Journal of Machine Learning Research 27 (2026) 1-39 Submitted 4/23; Revised 10/25; Published 1/26

Communication-efficient Distributed Statistical Inference for Massive Data with Heterogeneous Auxiliary Information

Miaomiao Yu MMYU@FEM.ECNU.EDU.CN
Key Laboratory of Advanced Theory and Application in Statistics and Data Science (MOE)
School of Statistics and Academy of Statistics and Interdisciplinary Sciences
East China Normal University
Shanghai, 200062, China

Zhongfeng Jiang JIANGZHONGFENG17@163.COM
Academy of Mathematics and System Sciences
Chinese Academy of Science
Beijing, 100864, China

Jiaxuan Li JIAXUANLEE0719@163.COM
Chengdu No. 7 High School
Chengdu, 610041, China

Yong Zhou * YZHOU@FEM.ECNU.EDU.CN
Key Laboratory of Advanced Theory and Application in Statistics and Data Science (MOE)
School of Statistics and Academy of Statistics and Interdisciplinary Sciences
East China Normal University
Shanghai, 200062, China

Editor: Silvia Villa

Abstract

Heterogeneous auxiliary information commonly arises in big data due to diverse study settings and privacy constraints. Excluding such indirect evidence often results in a substantial loss of statistical inference efficiency. This article proposes a novel framework for integrating a mixture of individual-level data and multiple external heterogeneous summary statistics by multiplying likelihood functions and confidence densities. Theoretically, we show that the proposed method possesses desirable properties and can achieve statistical efficiency comparable to that of the individual participant data (IPD) estimator, which uses all available individual-level data. Furthermore, we develop a communication-efficient distributed inference procedure for massive datasets with heterogeneous auxiliary information. We demonstrate that the proposed iterative algorithm achieves linear convergence under general conditions or generalized linear models. Finally, extensive simulations and real data applications are conducted to illustrate the performance of the proposed methods.

Keywords: heterogeneous auxiliary information, confidence density, massive data, communication efficiency, distributed inference

1. Introduction

The explosive growth of big datasets has presented significant challenges to traditional statistical inference due to the high calculation costs and substantial storage requirements of massive datasets. To accommodate this bottleneck, distributed statistical inference meth-

ods, such as the divide-and-conquer algorithm, have been proposed and developed in recent years. The main procedure of such distributed strategies involves partitioning the entire large dataset into multiple subsets, each stored on different local machines. After performing local computations and facilitating communication between the local machines and the central machine (which serves as the master node), the central machine aggregates the summary statistics received from the local machines to produce a global estimate of the parameters of interest. Moreover, based on the number of communication rounds, the distributed algorithms are classified into two categories in most literature. The first category, known as the "one-shot" approach, requires each local machine to compute estimators in parallel and transmit them to the central machine, which then aggregates the results into a global estimate through averaging. This approach has been intensively studied for a wide range of topics, for example, U-statistics (Lin and Xi, 2010; Xi and Lin, 2016), quantile regression (Chen and Zhou, 2020; Volgushev et al., 2019), high-dimensional sparse models (Chen and Xie, 2014; Tang et al., 2020; Lee et al., 2017; Battey et al., 2018), and nonparametric regression (Zhang et al., 2015; Zhao et al., 2016; Wang et al., 2019). However, this approach tends to be sub-optimal in most cases (Zhang et al., 2013). The second approach is an optimal iteration algorithm, which typically needs multiple rounds of communication and local calculations. As typical examples, Jordan et al. (2019) and Yu et al. (2026) developed a communication-efficient surrogate likelihood framework for solving distributed statistical inference problems. Motivated by the proximal point algorithm (Rockafellar, 1976), Fan et al. (2021) proposed two communication-efficient accurate statistical estimators to adapt to modern local sample sizes. In addition, Chen et al. (2022) proposed a new multi-round distributed estimation procedure that approximates the Newton step using only stochastic subgradient information. Given that communication cost is a major challenge in distributed frameworks, a key focus of such iterative approaches is to ensure convergence within a finite number of steps.

The aforementioned distributed algorithms are primarily based on the assumption that the data are independent and identically distributed (i.i.d.). In practice, however, data from different machines or sources often exhibit significant heterogeneity. In many evidence synthesis applications, parameter heterogeneity, where estimable parameters differ across studies or parameters of interest may be inestimable in certain studies, also frequently arises due to variations in populations, covariate designs, and outcome measures (Sutton and Higgins, 2008). For example, Liu et al. (2015) illustrated a broad range of parameter heterogeneity settings by considering K independent clinical trials:

$$Y_{ij} = \alpha_i + \beta_1 X_{ij} + \beta_2 Z_{ij} + \beta_3 Z_{ij} X_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, K, j = 1, \ldots, n_i, \tag{1}$$

where Y_{ij}, X_{ij}, Z_{ij} are the response, the treatment status (1/0 for treatment/control), and the covariate of interest (for example drug dosage), respectively, for the jth subject in the ith study. ϵ_{ij} is a normal error with variance σ_i^2, α_i is study-specific effect, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$ are the common parameters among all studies.

Case 1 (Heterogeneity in populations): Population heterogeneity, which usually refers to different geographic regions, age periods, disease states, or other characteristics of subjects across studies, may affect the effect size in practice and cause study-specific effects such as α_is in statistical modelling. Specific examples include the treatment

effect for certain sub-populations (Zarrouf et al., 2009) and panel regression models with individual effects (Hahn and Newey, 2004). Under this setting, it is easy to see that the study-specific effect α_i of the ith study is not estimable in other studies, and thus the other studies cannot be used directly to make inferences about α_i.

Case 2 (Heterogeneity in covariate designs): Since each study may have its own specific considerations and constraints, parameter heterogeneity can also arise from different covariate designs across studies (Sutton and Higgins, 2008; Liu et al., 2015). A special case is missing covariate designs (Simmonds and Higgins, 2007), that is, if Study i is not designed to examine the effect of the covariate Z_{ij}, then all values of Z_{ij} for subjects in that study are fixed at a constant value z_i. In this case, Model (1) simplifies to the following form:

$$Y_{ij} = (\alpha_1 + z_i\beta_2) + (\beta_1 + z_i\beta_3)X_{ij} + \epsilon_{ij}, \quad j = 1, \ldots, n_i.$$

Here, the estimable parameters become $\alpha_1 + z_i\beta_2$ and $\beta_1 + z_i\beta_3$, which are indeed different from the parameters in Model (1).

Case 3 (Heterogeneity in outcomes): Different types of outcome reports are another important reason for parameter heterogeneity (Parmigiani and Dominici, 2000; Liu et al., 2015). A typical example is provided by the report policies of blood loss in Whitehead et al. (1999). In that example, although the outcome "blood loss" was continuous, some studies reported only binary responses $d_{ij} = I(y_{ij} \geq \tau_i)$, indicating a "severe" or "not severe" state of blood loss, where τ_i denotes a pre-specified threshold. Under such circumstances, Model (1) reduces to a probit model as follows:

$$Pr(d_{ij} = 1) = \Phi\left(\frac{\alpha_i - \tau_i}{\sigma_i} + \frac{\beta_1}{\sigma_i}X_{ij} + \frac{\beta_2}{\sigma_i}Z_{ij} + \frac{\beta_3}{\sigma_i}X_{ij}Z_{ij}\right), \quad j = 1, \ldots, n_i.$$

Since σ_i is unknown, the parameters $(\alpha_i, \beta_1, \beta_2, \beta_3)^T$ are not estimable in the ith study, and the different types of outcome report actually cause the heterogeneity of parameters.

It is worth noting that, although the above three examples involve multiple different parameters and the parameter of interest may not be estimable in some studies, these heterogeneous parameters can indeed be linked through known mapping or some common parameters. Therefore, the information for one parameter may potentially impact the inference of other parameters, and excluding a subset of the studies may result in a non-negligible loss of information. In other words, such indirect evidence is also useful and can be used to increase the efficiency of statistical inference.

In addition to the aforementioned issue of parameter heterogeneity, another challenge arises from privacy constraints and related factors: individual-level data often cannot be directly transmitted between different machines, necessitating the exchange of aggregated or summarized information instead. For the above scenario, synthesizing information on parameters of interest across multiple studies, confidence distribution has emerged as a powerful tool due to its logistical convenience and statistical efficiency. A confidence distribution is typically defined as a sample-dependent distribution function that encapsulates

confidence intervals of all levels for the target parameters (Xie and Singh, 2013). Specifically, we obtain summary results, such as an estimator $\hat{\boldsymbol{\gamma}}$ of parameter $\boldsymbol{\gamma}$ and an estimated covariance matrix $\hat{\Sigma}$ for $\hat{\boldsymbol{\gamma}}$, from several studies. If we further assume that, theoretically, $\hat{\Sigma}^{-1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ converges in distribution to a standard normal distribution, then the multivariate normal distribution $MN(\hat{\boldsymbol{\gamma}}, \hat{\Sigma})$ can be regarded as a confidence distribution for $\boldsymbol{\gamma}$. Its corresponding density is commonly referred to as the confidence density. Building upon the framework of maximum likelihood inference, Xie et al. (2011) and Liu et al. (2015) also proposed a method of multiplying confidence densities to integrate multiple sources of summary information.

To address both of the aforementioned issues simultaneously, this article makes two main contributions. First, in scenarios where only partial individual-level data and external summary statistics from related studies are accessible, due to privacy constraints or data storage limitations common in big data and other fields, we develop a practical method for integrating a mixture of individual data and multiple heterogeneous summary information sources using confidence distributions. We establish the asymptotic properties of the proposed estimator and demonstrate that it effectively incorporates indirect evidence without efficiency loss compared to an estimator using all individual-level data. Second, we propose a communication-efficient distributed framework for statistical inference with massive heterogeneous auxiliary data. Specifically, we provide a rigorous theoretical analysis using the canonical generalized linear model as an illustrative example. We establish theoretical guarantees for the distributed algorithm under general conditions and support our findings through simulations and real data examples that validate both its theoretical properties and numerical performance.

The rest of this paper is organized as follows. Section 2 introduces a general framework for synthesizing the individual data and heterogeneous summary results, and establishes the corresponding asymptotic properties. Section 3 proposes a communication-efficient distributed statistical inference algorithm for massive data with heterogeneous auxiliary information, which can be applied to the distributed statistical inference of parameter-heterogeneous big data and even streaming data. The numerical simulation and real data applications are presented in Section 4 and 5, respectively. Section 6 concludes this paper and provides additional discussion.

2. Integrated analysis of a mixture of individual data and heterogeneous summary results

In this section, to address the issues mentioned above, we first introduce some notation to facilitate subsequent discussions. Then, we propose a general method that integrates individual-level data from the internal study and external heterogeneous summary statistics to improve estimation efficiency.

2.1 Setting and notation

Without loss of generality, assume multiple likelihood inference statistics are available from external studies or via published literature. Specifically, we obtain K independent external studies with n_k observations $\{\boldsymbol{v}_{kj}\}_{j=1}^{n_k}$ in the kth study, $k = 1, \cdots, K$. Let $h_k^*(\boldsymbol{v}, \boldsymbol{\gamma}_k)$ be the density function of the kth external model, where $\boldsymbol{\gamma}_k$ is a p_k-dimensional identifiable

parameter vector, such that $h_k^*(\boldsymbol{v}, \boldsymbol{\gamma}_k)$ cannot be represented by a lower-dimensional parameter vector. Moreover, the estimable $\boldsymbol{\gamma}_k$ is related to the target full parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ via a known smooth $R^p \to R^{p_k}$ mapping $\boldsymbol{\xi}_k$, that is, $\boldsymbol{\gamma}_k = \boldsymbol{\xi}_k(\boldsymbol{\theta})$. Here, we assume that the mapping function $\boldsymbol{\xi}_k$ is three times differentiable with respect to $\boldsymbol{\theta}$ for simplicity. Then for the kth study, by expressing the density function as $h_k(\boldsymbol{v}; \boldsymbol{\theta}) = h_k^*(\boldsymbol{v}; \boldsymbol{\gamma}_k)$, the corresponding likelihood function is given by

$$L_k(\boldsymbol{\theta}) = \prod_{j=1}^{n_k} h_k(\boldsymbol{v}_{kj}; \boldsymbol{\theta}) = \prod_{j=1}^{n_k} h_k^*(\boldsymbol{v}_{kj}; \boldsymbol{\gamma}_k) = L_k^*(\boldsymbol{\gamma}_k).$$

Let $\hat{\boldsymbol{\gamma}}_k = \arg\max_{\boldsymbol{\gamma}_k} L_k^*(\boldsymbol{\gamma}_k)$ be the maximum likelihood estimate of $\boldsymbol{\gamma}_k$ and denote as $\hat{\Sigma}_k$ the estimated covariance matrix of $\hat{\boldsymbol{\gamma}}_k$. From the theory of maximum likelihood estimation (MLE), this estimator is given by the inverse of the observed information matrix:

$$\hat{\Sigma}_k = [\Gamma_k(\hat{\boldsymbol{\gamma}}_k)]^{-1}, \text{ where } \Gamma_k(\hat{\boldsymbol{\gamma}}_k) = -\partial^2 \log L_k^*(\hat{\boldsymbol{\gamma}}_k)/\partial \boldsymbol{\gamma}_k \partial \boldsymbol{\gamma}_k^T.$$

In fact, the assumption on the estimator $\hat{\boldsymbol{\gamma}}_k$ can be relaxed: it suffices to consider any estimator $\check{\boldsymbol{\gamma}}_k$ satisfying $\|\check{\boldsymbol{\gamma}}_k - \hat{\boldsymbol{\gamma}}_k\|_1 = o_p(1/\sqrt{n_k})$. More generally, within the loss framework, one may also use any asymptotically normal and $\sqrt{n_k}$-consistent estimator. It is worth noting that we only assume that the auxiliary summary statistics $\hat{\boldsymbol{\gamma}}_k$ and $\hat{\Sigma}_k$ are available, rather than all individual data from various external studies. Following the asymptotic theory of MLE, we know that under some regularization conditions, $\sqrt{n_k}\,(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k) \to MN(0, I_k^{-1})$ in distribution and $\Gamma_k(\hat{\boldsymbol{\gamma}}_k)/n_k \to I_k$ in probability as $n_k \to \infty$, where $I_k = -\mathbb{E}(\partial^2 \log h_k^*(\boldsymbol{v}, \boldsymbol{\gamma}_k)/\partial \boldsymbol{\gamma}_k \partial \boldsymbol{\gamma}_k^T)$ is the $p_k \times p_k$ Fisher information matrix.

For the internal study, we assume that there are n independent individual samples $\{\boldsymbol{v}_j\}_{j=1}^n$ drawn from the population distribution $f^*(\boldsymbol{v}, \boldsymbol{\eta})$. Similar to the external model, the p_0-dimensional unknown parameter $\boldsymbol{\eta}$ is related to the common full parameter $\boldsymbol{\theta}$ via a known mapping $\boldsymbol{\xi}_0$, whose form is typically determined by the study design. Accordingly, the density $f^*(\boldsymbol{v}, \boldsymbol{\eta})$ can also be expressed as $f(\boldsymbol{v}, \boldsymbol{\theta})$. Throughout this paper, since an estimate of $\boldsymbol{\eta}$ can be directly derived from that of $\boldsymbol{\theta}$, we focus on statistical inference for the target parameter $\boldsymbol{\theta}$, which serves as a unifying framework, bridging estimable parameters across different studies through known mappings. Similarly, $I_0 = -\mathbb{E}(\partial^2 \log f^*(\boldsymbol{v}, \boldsymbol{\eta})/\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T)$ is the $p_0 \times p_0$ Fisher information matrix.

Remark 1 *Although the relationships between $\boldsymbol{\gamma}_k$ ($k = 1, 2, \cdots, K$), $\boldsymbol{\eta}$, and $\boldsymbol{\theta}$ are uniquely determined, the choices of the mappings $\boldsymbol{\xi}_k$ ($k = 0, 2, \cdots, K$) are not. This non-uniqueness allows us to select appropriate forms of $\{\boldsymbol{\xi}_k\}_{k=0}^K$ such that they are three times differentiable.*

To illustrate, consider $\boldsymbol{\xi}_0$ as an example. Suppose the relationship between $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ is non-smooth, for instance, $\boldsymbol{\eta} = \tilde{\boldsymbol{\xi}}_0(\boldsymbol{\theta})$, where $\tilde{\boldsymbol{\xi}}$ is a relu-type function. In such cases, we may define an augmented parameter vector $\breve{\boldsymbol{\theta}} = (\boldsymbol{\theta}^\top, \boldsymbol{\eta}^\top)^\top$ and construct a linear mapping:

$$\boldsymbol{\xi}_0(\breve{\boldsymbol{\theta}}) = (\underbrace{0, 0, \cdots, 0}_{p \text{ dimension}}, \underbrace{1, 1, \cdots, 1}_{p_0 \text{ dimension}})\breve{\boldsymbol{\theta}} = \boldsymbol{\eta}.$$

This constructed $\boldsymbol{\xi}_0$ satisfies the required smoothness assumption, and subsequent inference can be carried out in terms of $\breve{\boldsymbol{\theta}}$. Since $\boldsymbol{\theta}$ is a subvector of $\breve{\boldsymbol{\theta}}$, an estimator of $\boldsymbol{\theta}$ can be directly obtained from the corresponding components of the estimator of $\breve{\boldsymbol{\theta}}$.

2.2 Methodology

In this section, we propose combining a mixture of individual data and multiple heterogeneous summary information sources using the tool of confidence distribution. The confidence distribution (see Schweder and Hjort (2002), Singh et al. (2005), Xie and Singh (2013)), as a concept defined and interpreted under the frequentist framework, has made great progress in modern meta-analysis (Xie et al., 2011). Intuitively, it is often viewed as a sample-dependent distribution function that can represent confidence intervals of all levels for a parameter of interest. Following Liu et al. (2015), under the setting presented in Section 2.1, we can easily construct the confidence distribution $MN(\hat{\boldsymbol{\gamma}}_k, \hat{\Sigma}_k)$ for parameter $\boldsymbol{\gamma}_k$ based on the summary statistics $(\hat{\boldsymbol{\gamma}}_k, \hat{\Sigma}_k)$ of the kth external study. Then the density function of $MN(\hat{\boldsymbol{\gamma}}_k, \hat{\Sigma}_k)$ can be treated as a confidence density (the derivative of the confidence distribution) of $\boldsymbol{\gamma}_k$ in our research. Writing this confidence density as $d_k(\boldsymbol{\gamma}_k; \mathcal{S}_k)$, where $\mathcal{S}_k$ represents the sample in the kth external study,

$$d_k(\boldsymbol{\gamma}_k; \mathcal{S}_k) = \frac{1}{(2\pi)^{p_k/2}|\hat{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma}_k - \hat{\boldsymbol{\gamma}}_k)^T \hat{\Sigma}_k^{-1}(\boldsymbol{\gamma}_k - \hat{\boldsymbol{\gamma}}_k)\right\}, \quad k = 1, 2, \cdots, K. \tag{2}$$

In fact, under some mild conditions, these normal confidence densities are asymptotically proportional to the likelihood functions. For more detail, see Singh et al. (2007).

Leveraging the independence across studies, we combine individual-level data and auxiliary information by multiplying their confidence densities $\{d_k(\boldsymbol{\gamma}_k; \mathcal{S}_k)\}_{k=1}^K$ and likelihood function $L_0(\boldsymbol{\theta}) = \prod_{i=1}^n f(\boldsymbol{v}_i, \boldsymbol{\theta})$. More specifically, denote

$$L_A(\boldsymbol{\theta}) = L_0(\boldsymbol{\theta}) \times \prod_{k=1}^K d_k(\boldsymbol{\gamma}; \mathcal{S}_k) = \prod_{i=1}^n f(\boldsymbol{v}_i, \boldsymbol{\theta}) \times \prod_{k=1}^K d_k(\boldsymbol{\xi}(\boldsymbol{\theta}); \mathcal{S}_k).$$

By maximizing the optimization function $L_A(\boldsymbol{\theta})$, we can obtain an estimator $\tilde{\boldsymbol{\theta}}$ of parameter $\boldsymbol{\theta}$. This is called CD estimation in the following for simplicity. Equivalently,

$$\begin{aligned} \tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_A(\boldsymbol{\theta}) &= \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) - \frac{1}{2}\sum_{k=1}^K (\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)^T \hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^n \log f(\boldsymbol{v}_i, \boldsymbol{\xi}(\boldsymbol{\theta})) - \frac{1}{2}\sum_{k=1}^K (\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)^T \hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k), \end{aligned} \tag{3}$$

where the augmented log-likelihood function $\ell_A(\boldsymbol{\theta}) = \log L_A(\boldsymbol{\theta}) + c$ for some negligible constant c and $\ell(\boldsymbol{\theta}) = \log L_0(\boldsymbol{\theta})$ is the log-likelihood function of the individual data.

When all individual-level data of external models are available, we can also extract an MLE by maximizing the multiplied likelihood function $\prod_{k=0}^K L_k(\boldsymbol{\theta})$. Here, we call this the IPD estimator, $\hat{\boldsymbol{\theta}}_{IPD}$. Following the large-sample theory of MLE, we know that the IPD estimator is consistent and asymptotically normally distributed. This naturally raises a question: whether our CD estimator still retains the asymptotic properties and full efficiency of an IPD estimator. Theorem 2 demonstrates the desirable theoretical properties of the CD estimator $\tilde{\boldsymbol{\theta}}$. Here we define $\boldsymbol{\theta}^*$ as the true value of $\boldsymbol{\theta}$ for convenience.

Theorem 2 *Under the regularity conditions stated in Appendix B, the CD estimator $\tilde{\boldsymbol{\theta}}$ possesses the following properties: as $n \to \infty$,*

(a) *The CD estimator* $\tilde{\boldsymbol{\theta}}$ *is consistent and* $\sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ *asymptotically converges to a zero-mean normal distribution with limiting covariance matrix*

$$\left\{ \sum_{k=0}^{K} c_k J_k(\boldsymbol{\theta}^*)^T I_k J_k(\boldsymbol{\theta}^*) \right\}^{-1},$$

where $0 < c_k = \lim_{N\to\infty} n_k/N < 1$, $n_0 = n$, $N = \sum_{k=1}^{K} n_k + n$, $J_k(\boldsymbol{\theta}) = \partial \boldsymbol{\xi}_k(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ *is the Jacobian matrix of mapping* $\boldsymbol{\xi}_k$ *with respect to* $\boldsymbol{\theta}$, *and* $\boldsymbol{\theta}^*$ *is the true value of* $\boldsymbol{\theta}$. *Moreover, the asymptotic covariance matrix of* $\sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ *can be consistently estimated by*

$$N\{-\partial^2 \ell_A(\tilde{\boldsymbol{\theta}})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T\}^{-1}.$$

(b) *The CD estimator* $\tilde{\boldsymbol{\theta}}$ *is asymptotically as efficient as the IPD estimator* $\hat{\boldsymbol{\theta}}_{IPD}$. *For the special scenario* $\boldsymbol{\gamma}_k = (\alpha_k, \boldsymbol{\beta})$ *with study-specific parameters* α_k, $k = 0, \ldots, K$, *the CD estimator* $\tilde{\alpha}_k$ *is more asymptotically effective than the study-specific estimator* $\hat{\alpha}_k$ *from Study* k.

(c) *The CD approach is robust against misspecification of the covariance structure of parameter estimates in external studies. Specifically, if we use the working covariance matrix* $\hat{\Sigma}_{k,W}$ *in place of* $\hat{\Sigma}_k$ *in (3) for* $k = 1, \ldots, K$, *then the new estimator* $\tilde{\boldsymbol{\theta}}_W$ *of* $\boldsymbol{\theta}$ *remains consistent and asymptotically normally distributed with an adjusted limiting sandwich covariance matrix* $A^{-1}BA^{-1}$, *where* $A = c_0 J_0(\boldsymbol{\theta}^*)^T I_0 J_0(\boldsymbol{\theta}^*) + \sum_{k=1}^{K} c_k J_k(\boldsymbol{\theta}^*)^T A_k J_k(\boldsymbol{\theta}^*)$, $B = c_0 J_0(\boldsymbol{\theta}^*)^T I_0 J_0(\boldsymbol{\theta}^*) + \sum_{k=1}^{K} c_k J_k(\boldsymbol{\theta}^*)^T A_k I_k^{-1} A_k J_k(\boldsymbol{\theta}^*)$, *and the positive definite matrix* $A_k = \lim_{n_k\to\infty}(n_k\, \Sigma_{k,W})^{-1}$.

(d) *The asymptotic covariance matrix of* $\sqrt{N}(\tilde{\boldsymbol{\theta}}_W - \boldsymbol{\theta}^*)$ *can be consistently estimated by* $\hat{A}^{-1}\hat{B}\hat{A}^{-1}$, *where*

$$\begin{aligned}
\hat{A} =& \frac{1}{N}\sum_{i=1}^{n} \lambda(\boldsymbol{\nu}_i, \tilde{\boldsymbol{\theta}}_W)\lambda(\boldsymbol{\nu}_i, \tilde{\boldsymbol{\theta}}_W)^T + \frac{1}{N}\sum_{k=1}^{K} J_k(\tilde{\boldsymbol{\theta}}_W)^T \hat{\Sigma}_{k,W}^{-1} J_k(\tilde{\boldsymbol{\theta}}_W), \\
\hat{B} =& \frac{1}{N}\sum_{i=1}^{n} \lambda(\boldsymbol{\nu}_i, \tilde{\boldsymbol{\theta}}_W)\lambda(\boldsymbol{\nu}_i, \tilde{\boldsymbol{\theta}}_W)^T \\
&+ \frac{1}{N}\sum_{k=1}^{K} n_k^2 J_k(\tilde{\boldsymbol{\theta}}_W)^T \hat{\Sigma}_{k,W}^{-1} \left\{ J_k(\tilde{\boldsymbol{\theta}}_W) \left[\sum_{i=1}^{n} \lambda(\boldsymbol{\nu}_i, \tilde{\boldsymbol{\theta}}_W)\lambda(\boldsymbol{\nu}_i, \tilde{\boldsymbol{\theta}}_W)^T \right]^{-1} J_k(\tilde{\boldsymbol{\theta}}_W)^T \right\} \hat{\Sigma}_{k,W}^{-1} J_k(\tilde{\boldsymbol{\theta}}_W),
\end{aligned}$$

and

$$\lambda(\boldsymbol{\nu}, \boldsymbol{\theta}) = \frac{\partial \log f(\boldsymbol{\nu}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Furthermore, when $\hat{\Sigma}_{k,W}$ *is equal to* $\hat{\Sigma}_k$, *the asymptotic covariance matrix of* $\sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ *can be consistently estimated by*

$$\left[\frac{1}{N}\sum_{i=1}^{n} \lambda(\boldsymbol{\nu}_i, \tilde{\boldsymbol{\theta}})\lambda(\boldsymbol{\nu}_i, \tilde{\boldsymbol{\theta}})^T + \frac{1}{N}\sum_{k=1}^{K} J_k(\tilde{\boldsymbol{\theta}})^T \hat{\Sigma}_k^{-1} J_k(\tilde{\boldsymbol{\theta}}) \right]^{-1}.$$

Theorem 2 summarizes the asymptotic properties of the estimator $\tilde{\boldsymbol{\theta}}$ given in (3) and detailed proofs are provided in Appendix A. Part **(a)** shows that the CD estimator $\tilde{\boldsymbol{\theta}}$ is consistent and asymptotically normal, and proposes a consistent estimator for the limiting covariance matrix. Furthermore, using the Slutsky's theorem (Ferguson, 1996), the results in **(a)** can imply the large-sample theory for other parameters of interest, such as the $\boldsymbol{\eta}$ of the internal study. In fact, by following the proof of Theorem 2, it is easy to see that result **(a)** can also be extended to a general loss framework with external asymptotically normal $\sqrt{n_k}$-consistent estimation, but the asymptotic covariance matrix may be slightly different. Beyond its asymptotic properties, Theorem 2 demonstrates that our confidence distribution approach incorporates indirect evidence without efficiency loss compared to the IPD estimator $\hat{\boldsymbol{\theta}}_{\text{IPD}}$. As noted by Liu et al. (2015), the efficiency gain stems from two key features: (i) the reparameterization pools external indirect evidence via the mapping $\boldsymbol{\xi}_k$ between study-specific ($\boldsymbol{\gamma}_k$) and common ($\boldsymbol{\theta}$) parameters; and (ii) the confidence density (likelihood) captures within-study correlation, enabling information borrowing across studies. This cross-study integration refines common parameter estimates, thereby shrinking study-specific estimates toward their true values.

The robustness property established in the last statement of Theorem 2 greatly broadens the applicability of our approach and has important ramifications. In some practical situations, the covariance structure of parameter estimates may be misspecified or unknown. For example, some data publications may report only estimates of the individual variances, but not the full covariance matrix. In this scenario, our method is still valid. Moreover, Theorem 2 also provides us with some useful guidelines on how to choose a suitable and flexible working covariance matrix to achieve greater efficiency. More details on this can be found in Liang and Zeger (1986) or Liu et al. (2015). In fact, similar to the description of Liu et al. (2015), even naively using the identity matrix, the CD approach still gains efficiency via the incorporation of indirect evidence. This implies that if the estimated covariance matrix $\hat{\Sigma}_k$ is unavailable (though $\hat{\boldsymbol{\gamma}}_k$ is known), or if $\boldsymbol{\gamma}_k$ is high-dimensional, leading to difficulties in matrix inversion, we may take $\hat{\Sigma}_k$ as the $p_k \times p_k$ identity matrix. This significantly extends the applicability of (3) and improves computational efficiency in such special scenarios. Furthermore, Model (2) is theoretically founded on the assumption that $\boldsymbol{\gamma}_k$ is asymptotically distributed as multivariate normal. It is important to note, however, that this normality assumption is not strictly necessary for the validity of the proposed method. More generally, even when the distribution deviates from normality, the estimation procedure defined in (3) can still be interpreted within a generalized estimating equations (GEE) framework. This connection holds as long as $\hat{\boldsymbol{\gamma}}_k$ is an asymptotically unbiased estimator of $\boldsymbol{\gamma}_k$, under which condition Model (2) corresponds to a form of l_2 norm minimization. As established in GEE theory, such estimators maintain consistency and asymptotic normality under mild regularity conditions, thereby ensuring the theoretical robustness of the proposed approach.

3. Communication-efficient distributed statistical inference for massive data with heterogeneous auxiliary information

In this section, we extend the aforementioned method to distributed settings and establish corresponding theoretical guarantees. Subsequently, we illustrate the approach through a generalized linear model as a specific example to enhance readers' understanding.

3.1 General case

We now propose a computationally efficient distributed strategy to conduct statistical inference on parameter $\boldsymbol{\theta}$ (or further, the unknown parameter $\boldsymbol{\eta}$) of (3) for massive data, while taking into consideration the communication cost among the machines. It is evident that this framework can be extended to distributed statistical inference for big data with parameter heterogeneity, and even to streaming data settings. For simplicity, we assume that the full dataset $\mathcal{D}^{full} = \{\boldsymbol{v}_i\}_{i=1}^n$ is randomly and evenly divided into T disjoint subsets $\mathcal{D}_1, \ldots, \mathcal{D}_T$, each stored separately on T machines $\{\mathcal{M}_t\}_{t=1}^T$. That is, $\mathcal{D}^{full} = \cup_{t=1}^T \mathcal{D}_t$ and $\mathcal{D}_{t_1} \cap \mathcal{D}_{t_2} = \emptyset$ for any $t_1 \neq t_2$. Each subset has an equal sample size of $m = |\mathcal{D}_1| = |\mathcal{D}_2| = \cdots = |\mathcal{D}_T| = n/T$. Let $\mathcal{I}_t$ be the index set corresponding to the elements of $\mathcal{D}_t$ and $\mathcal{I}^{full} = \{1, \ldots, n\} = \cup_{t=1}^T \mathcal{I}_t$.

Define the local negative log-likelihood function $\ell_t(\boldsymbol{\theta})$ of the tth machine as

$$\ell_t(\boldsymbol{\theta}) = \sum_{i \in \mathcal{I}_t} \log f(\boldsymbol{v}_i, \boldsymbol{\xi}(\boldsymbol{\theta})).$$

Then the global log-likelihood function in Equation (3) equals $\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(\boldsymbol{v}_i, \xi(\boldsymbol{\theta})) = \sum_{t=1}^T \ell_t(\boldsymbol{\theta})$. This inspires us to approximate the $\ell(\boldsymbol{\theta})$ via a surrogate loss based on some local calculation results. By applying the Taylor expansion around any initial estimator $\bar{\boldsymbol{\theta}}$, we have

$$\frac{\ell(\boldsymbol{\theta})}{n} = \frac{\ell(\bar{\boldsymbol{\theta}})}{n} + \big\langle \frac{\nabla \ell(\bar{\boldsymbol{\theta}})}{n}, \boldsymbol{\theta} - \bar{\boldsymbol{\theta}} \big\rangle + \sum_{j=1}^{\infty} \frac{1}{j!} \frac{\nabla^j \ell(\bar{\boldsymbol{\theta}})}{n} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^{\otimes j},$$

where ∇ and ∇^j are the first and jth ($j \geq 2$) derivatives, respectively. Similar to Jordan et al. (2019), we can replace the global higher-order derivatives with local derivatives of any subdataset $\mathcal{D}_t$. That results in the following approximation of $\ell(\boldsymbol{\theta})$:

$$\frac{\ell(\boldsymbol{\theta})}{n} = \frac{\ell(\bar{\boldsymbol{\theta}})}{n} + \big\langle \frac{\nabla \ell(\bar{\boldsymbol{\theta}})}{n}, \boldsymbol{\theta} - \bar{\boldsymbol{\theta}} \big\rangle + \frac{\ell_t(\boldsymbol{\theta})}{m} - \frac{\ell_t(\bar{\boldsymbol{\theta}})}{m} - \big\langle \frac{\nabla \ell_t(\bar{\boldsymbol{\theta}})}{m}, \boldsymbol{\theta} - \bar{\boldsymbol{\theta}} \big\rangle. \tag{4}$$

Combining (3) and (4), and omitting some additive constants, we can construct a surrogate augmented log-likelihood function as

$$\tilde{\ell}_{A,t}(\boldsymbol{\theta}) = T \times \ell_t(\boldsymbol{\theta}) - \big\langle \boldsymbol{\theta}, T \times \nabla \ell_t(\bar{\boldsymbol{\theta}}) - \nabla \ell(\bar{\boldsymbol{\theta}}) \big\rangle - \frac{1}{2} \sum_{k=1}^{K} (\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)^T \hat{\Sigma}_k^{-1} (\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k), \tag{5}$$

for any $t = 1, \ldots, T$. Obviously, $\tilde{\boldsymbol{\theta}}$ is a fixed point of (5).

Using the communication-efficient surrogate likelihood (CSL) approximation on the first machine, Jordan et al. (2019) investigated a unified distributed statistical inference framework that treats the solution of the CSL on the first machine as a new estimate and iterates multiple steps until convergence. In their approach, only the first machine solved the optimization problem, while the other machines merely calculated local gradients and remained idle for most of the time. To fully use the power and effectiveness of a computer cluster, Fan et al. (2021) proposed a natural improvement in which all machines optimized

the corresponding objective functions in parallel, and the central processor then aggregated those results. Both of the aforementioned schemes required only the transmission of vector information between machines, making them communication-efficient. Moreover, motivated by the proximal point algorithm (Rockafellar, 1976), Fan et al. (2021) also developed communication-efficient accurate statistical estimators (CEASE) by adding a strict convex quadratic regularization function to adapt to the moderate local sample size and general initialization conditions. Similarly, we consider subtracting an adjusted regular term from the surrogate augmented log-likelihood function (5) to obtain a multi-step iterative estimator $\boldsymbol{\theta}_s$. More details are provided in Algorithm 1.

Algorithm 1 Communication-efficient distributed statistical inference with heterogeneous auxiliary information

Input: Initial estimator $\boldsymbol{\theta}_0$, number of iterations S, and regularization parameter α.

1: **for** $s = 0, 1, \ldots, S-1$ **do**

2: The central processor transmits the current iterate $\boldsymbol{\theta}_s$ to local machines $\{\mathcal{M}_t\}_{t=1}^T$;

3: Each machine computes the local gradient $\triangledown \ell_t(\boldsymbol{\theta}_s)$ at machine $\mathcal{M}_t$ and sends it to the central processor;

4: The central processor calculates the global gradient $\nabla \ell\big(\boldsymbol{\theta}_s\big) = \sum_{t=1}^T \nabla \ell_t\left(\boldsymbol{\theta}_s\right)$ and broadcasts it to each of the machines;

5: Each machine computes the maximizer $\boldsymbol{\theta}_{s+1,t}$ as

$$\boldsymbol{\theta}_{s+1,t} = \arg\max_{\boldsymbol{\theta}} \Big\{ \ell_t(\boldsymbol{\theta})/m - \big\langle\, \boldsymbol{\theta}, \triangledown \ell_t(\boldsymbol{\theta}_s)/m - \triangledown \ell(\boldsymbol{\theta}_s)/n \,\big\rangle - \frac{\alpha}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_s\|^2 - \frac{1}{2n}\sum_{k=1}^{K}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)^T \hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k) \Big\},$$

and sends it to the central processor.

6: The central processor computes $\boldsymbol{\theta}_{s+1} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{\theta}_{s+1,t}$.

7: **end for**

Output: Estimator $\boldsymbol{\theta}_S$.

In Algorithm 1, the regularization parameter α can be flexibly selected in practice. In each iteration, Algorithm 1 consists of two rounds of communication; one round involves parallel optimization, and the other is simply averaging. These two rounds of communication can make $\boldsymbol{\theta}_s$ closer to the global maximum $\tilde{\boldsymbol{\theta}}$ by a contraction factor. To demonstrate this, we first assume the following conditions for $\ell(\boldsymbol{\theta})/n$ or $F(\boldsymbol{\theta}) = \mathbb{E}[\log f(\boldsymbol{v}, \boldsymbol{\theta})]$.

Condition A (Homogeneity). For all $1 \le t \le T$ and $\boldsymbol{\theta} \in B(\tilde{\boldsymbol{\theta}}, R)$ with a positive constant R, there exists a universal constant $\delta > 0$ such that $\|\nabla^2 \ell_t(\boldsymbol{\theta})/m - \nabla^2 F(\boldsymbol{\theta})\|_2 \le \delta/2$ holds.

Condition A is a general measure of the similarity between $\nabla^2 F(\boldsymbol{\theta})$ and $\nabla^2 \ell_t(\boldsymbol{\theta})/m$. Following this condition, we know $\|\nabla^2 \ell(\boldsymbol{\theta})/n - \nabla^2 F(\boldsymbol{\theta})\|_2 \le \delta/2$, $\|\nabla^2 \ell_t(\boldsymbol{\theta})/m - \nabla^2 \ell(\boldsymbol{\theta})/n\|_2 \le \delta$, and $\|\nabla^2 \ell_{t_1}(\boldsymbol{\theta})/m - \nabla^2 \ell_{t_2}(\boldsymbol{\theta})/m\|_2 \le \delta$ for any $1 \le t_1, t_2 \le T$. Since a large value of m implies a small δ, this observation indicates that when m is sufficiently large, each local

empirical approximation of F should be as accurate as possible and concentrate closely around the mean. We thus refer to δ as a homogeneity parameter. In most cases of interest, the rate of δ can be explicitly determined and is primarily influenced by the local sample size.

Condition B (Strong convexity). The following two functions are ρ-strongly convex in $B(\tilde{\boldsymbol{\theta}}, R)$ for some positive constant ρ:

$$-\ell(\boldsymbol{\theta})/n + \frac{1}{2n}\sum_{k=1}^{K}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)^T\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k),$$

$$-F(\boldsymbol{\theta}) + \frac{1}{2n}\sum_{k=1}^{K}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)^T\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k).$$

Here, a convex function h defined on some convex open set $\Omega \subset R^p$ is said to be ρ-strongly convex if $h(\mathbf{y}) \geq h(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y}-\mathbf{x}\rangle + (\rho/2)\|\mathbf{y}-\mathbf{x}\|^2$ for $\forall \mathbf{x}$, $\mathbf{y} \in \Omega$ and all $\mathbf{g}$ in subdifferential set $\nabla h(\mathbf{x})$. There also exist other equivalent definitions such as $\langle \mathbf{g_y} - \mathbf{g_x}, \mathbf{y}-\mathbf{x}\rangle \geq \rho\,\|\mathbf{y}-\mathbf{x}\|_2$ for $\forall \mathbf{x}$, $\mathbf{y} \in \Omega$, $\mathbf{g_x} \in \nabla h(\mathbf{x})$, and $\mathbf{g_y} \in \nabla h(\mathbf{y})$. For more details, see, for example, Nesterov (2014). Under the homogeneity condition, this strong convexity assumption can be relaxed to the assumption that function $-F(\boldsymbol{\theta}) + \frac{1}{2n}\sum_{k=1}^{K}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)^T\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)$ (or $-\ell(\boldsymbol{\theta})/n + \frac{1}{2n}\sum_{k=1}^{K}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)^T\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)$) is ρ_1-strongly convex in $B(\tilde{\boldsymbol{\theta}}, R)$ with constant $\rho_1 > 0$. It is easy to see $\rho \geq \rho_1 - \delta/2$ in that scenario, and when n is large, the difference between ρ and ρ_1 is basically negligible. Moreover, note that the Hessian matrix of $\nabla^2[-F(\boldsymbol{\theta}) + \frac{1}{2n}\sum_{k=1}^{K}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)^T\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \succeq \lambda I_{p\times p}$ for some constant $\lambda > 0$ in most interesting problems, where $\succeq$ is used to denote positive semidefinite ordering (that is, $A \succeq B$ means $A - B$ is positive semidefinite). This local strong convexity condition is mild and easily met in practical situations.

Denote

$$\rho_0 = \sup\Bigg\{c \in [0, \rho] : \Big\{-\ell_t(\boldsymbol{\theta})/m + \frac{1}{2n}\sum_{k=1}^{K}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)^T\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)\Big\}_{t=1}^{T} \text{ are } c\text{-strongly convex in } B(\tilde{\boldsymbol{\theta}}, R)\Bigg\}.$$

According to both assumptions above, it is easy to see that $\max\{\rho - \delta, 0\} \leq \rho_0 \leq \rho$. Under Conditions A and B, we can present a contraction rate of optimization errors. To obtain a much stronger result, we assume the following additional condition, which can significantly enhance accuracy.

Condition C (Smoothness). There exists a positive constant M such that

$$\|\nabla^2 F(\boldsymbol{\theta}_1) - \nabla^2 F(\boldsymbol{\theta}_2)\|_2 \leq M\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2,\ \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in B(\tilde{\boldsymbol{\theta}}, R).$$

Under the assumptions above, the following Theorem 3 gives contraction guarantees for Algorithm 1.

Theorem 3 *Under the Conditions A and B, if the initial value $\boldsymbol{\theta}_0 \in B(\tilde{\boldsymbol{\theta}}, R/2)$ and $[\delta/(\rho_0+\alpha)]^2 < \rho/(\rho+2\alpha)$ hold, then for iterates $\{\boldsymbol{\theta}_s\}_{s=1}^S$ generated by Algorithm 1, we have*

$$\|\boldsymbol{\theta}_{s+1} - \tilde{\boldsymbol{\theta}}\|_2 \leq \|\boldsymbol{\theta}_s - \tilde{\boldsymbol{\theta}}\|_2 \cdot \frac{\frac{\delta}{\rho_0+\alpha} \cdot \sqrt{\rho^2+2\alpha\rho} + \alpha}{\rho+\alpha}, \quad 0 \leq s \leq S-1. \tag{6}$$

Furthermore, if Condition C holds, Algorithm 1 has a much stronger contraction property

$$\|\boldsymbol{\theta}_{s+1} - \tilde{\boldsymbol{\theta}}\|_2 \leq \|\boldsymbol{\theta}_s - \tilde{\boldsymbol{\theta}}\|_2 \cdot \frac{\frac{\delta}{\rho_0+\alpha} \cdot \sqrt{\rho^2+2\alpha\rho} \cdot \min\{1, \frac{\delta}{\rho+\alpha}(1+\frac{M}{\rho_0+\alpha})\|\boldsymbol{\theta}_s - \tilde{\boldsymbol{\theta}}\|_2\} + \alpha}{\rho+\alpha}. \tag{7}$$

Here, both the multiplicative factors in (6) and (7) are strictly less than 1.

The proof of Theorem 3 is provided in Appendix A.2. Under the general assumptions $\boldsymbol{\theta}_0 \in B(\tilde{\boldsymbol{\theta}}, R/2)$ and $[\delta/(\rho_0+\alpha)]^2 < \rho/(\rho+2\alpha)$, Theorem 3 guarantees Q-linear convergence (Nocedal and Wright, 2006) for Algorithm 1, that is, for sufficiently large s, there exists $r \in (0,1)$ such that $\|\boldsymbol{\theta}_{s+1} - \tilde{\boldsymbol{\theta}}\|_2 \leq r\|\boldsymbol{\theta}_s - \tilde{\boldsymbol{\theta}}\|_2$ for sequence $\{\boldsymbol{\theta}_s\}_{s=1}^S$ generated by Algorithm 1. Note that the condition $[\delta/(\rho_0+\alpha)]^2 < \rho/(\rho+2\alpha)$ is easy to enforce by choosing sufficiently large α; therefore, Algorithm 1 can be widely used in practice. Moreover, we only assume that the initial value $\boldsymbol{\theta}_0$ satisfies $\boldsymbol{\theta}_0 \in B(\tilde{\boldsymbol{\theta}}, R/2)$, which implies that the initial estimator need not be consistent. However, the accuracy of the initial estimator $\boldsymbol{\theta}_0$ can help in reducing the number of iterations.

The choices of the regularization parameter α and the number of iterations S are of importance in reality. Note that on the one hand, under the assumptions of Theorem 3, Fan et al. (2021) showed that if $\alpha \geq 4\delta^2/\rho$, the contraction factors in Theorem 3 are bounded by $\big(1-\rho/[10(\alpha+\rho)]\big)$, and thus finite iteration steps $S = O\big((1+\alpha/\rho)\log(\|\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}\|_2/\epsilon)\big)$ suffice to reach a statistical ϵ-accuracy. On the other hand, when δ/ρ is sufficiently small and $\alpha \leq C\delta^2/\rho$ for some constant C, the contraction factors in (6) and (7) become δ/ρ and $(\delta/\rho)^2$, respectively, which are of the same order of the contraction rates for the unregularized version (that is, $\alpha = 0$) in Fan et al. (2021). Moreover, we can also show that we only need at most $S = O\big(\log(\|\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}\|_2/\epsilon)/\log(\rho/\delta)\big)$ iterations to achieve ϵ-accuracy. By combining the facts above, we conclude that $\alpha \asymp \delta^2/\rho$ is a good choice to ensure Algorithm 1 converges quickly and robustly. The same statistical accuracy as for the estimators performed on the whole dataset $\mathcal{D}^{\text{full}}$, that is, ϵ-accuracy $\epsilon = 1/n^{1/2+\epsilon_0}$ (or equivalently $1/N^{1/2+\epsilon_0}$) with some small constant $\epsilon_0 > 0$, can be achieved within $S = O\big(\max(1, \delta^2/\rho^2)\cdot\log(\|\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}\|_2/\epsilon)\big)$ rounds of communication. In practice, S can also be chosen adaptively using more discriminative criteria, such as checking if the difference between successive iterates falls below a preset threshold.

Corollary 4 *The communication-efficient estimator $\boldsymbol{\theta}_S$ is consistent, and $\sqrt{N}(\boldsymbol{\theta}_S - \boldsymbol{\theta}^*)$ asymptotically converges to a zero-mean normal distribution with limiting covariance matrix $\tilde{A}^{-1}\tilde{B}\tilde{A}^{-1}$, where*

$$\tilde{A} = c_0 J_0^T(\boldsymbol{\theta}^*) I_0 J_0(\boldsymbol{\theta}^*) + c_0\alpha I + \sum_{k=1}^K c_k J_k^T(\boldsymbol{\theta}^*)\tilde{A}_k J_k(\boldsymbol{\theta}^*),$$

$$\tilde{B} = c_0 J_0^T(\boldsymbol{\theta}^*) I_0 J_0(\boldsymbol{\theta}^*) + \sum_{k=1}^K c_k J_k^T(\boldsymbol{\theta}^*)\tilde{A}_k J_k(\boldsymbol{\theta}^*),$$

where $\tilde{A}_k = \lim_{n_k\to\infty}(\hat{\Sigma}_k n_k)^{-1}$. *The consistent estimators* $\hat{\tilde{A}}$ *and* $\hat{\tilde{B}}$ *are similar to that in the Proof of Theorem 2:*

$$\hat{\tilde{A}} = \frac{1}{N}\sum_{t=1}^{T}\sum_{i\in\mathcal{I}_t}\lambda(\boldsymbol{\nu}_i,\boldsymbol{\theta}_S)\lambda(\boldsymbol{\nu}_i,\boldsymbol{\theta}_S)^T + c_o\alpha I + \frac{1}{N}\sum_{k=1}^{K} J_k(\boldsymbol{\theta}_S)^T\hat{\Sigma}_k^{-1}J_k(\boldsymbol{\theta}_S),$$
$$\hat{\tilde{B}} = \frac{1}{N}\sum_{t=1}^{T}\sum_{i\in\mathcal{I}_t}\lambda(\boldsymbol{\nu}_i,\boldsymbol{\theta}_S)\lambda(\boldsymbol{\nu}_i,\boldsymbol{\theta}_S)^T + \frac{1}{N}\sum_{k=1}^{K} J_k(\boldsymbol{\theta}_S)^T\hat{\Sigma}_k^{-1}J_k(\boldsymbol{\theta}_S).$$

From Corollary 4, we can see that when the $\hat{\Sigma}_k$ is replaced by any working matrix $\hat{\Sigma}_{k,W}$ with the same dimension, the consistency and asymptotic normality both hold, except that the limiting covariance matrix is $\tilde{\tilde{A}}^{-1}\tilde{\tilde{B}}\tilde{\tilde{A}}^{-1}$, where $\tilde{\tilde{A}} = c_0 J_0(\boldsymbol{\theta}^*)^T I_0 J_0(\boldsymbol{\theta}^*) + c_0\alpha I + \sum_{k=1}^{K} c_k J_k(\boldsymbol{\theta}^*)^T A_k J_k(\boldsymbol{\theta}^*)$ and $\tilde{\tilde{B}} = c_0 J_0(\boldsymbol{\theta}^*)^T I_0 J_0(\boldsymbol{\theta}^*) + \sum_{k=1}^{K} c_k J_k(\boldsymbol{\theta}^*)^T A_k I_k^{-1} A_k J_k(\boldsymbol{\theta}^*)$.

3.2 A special example of an internal study: a generalized linear model

In most statistical examples, the contraction factor in Algorithm 1 can be explicitly determined by finding the rate of convergence for the homogeneity parameter δ of the internal study. As illustrated in the following, we consider a classical generalized linear model

$$f(y_i;\boldsymbol{x}_i,\boldsymbol{\eta}^*) = \exp\left[\frac{y_i(\boldsymbol{x}_i^T\boldsymbol{\eta}^*) - b(\boldsymbol{x}_i^T\boldsymbol{\eta}^*)}{\phi} - c(y_i,\phi)\right], \quad i = 1,\ldots,n, \tag{8}$$

where y_i is the response, $\boldsymbol{x}_i$ is the p_0-dimensional covariate vector, $\boldsymbol{\eta}^*$ is the true value of the unknown p_0-dimensional parameter $\boldsymbol{\eta}$, b is some known convex function, for example, $b(t) = \log(1+e^t)$ in logistic regression and $b(t) = e^t$ in Poisson regression, $c(y_i,\phi)$ is a known normalized function, and ϕ is the dispersion parameter, which is simply assumed to be a constant or some irrelevant nuisance parameter in this section. Let $J_0(\boldsymbol{\theta}) = \partial\boldsymbol{\eta}/\partial\boldsymbol{\theta}$ be the $p_0 \times p$ Jacobian matrix of mapping $\boldsymbol{\eta} = \boldsymbol{\xi}(\boldsymbol{\theta})$ with respect to the p-dimensional $\boldsymbol{\theta}$. Simple algebra then yields

$$\frac{\partial \log f(y_i;\boldsymbol{x}_i,\boldsymbol{\eta})}{\partial\boldsymbol{\theta}} = J_0^T(\boldsymbol{\theta})\frac{\big(y_i - \nabla b(\boldsymbol{x}_i^T\boldsymbol{\eta})\big)\boldsymbol{x}_i}{\phi},$$
$$\frac{\partial^2 \log f(y_i;\boldsymbol{x}_i,\boldsymbol{\eta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = J_0^T(\boldsymbol{\theta})\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T\nabla^2 b(\boldsymbol{x}_i^T\boldsymbol{\eta})}{\phi}J_0(\boldsymbol{\theta}) + \frac{\partial\{J_0^T(\boldsymbol{\theta})\boldsymbol{x}_i\}}{\partial\boldsymbol{\theta}^T}\cdot\frac{\big(y_i - \nabla b(\boldsymbol{x}_i^T\boldsymbol{\eta})\big)}{\phi}, \tag{9}$$

and

$$\frac{\partial\{J_0^T(\boldsymbol{\theta})\boldsymbol{x}_i\}}{\partial\boldsymbol{\theta}^T} = \begin{bmatrix} \sum_{j=1}^{p_0}\nabla J_0^{(j,1)}(\boldsymbol{\theta})\boldsymbol{x}_i^{(j)} \\ \cdots \\ \sum_{j=1}^{p_0}\nabla J_0^{(j,p)}(\boldsymbol{\theta})\boldsymbol{x}_i^{(j)} \end{bmatrix}_{p\times p}.$$

Here, we use the notation $A^{(j,k)}$ to represent the (j,k)th component of the matrix A, and use similar definitions for the vectors.

Without loss of generality, we also assume that the covariate vectors $\boldsymbol{x}_i = (1, \boldsymbol{u}_i^T)^T \in \mathbb{R}^{p_0}$, where $\{\boldsymbol{u}_i\}_{i=1}^n \subseteq \mathbb{R}^{p_0-1}$ are n i.i.d. random samples with zero mean and covariance matrix D. To explicitly determine the rate for δ in Condition A, we apply some extra regularity assumptions.

Condition A* For a generalized linear model (8) and Jacobian matrix $J_0(\boldsymbol{\theta})$, assume that the following hold:

- **(a)** Suppose $C_1 \le \|D\|_2 \le C_2$ for some positive constants C_1 and C_2, and $\{D^{-1/2}\boldsymbol{u}_i\}_{i=1}^n$ are i.i.d. sub-Gaussian random vectors.
- **(b)** There exists a universal positive constant C_3 such that $|\nabla b(\cdot)|$, $|\nabla^2 b(\cdot)|$, and $|\nabla^3 b(\cdot)|$ can be bounded by C_3. The true values $\boldsymbol{\theta}^*$, $\boldsymbol{\eta}^*$ are also bounded.
- **(c)** For all $\boldsymbol{\theta} \in B(\tilde{\boldsymbol{\theta}}, R)$, the Jacobian matrices $J_0(\boldsymbol{\theta})$ satisfy that $\max_{j,k} |J_0^{(j,k)}(\boldsymbol{\theta})| \le C_4$ and $\max_{j,k} \|\nabla J_0^{(j,k)}(\boldsymbol{\theta})\|_\infty \le C_4$ for some constant C_4.

It is noteworthy that the covariate conditions (a) and (b) are both common regularity assumptions; see, for example, Fan et al. (2021). Condition (a) requires the sub-Gaussian random design to guarantee the sub-exponential tails of some quantity in (9); similar conditions are also given in Jordan et al. (2019), Fan et al. (2021), and other literature. Condition (b) is a mild condition for a generalized linear model, and assumption (c) is also easily met in practice.

Moreover, if we assume that $m \ge n^a$ with $0 < a \le 1$, we will show in the Appendix A.4 that

$$\sup_{1\le t\le T} \sup_{\boldsymbol{\theta}\in B(\tilde{\boldsymbol{\theta}},R)} \|\nabla^2 \ell_t(\boldsymbol{\theta})/m - \nabla^2 F(\boldsymbol{\theta})\|_2 \le O_p\Big(\sup_{1\le t\le T} \Big\| \frac{1}{m} \sum_{i\in\mathcal{I}_t} \boldsymbol{x}_i y_i - \mathbb{E}\big[\boldsymbol{x}_i y_i\big] \Big\|_2 \Big) \times \left[\sup_{\boldsymbol{\theta}\in B(\tilde{\boldsymbol{\theta}},R)} \max_{j,k} \big\| \nabla J_0^{(j,k)}(\boldsymbol{\theta}) \big\|_1 \right] + O_p\Big(\sqrt{\frac{\log n}{m}} \Big), \tag{10}$$

where $\sup_{\boldsymbol{\theta}\in B(\tilde{\boldsymbol{\theta}},R)} \max_{j,k} \big\| \nabla J_0^{(j,k)}(\boldsymbol{\theta}) \big\|_1$ is a bounded constant.

The general rate (10) of the homogeneity parameter δ tends to be $O_p(\sqrt{\log n/m})$ in many special situations. For example, in settings such as linear regression with normal noise or logistic regression models, the response variable is sub-Gaussian, thereby ensuring that the first term in (10) achieves a convergence rate $O_p(\sqrt{\log n/m})$. Another common scenario is the linear transformation between $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$, which means that the first quantity of (10) equals zero. Under those special situations and regularity condition B, we know the contraction factor of Algorithm 1 (taking $\alpha = O(\delta^2/\rho)$ or 0) becomes $O_p(\sqrt{\log n/m})$ or $O_p(\log n/m)$ for large m, and after only a finite $S = O\big(\frac{\log(\|\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}\|_2 \sqrt{n})}{\log(m/\log n)}\big)$ steps, the optimization error $\|\boldsymbol{\theta}_S - \tilde{\boldsymbol{\theta}}\|_2$ can become negligible compared to the statistical error $O_p(n^{-1/2})$. That is, the multi-step estimator $\boldsymbol{\theta}_S$ achieves the statistical efficiency of $\tilde{\boldsymbol{\theta}}$ and can be used for the inference of $\boldsymbol{\theta}$ or $\boldsymbol{\eta}$.

4. Simulation

This section comprises three numerical studies. The first two (Subsections 4.1 and 4.2) assess the empirical performance of our proposed approaches, while the third (Subsection 4.3) examines the impact of misspecification in the link mappings $\{\boldsymbol{\xi}_k\}_{k=1}^K$.

4.1 The performance of Model (3)

In the first example, we numerically verify the theoretical properties of our proposed CD approach and provide a comparison of several popular methods. We generate an $n_1 = 300$ internal-study sample $\left\{\left(y_i^{(1)}, \mathbf{X}_i^{(1)} = (X_{1,i}^{(1)}, X_{2,i}^{(1)}, X_{3,i}^{(1)})\right)^T\right\}_{i=1}^{n_1}$ from the fixed effect model

$$y_i^{(1)} = \alpha_1 + \beta_1 X_{1,i}^{(1)} + \beta_2 X_{2,i}^{(1)} + \beta_3 X_{3,i}^{(1)} + \epsilon_i^{(1)}, \quad i = 1, \ldots, n_1,$$

where $X_{1,i}^{(1)} \sim Ber(1, 0.5)$, $X_{2,i}^{(1)} \sim U(0,1)$, $X_{3,i}^{(1)} \sim N(2,1)$, and $\epsilon_i^{(1)} \sim N(0,3)$. Furthermore, consider that we can also derive some summary statistics $\left\{\left(\hat{\alpha}_2, \hat{\boldsymbol{\beta}}_2 = (\hat{\beta}_{1,2}, \hat{\beta}_{2,2}, \hat{\beta}_{3,2})\right)^T\right\}$ and corresponding covariance matrix estimation from the external Study 2:

$$y_i^{(2)} = \alpha_2 + \beta_1 X_{1,i}^{(2)} + \beta_2 X_{2,i}^{(2)} + \beta_3 X_{3,i}^{(2)} + \epsilon_i^{(2)}, \quad i = 1, \ldots, n_w,$$

where $n_2 = 300$, $X_{1,i}^{(2)} \sim Ber(1, 0.5)$, $X_{2,i}^{(2)} \sim U(0,1)$, $X_{3,i}^{(2)} \sim N(2,1)$, but $\epsilon_i^{(2)} \sim N(0,1)$. Similarly, we can get the common parameter results from the external Study 3:

$$y_i^{(3)} = \alpha_3 + \beta_1 X_{1,i}^{(3)} + \beta_2 X_{2,i}^{(3)} + \beta_3 X_{3,i}^{(3)} + \epsilon_i^{(2)}, \quad i = 1, \ldots, n_3,$$

where $n_3 = 300$, $X_{1,i}^{(3)} \sim Ber(1, 0.5)$, $\epsilon_i^{(3)} \sim N(0,3)$, but $X_{2,i}^{(2)} \sim U(1/3, 2/3)$, $X_{3,i}^{(2)} \sim N(3,2)$. Moreover, suppose that we can also obtain auxiliary summary estimation information from another Study 4 with fixed $X_{1,i}^{(4)} = 1$, that is,

$$y_i^{(4)} = (\alpha_1 + \beta_1) + \beta_2 X_{2,i}^{(4)} + \beta_3 X_{3,i}^{(4)} + \epsilon_i^{(4)}, \quad i = 1, \ldots, n_4.$$

Here we also let $X_{2,i}^{(4)} \sim U(0,1)$, $X_{3,i}^{(4)} \sim N(2,1)$, but $\epsilon_i^{(4)} \sim N(0, 3/2)$ and $n_4 = 200$. Taking $\alpha_1 = 1$, $\alpha_2 = 2$, $\alpha_3 = -1$, and $\beta_1 = 1$, $\beta_2 = 2$, $\beta_3 = -1$, the simulation results based on 1000 replications are summarized in Table 3.

Table 3 reports the average parameter estimates (Mean), empirical standard error (SE), and the average standard error estimates (SEE) of several popular methods. Here, the Individual method only involves data sources from Study 1; the IPD method uses all individual-level data to make statistical inferences for all parameters $(\alpha_1, \alpha_2, \alpha_3, \boldsymbol{\beta}^T)^T$; the proposed CD method is the approach developed in Section 2; and Liu's CD method is the heterogeneous multivariate meta-analysis described in Liu et al. (2015). Specifically, Liu et al. (2015) first obtained the summary statistics from each data source based on likelihood inference, and then maximized the product confidence density function to obtain an estimate of the parameter of interest. It can be found directly that our proposed CD method performs well. In particular, the point estimates are nearly unbiased, and the SEE agree with SE quite well. Table 3 also provides a comparison of various approaches. The numerical performance of our proposed CD method is quite close to that of the IPD method, even for moderate

sample sizes, and our proposed CD estimates have smaller variance than those of the Individual method. We also note that Liu's CD method provides similar numerical performance to our proposed CD method; the main differences between these two approaches are that our developed framework can be applied to communication-efficient iterative distributed statistical inference.

Table 1: Integrated analysis for individual data and heterogeneous summary results.

	Individual method			Proposed CD method			IPD method			Liu's CD method		
Parameters	Mean	SE	SEE	Mean	SE	SEE	Mean	SE	SEE	Mean	SE	SEE
$\alpha_1 = 1$	0.9990	0.3020	0.2994	0.9976	0.1395	0.1354	0.9982	0.1368	0.1356	0.9982	0.1367	0.1351
$\alpha_2 = 2$	NA	NA	NA	1.9984	0.1257	0.1215	1.9983	0.1245	0.1218	1.9986	0.1245	0.1214
$\alpha_3 = -1$	NA	NA	NA	-1.0016	0.1675	0.1632	-1.0015	0.1654	0.1637	-1.0015	0.1654	0.1631
$\beta_1 = 1$	1.0018	0.2009	0.1992	1.0008	0.0860	0.0843	1.0007	0.0867	0.0845	1.0007	0.0866	0.0842
$\beta_2 = 2$	1.9977	0.3484	0.3457	2.0039	0.1509	0.1476	2.0034	0.1492	0.1480	2.0034	0.1492	0.1474
$\beta_3 = -1$	-0.9985	0.1022	0.0999	-1.0001	0.0328	0.0326	-1.0001	0.0332	0.0327	-1.0000	0.0332	0.0325

Furthermore, we examine the effect of sample size on the performance of the method. For convenience, in the four studies mentioned above, we set $n_1 = n_2 = n_3 = n_4 = n$ here, while keeping all other settings unchanged. Since the results across different dimensions show a convergent trend, we only present the outcomes for β_2. Results for other parameters are available upon request. As illustrated in Figure 1, the latter three methods consistently demonstrate superior performance over the Individual method across all sample sizes. This superiority is quantitatively supported by multiple metrics: a lower SE reflecting reduced estimation variance, mean estimates closer to the true parameter values, and a closer agreement between the SEE and SE, indicating enhanced estimation accuracy. These results lead to the conclusion that incorporating additional information, whether through auxiliary summary statistics or direct integration of external individual-level data, consistently improves statistical efficiency in practical applications, irrespective of sample size. Furthermore, the improvement in statistical efficiency achieved by the three methods above gradually diminishes as the sample size grows. This trend highlights the practical advantage of our method in real-world settings where internal data are scarce: under such information-limited conditions, our approach offers significantly enhanced decision reliability. Furthermore, the performance of our proposed CD method remains highly consistent with that of the method using full external individual-level data, regardless of the sample size. This robustness to varying sample sizes confirms the practical utility of our method and supports its broader application in data-sensitive scenarios.

4.2 The performance in the distributed cases

A second experiment was designed to evaluate the numerical performance of the proposed distributed statistical inference algorithm and the effect of local sample size. For the internal study, we generate the response y_i given the covariates $\boldsymbol{x}_i$ from the logistic regression:

$$\mathbb{P}(y_i = 1) = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\eta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\eta}}}, \quad i = 1, \ldots, n = 10000,$$

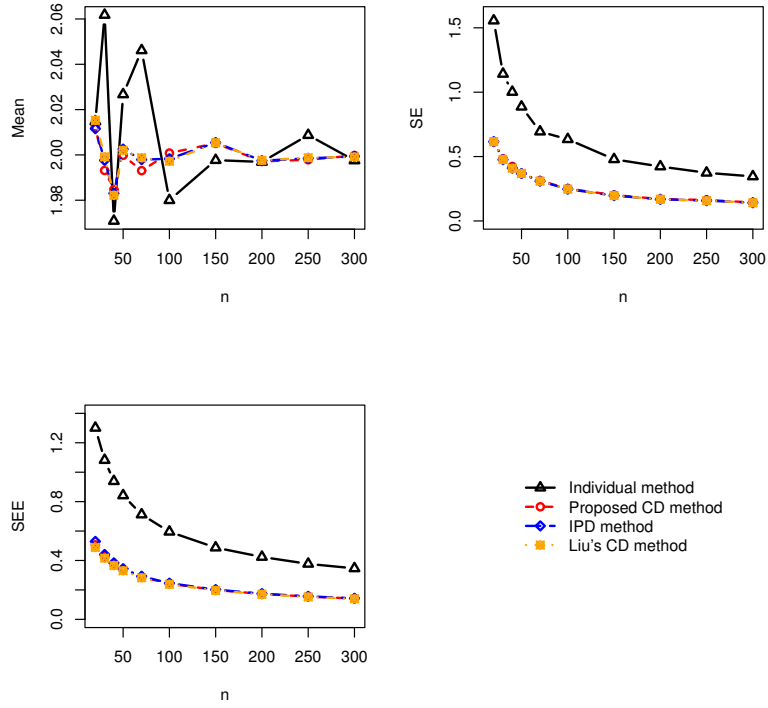

Figure 1: The effects of sample size on the performance of the method.

where $\boldsymbol{\eta} = (\alpha_0, \boldsymbol{\beta}^T)^T$ is a six-dimensional parameter vector, $\boldsymbol{x}_i = (1, \boldsymbol{u}_i^T)^T$, and $\boldsymbol{u}_i$ are generated from the multivariate normal population

$$MN\left(\mathbf{0}, \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{bmatrix}\right).$$

There also exist two types of heterogeneous external auxiliary information. For the first type, we consider three study-specific datasets, and the $n_k = 10000$ i.i.d. random samples of the kth dataset are generated as follows: $\boldsymbol{x}_{k,i} = (1, \boldsymbol{u}_{k,i}^T)^T$ with $\boldsymbol{u}_{k,i} \sim MN(\mathbf{0}, k \times \mathrm{I}_{5\times 5})$, and

$$\mathbb{P}(y_{k,i} = 1) = \frac{e^{\boldsymbol{x}_{k,i}^T \boldsymbol{\gamma}_k}}{1 + e^{\boldsymbol{x}_{k,i}^T \boldsymbol{\gamma}_k}}, \quad i = 1, \ldots, n_k;\ k = 1, \ldots, 3,$$

where $\boldsymbol{\gamma}_k = (\alpha_k, \boldsymbol{\beta}^T)^T$ is a six-dimensional vector with scalar quantity α_k for $k = 1, \ldots, 3$; and $\mathrm{I}_{5\times 5}$ stands for the 5×5 identity matrix. Under the logistic regression model of the internal study, we further construct three heterogeneous studies with different covariate designs as the second type of external model. Specifically, we assume for Study 4, the variable $x^{(1)}$ is fixed at 0.5. Then, the corresponding true model in the case of $n_4 = 5000$

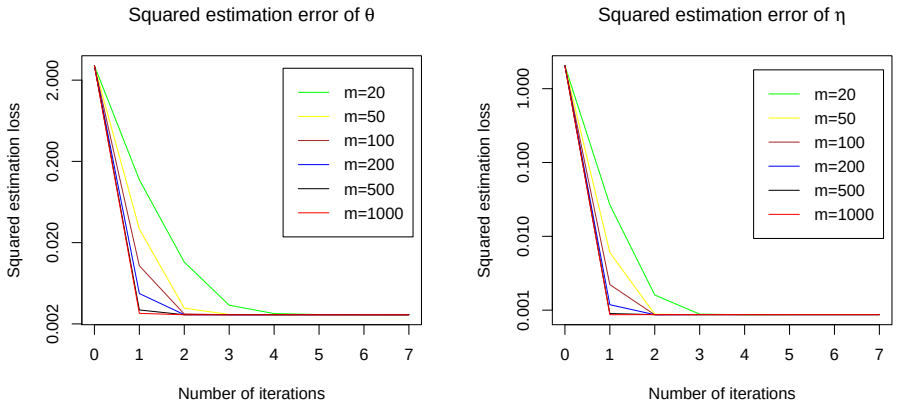

Figure 2: The squared estimation error $\|\boldsymbol{\theta}_s - \boldsymbol{\theta}^*\|_2^2$ and $\|\boldsymbol{\eta}_s - \boldsymbol{\eta}^*\|_2^2$ versus the number of iterations for logistic regression. The x-axis and y-axis are the number of iterations and the averaged squared estimation loss of 100 independent simulations. With a fixed total sample size $n = 10000$ and initial value $\boldsymbol{\theta}_0 = \mathbf{0}$, the dashed lines show the errors under various local sample sizes m.

becomes

$$\mathbb{P}(y_i = 1) = \frac{e^{(\alpha_0 + 0.5 * \beta^{(1)}) + \beta^{(2)} x_i^{(2)} + \beta^{(3)} x_i^{(3)} + \beta^{(4)} x_i^{(4)} + \beta^{(5)} x_i^{(5)}}}{1 + e^{(\alpha_0 + 0.5 * \beta^{(1)}) + \beta^{(2)} x_i^{(2)} + \beta^{(3)} x_i^{(3)} + \beta^{(4)} x_i^{(4)} + \beta^{(5)} x_i^{(5)}}}.$$

Similarly, suppose $x^{(2)}$ is fixed at 0.2 for Study 5 with $n_5 = 5000$ samples, and $x^{(1)} \equiv 0.5$, $x^{(2)} \equiv 0.2$ for Study 6 with $n_6 = 5000$ samples. The other settings of those external studies with heterogeneous covariate designs are the same as in the internal study.

Taking $\mathbf{0}$ as the initial value and $\alpha = 0.15p/m$ with dimension $p = 9$, we run 100 independent trials to estimate the full parameter $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \alpha_3, \alpha_0, \boldsymbol{\beta}^T)^T$ under the previous simulation setup. For each replication of the simulation, we choose the local sample size m to be $20, 50, 100, 200, 500$, or 1000, and sample the true value of $\boldsymbol{\theta}$ uniformly from the nine-dimensional unit cube $[0, 1]^9$. The averaged results of the squared estimation error and the effects of the local sample size are shown in Figure 2. As we can see, our proposed algorithm performs well and converges rapidly even with medium or small local sample sizes. We also compare our proposed estimate with heterogeneous external results and the CEASE estimate (Fan et al., 2021) without any auxiliary information in Table 2. As expected, the heterogeneous auxiliary results do improve the statistical accuracy and further accelerate the convergence of the CEASE algorithm, mainly because the quadratic term of the heterogeneity part enhances the strong convexity of the optimization function.

4.3 The impacts of potential misspecification in link mappings

Although we assume that the link mappings are known and explicit, in practice, misspecification may arise due to misunderstandings during information exchange or intentional concealment by external information owners for reasons such as competitive advantage. In

Table 2: The squared estimation error results for logistic regression

	Estimation with heterogeneous results				Estimation without auxiliary information			
Iterations	$m = 20$	$m = 50$	$m = 200$	$m = 1000$	$m = 20$	$m = 50$	$m = 200$	$m = 1000$
0	2.035903	2.035903	2.035903	2.035903	2.035903	2.035903	2.035903	2.035903
1	0.025974	0.006061	0.001287	0.000915	0.139612	0.103122	0.016164	0.003717
2	0.001691	0.000934	0.000894	0.000893	0.010130	0.004930	0.003193	0.003144
3	0.000933	0.000894	0.000894	0.000894	0.003606	0.003204	0.003143	0.003142
4	0.000896	0.000894	0.000894	0.000894	0.003227	0.003150	0.003143	0.003142
5	0.000894	0.000894	0.000894	0.000894	0.003157	0.003143	0.003142	0.003142
6	0.000894	0.000894	0.000894	0.000894	0.003148	0.003143	0.003142	0.003142
7	0.000894	0.000894	0.000894	0.000894	0.003144	0.003142	0.003142	0.003142
8	0.000894	0.000894	0.000894	0.000894	0.003143	0.003142	0.003142	0.003142
9	0.000894	0.000894	0.000894	0.000894	0.003142	0.003142	0.003142	0.003142

this subsection, we examine the impact of misspecification in the link mappings and discuss the applicability of the method under such conditions.

The experimental setup in this subsection aligns closely with that described in Section 4.1. Study 1 is designated as the internal study, while Study 2 is incorporated as an external source, utilizing the link mapping $\boldsymbol{\xi}_2$. We evaluate and compare three methodologies: the Individual method, the proposed CD method, and the IPD method. In the case of the proposed CD method, the following three scenarios are considered:

- **Scenario I**: The link mapping $\boldsymbol{\xi}_2$ is correctly specified.
- **Scenario II**: The model in Study 2 is misspecified as follows:

$$\tilde{y}_i^{(2)} = \alpha_2 + \tilde{\beta}_1 X_{1,i}^{(2)} + \beta_2 X_{2,i}^{(2)} + \beta_3 X_{3,i}^{(2)} + \epsilon_i^{(2)}, \quad i = 1, \ldots, n_w.$$

  That is, β_1 is mistaken for $\tilde{\beta}_1$, and thus the covariate corresponding to it is omitted in the link mapping $\boldsymbol{\xi}_2$, while all other components remain correctly specified.

- **Scenario III**: The model in Study 2 is misspecified as:

$$\tilde{y}_i^{(2)} = \alpha_2 + (\alpha_1 + \beta_2) X_{1,i}^{(2)} + \beta_2 X_{2,i}^{(2)} + \beta_3 X_{3,i}^{(2)} + \epsilon_i^{(2)}, \quad i = 1, \ldots, n_w.$$

  Here, the coefficient for $X_1^{(2)}$ is misdefined as $\alpha_1 + \beta_1$ instead of β_1, while the rest of the model remains complete.

The Mean and SEE are shown in Table 3. A closer examination of the misspecification effects reveals that, despite errors in some dimensions, the estimates of correctly specified parameters, such as β_2 and β_3, remain consistent. Specifically, compared to the Individual method, our proposed method achieves a lower SEE in the correctly specified dimensions, regardless of whether the link mappings are specified correctly or incorrectly. This indicates that even partially misspecified external information can still enhance statistical inference. Furthermore, when compared to the IPD method, the performance of our method under Scenario I is comparable. Although it performs slightly worse in the other two scenarios, the differences are generally modest, demonstrating the robustness of the CD-based approach.

A more detailed analysis of the misspecification effects is now discussed. When a subset of external information is omitted (as in Scenario II), the improvement in estimation for the remaining correctly specified dimensions remains consistent with that under full correct specification. This implies that when the dimensions are independent, omitting partial information does not adversely affect the performance of the remaining dimensions. However, when the relationship between dimensions is misspecified (that is, β_1 is incorrectly defined as $\alpha_1 + \beta_1$), these dimensions become interdependent, leading to deteriorated estimation performance for both. In such cases, the Mean result may be even worse than that of the Individual method. These simulation results highlight that external information must originate from reliable sources. When uncertainty exists, it is preferable to discard questionable information and use only verified data to enhance estimation accuracy.

Table 3: The effects of potential misspecification in link mapping.

	Individual method		Proposed CD method (Scenario I)		Proposed CD method (Scenario II)		Proposed CD method (Scenario III)		IPD method	
Parameters	Mean	SEE	Mean	SEE	Mean	SEE	Mean	SEE	Mean	SEE
$\alpha_1 = 1$	0.9891	0.2907	0.9966	0.1806	0.9928	0.1956	0.7269	0.1781	0.9957	0.1739
$\alpha_2 = 2$	NA	NA	1.9956	0.1591	1.9963	0.1620	1.9420	0.1632	1.9946	0.1523
$\beta_1 = 1$	1.0081	0.1943	1.0006	0.1009	1.0080	0.1939	0.4488	0.1836	1.0015	0.0984
$\beta_2 = 2$	2.0096	0.3319	2.0056	0.1729	2.0055	0.1731	2.0034	0.1734	2.0059	0.1676
$\beta_3 = -1$	-0.9984	0.0983	-0.9993	0.0529	-0.9993	0.0529	-0.9233	0.0474	-0.9991	0.0508

5. Real data application

In this section, three real examples are shown to illustrate the feasibility of our proposed algorithm. The first is a medical case with accessible individual-level data, while the second and the third concern climate monitoring and survival effects with data stored on different machines by year.

5.1 A medical case

Bronchial asthma caused by genetic and environmental factors is a common and frequently occurring disease that seriously impacts daily life through missed work, restricted activities, economic burdens, and other mechanisms. An asthma test allows patients to avoid allergens, including smoke, perfume, paint, and pets, preventing acute and potentially fatal asthma attacks. Thus, it is necessary to determine if a person has asthma based on a limited set of indicators.

The individual-level data are drawn from medical records, in which the response variable is equal to one if a person has asthma and zero otherwise. There are four types of asthma: eosinophilic asthma (EA), neutrophilic asthma (NA), paucigranulocytic asthma (PA), and mixed granulocytic asthma (MA). The initial experiment focuses on the first two types of asthma to study how the level of interleukin 5 (IL-5) influences asthma. Then, the percentage of the sputum eosinophils (PE) is also considered in studying PA. The sample sizes of the three experiments for EA, NA, and PA are 37, 26, and 24, while the sample size of the control group is 30.

Because of the heterogeneity between different types of asthma, we consider the following logistic model to combine the studies from EA, NA, and PA:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_i + \beta_1 x_{1ij} + \beta_2 x_{2ij},$$

where $i = 1, 2, 3$, $j = 1, 2, \cdots, n_i$. Here, p_{ij} is the probability of asthma for the jth observation in the ith dataset ($i = 1$, 2, 3 for PA, EA, and NA, respectively), and x_{1ij} and x_{2ij} are the performance measures of IL-5 and PE. Four methods mentioned in the first simulation experiment are compared here, and the estimation results are shown in Table 4. First, Liu's method performs the worst, as it exhibits an excessively large standard deviation in estimation, and all five of its estimated variables are statistically non-significant, a finding that contradicts both the underlying facts and the results from other methods. Second, both our proposed method and the IPD method outperform the individual-level method, as they yield smaller standard deviations. This improvement can be attributed to the incorporation of additional information in both approaches. Finally, the performance of our proposed method is comparable to that of the IPD method, indicating that the summary information is used appropriately in our approach, thereby achieving a similar effect as using individual-level data. All these empirical findings align well with our theoretical conclusions.

Table 4: The estimation results in the asthma case.

Parameters	Individual method	Proposed CD method	IPD method	Liu's CD method
α_1	-8.32(3.16)	-6.87(2.37)	-7.50(2.56)	-8.32 (3667.42)
α_2	NA	-6.36(2.29)	-6.99(2.47)	-9.39(1753.70)
α_3	NA	-5.77(1.88)	-6.37(2.05)	-3.79(1506.51)
β_1	0.85(0.66)	0.85(0.87)	0.85(0.89)	0.85(3795.78)
β_2	0.35(0.15)	0.28(0.09)	0.31(0.10)	1.29(128.93)

The prediction accuracy in the test dataset, comprising 10 patients with PA and 10 healthy individuals, is illustrated in Table 5 to demonstrate the model's performance. Three criteria are considered here, and their specific definitions are as follows:

1. accuracy= $\frac{TP+TN}{TP+TN+FP+FN}$,
2. precision= $\frac{TP}{TP+FP}$,
3. recall= $\frac{TP}{TP+FN}$,

where TP, FP, FN, and TN are listed in the following table. As expected, the performance

	Actual		
Predicted		Positive	Negative
	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

of our proposed method is nearly identical to that of the IPD method, and both yield more accurate prediction results than the Individual method. In contrast, Liu's method incorrectly classifies the entire population as patients, rendering its predictions unreliable. These results demonstrate that appropriately incorporating summarized heterogeneous information can significantly enhance the performance of individual methods.

Table 5: The accuracy of the four methods in the test dataset.

Criterion	Individual method	Proposed CD method	IPD method	Liu's CD method
Accuracy	0.7	0.75	0.75	0.5
Precision	0.67	0.73	0.73	0.5
Recall	0.8	0.8	0.8	1

5.2 A climate case

For illustration purposes, historical climate data in Canada is considered in this subsection. The data are available at https://climate.weather.gc.ca/. The dataset contains detailed hourly weather conditions from 2005 to 2015, and its sample size is larger than 8000 per year. For privacy reasons and actual data management, the data is usually stored on different machines by year; that is to say, Algorithm 1 should be used for exploratory analytic work.

Our goal is to predict the humidity (y) given the other three continuous covariate variables: temperature (X_1), atmospheric pressure (X_2), and wind speed (X_3). Analysis results from the observed sample are reported by studies according to the location of the sensors. For simplicity, we illustrate the application of our approach to this project using only three studies. The heterogeneous information is derived from the summary statistics of Wiarton and Thunder Bay, while the individual-level data pertains to climate information from Barrie. To combine the three studies, we consider the following linear model:

$$y_{ij} = \alpha_i + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \epsilon_{ij},$$

where $i = 1, 2, 3$ ($i = 1$, 2, 3 for Barrie, Wiarton, and Thunder Bay, respectively), $j = 1, 2, \cdots, n_i$. A total of 8000 samples per year are randomly chosen to construct the training set, that is to say, $T = 11$ and $m = 8000$ in Algorithm 1, while the rest of the samples are reserved for the test set. We compare our proposed methods with two counterparts: the traditional Individual method and the conventional communication-efficient distributed statistical inference (CDS) method, which are applied to individual-level data only. Table 6 illustrates the performance of the above linear model, in which the adjusted R^2 ($R^2 \in [0, 1]$) is applied to show how much of the variance of the data is explained by the model. From Table 6, our proposed method is superior in two aspects. First, the prediction performance of our proposed method (S=5) is better than that of the other two methods since it has the largest value of the adjusted R^2. Second, in the distributed case, our proposed method converges more rapidly than CDS. In more detail, the value of the adjusted R^2 in our proposed method increases with the number of iterations. Furthermore, the difference in

Table 6: The performance of three methods in the climate case.

Parameters	Individual method	CDS (S=1)	CDS (S=3)	CDS (S=5)	Proposed method (S=1)	Proposed method (S=3)	Proposed method (S=5)
α_1	849.16	736.29	823.74	868.43	724.19	850.67	865.96
α_2	NA	NA	NA	NA	714.57	858.80	878.69
α_3	NA	NA	NA	NA	694.364	856.36	875.27
β_1	-0.32	0.80	-0.40	-0.44	1.04	-0.38	-0.48
β_2	-0.79	-0.56	-0.79	-0.76	-0.69	-0.80	-0.89
β_3	-7.77	-6.36	-7.54	-8.00	-6.36	-7.82	-7.94
Adjusted R^2	0.66	0.06	0.66	0.59	0.05	0.65	0.69

Table 7: Variable description for the Medical Birth Registry of Norway data.

Variable	Description	Variable used in the model
y	The time between the first and second births.	Used as the response variable
X_1	The mother's age at first birth.	Used as a numerical variable
X_2	The baby's sex. 1 for male; 0 for female.	Converted to a dummy variable
ϖ	The censored status. 0 for censored data; 1 otherwise.	Converted to a dummy variable

our proposed method between the two cases, which are $S = 3$ and $S = 5$, is smaller than that in CDS. In short, the effective use of heterogeneous information is beneficial in enhancing the effectiveness of statistical inference.

5.3 A censored data case

In this subsection, we use a dataset obtained from the Medical Birth Registry of Norway to examine how a mother's age at first birth influences the time interval between the first and second birth (Sit and Xing, 2023). The dataset comprises 53,296 samples, all of which correspond to mothers whose first child was still alive at the time of the second birth. Each observation represents one mother and includes a continuous response variable measuring the duration from the first to the second birth, along with covariates such as the mother's age, the baby's sex, and a binary censoring indicator. A detailed description of the variables is provided in Table 7.

In survey studies, the Cox proportional hazards model is commonly employed to analyze time-to-event data. This model takes the form:

$$h(t) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2),$$

where the hazard function $h(t)$ is defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le y \le t + \Delta t | y \ge t)}{\Delta t}.$$

Parameters are estimated using partial maximum likelihood estimation, with the likelihood contribution for each observation given by:

$$f(\boldsymbol{\nu}_i, \boldsymbol{\beta}) = \left[\frac{\exp(\boldsymbol{X}_i^T \boldsymbol{\beta})}{\sum_{j: y_j \ge y_i} \exp(\boldsymbol{X}_j^T \boldsymbol{\beta})} \right]^{\varpi_i}.$$

In this framework, only uncensored observations contribute to the parameter estimation.

In our dataset, although the sample size is substantial, a significant proportion of the observations, 69.63%, are censored. Relying solely on the individual-level data would therefore lead to considerable information loss during estimation and reduce the accuracy of our results. Moreover, we have access to valuable external information: parameter estimates derived from a similar dataset comprising 262 observations, which includes mothers whose first child was not alive at the time of the second birth. This supplementary dataset provides a unique opportunity to 'borrow strength' from a related population, thereby enriching our analytical framework. By incorporating this auxiliary information, we can compensate for the high censoring rate, improve the precision of our estimates, and achieve a more robust understanding of the underlying processes. The external and internal datasets share the same covariate effects, but differ in their baseline hazard functions $\lambda_0(t)$.

As suggested by Sit and Xing (2023), the individual-level data are randomly partitioned across 100 machines, with the pre-specified parameter α set to the same value as in Section 4.2. The coefficient estimates and corresponding standard errors for the three methods described in Section 5.2 are presented in Table 8. As S increases, the performance of the CDS method becomes increasingly similar to that of the Individual method, both in terms of the magnitude of the estimates and the size of the SE. For the proposed method, which incorporates external information, the SE is consistently smaller than that of CDS at the same value of S. This example clearly demonstrates that when external information is used, the variance of the estimator is reduced, which aligns with the theoretical conclusions presented in the paper.

Table 8: The performance of three methods in the survey case (with standard errors in parentheses, 10^{-3}).

Parameters	Individual method	CDS (S=1)	CDS (S=3)	CDS (S=5)	Proposed method (S=1)	Proposed method (S=3)	Proposed method (S=5)
β_1	0.0891	0.0907	0.0897	0.0897	0.0899	0.0889	0.0889
	(3.6312)	(3.6828)	(3.6829)	(3.6829)	(3.6686)	(3.6686)	(3.6686)
β_2	-0.0175	-0.0173	-0.0169	-0.0169	-0.0176	-0.0172	-0.0172
	(15.7348)	(15.5027)	(15.5028)	(15.5028)	(15.4753)	(15.4753)	(15.4753)

6. Discussion and conclusions

In this paper, we have proposed a feasible and efficient framework for integrating individual-level data with heterogeneous external summary information, supported by both theoretical guarantees and numerical validation. The core idea of our method is to convert heterogeneous summary statistics into a confidence distribution, whose density is then used to approximate the likelihood function of the external model. This enables the combined use of individual data and auxiliary summary information by multiplying the corresponding likelihood and confidence density functions. It is worth emphasizing that the proposed estimator achieves efficiency comparable to that of the ideal IPD estimator that uses all individual-level data, while remaining robust to potential misspecification of the covariance

structure of parameter estimates from external studies. Moreover, if the sample size is small or medium, our proposed approach is equivalent to the heterogeneous meta-analysis developed by Liu et al. (2015). However, when the sample size of the internal study is too large, the internal estimator cannot usually be calculated on a single machine. Although we can use distributed algorithms such as the "one-shot procedure" to get an estimation of our internal study and then obtain the final estimator by applying meta analysis, this naive estimator is often sub-optimal, and the number of machines T has to be $o(\sqrt{n})$, where n is the total sample size of the internal study. Unlike Liu et al. (2015), our developed framework can be adapted to the communication-efficient iterative distributed statistical inference for massive data and successively relax the restriction on the number of machines.

We have also developed a communication-efficient distributed statistical inference method for massive data with heterogeneous auxiliary information, and have shown that the proposed distributed algorithm enjoys linear convergence under some general conditions and that statistical efficiency can be achieved within a finite number of rounds of communication. A specific example of the generalized linear model for an internal study and extensive simulations also demonstrate the superior performance of our proposed iterative algorithm. In fact, our proposed distributed algorithm can also be applied to the distributed statistical inference of parameter-heterogeneous big data and even streaming data.

Several aspects of this work are worth extending. Although the main contribution of this article was developed under the likelihood inference framework, the proposed approaches can also be applied to general loss scenarios by replacing the averaged negative log-likelihood function with the empirical loss risk. The theoretical properties, such as consistency and asymptotic normality in Section 2, and the contraction of optimization errors in Section 3, still hold under some regularity conditions. Moreover, we can generalize our method to high-dimensional sparse models. Based on some summary results of high-dimensional regression, we can also construct asymptotic confidence densities using the de-biased lasso procedure (van de Geer et al., 2014; Zhang and Zhang, 2014; Javanmard and Montanari, 2014). By subtracting a penalty term in the optimization function (3) and (5) (see He et al. (2016) and Cai et al. (2022)), we can synthesize evidence from individual data and some heterogeneous summary results for variable selection, and perform communication-efficient distributed statistical inference for high-dimensional massive datasets. This scalability and effectiveness enhance the applicability of our proposed method.

Acknowledgments

We are grateful to two anonymous reviewers for their insightful comments and suggestions that have helped improve the presentation of this paper. This work is supported in part by funds from the National Key R&D Program of China (2021YFA1000100 and 2021YFA1000101), Program of National Natural Science Foundation of China (72301108), State Key Program of National Natural Science Foundation of China (72331005 and 72531003), Shanghai Pujiang Program (23PJC040), Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM904), Shanghai Pilot Program for Basic Research (TQ20240201), Natural Science Foundation of Shanghai (23JS1400500), Shanghai Municipal Education Commission (2024AI01002).

Appendix A. Proofs

This section is organized as follows. We first present the detailed proofs of Theorem 2, Theorem 3, and Corollary 4. The final subsection then provides the justification for the specific examples of the generalized linear model.

A.1 Proof of Theorem 2

Denote the true value of $\boldsymbol{\gamma}_k$ by $\boldsymbol{\gamma}_k^*$. Recall that $\hat{\boldsymbol{\theta}}_{\text{IPD}}$ is the maximum likelihood estimator of the multiplied likelihood function $L(\boldsymbol{\theta}) = \prod_{k=0}^{K} L_k(\boldsymbol{\theta})$ when all individual-level data are available. Following the large-sample theory of MLE, it is easy to see that the IPD estimator $\hat{\boldsymbol{\theta}}_{\text{IPD}}$ is $\sqrt{N}$-consistent under the regularity conditions stated in Appendix B. Moreover, the first-order Taylor expansion of $\partial \log L(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ gives

$$0 = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\hat{\boldsymbol{\theta}}_{\text{IPD}}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}^*) + \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log L(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_{\text{IPD}} - \boldsymbol{\theta}^*) + O_p(1).$$

After some simple algebraic manipulations, we obtain

$$\begin{aligned}
\sqrt{N}(\hat{\boldsymbol{\theta}}_{\text{IPD}} - \boldsymbol{\theta}^*) &= \left[\frac{1}{N} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log L(\boldsymbol{\theta}^*)\right]^{-1} \left[\frac{1}{\sqrt{N}} \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}^*)\right] \\
&= \left[\sum_{k=0}^{K} J_k(\boldsymbol{\theta}^*)^T \left\{\frac{n_k}{N} \cdot \frac{1}{n_k} \frac{\partial^2}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_k^T} \log L_k^*(\boldsymbol{\gamma}_k^*) J_k(\boldsymbol{\theta}^*)\right\} + o_p(1)\right]^{-1} \\
&\quad \times \left[\sum_{k=0}^{K} J_k(\boldsymbol{\theta}^*)^T \left\{\sqrt{\frac{n_k}{N}} \cdot n_k^{-1/2} \cdot \frac{\partial}{\partial \boldsymbol{\gamma}_k} \log L_k^*(\boldsymbol{\gamma}_k^*)\right\}\right] + o_p(1).
\end{aligned}$$

Then we have that $\sqrt{N}(\hat{\boldsymbol{\theta}}_{\text{IPD}} - \boldsymbol{\theta}^*)$ asymptotically converges to a zero-mean normal distribution with covariance matrix

$$\left\{\sum_{k=0}^{K} c_k J_k(\boldsymbol{\theta}^*)^T I_k J_k(\boldsymbol{\theta}^*)\right\}^{-1}.$$

Now, we establish the asymptotic properties of the CD estimator $\tilde{\boldsymbol{\theta}}$. Let $B_\delta(\boldsymbol{\theta}^*) = \{\boldsymbol{\theta} \,|\, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \le \delta\}$ be a closed δ-ball neighborhood of $\boldsymbol{\theta}^*$, for any $\boldsymbol{\theta} \in \partial B_\delta(\theta^*)$, sufficiently large N or n, and some constants C_1 and C_2, the following holds

$$\begin{aligned}
&\frac{\ell_A(\boldsymbol{\theta}) - \ell_A(\boldsymbol{\theta}^*)}{N} \\
\le &\frac{1}{N} \frac{\partial \log L_0(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^T}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{1}{2N}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \frac{\partial^2 \log L_0(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + C_1 \delta^3 \\
&- \frac{1}{2N} \sum_{k=1}^{K} (\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)^T \hat{\Sigma}_k^{-1} (\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k) + \frac{1}{2N} \sum_{k=1}^{K} (\boldsymbol{\xi}_k(\boldsymbol{\theta}^*) - \hat{\boldsymbol{\gamma}}_k)^T \hat{\Sigma}_k^{-1} (\boldsymbol{\xi}_k(\boldsymbol{\theta}^*) - \hat{\boldsymbol{\gamma}}_k) \\
\le &\frac{1}{2N} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \Big[\frac{\partial^2 \log L_0(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} - J_k^T(\boldsymbol{\theta}^*) \hat{\Sigma}_k^{-1} J_k(\boldsymbol{\theta}^*)\Big](\boldsymbol{\theta} - \boldsymbol{\theta}^*) + C_2 \delta^3 + o_p(1),
\end{aligned}$$

where the first inequality makes use of the boundedness condition of the third derivative of $\log f(\boldsymbol{\theta})$. Note that

$$\frac{1}{n}\frac{\partial^2 \log L_0(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = -J_0(\boldsymbol{\theta}^*)^T I_0 J_0(\boldsymbol{\theta}^*) + o_p(1), \quad n \to \infty.$$

Taking any

$$\delta \leq \frac{1}{2C_2\,\lambda_{\min}\Big(c_0 J_0(\boldsymbol{\theta}^*)^T I_0 J_0(\boldsymbol{\theta}^*) + \sum_{k=1}^K c_k J_k(\boldsymbol{\theta}^*)^T \big[\lim \hat{\Sigma}_k^{-1}/n_k\big] J_k(\boldsymbol{\theta}^*)\Big)},$$

we have for any $\boldsymbol{\theta} \in \partial B_\delta(\boldsymbol{\theta}^*)$, $P(\ell_A(\boldsymbol{\theta})/N < \ell_A(\boldsymbol{\theta}^*)/N) \to 1$ as $N \to \infty$. Since $\ell_A(\boldsymbol{\theta})$ is continuous in $B_\delta(\boldsymbol{\theta}^*)$, there exists a maximum value inside the ball $B_\delta(\boldsymbol{\theta}^*)$ with probability 1 as $N \to \infty$; thus, the weak consistency follows from the arbitrariness of δ. To show asymptotic normality, we apply the Taylor expansion and obtain

$$\begin{aligned}
0 &= \frac{1}{N}\frac{\partial}{\partial \boldsymbol{\theta}} \log L_0(\tilde{\boldsymbol{\theta}}) + \frac{1}{N}\sum_{k=1}^K \frac{\partial \log d_k(\boldsymbol{\xi}_k(\tilde{\boldsymbol{\theta}}))}{\partial \boldsymbol{\theta}} \\
&= \frac{1}{N}\frac{\partial}{\partial \boldsymbol{\theta}} \log L_0(\boldsymbol{\theta}^*) + \frac{1}{N}\sum_{k=1}^K \frac{\partial \log d_k(\boldsymbol{\xi}_k(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\theta}} \\
&\quad + \frac{1}{N}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \log L_0(\boldsymbol{\theta}^*) + \sum_{k=1}^K \frac{\partial^2 \log d_k(\boldsymbol{\xi}_k(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right](\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + O_p(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2).
\end{aligned}$$

It is straightforward to verify that

$$\begin{aligned}
\frac{\partial \log d_k(\boldsymbol{\xi}_k(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\theta}} &= J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1}(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\xi}_k(\boldsymbol{\theta}^*)), \\
\frac{\partial^2 \log d_k(\boldsymbol{\xi}_k(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} &= J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1} J_k(\boldsymbol{\theta}^*) + O_p(N^{-\frac{1}{2}}),
\end{aligned}$$

and

$$\frac{\partial \log L_k^*(\boldsymbol{\gamma}_k^*)}{\partial \boldsymbol{\gamma}_k} = \frac{\partial \log L_k^*(\hat{\boldsymbol{\gamma}}_k)}{\partial \boldsymbol{\gamma}_k} + \frac{\partial^2 \log L_k^*(\hat{\boldsymbol{\gamma}}_k)}{\partial \boldsymbol{\gamma}_k \partial \boldsymbol{\gamma}_k^T}(\boldsymbol{\gamma}_k^* - \hat{\boldsymbol{\gamma}}_k) + O_p(1) = \hat{\Sigma}_k^{-1}(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\xi}_k^*(\boldsymbol{\theta})) + O_p(1).$$

So we have

$$\frac{\partial \log d_k(\boldsymbol{\xi}_k(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\theta}} = \frac{\partial \log L_k(\boldsymbol{\xi}_k(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\theta}} + O_p(1), \tag{A.1.1}$$

and further, $\sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ asymptotically converges to a zero-mean normal distribution with the same limiting covariance matrix as $\sqrt{N}(\hat{\boldsymbol{\theta}}_{\text{IPD}} - \boldsymbol{\theta}^*)$; that is, the estimator $\tilde{\boldsymbol{\theta}}$ is asymptotically as efficient as the IPD estimator $\hat{\boldsymbol{\theta}}_{\text{IPD}}$. Moreover, since

$$\frac{1}{N}\frac{\partial^2 \log d_k(\boldsymbol{\xi}_k(\tilde{\boldsymbol{\theta}}))}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T} = \frac{1}{N}\frac{\partial^2 \log L_k(\boldsymbol{\xi}_k(\tilde{\boldsymbol{\theta}}))}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T} + o_p(1),$$

it is easy to see that the asymptotic covariance matrix of $\sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ can be consistently estimated by $N\{-\partial^2 \ell_A(\tilde{\boldsymbol{\theta}})/\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T\}^{-1}$.

Similar to the proofs in part (**a**) of Theorem 3 in Liu et al. (2015), we can obtain the asymptotic properties of $\tilde{\boldsymbol{\theta}}_W$ in part (**c**). The second result in part (**b**) can also be directly obtained by following Corollary 1 in Liu et al. (2015) or Shen et al. (2020), whose proofs are essentially comparisons of the two different estimators' asymptotic variances using some matrix algebraic inequalities. Finally, given that both $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}_W$ are consistent estimators of $\boldsymbol{\theta}$, and that the sample mean is a consistent estimator of the expectation, it follows that the plug-in estimator of the covariance matrix is also consistent. By applying Slutsky's theorem, the second result in parts (**a**) and (**d**) holds. This completes the proof of Theorem 2.

A.2 Proof of Theorem 3

This proof is similar to the proof of Lemma E.4 in Fan et al. (2021). Let

$$\boldsymbol{\theta}_s^+ = \arg\max\left\{\ell(\boldsymbol{\theta})/n - \frac{\alpha}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_s\|^2 - \frac{1}{2n}\sum_{k=1}^{K}(\boldsymbol{\xi}_k(\boldsymbol{\theta})-\hat{\boldsymbol{\gamma}}_k)^T\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta})-\hat{\boldsymbol{\gamma}}_k)\right\}.$$

First, under Conditions A and B, we show that $\|\boldsymbol{\theta}_{s+1}-\boldsymbol{\theta}_s^+\| \le \delta/(\rho_0+\alpha)\|\boldsymbol{\theta}_s-\boldsymbol{\theta}_s^+\|$ if we have the prior information that $\boldsymbol{\theta}_s \in B(\tilde{\boldsymbol{\theta}}, R/2)$. It follows from Lemma F.3 in Fan et al. (2021) that $\|\boldsymbol{\theta}_s^+-\tilde{\boldsymbol{\theta}}\| \le \|\boldsymbol{\theta}_s-\tilde{\boldsymbol{\theta}}\|_2$. Combining this fact and the condition $\|\boldsymbol{\theta}_s-\tilde{\boldsymbol{\theta}}\|_2 < R/2$, we know $B(\boldsymbol{\theta}_s^+, R/2) \subset B(\tilde{\boldsymbol{\theta}}, R)$. We fix $\boldsymbol{\zeta} \in B(\tilde{\boldsymbol{\theta}}, R)$, and for any $t = 1, \ldots, T$ denote

$$g(\boldsymbol{\theta}) = -\frac{\alpha}{2}\|\boldsymbol{\theta}-\boldsymbol{\zeta}\|^2 - \frac{1}{2n}\sum_{k=1}^{K}(\boldsymbol{\xi}_k(\boldsymbol{\theta})-\hat{\boldsymbol{\gamma}}_k)^T\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta})-\hat{\boldsymbol{\gamma}}_k),$$

and

$$\psi_t(\boldsymbol{\zeta}) = \arg\max_{\boldsymbol{\theta}}\left\{\ell_t(\boldsymbol{\theta})/m - \left\langle\, \boldsymbol{\theta}, \triangledown\ell_t(\boldsymbol{\zeta})/m - \triangledown\ell(\boldsymbol{\zeta})/n \,\right\rangle - \frac{\alpha}{2}\|\boldsymbol{\theta}-\boldsymbol{\zeta}\|^2 - \frac{1}{2n}\sum_{k=1}^{K}(\boldsymbol{\xi}_k(\boldsymbol{\theta})-\hat{\boldsymbol{\gamma}}_k)^T\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta})-\hat{\boldsymbol{\gamma}}_k)\right\}.$$

Then, the first-order condition of $\psi_t(\boldsymbol{\zeta})$ yields

$$\triangledown\ell_t(\boldsymbol{\zeta})/m - \triangledown\ell(\boldsymbol{\zeta})/n = \triangledown\ell_t(\psi_t(\boldsymbol{\zeta}))/m + \triangledown g(\psi_t(\boldsymbol{\zeta})).$$

Moreover, since $\boldsymbol{\theta}_s^+$ is a fixed point of $\psi_t(\cdot)$, we have

$$\triangledown\ell_t(\boldsymbol{\theta}_s^+)/m - \triangledown\ell(\boldsymbol{\theta}_s^+)/n = \triangledown\ell_t(\boldsymbol{\theta}_s^+))/m + \triangledown g(\psi_t(\boldsymbol{\theta}_s^+)).$$

From the Taylor expansion, we know

$$\begin{aligned}
&\left\|\left[\triangledown\ell_t(\psi_t(\boldsymbol{\theta}_s))/m + \triangledown g(\psi_t(\boldsymbol{\theta}_s))\right] - \left[\triangledown\ell_t(\boldsymbol{\theta}_s^+)/m + \triangledown g(\boldsymbol{\theta}_s^+)\right]\right\|_2 \\
&= \left\|\left[\triangledown\ell_t(\boldsymbol{\theta}_s)/m - \triangledown\ell(\boldsymbol{\theta}_s)/n\right] - \left[\triangledown\ell_t(\boldsymbol{\theta}_s^+)/m - \triangledown\ell(\boldsymbol{\theta}_s^+)/n\right]\right\|_2 \\
&\le \left\|\triangledown^2\ell_t(\check{\boldsymbol{\theta}})/m - \triangledown^2\ell(\check{\boldsymbol{\theta}})/n\right\|_2 \left\|\boldsymbol{\theta}_s-\boldsymbol{\theta}_s^+\right\|_2 \\
&\le \delta\left\|\boldsymbol{\theta}_s-\boldsymbol{\theta}_s^+\right\|_2 \le (\rho_0+\alpha)R,
\end{aligned}$$

where $\check{\boldsymbol{\theta}} = \xi\boldsymbol{\theta}_s + (1-\xi)\boldsymbol{\theta}_s^+ \in B(\tilde{\boldsymbol{\theta}}, R)$ with $\xi \in (0,1)$. Here we have used the homogeneity condition $\|\nabla^2\ell_t(\check{\boldsymbol{\theta}})/m - \nabla^2\ell(\check{\boldsymbol{\theta}})/n\|_2 \leq \delta$ for $\check{\boldsymbol{\theta}} \in B(\tilde{\boldsymbol{\theta}}, R)$ and the condition $[\delta/(\rho_0+\alpha)]^2 < \rho/(\rho+2\alpha)$, which ensures $\delta < \rho_0+\alpha$. Furthermore, using Lemma F.2 of Fan et al. (2021), we can determine that $\|\psi_t(\boldsymbol{\theta}_s) - \boldsymbol{\theta}_s^+\|_2 \leq (\delta/(\alpha+\rho_0))\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^+\|_2$. From the triangle inequality, we directly obtain

$$\|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s^+\|_2 = \|\frac{1}{T}\sum_{t=1}^{T}\psi_t(\boldsymbol{\theta}_s) - \boldsymbol{\theta}_s^+\|_2 \leq \frac{\delta}{\alpha+\rho_0}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^+\|_2.$$

Furthermore, if Condition C holds, we can also show

$$\|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s^+\|_2 = \|\frac{1}{T}\sum_{t=1}^{T}\psi_t(\boldsymbol{\theta}_s) - \boldsymbol{\theta}_s^+\|_2 \leq \frac{\delta^2}{(\alpha+\rho_0)(\alpha+\rho)}\Big(1 + \frac{M}{\alpha+\rho_0}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^+\|_2\Big)\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^+\|_2.$$

We recall that the first-order condition of $\psi_t(\boldsymbol{\zeta})$ yields

$$[\nabla\ell_t(\psi_t(\boldsymbol{\theta}_s))/m + \nabla g(\psi_t(\boldsymbol{\theta}_s))] - \nabla\ell_t(\boldsymbol{\theta}_s)/m + \nabla\ell(\boldsymbol{\theta}_s)/n = 0 = [\nabla\ell(\boldsymbol{\theta}_s^+)/n + \nabla g(\boldsymbol{\theta}_s^+)].$$

and the fixed-point property $\psi_t(\boldsymbol{\theta}_s^+) = \boldsymbol{\theta}_s^+$. Then we have

$$\begin{aligned}
&\Big[\nabla\ell_t(\psi_t(\boldsymbol{\theta}_s))/m + \nabla g(\psi_t(\boldsymbol{\theta}_s))\Big] - \Big[\nabla\ell_t(\boldsymbol{\theta}_s^+)/m + \nabla g(\boldsymbol{\theta}_s^+)\Big] \\
&= \Big[\big(\nabla\ell_t(\boldsymbol{\theta}_s)/m + \nabla g(\boldsymbol{\theta}_s)\big) - \big(\nabla\ell_t(\boldsymbol{\theta}_s^+)/m + \nabla g(\boldsymbol{\theta}_s^+)\big)\Big] \\
&\quad - \Big[\big(\nabla\ell(\boldsymbol{\theta}_s)/n + \nabla g(\boldsymbol{\theta}_s)\big) - \big(\nabla\ell(\boldsymbol{\theta}_s^+)/n + \nabla g(\boldsymbol{\theta}_s^+)\big)\Big] \\
&= H_t \cdot (\psi_t(\boldsymbol{\theta}_s) - \boldsymbol{\theta}_s^+) = \hat{H} \cdot (\psi_t(\boldsymbol{\theta}_s) - \boldsymbol{\theta}_s^+) + (H_t - \hat{H}) \cdot (\psi_t(\boldsymbol{\theta}_s) - \boldsymbol{\theta}_s^+),
\end{aligned}$$

where $H_t = \nabla^2\ell_t(\check{\boldsymbol{\theta}}_s)/m + \nabla^2 g(\check{\boldsymbol{\theta}}_s)$ with $\check{\boldsymbol{\theta}}_s = \xi\,\psi_t(\boldsymbol{\theta}_s) + (1-\xi)\,\boldsymbol{\theta}_s^+$ and some $\xi \in (0,1)$, and $\hat{H} = -\nabla^2 F(\boldsymbol{\theta}_s^+)/n + \nabla^2 g(\boldsymbol{\theta}_s^+)$. Note that the average of the terms

$$\Big[\big(\nabla\ell_t(\boldsymbol{\theta}_s)/m + \nabla g(\boldsymbol{\theta}_s)\big) - \big(\nabla\ell_t(\boldsymbol{\theta}_s^+)/m + \nabla g(\boldsymbol{\theta}_s^+)\big)\Big] - \Big[\big(\nabla\ell(\boldsymbol{\theta}_s)/n + \nabla g(\boldsymbol{\theta}_s)\big) - \big(\nabla\ell(\boldsymbol{\theta}_s^+)/n + \nabla g(\boldsymbol{\theta}_s^+)\big)\Big]$$

over $t = 1, \ldots T$ is 0, so we have

$$\begin{aligned}
0 &= \frac{1}{T}\sum_{t=1}^{T}\Big\{\hat{H}\,(\psi_t(\boldsymbol{\theta}_s) - \boldsymbol{\theta}_s^+) + (H_t - \hat{H})\,(\psi_t(\boldsymbol{\theta}_s) - \boldsymbol{\theta}_s^+)\Big\} \\
&= \hat{H}\,(\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s^+) + \frac{1}{T}\sum_{t=1}^{T}(H_t - \hat{H})\,(\psi_t(\boldsymbol{\theta}_s) - \boldsymbol{\theta}_s^+)\Big\},
\end{aligned}$$

and

$$\begin{aligned}
\|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s^+\|_2 &= \|\frac{1}{T}\sum_{t=1}^{T}\hat{H}^{-1}(H_t - \hat{H})(\psi(\boldsymbol{\theta}_s) - \boldsymbol{\theta}_s^+)\|_2 \\
&\leq \|\hat{H}^{-1}\|_2 \max_{t=1,\ldots,T}\|H_t - \hat{H}\|_2 \max_{t=1,\ldots,T}\|(\psi_t(\boldsymbol{\theta}_s) - \boldsymbol{\theta}_s^+)\|_2 \\
&\leq \frac{1}{\rho+\alpha} \cdot (\delta + \frac{M\delta}{\alpha+\rho_0}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^+\|_2) \cdot \frac{\delta}{\alpha+\rho_0}.
\end{aligned}$$

Here, the last inequality follows from the fact

$$\begin{aligned}\|H_t - \hat{H}\|_2 &= \|\big[\nabla^2\ell_t(\check{\boldsymbol{\theta}}_s)/m + \nabla^2 g(\check{\boldsymbol{\theta}}_s)\big] - \big[-\nabla^2 F(\boldsymbol{\theta}_s^+)/n + \nabla^2 g(\boldsymbol{\theta}_s^+)\big]\|_2 \\ &\le \|\big[\nabla^2\ell_t(\check{\boldsymbol{\theta}}_s)/m + \nabla^2 g(\check{\boldsymbol{\theta}}_s)\big] - \big[-\nabla^2 F(\check{\boldsymbol{\theta}}_s)/n + \nabla^2 g(\check{\boldsymbol{\theta}}_s)\big] \\ &\quad + \|\big[-\nabla^2 F(\check{\boldsymbol{\theta}}_s)/n + \nabla^2 g(\check{\boldsymbol{\theta}}_s)\big] - \big[-\nabla^2 F(\boldsymbol{\theta}_s^+)/n + \nabla^2 g(\boldsymbol{\theta}_s^+)\big] \\ &\le \delta + M\|\psi_t(\boldsymbol{\theta}_s) - \boldsymbol{\theta}_s^+\|_2 \le \delta + \frac{M\delta}{\alpha+\rho_0}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^+\|_2.\end{aligned}$$

In summary, under Conditions A and B, we define

$$\kappa_s = \begin{cases} \frac{\delta}{\alpha+\rho_0} \cdot \min\{1, \frac{\delta}{\alpha+\rho}(1 + \frac{M}{\alpha+\rho_0}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^+\|_2)\}, & \text{if Condition C holds,} \\ \frac{\delta}{\alpha+\rho_0}, & \text{otherwise} \end{cases}$$

for $s = 1, \ldots, S$. It is worth noting that we have shown that $\|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s^+\| \le \kappa_s\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^+\|$ for $\boldsymbol{\theta}_s \in B(\tilde{\boldsymbol{\theta}}, R/2)$.

Now we return to the main proof of Theorem 2. By applying the triangle inequality, we know

$$\|\boldsymbol{\theta}_{s+1} - \tilde{\boldsymbol{\theta}}\|_2 \le \|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s^+\|_2 + \|\boldsymbol{\theta}_s^+ - \tilde{\boldsymbol{\theta}}\|_2.$$

It follows from the proximal mapping Lemma F.3 of Fan et al. (2021) that

$$\|\boldsymbol{\theta}_s^+ - \tilde{\boldsymbol{\theta}}\| \,/\|\boldsymbol{\theta}_s - \tilde{\boldsymbol{\theta}}\|_2 \le \alpha/(\alpha+\rho) \le 1, \tag{A.2.1}$$

and

$$\begin{aligned}\|\boldsymbol{\theta}_{s+1} - \tilde{\boldsymbol{\theta}}\|_2 &\le \kappa_s\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_s^+\|_2 + \|\boldsymbol{\theta}_s^+ - \tilde{\boldsymbol{\theta}}\|_2 \\ &\le \kappa_s\Big(\|\boldsymbol{\theta}_s - \tilde{\boldsymbol{\theta}}\|_2^2 - \|\boldsymbol{\theta}_s^+ - \tilde{\boldsymbol{\theta}}\|_2^2\Big)^{1/2} + \|\boldsymbol{\theta}_s^+ - \tilde{\boldsymbol{\theta}}\|_2 \\ &\le \|\boldsymbol{\theta}_s - \tilde{\boldsymbol{\theta}}\|_2 \times r\big(\|\boldsymbol{\theta}_s^+ - \tilde{\boldsymbol{\theta}}\|_2 \,/\|\boldsymbol{\theta}_s - \tilde{\boldsymbol{\theta}}\|_2\big),\end{aligned} \tag{A.2.2}$$

where $r(x) = \kappa_s\sqrt{1-x^2} + x$, $\forall x \in [0,1]$. Note that a simple derivative calculation derives $r'(x) \ge 0$ on $[0, 1/\sqrt{1+\kappa_s^2}]$, and under the condition $[\delta/(\rho_0+\alpha)]^2 < \rho/(\rho+2\alpha)$, it is easy to see $\kappa_s^2 \le \rho/(\rho+2\alpha)$ and

$$\frac{1}{\sqrt{1+\kappa_s^2}} > \frac{1}{\sqrt{1+\rho/(\rho+2\alpha)}} \ge \frac{\rho/2+\alpha}{\rho+\alpha} \ge \frac{\alpha}{(\rho+\alpha)}.$$

Thus, it holds that $r'(x) \ge 0$ on $[0, \alpha/(\rho+\alpha)]$. Combining this fact with equations (A.2.1) and (A.2.2), we obtain that

$$\frac{\|\boldsymbol{\theta}_{s+1} - \tilde{\boldsymbol{\theta}}\|_2}{\|\boldsymbol{\theta}_s - \tilde{\boldsymbol{\theta}}\|_2} \le r\bigg(\frac{\|\boldsymbol{\theta}_s^+ - \tilde{\boldsymbol{\theta}}\|_2}{\|\boldsymbol{\theta}_s - \tilde{\boldsymbol{\theta}}\|_2}\bigg) \le r\bigg(\frac{\alpha}{\rho+\alpha}\bigg) = \frac{\kappa_s\sqrt{\rho^2+2\rho\alpha}+\alpha}{\rho+\alpha},$$

and since

$$\frac{\kappa_s\sqrt{\rho^2+2\rho\alpha}+\alpha}{\rho+\alpha} < \frac{\sqrt{\rho/(\rho+2\alpha)}\sqrt{\rho^2+2\rho\alpha}+\alpha}{\rho+\alpha} = 1,$$

we obtain the last statement of Theorem 2 and complete the proof.

A.3 Proof of Corollary 4

Theorem 3 tells us that $S = O\big(\log(\|\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}\|_2/\epsilon)/\log(\rho/\delta)\big)$, where $\epsilon = O(N^{-1/2-\epsilon_0})$. Then, for any $\pi > 0$ and $N \to \infty$, we have

$$P(\|\boldsymbol{\theta}_S - \tilde{\boldsymbol{\theta}}\|_2 \geq \pi) \leq \frac{\left(\frac{\frac{\delta}{\rho_0+\alpha}\cdot\sqrt{\rho^2+2\alpha\rho}+\alpha}{\rho+\alpha}\right)^S \mathbb{E}(\|\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}\|_2)}{\pi} \to 0.$$

The latter comes from the fact that $\frac{\frac{\delta}{\rho_0+\alpha}\cdot\sqrt{\rho^2+2\alpha\rho}+\alpha}{\rho+\alpha} < 1$ and $S \to \infty$ as $N \to \infty$. Because $\tilde{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$ from Theorem 2, the following equation holds:

$$P(\|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|_2 \geq 2\pi) \leq P(\|\boldsymbol{\theta}_S - \tilde{\boldsymbol{\theta}}\|_2 \geq \pi) + P(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \geq \pi) \to 0.$$

That is, $\boldsymbol{\theta}_S$ is a consistent estimator of $\boldsymbol{\theta}$. Next, we establish its asymptotic normality. Let

$$\begin{aligned}\tilde{\ell}_t(\boldsymbol{\theta}) = \ell_t(\boldsymbol{\theta})/m - \big\langle\, \boldsymbol{\theta}, \triangledown\ell_t(\boldsymbol{\theta}_s)/m - \triangledown\ell(\boldsymbol{\theta}_s)/n \,\big\rangle - \frac{\alpha}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_s\|^2 \\ - \frac{1}{2n}\sum_{k=1}^{K}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k)^T\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\boldsymbol{\gamma}}_k).\end{aligned}$$

Then, similarly to Equation (A.1.1), we have

$$\begin{aligned}\frac{1}{N}\nabla\tilde{\ell}_t(\boldsymbol{\theta}) =& \frac{1}{Nm}\nabla\ell_t(\boldsymbol{\theta}) - \frac{1}{N}\left(\frac{\nabla\ell_t(\boldsymbol{\theta}_s)}{m} - \frac{\nabla\ell(\boldsymbol{\theta}_s)}{n}\right) \\ & - \frac{\alpha}{N}(\boldsymbol{\theta} - \boldsymbol{\theta}_s) - \frac{1}{Nn}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta})\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}) - \hat{\gamma}_k), \\ \frac{1}{N}\nabla^2\tilde{\ell}_t(\boldsymbol{\theta}) =& \frac{1}{Nm}\nabla^2\ell_t(\boldsymbol{\theta}) - \frac{\alpha}{N}I - \frac{1}{Nn}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta})\hat{\Sigma}_k^{-1}J_k(\boldsymbol{\theta}) + o_p(1) \\ =& \frac{-1}{N}J_0^T(\boldsymbol{\theta})I_0J_0(\boldsymbol{\theta}) - \frac{\alpha}{N}I - \frac{1}{Nn}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta})\hat{\Sigma}_k^{-1}J_k(\boldsymbol{\theta}) + o_p(1),\end{aligned}$$

where I is the identity matrix. Since $\nabla\tilde{\ell}_t(\boldsymbol{\theta}_{s+1,t}) = \mathbf{0}$, substituting this into a Taylor expansion around $\boldsymbol{\theta}^*$ yields $0 = \nabla\tilde{\ell}_t(\boldsymbol{\theta}^*) + \nabla^2\tilde{\ell}_t(\boldsymbol{\theta}^*)(\boldsymbol{\theta}_{s+1,t} - \boldsymbol{\theta}^*) + O_p(\|\boldsymbol{\theta}_{s+1,t} - \boldsymbol{\theta}^*\|^2)$. That is,

$$\begin{aligned}\boldsymbol{\theta}_{s+1,t} - \boldsymbol{\theta}^* &= \left[-\frac{1}{N}\nabla^2\tilde{\ell}_t(\boldsymbol{\theta}^*)\right]^{-1}\left[\frac{1}{N}\nabla\tilde{\ell}_t(\boldsymbol{\theta}^*)\right] \\ &= \left[\frac{1}{n}\left(c_0J_0^T(\boldsymbol{\theta}^*)I_0J_0(\boldsymbol{\theta}^*) + \frac{n\alpha}{N}I + \frac{1}{N}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1}J_k(\boldsymbol{\theta}^*)\right)\right]^{-1}\left[\frac{1}{N}\nabla\tilde{\ell}_t(\boldsymbol{\theta}^*)\right].\end{aligned}$$

From $\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^* = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\theta}_{s+1,t} - \boldsymbol{\theta}^*)$, we can obtain that

$$\begin{aligned}
\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^* &= \left[\frac{1}{n}\left(c_0 J_0^T(\boldsymbol{\theta}^*) I_0 J_0(\boldsymbol{\theta}^*) + c_0\alpha I + \frac{1}{N}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1} J_k(\boldsymbol{\theta}^*)\right)\right]^{-1} \\
&\quad \times \left[\frac{1}{TN}\sum_{t=1}^{T}\nabla\tilde{\ell}_t(\boldsymbol{\theta}^*)\right] \\
&= \left[\frac{1}{n}\left(c_0 J_0^T(\boldsymbol{\theta}^*) I_0 J_0(\boldsymbol{\theta}^*) + c_0\alpha I + \frac{1}{N}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1} J_k(\boldsymbol{\theta}^*)\right)\right]^{-1} \\
&\quad \times \left[\frac{\nabla\ell(\boldsymbol{\theta}^*)}{nN} - \frac{\alpha}{N}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_s) - \frac{1}{Nn}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}^*) - \hat{\gamma}_k)\right].
\end{aligned}$$

By rearranging terms, we get

$$\begin{aligned}
&\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}^* - \left[\frac{1}{n}\left(c_0 J_0^T(\boldsymbol{\theta}^*) I_0 J_0(\boldsymbol{\theta}^*) + c_0\alpha I + \frac{1}{N}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1} J_k(\boldsymbol{\theta}^*)\right)\right]^{-1} \frac{\alpha}{N}(\boldsymbol{\theta}_s - \boldsymbol{\theta}^*) \\
&= \left[\left(c_0 J_0^T(\boldsymbol{\theta}^*) I_0 J_0(\boldsymbol{\theta}^*) + c_0\alpha I + \frac{1}{N}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1} J_k(\boldsymbol{\theta}^*)\right)\right]^{-1} \\
&\quad \times \left[\frac{\nabla\ell(\boldsymbol{\theta}^*)}{N} - \frac{1}{N}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}^*) - \hat{\gamma}_k)\right].
\end{aligned}$$

We replace the $\boldsymbol{\theta}_s - \boldsymbol{\theta}^*$ with $\boldsymbol{\theta}_{s-1} - \boldsymbol{\theta}^*$ and iterate continuously to obtain

$$\begin{aligned}
&\boldsymbol{\theta}_S - \boldsymbol{\theta}^* - \left[\frac{1}{n}\left(c_0 J_0^T(\boldsymbol{\theta}^*) I_0 J_0(\boldsymbol{\theta}^*) + n\alpha + \frac{1}{N}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1} J_k(\boldsymbol{\theta}^*)\right)\right]^{-S} \left(\frac{\alpha}{N}\right)^S (\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) - o_p\left(\frac{1}{\sqrt{N}}\right) \\
&= \left[c_0 J_0^T(\boldsymbol{\theta}^*) I_0 J_0(\boldsymbol{\theta}^*) + n\alpha + \frac{1}{N}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1} J_k(\boldsymbol{\theta}^*)\right]^{-1} \times \left[\frac{\nabla\ell(\boldsymbol{\theta}^*)}{N} - \frac{1}{N}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}^*) - \hat{\gamma}_k)\right].
\end{aligned}$$

Since $\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*$ is bounded by $\frac{R}{2} + O_p(\frac{1}{\sqrt{N}})$ and $0 < c)0 < 1$, we have

$$\begin{aligned}
&\left\|\left[\frac{1}{n}\left(c_0 J_0^T(\boldsymbol{\theta}^*) I_0 J_0(\boldsymbol{\theta}^*) + n\alpha + \frac{1}{N}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1} J_k(\boldsymbol{\theta}^*)\right)\right]^{-S} \left(\frac{\alpha}{N}\right)^S (\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)\right\|_2 \\
&\leq \left(\frac{c_0\alpha}{\rho + \alpha}\right)^S \|(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*)\|_2 = o_p(\frac{1}{\sqrt{N}}).
\end{aligned}$$

As a result,

$$\begin{aligned}
\sqrt{N}(\boldsymbol{\theta}_S - \boldsymbol{\theta}^*) &= \left[c_0 J_0^T(\boldsymbol{\theta}^*) I_0 J_0(\boldsymbol{\theta}^*) + c_0\alpha I + \frac{1}{N}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1} J_k(\boldsymbol{\theta}^*)\right]^{-1} \\
&\quad \times \left[\frac{\nabla\ell(\boldsymbol{\theta}^*)}{\sqrt{N}} - \frac{1}{\sqrt{N}}\sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*)\hat{\Sigma}_k^{-1}(\boldsymbol{\xi}_k(\boldsymbol{\theta}^*) - \hat{\gamma}_k)\right].
\end{aligned}$$

Based on the assumption that $\hat{\gamma}_k$ is an unbiased or approximately unbiased estimator, the expectation of the right-hand side is asymptotically zero. Moreover, the central limit theorem implies that $\frac{\nabla \ell(\boldsymbol{\theta}^*)}{\sqrt{N}} - \frac{1}{\sqrt{N}} \sum_{k=1}^{K} J_k^T(\boldsymbol{\theta}^*) \hat{\Sigma}_k^{-1} (\boldsymbol{\xi}_k(\boldsymbol{\theta}^*) - \hat{\gamma}_k)$ converges in distribution to a normal distribution with mean zero. Then, from Slutsky's theorem, we have that $\sqrt{N}(\boldsymbol{\theta}_S - \boldsymbol{\theta}^*)$ asymptotically converges to a zero-mean normal distribution with limiting covariance matrix $\tilde{A}^{-1} \tilde{B} \tilde{A}^{-1}$, where

$$\tilde{A} = c_0 J_0^T(\boldsymbol{\theta}^*) I_0 J_0(\boldsymbol{\theta}^*) + c_0 \alpha I + \sum_{k=1}^{K} c_k J_k^T(\boldsymbol{\theta}^*) \tilde{A}_k J_k(\boldsymbol{\theta}^*),$$

$$\tilde{B} = c_0 J_0^T(\boldsymbol{\theta}^*) I_0 J_0(\boldsymbol{\theta}^*) + \sum_{k=1}^{K} c_k J_k^T(\boldsymbol{\theta}^*) \tilde{A}_k J_k(\boldsymbol{\theta}^*),$$

where $\tilde{A}_k = \lim_{n_k \to \infty} (\hat{\Sigma}_k n_k)^{-1}$. The consistent estimators $\hat{\tilde{A}}$ and $\hat{\tilde{B}}$ are similar to those in the proof of Theorem 2:

$$\hat{\tilde{A}} = \frac{1}{N} \sum_{t=1}^{T} \sum_{i \in \mathcal{I}_t} \lambda(\boldsymbol{\nu}_i, \boldsymbol{\theta}_S) \lambda(\boldsymbol{\nu}_i, \boldsymbol{\theta}_S)^T + c_o \alpha I + \frac{1}{N} \sum_{k=1}^{K} J_k(\boldsymbol{\theta}_S)^T \hat{\Sigma}_k^{-1} J_k(\boldsymbol{\theta}_S),$$

$$\hat{\tilde{B}} = \frac{1}{N} \sum_{t=1}^{T} \sum_{i \in \mathcal{I}_t} \lambda(\boldsymbol{\nu}_i, \boldsymbol{\theta}_S) \lambda(\boldsymbol{\nu}_i, \boldsymbol{\theta}_S)^T + \frac{1}{N} \sum_{k=1}^{K} J_k(\boldsymbol{\theta}_S)^T \hat{\Sigma}_k^{-1} J_k(\boldsymbol{\theta}_S).$$

A.4 Justification for the special examples of the generalized linear model

Denote the covariance matrix of $\boldsymbol{x}_i = (1, \boldsymbol{u}_i^T)^T$ by $D^* = \mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i^T) = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & D \end{pmatrix}$. Under Condition B*, the objective of the subsequent proof is to establish that

$$\sup_{1 \le t \le T} \sup_{\boldsymbol{\theta} \in B(\tilde{\boldsymbol{\theta}}, R)} \|\nabla^2 \ell_t(\boldsymbol{\theta})/m - \nabla^2 F(\boldsymbol{\theta})\|_2 \le O_p\Big(\sup_{1 \le t \le T} \Big\| \frac{1}{m} \sum_{i \in \mathcal{I}_t} \boldsymbol{x}_i y_i - \mathbb{E}\big[\boldsymbol{x}_i y_i\big] \Big\|_2 \Big) \times \Bigg[\sup_{\boldsymbol{\theta} \in B(\tilde{\boldsymbol{\theta}}, R)} \max_{j,k} \Big\| \frac{\partial J_0^{(j,k)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \Big\|_1 \Bigg] + O_p\Big(\sqrt{\frac{\log n}{m}} \Big), \quad \text{(A.4.1)}$$

where $\sup_{\boldsymbol{\theta} \in B(\tilde{\boldsymbol{\theta}}, R)} \max_{j,k} \big\| \frac{\partial J_0^{(j,k)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \big\|_1$ is a bounded constant due to Condition B*.

Since

$$\frac{\partial^2 \log f(y_i; \boldsymbol{x}_i, \boldsymbol{\eta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = J_0^T(\boldsymbol{\theta}) \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T \nabla^2 b(\boldsymbol{x}_i^T \boldsymbol{\eta})}{\phi} J_0(\boldsymbol{\theta}) + \frac{\partial \{J_0^T(\boldsymbol{\theta}) \boldsymbol{x}_i\}}{\partial \boldsymbol{\theta}^T} \cdot \frac{\big(y_i - \nabla b(\boldsymbol{x}_i^T \boldsymbol{\eta})\big)}{\phi},$$

and

$$\frac{\partial \{J_0^T(\boldsymbol{\theta}) \boldsymbol{x}_i\}}{\partial \boldsymbol{\theta}^T} \cdot \frac{\big(y_i - \nabla b(\boldsymbol{x}_i^T \boldsymbol{\eta})\big)}{\phi} = \begin{bmatrix} \sum_{j=1}^{p_0} \frac{\partial J_0^{(j,1)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \cdot \frac{x_i^{(j)} \big(y_i - \nabla b(\boldsymbol{x}_i^T \boldsymbol{\eta})\big)}{\phi} \\ \cdots \\ \sum_{j=1}^{p_0} \frac{\partial J_0^{(j,p)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \cdot \frac{x_i^{(j)} \big(y_i - \nabla b(\boldsymbol{x} i^T \boldsymbol{\eta})\big)}{\phi} \end{bmatrix}_{p \times p}.$$

Then following some basic norm inequalities and the equivalence of matrix norms, for some constant C_1' we have

$$\begin{aligned}\phi \cdot \|\nabla^2 \ell_t(\boldsymbol{\theta})/m - \nabla^2 F(\boldsymbol{\theta})\|_2 \leq \|J_0(\boldsymbol{\theta})\|_2^2 \cdot \Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t} \boldsymbol{x}_i\boldsymbol{x}_i^T \nabla^2 b(\boldsymbol{x}_i^T\boldsymbol{\eta}) - \mathbb{E}\big[\boldsymbol{x}_i\boldsymbol{x}_i^T \nabla^2 b(\boldsymbol{x}_i^T\boldsymbol{\eta})\big]\Big\|_2 \\ + C_1' p_0 \max_{j,k}\Big\|\frac{\partial J_0^{(j,k)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\Big\|_1 \cdot \Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t} \boldsymbol{x}_i\big(y_i - \nabla b(\boldsymbol{x}_i^T\boldsymbol{\eta})\big) - \mathbb{E}\big[\boldsymbol{x}_i\big(y_i - \nabla b(\boldsymbol{x}_i^T\boldsymbol{\eta})\big)\big]\Big\|_2.\end{aligned} \tag{A.4.2}$$

Moreover, note that the equivalence of matrix norms and the boundedness of the Jacobian matrix also imply $\{\boldsymbol{\theta} \in B(\tilde{\boldsymbol{\theta}}, R)\} \subset \{\boldsymbol{\eta} \in B(\tilde{\boldsymbol{\eta}}, R_1)\}$ with $\tilde{\boldsymbol{\eta}} = \boldsymbol{\xi}(\tilde{\boldsymbol{\theta}})$ and $\boldsymbol{\eta} = \boldsymbol{\xi}(\boldsymbol{\theta})$ for some $R_1 > 0$. The problem "$\sup_{\boldsymbol{\theta}\in B(\tilde{\boldsymbol{\theta}},R)}$" can be transformed into "$\sup_{\boldsymbol{\eta}\in B(\tilde{\boldsymbol{\eta}},R_1)}$".

For the supremum of the first term on the right-hand side of inequality (A.4.2), it directly follows from the proof of Lemma E.5 in Fan et al. (2021) that

$$\sup_{1\leq t\leq T}\sup_{\boldsymbol{\eta}\in B(\tilde{\boldsymbol{\eta}},R_1)}\Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t} \boldsymbol{x}_i\boldsymbol{x}_i^T \nabla^2 b(\boldsymbol{x}_i^T\boldsymbol{\eta}) - \mathbb{E}\big[\boldsymbol{x}_i\boldsymbol{x}_i^T \nabla^2 b(\boldsymbol{x}_i^T\boldsymbol{\eta})\big]\Big\|_2 = O_p\Big(\sqrt{\frac{\log n}{m}}\Big).$$

Now we bound the supremum of the second term in inequality (A.4.2). Note that

$$\begin{aligned}&\sup_{1\leq t\leq T}\sup_{\boldsymbol{\eta}\in B(\tilde{\boldsymbol{\eta}},R_1)}\Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t} \boldsymbol{x}_i\big(y_i - \nabla b(\boldsymbol{x}_i^T\boldsymbol{\eta})\big) - \mathbb{E}\big[\boldsymbol{x}_i\big(y_i - \nabla b(\boldsymbol{x}_i^T\boldsymbol{\eta})\big)\big]\Big\|_2 \\ \leq\ &\sup_{1\leq t\leq T}\Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t} \boldsymbol{x}_i y_i - \mathbb{E}\big[\boldsymbol{x}_i y_i\big]\Big\|_2 \sup_{1\leq t\leq T}\sup_{\boldsymbol{\eta}\in B(\tilde{\boldsymbol{\eta}},R_1)}\Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t} \boldsymbol{x}_i\nabla b(\boldsymbol{x}_i^T\boldsymbol{\eta}) - \mathbb{E}\big[\boldsymbol{x}_i\nabla b(\boldsymbol{x}_i^T\boldsymbol{\eta})\big]\Big\|_2.\end{aligned} \tag{A.4.3}$$

So we only need to bound the second quantity of (A.4.3) by $O_p(\sqrt{\log n/m})$ to complete the proof of result (A.4.1). Let $\check{\boldsymbol{x}}_i = (D^*)^{-1/2}\boldsymbol{x}_i$ and $\check{\boldsymbol{\eta}} = (D^*)^{1/2}\boldsymbol{\eta}$. Then,

$$\begin{aligned}&\sup_{1\leq t\leq T}\sup_{\boldsymbol{\eta}\in B(\tilde{\boldsymbol{\eta}},R_1)}\Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t} \boldsymbol{x}_i\nabla b(\boldsymbol{x}_i^T\boldsymbol{\eta}) - \mathbb{E}\big[\boldsymbol{x}_i\nabla b(\boldsymbol{x}_i^T\boldsymbol{\eta})\big]\Big\|_2 \\ &\leq \|D^*\|_2^{1/2}\sup_{1\leq t\leq T}\sup_{\check{\boldsymbol{\eta}}\in B\left((D^*)^{1/2}\tilde{\boldsymbol{\eta}},\,\|D^*\|_2^{1/2}R_1\right)}\Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t} \check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}) - \mathbb{E}\big[\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}})\big]\Big\|_2 \\ &\triangleq \|D^*\|_2^{1/2}\,\Delta_m.\end{aligned}$$

Here, the notation "$\triangleq$" represents a definition. For a fixed $\check{\boldsymbol{\eta}} \in B\big((D^*)^{1/2}\tilde{\boldsymbol{\eta}}, \|D^*\|_2^{1/2}R_1\big)$, since the product of sub-Gaussian variables is sub-exponential, the Bernstein inequality (see Vershynin (2018)) gives that for any $\epsilon > 0$ and some universal constants C_2' the following holds:

$$\mathbb{P}\left(\Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t} \check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}) - \mathbb{E}\big[\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}})\big]\Big\|_2 \geq \epsilon\right) \leq 2pe^{-C_2' m\cdot\min\{\frac{\epsilon}{\sqrt{p_0}},\frac{\epsilon^2}{p_0}\}}.$$

Next, we try to bound Δ_m by using the covering of $B\big((D^*)^{1/2}\tilde{\boldsymbol{\eta}}, \|D^*\|_2^{1/2}R_1\big)$. Denote the event $E_q \overset{\triangle}{=} \{\max_{i=1}^n \|\check{\boldsymbol{x}}_i\|_2^2 \le 8q\mathbb{E}\|\check{\boldsymbol{x}}_i\|_2^2\}$ for $q \ge 1$. Then, it follows from Theorem 2.1 in Hsu et al. (2012) that $\mathbb{P}(E_q^c) \le ne^{-p_0 q}$. Under event E_q, we have

$$\begin{aligned}
&\Big\|\Big\{\frac{1}{m}\sum_{i\in\mathcal{I}_t}\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}_1) - \mathbb{E}\big[\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}_1)\big]\Big\} - \Big\{\frac{1}{m}\sum_{i\in\mathcal{I}_t}\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}_2) - \mathbb{E}\big[\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}_2)\big]\Big\}\Big\|_2 \\
&\le \Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t}\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}_1) - \frac{1}{m}\sum_{i\in\mathcal{I}_t}\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}_2)\Big\|_2 + \Big\|\mathbb{E}\big[\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}_1)\big] - \mathbb{E}\big[\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}_2)\big]\Big\|_2 \\
&\le C_3'\,\|\check{\boldsymbol{\eta}}_1 - \check{\boldsymbol{\eta}}_2\|_2\Big(\frac{1}{m}\sum_{i\in\mathcal{I}_t}\|\check{\boldsymbol{x}}_i\|_2^2 + \mathbb{E}\|\check{\boldsymbol{x}}_i\|_2^2\Big) \le 9\,C_3'\,p_0\,q\,\|\check{\boldsymbol{\eta}}_1 - \check{\boldsymbol{\eta}}_2\|_2,
\end{aligned}$$

where C_3' is a constant only depending on the bound of $\nabla^2 b(\cdot)$. Denote as $\mathcal{N}_\delta$ a δ-covering of $B\big((D^*)^{1/2}\tilde{\boldsymbol{\eta}}, \|D^*\|_2^{1/2}R_1\big)$. Note that the covering number of the net $\mathcal{N}_\delta$ is less than the δ-packing number (van der Vaart and Wellner, 1996), which is the maximum number of δ-separated points (the distance between each pair of points is strictly larger than δ). Write the corresponding collection of δ-separated points in $B\big((D^*)^{1/2}\tilde{\boldsymbol{\eta}}, \|D^*\|_2^{1/2}R_1\big)$ as $\{\check{\boldsymbol{\eta}}_g\}_{g=1}^G$. Because the sum of the volumes of the balls $\{B(\check{\boldsymbol{\eta}}_g, \delta)\}_{g=1}^G$ is less than the volume of $B\big((D^*)^{1/2}\tilde{\boldsymbol{\eta}}, \|D^*\|_2^{1/2}R_1 + \delta\big)$, it is easy to see that the covering number $|\mathcal{N}_\delta| \le (1 + \|D^*\|_2^{1/2}R_1/\delta)^{p_0}$. Take $\delta = \epsilon/(9\,C_3'\,p_0\,q)$. Then, under events E_q, we have

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{\check{\boldsymbol{\eta}}\in B\left((D^*)^{1/2}\tilde{\boldsymbol{\eta}},\, \|D^*\|_2^{1/2}R_1\right)}\Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t}\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}) - \mathbb{E}\big[\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}})\big]\Big\|_2 \ge 2\epsilon\Big) \\
&\qquad \le \mathbb{P}\Big(\sup_{\check{\boldsymbol{\eta}}\in\mathcal{N}_\delta}\Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t}\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}) - \mathbb{E}\big[\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}})\big]\Big\|_2 \ge \epsilon\Big).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{P}(\Delta \ge 2\epsilon) &\le \mathbb{P}(E_q^c) + T\,\mathbb{P}\Big(E_q \cap \Big\{\sup_{\check{\boldsymbol{\eta}}\in\mathcal{N}_\delta}\Big\|\frac{1}{m}\sum_{i\in\mathcal{I}_t}\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}}) - \mathbb{E}\big[\check{\boldsymbol{x}}_i\nabla b(\check{\boldsymbol{x}}_i^T\check{\boldsymbol{\eta}})\big]\Big\|_2 \ge \epsilon\Big\}\Big) \\
&\le ne^{-p_0 q} + C_4'\,T\,e^{p_0\log\frac{qp_0}{\epsilon} - C_2'\,\min\{\frac{\epsilon}{\sqrt{p_0}}, \frac{\epsilon^2}{p_0}\}\,m}.
\end{aligned}$$

Choosing $q = m^{1/3} \to \infty$ and $\epsilon = \sqrt{\log n/m}$, we conclude that $\Delta_m = O_p(\sqrt{\log n/m})$ under the setting $m = n^a$ with $a \in (0, 1]$.

Appendix B. Regularity conditions

some regularity conditions discussed in the paper are presented here. For notational convenience, we write $n_0 = n$, $\boldsymbol{\gamma}_0 = \boldsymbol{\eta}$, $h_0^*(\boldsymbol{v}, \boldsymbol{\gamma}_0) = f^*(\boldsymbol{v}, \boldsymbol{\eta})$, and $\boldsymbol{\xi}_0 = \boldsymbol{\xi}$.

B1 The parameter space Θ_k of $\boldsymbol{\gamma}_k$ contains an open set $\Theta_k^o \in R^{p_k}$, and the true value $\boldsymbol{\gamma}_k^*$ is an interior point of Θ_k^o, for $k = 0, \ldots, K$.

B2 The density function $h_k^*(\boldsymbol{\gamma}_k)$ is identifiable, that is, $h_k^*(\boldsymbol{\gamma}_{k,1}) = h_k^*(\boldsymbol{\gamma}_{k,2}) \Rightarrow \boldsymbol{\gamma}_{k,1} = \boldsymbol{\gamma}_{k,2}$. The support $\{\boldsymbol{v}\big|h_k^*(\boldsymbol{v}, \boldsymbol{\gamma}_k) > 0\}$ is independent of the unknown parameter $\boldsymbol{\gamma}_k$.

B3 The density function $h_k^*(\boldsymbol{\gamma}_k)$ admits third derivatives for $\boldsymbol{\gamma}_k \in \Theta_k^o$.

B4 The logarithm of the density $h_k^*(\boldsymbol{\gamma}_k)$ satisfies the unbiased condition $\mathbb{E}\{\partial \log h_k^*(\boldsymbol{v}, \boldsymbol{\gamma}_k^*)/\partial \boldsymbol{\gamma}_k\} = 0$, and the Fisher information matrix

$$I_k(\boldsymbol{\gamma}_k) = \mathbb{E}\{\partial \log h_k^*(\boldsymbol{v}, \boldsymbol{\gamma}_k)/\partial \boldsymbol{\gamma}_k \cdot \partial \log h_k^*(\boldsymbol{v}, \boldsymbol{\gamma}_k)/\partial \boldsymbol{\gamma}_k^T\} = -\mathbb{E}\{\partial^2 \log h_k^*(\boldsymbol{v}, \boldsymbol{\gamma}_k)/\partial \boldsymbol{\gamma}_k \partial \boldsymbol{\gamma}_k^T\}$$

is positive definite for $\boldsymbol{\gamma}_k \in \Theta_k^o$, $k = 1, \ldots, K$.

B5 There exist integrable functions $G_{i,j,l}(\boldsymbol{v})$ such that the third partial derivatives

$$\left| \frac{\partial^3 \log h_k^*(\boldsymbol{v}, \boldsymbol{\gamma}_k)}{\partial \gamma_k^i \, \partial \gamma_k^j \, \partial \gamma_k^l} \right| \leq G_{i,j,l}(\boldsymbol{v})$$

for all $\boldsymbol{\gamma}_k \in \Theta_k^o$ and $i, j, l = 1, \ldots, p_k$. Here, the notation $\boldsymbol{\gamma}^i$ denotes the ith component of vector $\boldsymbol{\gamma}$.

B6 The matrix $\sum_{k=0}^K c_k J_k(\boldsymbol{\theta})^T I_k J_k(\boldsymbol{\theta})$ is positive definite for $\boldsymbol{\theta} \in \Theta$, where Θ is the parameter space that contains the interior point $\boldsymbol{\theta}^*$, which is the true value of $\boldsymbol{\theta}$, and $J_k(\boldsymbol{\theta}) = \partial \boldsymbol{\gamma}_k / \partial \boldsymbol{\theta}$ is the Jacobian matrix of the mapping $\boldsymbol{\xi}_k$ with respect to $\boldsymbol{\theta}$.

References

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3):1352 – 1382, 2018.

Tianxi Cai, Molei Liu, and Yin Xia. Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *Journal of the American Statistical Association*, 117(540):2105–2119, 2022.

Lanjue Chen and Yong Zhou. Quantile regression in big data: A divide and conquer based strategy. *Computational Statistics & Data Analysis*, 144:106892, 2020.

Xi Chen, Weidong Liu, and Yichen Zhang. First-order Newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, 117 (540):1858–1874, 2022.

Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24:1655–1684, 2014.

Jianqing Fan, Yongyi Guo, and Kaizheng Wang. Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*, 2021. doi: 10.1080/01621459.2021.1969238.

Thomas S. Ferguson. *A Course in Large Sample Theory*. London: Chapman and Hall, 1996.

Jinyong Hahn and Whitney Newey. Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, 72(4):1295–1319, 2004.

Qianchuan He, Hao Helen Zhang, Christy L. Avery, and D. Y. Lin. Sparse meta-analysis with high-dimensional data. *Biostatistics*, 17(2):205–220, 2016.

Daniel J. Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 16:1–6, 2012.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(82):2869–2909, 2014.

Michael I. Jordan, Jason D. Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

Jason D. Lee, Qiang Liu, Yuekai Sun, and Jonathan E. Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

Kung Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

N. Lin and R. Xi. Fast surrogates of u-statistics. *Computational Statistics and Data Analysis*, 54(1):16–24, 2010.

Dungang Liu, Regina Y. Liu, and Minge Xie. Multivariate meta-analysis of heterogeneous studies using only summary statistics: Efficiency and robustness. *Journal of the American Statistical Association*, 110(509):326–340, 2015.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1st edition, 2014.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. New York: Springer, 2006.

Giovanni Parmigiani and Francesca Dominici. Combining studies with continuous and dichotomous responses: A latent-variables approach. *Meta-Analysis in Medicine and Health Policy*, page 105–125, 2000.

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

Tore Schweder and Nils Lid Hjort. Confidence and likelihood. *Scandinavian Journal of Statistics*, 29(2):309–332, 2002.

Jieli Shen, Regina Y. Liu, and Min ge Xie. iFusion: Individualized fusion learning. *Journal of the American Statistical Association*, 115(531):1251–1267, 2020.

M. C. Simmonds and J. P. T. Higgins. Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. *Statistics in Medicine*, 26(15):2982–2999, 2007.

Kesar Singh, Min-ge Xie, and William E. Strawderman. Combining information from independent sources through confidence distributions. *Annals of Statistics*, 33:159–183, 2005.

Kesar Singh, Minge Xie, and William E. Strawderman. Confidence distribution (CD): Distribution estimator of a parameter. *Lecture Notes-Monograph Series*, 54:132–150, 2007.

Tony Sit and Yue Xing. Distributed censored quantile regression. *Journal of Computational and Graphical Statistics*, 32(4):1685–1697, 2023.

Alexander J. Sutton and Julian P. T. Higgins. Recent developments in meta-analysis. *Statistics in Medicine*, 27(5):625–650, 2008.

Lu Tang, Ling Zhou, and Peter X.-K. Song. Distributed simultaneous inference in generalized linear models via confidence distribution. *Journal of Multivariate Analysis*, 176, 2020.

Sara A. van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202, 2014.

Aad van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer-Verlag, 1996.

Stanislav Volgushev, Shih-Kang Chao, and Guang Cheng. Distributed inference for quantile regression processes. *The Annals of Statistics*, 47(3):1634 – 1662, 2019.

Binhuan Wang, Yixin Fang, Heng Lian, and Hua Liang. Additive partially linear models for massive heterogeneous data. *Electronic Journal of Statistics*, 13(1):391 – 431, 2019.

Anne Whitehead, Andrea J. Bailey, and Diana Elbourne. Combining summaries of binary outcomes with those of continuous outcomes in a meta-analysis. *Journal of Biopharmaceutical Statistics*, 9(1):1–16, 1999.

Ruibin Xi and Nan Lin. Direct regression modelling of high-order moments in big data. *Statistics and Its Interface*, 9(4):445–452, 2016.

Min-ge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1):3–39, 2013.

Minge Xie, Kesar Singh, and William E. Strawderman. Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, 106(493):320–333, 2011.

Miaomiao Yu, Jiaxuan Li, and Yong Zhou. Enhancements of communication-efficient distributed statistical inference and its privacy preservation. *Journal of Econometrics*, 253: 106125, 2026. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2025.106125.

Fahd Aziz Zarrouf, Steven Artz, James Griffith, Cristian Sirbu, and Martin Kommor. Testosterone and depression: systematic review and meta-analysis. *Journal of psychiatric practice*, 15(4):289—305, 2009.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):217–242, 2014.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(1): 3321–3363, 2013.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

Tianqi Zhao, Guang Cheng, and Han Liu. A partially linear framework for massive heterogeneous data. *The Annals of Statistics*, 44(4):1400 – 1437, 2016.