be estimated allows the iterations in (4.5) to be rewritten as

$$\begin{cases} \beta & \text{step}: \beta^{(t+1)} = \arg\min\limits_{\beta \in \mathcal{B}} \Big[\left\langle \nabla \widetilde{L}_N(\beta^{(t)}) - \nabla \widetilde{L}_1(\beta^{(t)}), \beta \right\rangle \\ & \quad - \left\langle \theta^{(t)}, X\beta \right\rangle + \frac{\rho}{2}\|X\beta + Z^{(t)} - Y\|_2^2 \Big], \\ Z & \text{step}: Z^{(t+1)} = \arg\min\limits_{Z \in \mathbb{R}^n} \Big[G_\tau(Z) - \left\langle \theta^{(t)}, Z \right\rangle \\ & \quad + \frac{\rho}{2}\|X\beta^{(t+1)} + Z - Y\|_2^2 \Big], \\ \theta & \text{step}: \theta^{(t+1)} = \theta^{(t)} - \rho(X\beta^{(t+1)} + Z^{(t+1)} - Y). \end{cases} \quad (4.6)$$

Simple calculations lead to the following explicit solution of the β step in (4.6):

$$\beta^{(t+1)} = (X^{\mathrm{T}}X)^{-1} \Big[X^{\mathrm{T}}(\rho^{-1}\theta^{(t)} - Z^{(t)} + Y) - \rho^{-1}(\nabla L_N(\beta^{(t)}) - \nabla L_1(\beta^{(t)})) \Big]. \quad (4.7)$$

For solving the Z step in (4.6), the corresponding optimization function is "parametric separation", i.e., the update of $Z^{(t+1)}$ can be performed component-wisely. For $i = 1, 2, \cdots, n$, we have

$$\begin{aligned} Z_i^{(t+1)} &= \arg\min_{Z_i \in \mathbb{R}} \left[\frac{1}{n} w_i \rho_\tau(Z_i) - \theta_i^{(t)} Z_i + \frac{\rho}{2}(Z_i + X_i^{\mathrm{T}}\beta^{(t+1)} - Yi)^2 \right] \\ &= \arg\min_{Z_i \in \mathbb{R}} \left[\rho_\tau(Z_i) + \frac{n\rho}{2w_i} \left\{ Z_i - \left(Y_i - X_i^{\mathrm{T}}\beta^{(t+1)} + \frac{\theta_i^{(t)}}{\rho} \right) \right\}^2 \right]. \end{aligned} \quad (4.8)$$

For solving the univariate minimization problems (4.8), we define

$$\mathrm{Prox}_{\rho_\tau}[\alpha, \beta] = \arg\min_{u \in \mathbb{R}} \left[\rho_\tau(u) + \frac{\beta}{2}(u - \alpha)^2 \right].$$

We call operator $\mathrm{Prox}_{\rho_\tau}$ the proximal mapping of ρ_τ. Given $\tau \in (0, 1)$ and $\beta > 0$, the mapping has the following explicit expression:

$$\mathrm{Prox}_{\rho_\tau}[\alpha, \beta] = \begin{cases} \dfrac{\alpha\beta}{2(1-\tau) + \beta}, & \alpha \leq 0 \\ \dfrac{\alpha\beta}{2\tau + \beta}, & \alpha > 0. \end{cases}$$

Applying the proximal mapping formula to the Z step, for $i = 1, 2, \cdots, n$, we obtain

$$Z_i^{(t+1)} = \mathrm{Prox}_{\rho_\tau} \left[Y_i - X_i^{\mathrm{T}}\beta^{(t+1)} + \frac{\theta_i^{(t)}}{\rho}, \frac{n\rho}{w_i} \right]. \quad (4.9)$$

Equations (4.7) and (4.9) complete the algorithm for the proposed estimation in a linear expectile model. Note that we add the proximal mapping of ρ_τ in the Z step, so we call the algorithm the proximal ADMM algorithm, summarized as follows:

Note that the convergence of the proximal ADMM algorithm can be established similarly to Section 3.3 in Gu *et al.* [5]. As discussed in Gu *et al.* [5], the worst case convergence rate of the proximal ADMM algorithm is at least of order $1/t$ at each communication round, where t is the iteration number.

V. SIMULATION STUDIES

We conduct simulation studies to evaluate the finite-sample performance of the proposed method. We compare the proposed communication-efficient distributed optimization method (labeled "Proposed") that entails using the proximal ADMM algorithm to solve problem (2.10) with the optimal global method (labeled "Oracle") that entails using the classic ADMM algorithm to solve problem (2.8).

We consider the following linear expectile regression model:

$$Y_i = \beta_1 + \beta_2 X_i^1 + \beta_3 X_i^2 + \beta_4 X_i^3 + \epsilon_i, \quad i = 1, 2, \cdots, N, \quad (5.1)$$

where the true parameter $\beta_0 = (\beta_1, \beta_2, \beta_3, \beta_4)^{\mathrm{T}} = (-3, 1, -1, 1)^{\mathrm{T}}$. Covariates $X_i^1 \sim \mathcal{N}(0, 1)$, $X_i^2 \sim \mathcal{N}(0, 1)$ and $X_i^3 \sim$ Bernoulli(0.5) are independent. We assume that

Algorithm 1 Proximal ADMM Algorithm for Large-Scale Expectile Regression Model With Covariates Missing at Random

Initialize $\beta^{(0)} = \arg\min \widetilde{L}_1(\beta), Z^{(0)}, \gamma^{(0)}$.

1: **for** $m = 0, 1, 2, \ldots, M-1$ **do**
2: Transmit the current iteration's value $\beta^{(m)}$ to machines $\{\mathcal{M}_k\}_{k=1}^K$;
3: **for** $k = 1, 2, \ldots, K$ **do**
4: Calculate the gradient $\nabla\widetilde{L}_k(\beta^{(m)})$ at machine $\mathcal{M}_k$;
5: Send the gradient $\nabla\widetilde{L}_k(\beta^{(m)})$ to master machine $\mathcal{M}_1$;
6: **end for**
7: Compute the gradient $\nabla\widetilde{L}_N(\beta^{(m)}) = K^{-1}\sum_{k=1}^{K}\nabla\widetilde{L}_N(\beta^{(m)})$ at master machine $\mathcal{M}_1$;
8: Compute the CSL function $\widetilde{L}^{(m)}(\beta) = \widetilde{L}_1(\beta) + \langle\nabla\widetilde{L}_N(\beta^{(m)}) - \nabla\widetilde{L}_1(\beta^{(m)}), \beta\rangle$;
9: Execute the following iteration on master machine $\mathcal{M}_1$:
10: **for** $t = 0, 1, 2, \ldots, T-1$ **do**
11: Update

$$\beta^{(t+1)} = (X^{\mathrm{T}}X)^{-1}\Big[X^{\mathrm{T}}(\rho^{-1}\theta^{(t)} - Z^{(t)} + Y) - \rho^{-1}(\nabla L_N(\beta^{(t)}) - \nabla L_1(\beta^{(t)}))\Big]$$

12: Update

$$Z^{(t+1)} = \left(\mathrm{Prox}_{\rho_\tau}\left[Y_i - X_i^{\mathrm{T}}\beta^{(t+1)} + \frac{\theta_i^{(t)}}{\rho}, \frac{n\rho}{w_i}\right]\right)_{1\le i\le n}$$

13: Update $\theta^{(t+1)} = \theta^{(t)} - \rho(X\beta^{(t+1)} + Z^{(t+1)} - Y)$;
14: **end for**
15: Update $\beta^{(m+1)} = \beta^{(T)}, Z^{(m+1)} = Z^{(T)}, \theta^{(m+1)} = \theta^{(T)}$
16: **end for**

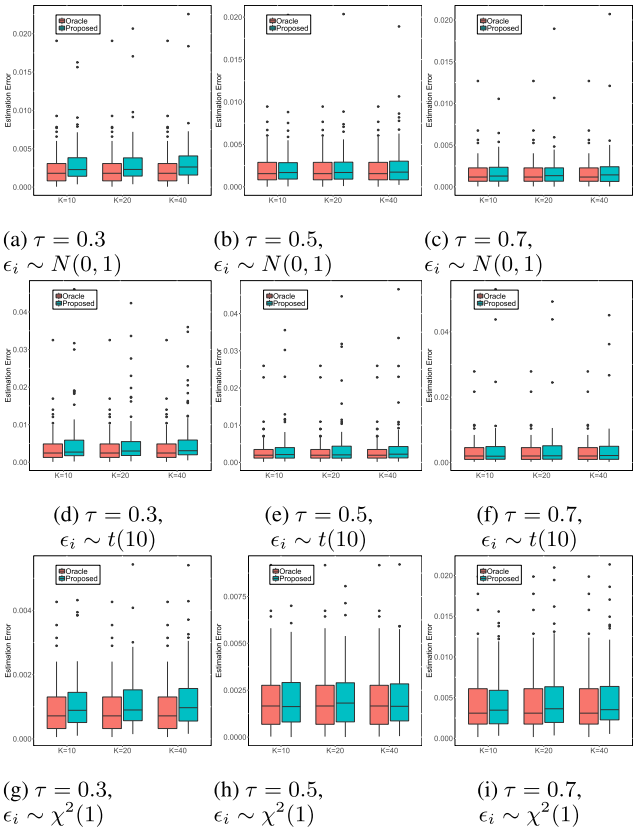
Return $\widehat{\beta} = \beta^{(M)}$.

X_i^2 are always observed, while (X_i^1, X_i^3) are missing at random. We consider different distributions of random errors ϵ_i: (I) homogeneous errors $\epsilon_i \sim \mathcal{N}(0,1), \epsilon_i \sim t(10)$ and $\epsilon_i \sim \chi^2(1)$, and (II) heterogeneous errors $\epsilon_i \sim (1 + X_i^3)\mathcal{N}(0,1), \epsilon_i \sim (1+X_i^3)t(10)$ and $\epsilon_i \sim (1+X_i^3)\chi^2(1)$.

Let R_i be a binary variable indicating whether (X_i^1, X_i^3) is fully observed. We consider R_i arising from the following logistic regression model:

$$\mathrm{logit}(P(R_i=1|X_i, Y_i)) = 4 + Y_i + X_i^2, \quad i = 1, 2, \cdots, N, \quad (5.2)$$

where $\mathrm{logit}(t) = \log(t/(1-t))$. We generate $N = 10000$ samples under model (5.1) with scheme (5.2) used to indicate missing data. For various distributions of ϵ_i, the above missing data mechanism produces on average the rate of missing data ranging from 20% to 43%. We randomly partition the data on $K = 10$, 20 and 40 machines. Thus, the local sample size n on each machine is 1000, 500 and 250, respectively. We set $\tau = 0.3, 0.5$ and 0.7. For each configuration, the results presented below are obtained from 100 independently generated datasets.



(a) $\tau = 0.3$, $\epsilon_i \sim N(0,1)$ (b) $\tau = 0.5$, $\epsilon_i \sim N(0,1)$ (c) $\tau = 0.7$, $\epsilon_i \sim N(0,1)$
(d) $\tau = 0.3$, $\epsilon_i \sim t(10)$ (e) $\tau = 0.5$, $\epsilon_i \sim t(10)$ (f) $\tau = 0.7$, $\epsilon_i \sim t(10)$
(g) $\tau = 0.3$, $\epsilon_i \sim \chi^2(1)$ (h) $\tau = 0.5$, $\epsilon_i \sim \chi^2(1)$ (i) $\tau = 0.7$, $\epsilon_i \sim \chi^2(1)$

FIGURE 1. Boxplots for estimation error $\|\widehat{\beta} - \beta_0\|_2^2$ versus the number of machines under homogeneous errors.

Tables 1 and 2 summarize the results of estimationg β. The tables include the means of squared estimation error $\|\widehat{\beta} - \beta_0\|_2^2$ (ER), and the sample standard derivations of squared estimation error (SD).

We present boxplots to show how the squared estimation error varies over different methods and the number of machines. Figures 1 and 2 display the results. We also show how the squared estimation error varies for the proposed method versus the number of rounds of communication. Figures 3 and 4 summarize the results.

In all cases considered here, results reported in Table 1, Table 2, Figure 1 and Figure 2 indicate that the oracle method gives us the best estimates and statistical performance. Our proposed approach produces estimates that can be competitive with those of the oracle method. The results reported in Figures 3 and 4 indicate that for our proposed communication-efficient distributed optimization method, the squared estimation error declines to SE of the oracle method within a few rounds of communication.

In addition, results reported in Figures 3 and 4 also show that the smaller the machine count K is, the faster ER of the proposed estimator converges to that of the "oracle case". This phenomenon is reasonable, because the local sample

TABLE 1. Results of estimation of β under homogeneous errors.

			$K=10$		$K=20$		$K=40$	
	τ	Method	ER	SD	ER	SD	ER	SD
$\epsilon_i \sim N(0,1)$	0.3	Oracle	0.0024	0.0025	0.0024	0.0025	0.0024	0.0025
		Proposed	0.0030	0.0026	0.0031	0.0029	0.0033	0.0030
	0.5	Oracle	0.0021	0.0018	0.0021	0.0018	0.0021	0.0018
		Proposed	0.0023	0.0025	0.0023	0.0025	0.0024	0.0026
	0.7	Oracle	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017
		Proposed	0.0019	0.0024	0.0018	0.0023	0.0019	0.0025
$\epsilon_i \sim t(10)$	0.3	Oracle	0.0039	0.0044	0.0039	0.0044	0.0039	0.0044
		Proposed	0.0054	0.0069	0.0054	0.0068	0.0055	0.0065
	0.5	Oracle	0.0030	0.0037	0.0030	0.0037	0.0030	0.0037
		Proposed	0.0041	0.0066	0.0041	0.0066	0.0042	0.0066
	0.7	Oracle	0.0033	0.0041	0.0033	0.0041	0.0033	0.0041
		Proposed	0.0041	0.0073	0.0041	0.0070	0.0040	0.0064
$\epsilon_i \sim \chi^2(1)$	0.3	Oracle	0.0009	0.0008	0.0009	0.0008	0.0009	0.0008
		Proposed	0.0011	0.0008	0.0011	0.0009	0.0012	0.0009
	0.5	Oracle	0.0020	0.0017	0.0020	0.0017	0.0020	0.0017
		Proposed	0.0021	0.0017	0.0021	0.0016	0.0021	0.0017
	0.7	Oracle	0.0045	0.0039	0.0045	0.0039	0.0045	0.0039
		Proposed	0.0048	0.0041	0.0049	0.0041	0.0051	0.0043

TABLE 2. Results of estimation of β under heterogeneous errors.

			$K=10$		$K=20$		$K=40$	
	τ	Method	PE	SD	PE	SD	PE	SD
$\epsilon_i \sim (1+X_i^3)N(0,1)$	0.3	Oracle	0.0093	0.0075	0.0093	0.0075	0.0093	0.0075
		Proposed	0.0124	0.0094	0.0128	0.0096	0.0129	0.0099
	0.5	Oracle	0.0101	0.0100	0.0101	0.0100	0.0101	0.0100
		Proposed	0.0112	0.0114	0.0113	0.0113	0.0113	0.0104
	0.7	Oracle	0.0105	0.0150	0.0105	0.0150	0.0105	0.0150
		Proposed	0.0112	0.0169	0.0112	0.0159	0.0111	0.0152
$\epsilon_i \sim (1+X_i^3)t(10)$	0.3	Oracle	0.0291	0.0659	0.0291	0.0659	0.0291	0.0659
		Proposed	0.0310	0.0648	0.0335	0.0833	0.0337	0.0709
	0.5	Oracle	0.0234	0.0438	0.0234	0.0438	0.0234	0.0438
		Proposed	0.0250	0.0504	0.0256	0.0541	0.0263	0.0582
	0.7	Oracle	0.0244	0.0620	0.0244	0.0620	0.0244	0.0620
		Proposed	0.0248	0.0678	0.0253	0.0667	0.0265	0.0665
$\epsilon_i \sim (1+X_i^3)\chi^2(1)$	0.3	Oracle	0.0017	0.0013	0.0017	0.0013	0.0017	0.0013
		Proposed	0.0022	0.0016	0.0022	0.0015	0.0024	0.0019
	0.5	Oracle	0.0042	0.0034	0.0042	0.0034	0.0042	0.0034
		Proposed	0.0043	0.0035	0.0043	0.0035	0.0046	0.0035
	0.7	Oracle	0.0099	0.0080	0.0099	0.0080	0.0099	0.0080
		Proposed	0.0101	0.0080	0.0102	0.0082	0.0105	0.0080

size n increases as the number of machines K varies from 40 to 10, thus, the proposed estimator exhibits faster convergence of ER to that of the oracle estimator as the number of samples stored on each machine increases.

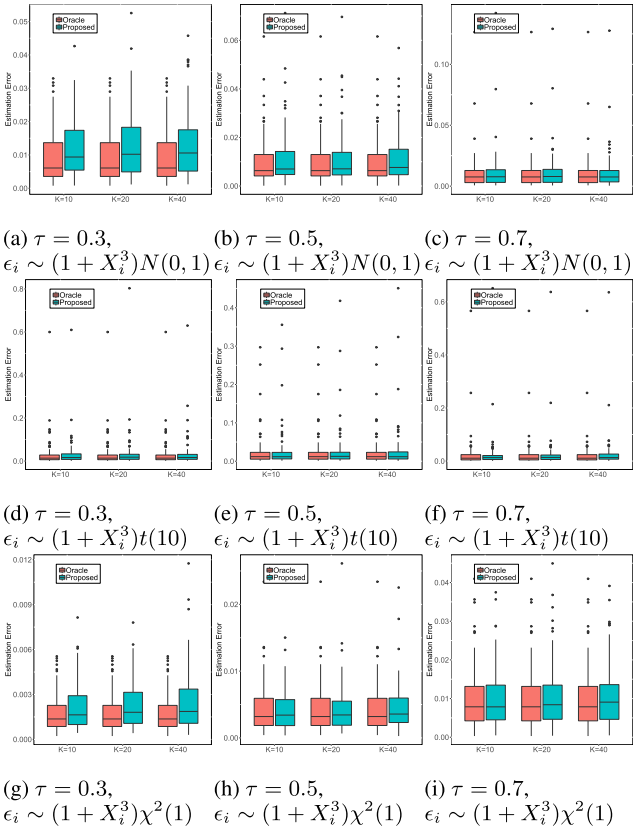
VI. REAL-WORLD DATA ANALYSIS

To illustrate the application of our proposed method, we perform an empirical study using data from 2017 Michigan BRFSS. The latter is a collaborative project of Centers for Disease Control and Prevention. In the analysis, we drop cases with missing, "don't know" and "refused" responses to covariates "Race", "Sex", "Age" and "Internet". We view variable "Income" as the covariate missing at random. Table 3 provides the demographic characteristics for the five considered covariates. The outcome of interest is "Weight", which represents the log-transformed value for every individual. After the above treatment, 411345 individuals are available for the study. The rate of missing data is approximately 13.91%.

We consider the linear expectile regression model to assess the effect of "Income", "Race", "Sex", "Age" and "Internet" on individuals' "Weight". Through the study, we demonstrate that our method provides new perspectives on this data. With a small percentage of individuals accounting

TABLE 3. Demographic characteristics of BRFSS data.

Variables	Description
Income	Income level (1=less than $\$10,000$; 2=$\$10,000$ to less than $\$15,000$; 3=$\$15,000$ to less than $\$20,000$; 4=$\$20,000$ to less than $\$25,000$; 5=$\$25,000$ to less than $\$35,000$; 6=$\$35,000$ to less than $\$50,000$; 7=$\$50,000$ to less than $\$75,000$; 8=$\$75,000$ or more; NA=Unknown/Unsure/Refused to answer/Not asked or Missing)
Race	Computed Race-Ethnicity grouping (1=White only; 2=Black only; 3=American Indian or Alaskan Native only; 4=Asian only; 5=Native Hawaiian or other Pacific Islander only; 6=Other race only; 7=Multiracial; 8=Hispanic)
Sex	Respondent's Sex (1=Male; 2=Female)
Age	Imputed age, divided into in six groups (1:18 - 24; 2: 25 - 34; 3: 35 - 44; 4: 45 - 54; 5: 55 - 64; 6: 65 and above)
Internet	Internet use in the past 30 days (1=Yes; 2=No)

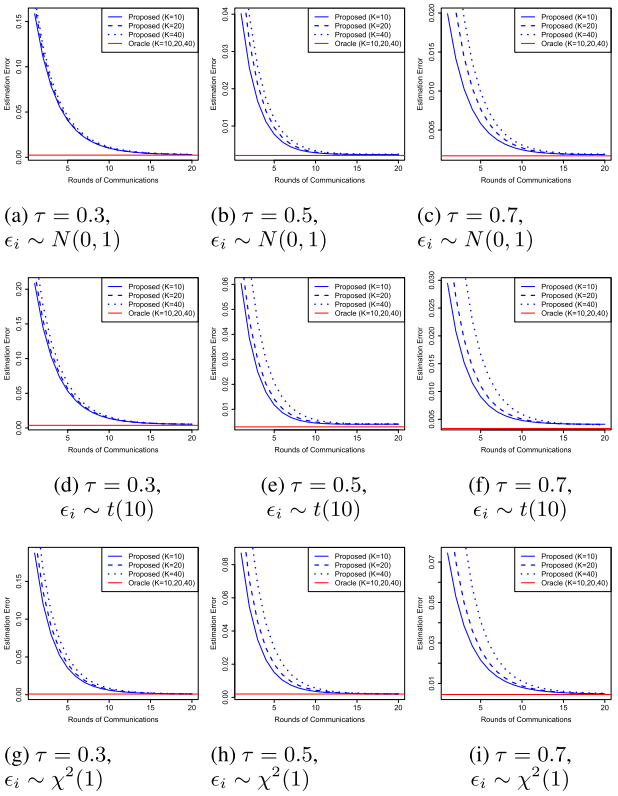


(a) $\tau = 0.3$, $\epsilon_i \sim (1 + X_i^3)N(0,1)$ (b) $\tau = 0.5$, $\epsilon_i \sim (1 + X_i^3)N(0,1)$ (c) $\tau = 0.7$, $\epsilon_i \sim (1 + X_i^3)N(0,1)$

(d) $\tau = 0.3$, $\epsilon_i \sim (1 + X_i^3)t(10)$ (e) $\tau = 0.5$, $\epsilon_i \sim (1 + X_i^3)t(10)$ (f) $\tau = 0.7$, $\epsilon_i \sim (1 + X_i^3)t(10)$

(g) $\tau = 0.3$, $\epsilon_i \sim (1 + X_i^3)\chi^2(1)$ (h) $\tau = 0.5$, $\epsilon_i \sim (1 + X_i^3)\chi^2(1)$ (i) $\tau = 0.7$, $\epsilon_i \sim (1 + X_i^3)\chi^2(1)$

FIGURE 2. Boxplots for estimation error $\|\widehat{\beta} - \beta_0\|_2^2$ versus the number of machines under heterogeneous errors.



(a) $\tau = 0.3$, $\epsilon_i \sim N(0,1)$ (b) $\tau = 0.5$, $\epsilon_i \sim N(0,1)$ (c) $\tau = 0.7$, $\epsilon_i \sim N(0,1)$

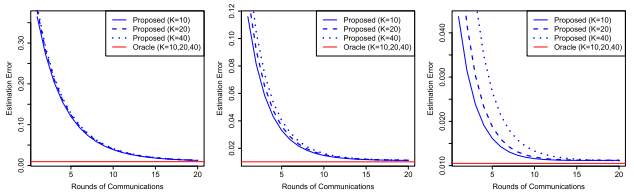
(d) $\tau = 0.3$, $\epsilon_i \sim t(10)$ (e) $\tau = 0.5$, $\epsilon_i \sim t(10)$ (f) $\tau = 0.7$, $\epsilon_i \sim t(10)$

(g) $\tau = 0.3$, $\epsilon_i \sim \chi^2(1)$ (h) $\tau = 0.5$, $\epsilon_i \sim \chi^2(1)$ (i) $\tau = 0.7$, $\epsilon_i \sim \chi^2(1)$

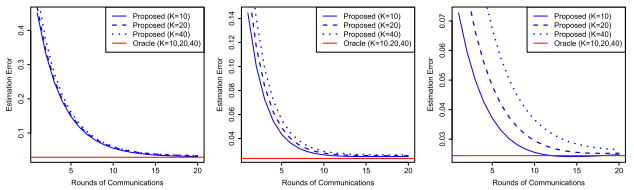
FIGURE 3. Graphs of estimation error $\|\widehat{\beta} - \beta_0\|_2^2$ versus the number of rounds of communication, assuming three machines and homogeneous errors.

for most of the problem, it is interesting to consider the people with different weights, i.e., different conditional expectiles, such as $\tau = 0.3$, 0.5 and 0.7.

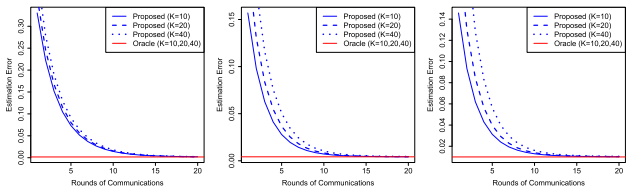
We create a random partition of the data. A total of 300000 individuals are randomly selected as the training set (D_{train}), and the remaining 111345 people constitute the testing set (D_{test}). For the training dataset, we set the machine count K to 50, 100 and 200; i.e., the local sample size n on each machine is 6000, 3000 and 1500, respectively. We model the missing data mechanism by fitting a logistic regression model; the missing data indicator is viewed as the response variable, and variable "Weight" of interest and the remaining



(a) $\tau = 0.3$, $\epsilon_i \sim (1 + X_i^3)N(0, 1)$ (b) $\tau = 0.5$, $\epsilon_i \sim (1 + X_i^3)N(0, 1)$ (c) $\tau = 0.7$, $\epsilon_i \sim (1 + X_i^3)N(0, 1)$



(d) $\tau = 0.3$, $\epsilon_i \sim (1 + X_i^3)t(10)$ (e) $\tau = 0.5$, $\epsilon_i \sim (1 + X_i^3)t(10)$ (f) $\tau = 0.7$, $\epsilon_i \sim (1 + X_i^3)t(10)$



(g) $\tau = 0.3$, $\epsilon_i \sim (1 + X_i^3)\chi^2(1)$ (h) $\tau = 0.5$, $\epsilon_i \sim (1 + X_i^3)\chi^2(1)$ (i) $\tau = 0.7$, $\epsilon_i \sim (1 + X_i^3)\chi^2(1)$

FIGURE 4. Graphs of estimation error $\|\widehat{\beta} - \beta_0\|_2^2$ versus the number of rounds of communication, assuming three machines and heterogeneous errors.

TABLE 4. Results of analysis of BRFSS data.

		$K = 50$		$K = 100$		$K = 200$	
τ	Method	PE	SD	PE	SD	PE	SD
0.3	Oracle	0.3161	0.0016	0.3161	0.0016	0.3161	0.0016
	Proposed	0.3166	0.0037	0.3170	0.0047	0.3181	0.0074
0.5	Oracle	0.3514	0.0015	0.3514	0.0015	0.3514	0.0015
	Proposed	0.3517	0.0015	0.3520	0.0016	0.3527	0.0020
0.7	Oracle	0.3868	0.0021	0.3868	0.0021	0.3868	0.0021
	Proposed	0.3867	0.0037	0.3869	0.0049	0.3870	0.0072

four covariates are predictors. Using the estimated weights, we then perform the coefficient estimation using the training dataset, and evaluate its prediction error on the testing dataset by calculating $(1/111345)\sum_{i\in D_{\text{test}}} \rho_\tau(Y_i - \widehat{Y}_i)$.

We use the proposed proximal ADMM algorithm to calculate the average prediction error (PE) and the standard derivation (SD) of PE. The results are shown in rows labeled "Proposed" in Table 4. We also consider the ADMM algorithm based on the global loss function $\widetilde{L}_N(\beta)$. The results are listed in the rows labeled "Oracle" in Table 4.

The results suggest that PE and SD obtained by the proposed method are competitive with those of the oracle method. In addition, results show that PE of our method increases very little and is close to that of the oracle method as the number of machines increases from 50 to 200.

VII. DISCUSSION

We study an efficient approach in an expectile regression framework for analyzing large-scale data with covariates missing at random. The expectile regression approach is particularly suitable for analyzing highly skewed and heteroscedastic data. To overcome the difficulties caused by large-scale data, a CSL function is constructed. We show that the approach of constructing a surrogate loss function (to the global loss function) based on the subsample of observations is very useful in solving the large-scale data problem where the sample size is so large that the calculation of the global loss function is prohibitive.

An estimator is obtained by solving an optimization problem using the CSL function as the objective function. The asymptotic properties of the proposed estimator are established. A proximal ADMM algorithm is proposed for the calculation of the proposed estimator. Simulation studies suggest that the proposed method performs well for a finite sample size. The easy implementation of the proposed method is demonstrated using real-world data examples.

The procedure described in the paper is most suitable if the frequency of missing data is low or medium. If it is overly high, applying an augmented estimating equation method [14] to use the partial information, and an imputation method [3] to fill in some of the missing values can be explored based on our distributed framework. Further studies also discuss exploring the ideas presented to reduce the computational cost of communication-efficient distributed multitask learning with shared support [19].

APPENDIX

Proof of Theorem 1: For simplicity, the dataset stored on the first machine is denoted by $\{X_{1i}, Y_{1i}\} \triangleq \{X_i, Y_i\}_{i=1}^n$. Define

$$\epsilon_i = Y_i - X_i^{\mathrm{T}}\beta_0, \quad \overline{\epsilon}_i = Y_i - X_i^{\mathrm{T}}\overline{\beta},$$

$$\Sigma_0 = \mathrm{E}\left[X_i X_i^{\mathrm{T}} \frac{1}{\pi(V_i, \gamma)} \psi_\tau^2(\epsilon_i)\right],$$

$$\overline{\Sigma}_0 = \mathrm{E}\left[X_i X_i^{\mathrm{T}} \frac{1}{\pi(V_i, \gamma)} \psi_\tau^2(\overline{\epsilon}_i)\right],$$

$$\overline{\overline{\Sigma}}_0 = \mathrm{E}\left[X_i X_i^{\mathrm{T}} \frac{1}{\pi(V_i, \gamma)} \psi_\tau(\epsilon_i)\psi_\tau(\overline{\epsilon}_i)\right],$$

$$\Sigma_2 = \mathrm{E}\left[(1 - \pi(V_i, \gamma))V_i X_i^{\mathrm{T}} \psi_\tau(\epsilon_i)\right],$$

$$\overline{\Sigma}_2 = \mathrm{E}\left[(1 - \pi(V_i, \gamma))V_i X_i^{\mathrm{T}} \psi_\tau(\overline{\epsilon}_i)\right],$$

$$I(\gamma) = \mathrm{E}\left[V_i V_i^{\mathrm{T}} \pi(V_i, \gamma)(1 - \pi(V_i, \gamma))\right],$$

$$I_1(\delta) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{R_i}{\pi(V_i, \widehat{\gamma})} X_i^{\mathrm{T}} \delta \psi_\tau(\epsilon_i),$$

$$I_2(\delta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i}{\pi(V_i,\widehat{\gamma})}X_i^{\mathrm{T}}\delta\psi_\tau(\overline{\epsilon}_i),$$

$$I_3(\delta) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{R_{ki}}{\pi(V_{ki},\widehat{\gamma})}X_{ki}^{\mathrm{T}}\delta\psi_\tau(\overline{\epsilon}_{ki})\right],$$

$$\mathrm{E}_n(\delta) = I_1(\delta) + I_2(\delta) + I_3(\delta).$$

Lemma 1: If Conditions (C1)-(C5) hold, then

$$E_n(\delta) \xrightarrow{d} \mathcal{N}(0, \delta^T\widetilde{\Sigma}\delta), \quad n\to\infty,$$

where

$$\widetilde{\Sigma} = \Sigma_0 - \Sigma_2^T I^{-1}(\gamma)\Sigma_2 + \frac{K-1}{K}(\overline{\Sigma}_0 - \overline{\Sigma}_2^T I^{-1}(\gamma)\overline{\Sigma}_2) - \frac{2K-2}{K}(\overline{\overline{\Sigma}}_0 - \Sigma_2^T I^{-1}(\gamma)\overline{\Sigma}_2).$$

Proof of Lemma 1: Under Conditions (C1)-(C5), similarly to arguments in Lemma 1 of [16], we obtain

$$I_1(\delta) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i + o_p(1), \quad I_2(\delta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i + o_p(1),$$

$$I_3(\delta) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{K}\zeta_i + o_p(1),$$

where

$$\xi_i = \frac{R_i}{\pi(V_i,\gamma)}X_i^{\mathrm{T}}\delta\psi_\tau(\epsilon_i) - (R_i - \pi(V_i,\gamma))V_i^{\mathrm{T}}I^{-1}(\gamma)\Sigma_2\delta,$$

$$\eta_i = \frac{R_i}{\pi(V_i,\gamma)}X_i^{\mathrm{T}}\delta\psi_\tau(\overline{\epsilon}_i) - (R_i - \pi(V_i,\gamma))V_i^{\mathrm{T}}I^{-1}(\gamma)\overline{\Sigma}_2\delta,$$

$$\zeta_i = \sum_{k=1}^{K}\left[\frac{R_{ki}}{\pi(V_{ki},\gamma)}X_{ki}^{\mathrm{T}}\delta\psi_\tau(\overline{\epsilon}_{ki}) - (R_{ki} - \pi(V_{ki},\gamma))V_{ki}^{\mathrm{T}}I^{-1}(\gamma)\overline{\Sigma}_2\delta\right],$$

and $o_p(1)$ is dimensionless.

Since

$$\mathrm{E}\left[\frac{R_i}{\pi(V_i,\gamma)}X_i^{\mathrm{T}}\delta\psi_\tau(\epsilon_i)\right] = \mathrm{E}\left[X_i^{\mathrm{T}}\delta\mathrm{E}[\psi_\tau(\epsilon_i)|X_i]\right] = 0$$

and

$$\mathrm{E}\left[(R_i - \pi(V_i,\gamma))V_i^{\mathrm{T}}I^{-1}(\gamma)\Sigma_2\delta\right] = \mathrm{E}\left[\mathrm{E}[R_i - \pi(V_i,\gamma)|X_i, V_i]\, V_i^{\mathrm{T}}I^{-1}(\gamma)\Sigma_2\delta\right] = 0,$$

we have $\mathrm{E}[\xi_i] = 0$. Then,

$$\begin{aligned}\mathrm{Var}(\xi_i) = {}& \mathrm{Var}\left(\frac{R_i}{\pi(V_i,\gamma)}X_i^{\mathrm{T}}\delta\psi_\tau(\epsilon_i)\right) + \mathrm{Var}\left((R_i - \pi(V_i,\gamma))V_i^{\mathrm{T}}I^{-1}(\gamma)\Sigma_2\delta\right) \\ & - 2\mathrm{Cov}\left(\frac{R_i}{\pi(V_i,\gamma)}X_i^{\mathrm{T}}\delta\psi_\tau(\epsilon_i), (R_i - \pi(V_i,\gamma)) \times V_i^{\mathrm{T}}I^{-1}(\gamma)\Sigma_2\delta\right).\end{aligned}$$

By performing a simple calculation, we obtain

$$\begin{aligned}&\mathrm{Var}\left(\frac{R_i}{\pi(V_i,\gamma)}X_i^{\mathrm{T}}\delta\psi_\tau(\epsilon_i)\right) \\ &= \delta^{\mathrm{T}}\mathrm{E}\left[X_iX_i^{\mathrm{T}}\frac{1}{\pi(V_i,\gamma)}\psi_\tau^2(\epsilon_i)\right]\delta = \delta^{\mathrm{T}}\Sigma_0\delta, \\ &\mathrm{Var}\left((R_i - \pi(V_i,\gamma))V_i^{\mathrm{T}}I^{-1}(\gamma)\Sigma_2\delta\right) \\ &= \delta^{\mathrm{T}}\Sigma_2^{\mathrm{T}}I^{-1}(\gamma)\mathrm{E}\left[\pi(V_i,\gamma)(1-\pi(V_i,\gamma))V_iV_i^{\mathrm{T}}\right]I^{-1}(\gamma)\Sigma_2\delta \\ &= \delta^{\mathrm{T}}\Sigma_2^{\mathrm{T}}I^{-1}(\gamma)\Sigma_2\delta,\end{aligned}$$

and

$$\begin{aligned}&\mathrm{Cov}\left(\frac{R_i}{\pi(V_i,\gamma)}X_i^{\mathrm{T}}\delta\psi_\tau(\epsilon_i), (R_i - \pi(V_i,\gamma))V_i^{\mathrm{T}}I^{-1}(\gamma)\Sigma_2\delta\right) \\ &= \mathrm{E}\left[\frac{R_i}{\pi(V_i,\gamma)}X_i^{\mathrm{T}}\delta\psi_\tau(\epsilon_i)(R_i - \pi(V_i,\gamma))V_i^{\mathrm{T}}I^{-1}(\gamma)\Sigma_2\delta\right] \\ &= \delta^{\mathrm{T}}\mathrm{E}\left[(1-\pi(V_i,\gamma))X_i\psi_\tau(\epsilon_i)V_i^{\mathrm{T}}\right]I^{-1}(\gamma)\Sigma_2\delta \\ &= \delta^{\mathrm{T}}\Sigma_2^{\mathrm{T}}I^{-1}(\gamma)\Sigma_2\delta,\end{aligned}$$

which show that

$$\mathrm{Var}(\xi_i) = \delta^{\mathrm{T}}(\Sigma_0 - \Sigma_2^{\mathrm{T}}I^{-1}(\gamma)\Sigma_2)\delta. \tag{A.1}$$

As in the derivation of (A.1), we obtain

$$\begin{aligned}\mathrm{Var}(\eta_i) &= \delta^{\mathrm{T}}(\overline{\Sigma}_0 - \overline{\Sigma}_2^{\mathrm{T}}I^{-1}(\gamma)\overline{\Sigma}_2)\delta, \\ \mathrm{Var}(\zeta_i) &= K\mathrm{Var}(\eta_i), \\ \mathrm{Cov}(\xi_i,\eta_i) &= \delta^{\mathrm{T}}\left[\overline{\overline{\Sigma}}_0 - \Sigma_2^{\mathrm{T}}I^{-1}(\gamma)\overline{\Sigma}_2\right]\delta, \\ \mathrm{Cov}(\xi_i,\zeta_i) &= \mathrm{Cov}(\xi_i,\eta_i), \quad \mathrm{Cov}(\eta_i,\zeta_i) = \mathrm{Var}(\eta_i).\end{aligned}$$

Denote $M_i = -\xi_i + \eta_i - \frac{1}{K}\zeta_i$; then, $\mathrm{Var}(M_i) = \delta^{\mathrm{T}}\widetilde{\Sigma}\delta$, where

$$\widetilde{\Sigma} = \Sigma_0 - \Sigma_2^{\mathrm{T}}I^{-1}(\gamma)\Sigma_2 + \frac{K-1}{K}(\overline{\Sigma}_0 - \overline{\Sigma}_2^{\mathrm{T}}I^{-1}(\gamma)\overline{\Sigma}_2) - \frac{2K-2}{K}(\overline{\overline{\Sigma}}_0 - \Sigma_2^{\mathrm{T}}I^{-1}(\gamma)\overline{\Sigma}_2). \tag{A.2}$$

Define $W_n^2 = \sum_{i=1}^{n}\mathrm{Var}(M_i)$. Note that $\mathrm{E}\left[M_i^4\right] < \infty$ because all terms except for V_i are bounded and $\mathrm{E}\left[(V_i^{\mathrm{T}}a)^4\right] < \infty$ for any $a \in \mathbb{R}^{p+1}$. By Hölder's inequality and Chebyshev's inequality, for any $\lambda > 0$, there exists $c > 0$ s.t.

$$\begin{aligned}&\frac{1}{W_n^2}\sum_{i=1}^{n}\mathrm{E}\left[M_i^2 I(|M_i| > \lambda W_n)\right] \\ &\le \frac{1}{W_n^2}\sum_{i=1}^{n}\left[\mathrm{E}M_i^4\right]^{1/2}\left[\mathrm{E}\left(I(|M_i| > \lambda W_n)\right)^2\right]^{1/2} \\ &\le \frac{b}{W_n^2}\sum_{i=1}^{n}\left[P\left(|M_i| > \lambda W_n\right)\right]^{1/2} \\ &\le \frac{b}{W_n^2}\sum_{i=1}^{n}\left[\frac{\mathrm{Var}(M_i)}{(\lambda W_n)^2}\right]^{1/2} \le \frac{c}{W_n^2}\sum_{i=1}^{n}\frac{1}{\lambda|W_n|}.\end{aligned} \tag{A.3}$$

By (A.2), (A.3) and the Lindeberg-Feller central limit theorem, we have

$$\mathrm{E}_n(\delta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} M_i \xrightarrow{d} \mathcal{N}(0, \delta^{\mathrm{T}}\widetilde{\Sigma}\delta), \quad n \longrightarrow \infty.$$

Lemma 2: If Conditions (C1)-(C5) hold, then

$$n\left[\widetilde{L}(\beta_0 + \frac{\delta}{\sqrt{n}}) - \widetilde{L}(\beta_0)\right] = g(\tau)\delta^T\Sigma_1\delta + E_n(\delta) + o_p(1), \quad n \to \infty.$$

Proof of Lemma 2: By performing a simple calculation and using the definitions of $I_2(\delta)$ and $I_3(\delta)$, and denote

$$H_n(\delta) = \sum_{i=1}^{n} \frac{R_i}{\pi(V_i, \widehat{\gamma})}\left[\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) - \rho_\tau(\epsilon_i)\right],$$

we obtain

$$\begin{aligned}
&n\left[\widetilde{L}(\beta_0 + \frac{\delta}{\sqrt{n}}) - \widetilde{L}(\beta_0)\right] \\
&= n\left[\widetilde{L}_1(\beta_0 + \frac{\delta}{\sqrt{n}}) - \widetilde{L}_1(\beta_0) + \left\langle \nabla\widetilde{L}_N(\overline{\beta}) - \nabla\widetilde{L}_1(\overline{\beta}), \frac{\delta}{\sqrt{n}}\right\rangle\right] \\
&= \sum_{i=1}^{n} \frac{R_i}{\pi(V_i, \widehat{\gamma})}\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) - \sum_{i=1}^{n} \frac{R_i}{\pi(V_i, \widehat{\gamma})}\rho_\tau(\epsilon_i) \\
&\quad + \frac{\delta^{\mathrm{T}}}{\sqrt{n}}\left[\sum_{i=1}^{n} \frac{R_i}{\pi(V_i, \widehat{\gamma})} X_i\psi_\tau(\overline{\epsilon}_i)\right] \\
&\quad - \frac{\delta^{\mathrm{T}}}{\sqrt{n}K}\left[\sum_{i=1}^{n}\sum_{k=1}^{K} \frac{R_{ki}}{\pi(V_{ki}, \widehat{\gamma})} X_{ki}\psi_\tau(\overline{\epsilon}_{ki})\right] \\
&= H_n(\delta) + I_2(\delta) + I_3(\delta). \qquad \text{(A.4)}
\end{aligned}$$

As $\mathrm{E}[I_1(\delta)] = 0$, we obtain

$$\begin{aligned}
H_n(\delta) &= \mathrm{E}[H_n(\delta)] + \sum_{i=1}^{n} \frac{R_i}{\pi(V_i, \widehat{\gamma})}\left[\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) - \rho_\tau(\epsilon_i)\right] \\
&\quad - \mathrm{E}[H_n(\delta)] \\
&= \mathrm{E}[H_n(\delta)] + \left[\sum_{i=1}^{n} S_{i,n}(\delta) + I_1(\delta)\right] \\
&\quad - \mathrm{E}\left[\sum_{i=1}^{n} S_{i,n}(\delta) + I_1(\delta)\right] \\
&= \mathrm{E}[H_n(\delta)] + I_1(\delta) + \sum_{i=1}^{n}\left[S_{i,n}(\delta) - \mathrm{E}(S_{i,n}(\delta))\right], \qquad \text{(A.5)}
\end{aligned}$$

where $S_{i,n}(\delta) = \frac{R_i}{\pi(V_i,\widehat{\gamma})}\left[\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) - \rho_\tau(\epsilon_i) + \psi_\tau(\epsilon_i)\frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}\right]$.

By applying Conditions (C2) and (C3), we obtain

$$\begin{aligned}
&\mathrm{E}[H_n(\delta)] \\
&= \mathrm{E}\left[\sum_{i=1}^{n} \frac{R_i}{\pi(V_i, \gamma)}\left(\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) - \rho_\tau(\epsilon_i)\right)\right] \\
&\quad + \mathrm{E}\left[\sum_{i=1}^{n}\left(\frac{R_i}{\pi(V_i, \widehat{\gamma})} - \frac{R_i}{\pi(V_i, \gamma)}\right)\right. \\
&\quad \left.\times\left(\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) - \rho_\tau(\epsilon_i)\right)\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n} \frac{R_i}{\pi(V_i, \gamma)}\left(\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) - \rho_\tau(\epsilon_i)\right)\right] + o_p(1) \\
&= \mathrm{E}\left[\mathrm{E}\left\{\sum_{i=1}^{n} \frac{R_i}{\pi(V_i, \gamma)}\left(\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) - \rho_\tau(\epsilon_i)\right) | X_i, Y_i\right\}\right] \\
&\quad + o_p(1) \\
&= \mathrm{E}\left[\sum_{i=1}^{n}\left(\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) - \rho_\tau(\epsilon_i)\right)\right] + o_p(1) \\
&\triangleq \mathrm{E}\left[\widetilde{H}_n(\delta)\right] + o_p(1), \qquad \text{(A.6)}
\end{aligned}$$

where $\widetilde{H}_n(\delta) = \sum_{i=1}^{n}\left(\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) - \rho_\tau(\epsilon_i)\right)$. Denote $M(t) = \mathrm{E}[\rho_\tau(\epsilon_i - t) - \rho_\tau(\epsilon_i)]$; then, $M(0) = 0$; performing a simple calculation, we obtain

$$\begin{aligned}
\nabla_t M(0) &= -\mathrm{E}[2(1-\tau)\epsilon_i I(\epsilon_i \le 0) + 2\tau\epsilon_i I(\epsilon_i > 0)] \\
&= -\mathrm{E}[\psi_\tau(\epsilon_i)] = 0, \\
\nabla_t^2 M(0) &= 2(1-\tau)\mathrm{E}[I(\epsilon_i \le 0)] + 2\tau\mathrm{E}[I(\epsilon_i > 0)] \\
&= 2(1-\tau)F(0) + 2\tau(1 - F(0)) = 2g(\tau),
\end{aligned}$$

where $g(\tau) = (1-\tau)F(0) + \tau(1-F(0))$. Using the Taylor's theorem, we have $M(t) = g(\tau)t^2 + o(t^2)$. Hence, under Condition (C5), for large n we have

$$\begin{aligned}
\mathrm{E}\left[\widetilde{H}_n(\delta)\right] &= \sum_{i=1}^{n} M(\frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) \\
&= \sum_{i=1}^{n}\left[g(\tau)(\frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}})^2 + o\left((\frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}})^2\right)\right] \\
&= g(\tau)\delta^{\mathrm{T}}\Sigma_1\delta + o_p(1). \qquad \text{(A.7)}
\end{aligned}$$

We perform a second-order Taylor expansion of $\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}})$ around ϵ_i and obtain

$$\begin{aligned}
\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) &= \rho_\tau(\epsilon_i) + \nabla_{\epsilon_i}\rho_\tau(\epsilon_i)\left[-\frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}\right] \\
&\quad + \frac{\nabla_{\epsilon_i}^2\rho_\tau(\eta_i)}{2}\delta^{\mathrm{T}}\left[\frac{X_iX_i^{\mathrm{T}}}{n}\right]\delta + o_p(1), \qquad \text{(A.8)}
\end{aligned}$$

where η_i is between ϵ_i and $\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}$. We can rewrite (A.8) as

$$\begin{aligned}
&\rho_\tau(\epsilon_i - \frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}) - \rho_\tau(\epsilon_i) + \psi_\tau(\epsilon_i)\left[\frac{X_i^{\mathrm{T}}\delta}{\sqrt{n}}\right] \\
&\quad = \frac{\nabla_{\epsilon_i}^2\rho_\tau(\eta_i)}{2}\delta^{\mathrm{T}}\left[\frac{X_iX_i^{\mathrm{T}}}{n}\right]\delta + o_p(1). \qquad \text{(A.9)}
\end{aligned}$$

By (A.9), the definition of $S_{i,n}(\delta)$ and the inequality $|\nabla^2_{\epsilon_i}\rho_\tau(\eta_i)| \leq 2\max(\tau, 1-\tau)$, there exists a constant $b > 0$ s.t.

$$|S_{i,n}(\delta)| = \left| \frac{R_i}{\pi(V_i,\gamma)} \left[\frac{\nabla^2_{\epsilon_i}\rho_\tau(\eta_i)}{2} \delta^{\mathrm{T}} \left(\frac{X_i X_i^{\mathrm{T}}}{n} \right) \delta \right] + o_p(1) \right| \leq b\max(\tau, 1-\tau)\delta^{\mathrm{T}} \left[\frac{X_i X_i^{\mathrm{T}}}{n} \right] \delta.$$

Under Condition (C5),

$$\delta^{\mathrm{T}} \left[\frac{\sum_{i=1}^n X_i X_i^{\mathrm{T}}}{n} \right] \delta \to \delta^{\mathrm{T}} \Sigma_1 \delta, \quad \max_{1\leq i\leq n} \frac{\|X_i\|^2}{\sqrt{n}} \to 0, \quad \|\delta\|_2 = C,$$

there exists a constant $c = c(\tau) > 0$ s.t.

$$\sum_{i=1}^n \mathrm{E}\left[S_{i,n}(\delta)\right]^2 \leq c(\tau) \left[\delta^{\mathrm{T}} \left\{ \frac{\sum_{i=1}^n X_i X_i^{\mathrm{T}}}{n} \right\} \delta \right] \cdot \max_{1\leq i\leq n} \left(\frac{\|X_i\|^2}{n} \right) \cdot \|\delta\|^2 \to 0. \tag{A.10}$$

Using (A.10), due to the cancelation of cross-product terms, for a fixed δ we obtain

$$\mathrm{E}\left[\sum_{i=1}^n \left(S_{i,n}(\delta) - \mathrm{E}S_{i,n}(\delta) \right) \right]^2 = \sum_{i=1}^n \mathrm{E}\left[S_{i,n}(\delta) - \mathrm{E}S_{i,n}(\delta) \right]^2 \leq \sum_{i=1}^n \mathrm{E}\left[S_{i,n}(\delta)\right]^2 \to 0,$$

which implies that

$$\sum_{i=1}^n \left(S_{i,n}(\delta) - \mathrm{E}S_{i,n}(\delta) \right) = o_p(1). \tag{A.11}$$

From (A.5), (A.6), (A.7) and (A.11), it follows that

$$H_n(\delta) = g(\tau)\delta^{\mathrm{T}}\Sigma_1\delta + I_1(\delta) + o_p(1). \tag{A.12}$$

Due to (A.4) and (A.12), we obtain

$$n\left[\widetilde{L}(\beta_0 + \frac{\delta}{\sqrt{n}}) - \widetilde{L}(\beta_0) \right] = g(\tau)\delta^{\mathrm{T}}\Sigma_1\delta + \mathrm{E}_n(\delta) + o_p(1), \quad n\to\infty,$$

where

$$\mathrm{E}_n(\delta) = I_1(\delta) + I_2(\delta) + I_3(\delta).$$

Lemma 3: Let U be a symmetric and positive definite matrix, V be a random variable and $A_n(\delta)$ be a convex objective function with the minimum point α_n. If

$$A_n(\delta) = \frac{1}{2}\delta^T U\delta + V^T\delta + o_p(1),$$

then

$$\alpha_n \xrightarrow{d} -U^{-1}V.$$

The proof of Lemma 3 was given in Hjørt and Pollard [6], and is thus omitted.

Proof of Theorem 1: By Lemma 1 and Lemma 2, we obtain

$$n\left[\widetilde{L}(\beta_0 + \frac{\delta}{\sqrt{n}}) - \widetilde{L}(\beta_0) \right] = g(\tau)\delta^{\mathrm{T}}\Sigma_1\delta + \mathrm{E}_n(\delta) + o_p(1),$$
$$\mathrm{E}_n(\delta) \xrightarrow{d} \mathcal{N}(0, \delta^{\mathrm{T}}\widetilde{\Sigma}\delta), \quad n\to\infty.$$

Define $\widehat{\delta}_n = \sqrt{n}(\widehat{\beta} - \beta_0)$; then, $\widehat{\beta} = \beta_0 + \frac{\widehat{\delta}_n}{\sqrt{n}}$. Note that

$$\widehat{\delta}_n = \arg\min_{\delta\in\mathbb{R}^{p+q}} n\left[\widetilde{L}(\beta_0 + \frac{\delta}{\sqrt{n}}) - \widetilde{L}(\beta_0) \right]$$

and by Lemma 3, we obtain

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \frac{1}{4g^2(\tau)}\Sigma_1^{-1}\widetilde{\Sigma}\Sigma_1^{-1}), \quad n\to\infty. \tag{A.13}$$

Due to $N = nK$ and the result (A.13), $\widehat{\beta}$ has the following asymptotic property:

$$\sqrt{N}(\widehat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \frac{1}{4g^2(\tau)}\Sigma_1^{-1}\Sigma\Sigma_1^{-1}), \quad N\to\infty,$$

where $\Sigma = K(\Sigma_0 - \Sigma_2^{\mathrm{T}} I^{-1}(\gamma)\Sigma_2) + (K-1)(\overline{\Sigma}_0 - \overline{\Sigma}_2^{\mathrm{T}} I^{-1}(\gamma)\overline{\Sigma}_2) - (2K-2)(\overline{\Sigma}_0 - \Sigma_2^{\mathrm{T}} I^{-1}(\gamma)\overline{\Sigma}_2)$.

REFERENCES

[1] S. Boyd, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.
[2] R. Chan, H. Yang, and T. Zeng, "A two-stage image segmentation method for blurry images with poisson or multiplicative gamma noise," *SIAM J. Imag. Sci.*, vol. 7, no. 1, pp. 98–127, Jan. 2014.
[3] S. X. Chen and I. Van Keilegom, "Estimation in semiparametric models with missing data," *Ann. Inst. Stat. Math.*, vol. 65, no. 4, pp. 785–805, Aug. 2013.
[4] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Trans. Autom. Control*, vol. 57, no. 3, pp. 592–606, Mar. 2012.
[5] Y. Gu, J. Fan, L. Kong, S. Ma, and H. Zou, "ADMM for high-dimensional sparse penalized quantile regression," *Technometrics*, vol. 60, no. 3, pp. 319–331, Jul. 2018.
[6] N. L. Hjørt and D. Pollard, "Asymptotics for minimisers of convex processes," *Stat. Res. Rep.*, May 1993.
[7] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, Jan. 2016.
[8] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *J. Amer. Stat. Assoc.*, vol. 47, no. 260, pp. 663–685, Dec. 1952.
[9] Z.-F. Jin, Z. Wan, Y. Jiao, and X. Lu, "An alternating direction method with continuation for nonconvex low rank minimization," *J. Sci. Comput.*, vol. 66, no. 2, pp. 849–869, Feb. 2016.
[10] M. I. Jordan, J. D. Lee, and Y. Yang, "Communication-efficient distributed statistical inference," *J. Amer. Stat. Assoc.*, vol. 114, no. 526, pp. 668–681, 2019.
[11] D. J. Lee, Y. Sun, Q. Liu, and J. E. Taylor, "Communication-efficient sparse regression: A one-shot approach," 2015, *arXiv:1503.04337*. [Online]. Available: https://arxiv.org/abs/1503.04337
[12] R. J. Little and D. B. Rubin, *Statistical Analysis With Missing Data*. New York, NY, USA: Wiley, 1987.
[13] W. K. Newey and J. L. Powell, "Asymmetric least squares estimation and testing," *Econometrica*, vol. 55, no. 4, p. 819, Jul. 1987.
[14] J. M. Robins and A. Rotnitzky, "Semiparametric efficiency in multivariate regression models with missing data," *J. Amer. Stat. Assoc.*, vol. 90, no. 429, pp. 122–129, Mar. 1995.
[15] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *J. Amer. Stat. Assoc.*, vol. 89, no. 427, pp. 846–866, Sep. 1994.

[16] B. Sherwood, L. Wang, and X. H. Zhou, "Weighted quantile regression for analyzing health care cost data with missing covariates," *Statist. Med.*, vol. 32, no. 28, pp. 4967–4979, Dec. 2013.
[17] F. Sobotka, S. Schnabel, L. S. Waltrup, P. Eilers, T. Kneib, G. Kauermann, and M. F. Sobotka. (2014). *Expectreg: Expectile and Quantile Regression. R Package Expectreg Version 0.39*. [Online]. Available: http://cran.r-project.org/web/packages/expectreg/index.html
[18] C. Y. Wang, S. Wang, L.-P. Zhao, and S.-T. Ou, "Weighted semiparametric estimation in regression analysis with missing covariate data," *J. Amer. Stat. Assoc.*, vol. 92, no. 438, pp. 512–525, Jun. 1997.
[19] J. Wang, M. Kolar, and N. Srerbo, "Distributed multitask learning," 2016, *arXiv:1510.00633*. [Online]. Available: https://arxiv.org/abs/1510.00633
[20] Q. Wang and Z. Sun, "Estimation in partially linear models with missing responses at random," *J. Multivariate Anal.*, vol. 98, no. 7, pp. 1470–1493, Aug. 2007.
[21] Y. Yang and H. Zou, "Nonparametric multiple expectile regression via ER-Boost," *J. Stat. Comput. Simul.*, vol. 85, no. 7, pp. 1442–1458, May 2015.
[22] Y. Zhang, J. C. Duchi, and M. J. Wainwright, "Communication-efficient algorithms for statistical optimization," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3321–3363, 2013.
[23] J. Zhao and Y. Zhang, "Variable selection in expectile regression," *Commun. Statist.- Theory Methods*, vol. 47, no. 7, pp. 1731–1746, Apr. 2018.
[24] Y. Zhou, A. T. K. Wan, and X. Wang, "Estimating equations inference with missing data," *J. Amer. Stat. Assoc.*, vol. 103, no. 483, pp. 1187–1199, Sep. 2008.
[25] J. F. Ziegel, "Coherence and elicitability," *Math. Finance*, vol. 26, no. 4, pp. 901–918, Oct. 2016.

YINGLI PAN received the B.S. degree in mathematics from Henan Normal University, Henan, China, in 2011, the M.S. degree in mathematics from Central China Normal University, Wuhan, China, in 2014, and the Ph.D. degree in statistics from the Huazhong University of Science and Technology, Wuhan, in 2018.

She has been with the School of Mathematics and Statistics, Hubei University, since July 2018. Her main research interests include survival analysis, data analysis and statistical calculation, and distributed optimization method.

ZHAN LIU received the B.S. and M.S. degrees in mathematics from Hubei University, Wuhan, China, in 2004 and 2007, respectively, and the Ph.D. degree in statistics from the Renmin University of China, Beijing, China, in 2017.

From August 2016 to March 2017, she was a Visiting Ph.D. Student with the JPSM, University of Maryland, College Park, MD, USA. She has been an Associate Professor with the School of Mathematics and Statistics, Hubei University, since July 2017. From May 2018 to August 2018 and December 2018 to February 2019, she worked as a Research Associate with the Department of Statistics, The Chinese University of Hong Kong, Hong Kong. Her research interests include sampling inference, missing data, and distributed optimization method.

WEN CAI was born in Yancheng, Jiangsu, in January 1996. At the age of eight, she started primary school. In 2008, she entered Dingzhuang middle school in Jiangsu Province. In 2011, she entered Dongtai middle school in Jiangsu Province. She received the B.S. degree in statistics from Yancheng Normal University, in 2018. She is currently pursuing the master's degree in applied statistics with Hubei University.

She has good performance in statistics learning and research. Her research interests include distributed optimization method and sampling inference. Her ambition is to become a data analyst. She looks forward to serving the public and humanity in the future.