Journal of Machine Learning Research 21 (2020) 1-35 Submitted 12/19; Revised 6/20; Published 7/20

Doubly Distributed Supervised Learning and Inference with High-Dimensional Correlated Outcomes

Emily C. Hector EHECTOR@NCSU.EDU
Department of Statistics
North Carolina State University
Raleigh, NC 27695, USA

Peter X.-K. Song PXSONG@UMICH.EDU
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48104, USA

Editor: Zhihua Zhang

Abstract

This paper presents a unified framework for supervised learning and inference procedures using the divide-and-conquer approach for high-dimensional correlated outcomes. We propose a general class of estimators that can be implemented in a fully distributed and parallelized computational scheme. Modeling, computational and theoretical challenges related to high-dimensional correlated outcomes are overcome by dividing data at both outcome and subject levels, estimating the parameter of interest from blocks of data using a broad class of supervised learning procedures, and combining block estimators in a closed-form meta-estimator asymptotically equivalent to estimates obtained by Hansen (1982)'s generalized method of moments (GMM) that does not require the entire data to be reloaded on a common server. We provide rigorous theoretical justifications for the use of distributed estimators with correlated outcomes by studying the asymptotic behaviour of the combined estimator with fixed and diverging number of data divisions. Simulations illustrate the finite sample performance of the proposed method, and we provide an R package for ease of implementation.

Keywords: Divide-and-conquer, Generalized method of moments, Estimating functions, Parallel computing, Scalable computing

1. Introduction

Although the divide-and-conquer paradigm has been widely used in statistics and computer science, its application with correlated data has been little investigated in the literature. We provide a theoretical justification, with theoretical guarantees, for divide-and-conquer methods with correlated data through a general unified estimating function theory framework. In particular, in this paper we focus on the large sample properties of a class of distributed and integrated estimators for supervised learning and inference with high-dimensional correlated outcomes. We consider N independent observations $\{\boldsymbol{y}_i, \boldsymbol{X}_i\}_{i=1}^N$ where both the sample size N and the dimension M of the response vector $\boldsymbol{y}_i$ may be so big that a direct analysis of the data using conventional methodology is computationally intensive, or even prohibitive. Such data may arise, for example, from imaging measurements of brain

activity or from genomic data. Denote by $f(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\Gamma}_i)$ the M-variate joint parametric distribution of $\boldsymbol{Y}_i$ conditioned on $\boldsymbol{X}_i$, where $\boldsymbol{\theta}$ is the parameter of interest and $\boldsymbol{\Gamma}_i$ contains parameters, such as for high-order dependencies, that may be difficult to model or handle computationally.

Statistical inference with big data can be extremely challenging due to the high volume and high variety of these data, as noted recently by Secchi (2018). In the statistics literature, methodological efforts to date have primarily focused on high-dimensional covariates (i.e. high-dimensional $\boldsymbol{X}_i$) with univariate responses (corresponding to $M = 1$); see Johnstone and Titterington (2009) for an overview of the difficulties and methods in linear regression, and the citations therein for references to the extensive publications in this field. By contrast, little work has focused on high-dimensional correlated outcomes (corresponding to large M), which pose an entirely new and different set of methodological challenges stemming from a high-dimensional likelihood. The divide-and-combine paradigm holds promise in overcoming these challenges; see Mackey et al. (2011) and Zhang et al. (2015b) for early examples of the power of divide-and-combine algorithms. Some recent divide-and-combine methods for independent outcomes can be found in Singh et al. (2005), Lin and Zeng (2010), Lin and Xi (2011), Chen and Xie (2014), and Liu et al. (2015), among others.

More recently, Hector and Song (2020) proposed a Distributed and Integrated Method of Moments (DIMM), a divide-and-combine strategy for supervised learning and inference in a regression setting with high-dimensional correlated outcomes $\boldsymbol{Y}$. DIMM splits the M elements of $\boldsymbol{Y}$ into blocks of low-dimensional response subvectors, analyzes these blocks in a distributed and parallelized computational scheme using pairwise composite likelihood (CL), and combines block-specific results using a closed-form meta-estimator in a similar spirit to Hansen (1982)'s seminal generalized method of moments (GMM). DIMM overcomes computational challenges associated with high-dimensional outcomes by running block analyses in parallel and combining block-specific results via a computationally and statistically efficient closed-form meta-estimator. DIMM is easily implemented using MapReduce in the Hadoop framework (Khezr and Navimipour (2017)), where blocks of data are loaded only once and in parallel. DIMM presents a useful and natural extension of the classical GMM framework, which easily accounts for inter-block dependencies. DIMM also improves on the classical meta-estimation where results from blocks are routinely assumed to be independent. DIMM is still challenged, however, when estimating a homogeneous parameter in the presence of heterogeneous parameters. Additionally, it is also challenged computationally when the sample size N is large; the strategy of dividing high-dimensional vectors of correlated outcomes into blocks is insufficient to address the excessive computational demand, since the sample size remains large in the block analyses. Thus, another division at the subject level is inevitable to mitigate the computational burden arising from matrix inversions and iterative calculations in the block analyses.

This paper proposes a new doubly divided procedure to learn and perform inference for a homogeneous parameter of interest in the presence of heterogeneous parameters with a general class of supervised learning procedures. The double division at the response and subject levels further speeds up computations in comparison to DIMM and results in a double division of the data, visualized in Table 1: a division of the response $\boldsymbol{Y}$, and a random division of subjects into independent subject groups, resulting in blocks of data with a smaller sample of low-dimensional response subvectors. We consider a general class

of supervised learning procedures to analyze these blocks separately and in parallel that is substantially. Then we establish a GMM-type combination procedure that yields a meta-estimator of heterogeneous and homogeneous parameters. This proposed estimator is substantially more general than the DIMM estimator in Hector and Song (2020), which only considered pairwise composite likelihood estimation of homogeneous mean parameters, and thus appealing in many practical settings where analyzing data with both large M and N is challenging. We achieve a doubly divided learning and inference procedure implemented in a distributed and parallelized computational scheme. The proposed class of supervised learning procedures is very general, including many important estimation methods as special cases, such as Fisher's maximum likelihood, Wedderburn (1974)'s quasi-likelihood, Liang and Zeger (1986)'s generalized estimating equations, Huber (1964)'s M-estimation for robust inference, with possible extensions to semi-parametric and non-parametric models. We also provide a rigorous, well-defined and broad theoretical framework for the justification of divide-and-conquer schemes when the number of data divisions diverges, which was not considered in Hector and Song (2020).

Block \ Group	Subject 1	…	Subject n_1	…	…	Subject 1	…	Subject n_K
1	$y_{11,11}$	$\cdots$	$y_{n_1 1,11}$	$\cdots$	$\cdots$	$y_{11,1K}$	$\cdots$	$y_{n_K 1,1K}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m_1	$y_{1m_1,11}$	$\cdots$	$y_{n_1 m_1,11}$	$\cdots$	$\cdots$	$y_{1m_1,1K}$	$\cdots$	$y_{n_K m_1,1K}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	$y_{11,J1}$	$\cdots$	$y_{n_1 1,J1}$	$\cdots$	$\cdots$	$y_{11,JK}$	$\cdots$	$y_{n_K 1,JK}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m_J	$y_{1m_J,J1}$	$\cdots$	$y_{n_1 m_J,J1}$	$\cdots$	$\cdots$	$y_{11,JK}$	$\cdots$	$y_{n_K m_J,JK}$

Table 1: Double division of outcome data on both the dimension of responses (into blocks) and sample size (into groups).

The proposed Doubly Distributed and Integrated Method of Moments (DDIMM) not only provides a unified framework of various supervised learning procedures of parameters with heterogeneity under the divide-and-combine paradigm, but provides key theoretical guarantees for statistical inference, such as consistency and asymptotic normality, while offering significant computational gains when response dimension M and sample size N are large. These are useful and innovative contributions to the arsenal of tools for high-dimensional correlated data analysis, and to the collection of divide-and-combine algorithms, which have so far concentrated on independently sampled data. In this paper, we focus on the theoretical aspects of doubly distributed learning and inference, including a goodness-of-fit test based on a χ^2 statistic. We also study consistency and asymptotic normality of the proposed estimator as the number of data divisions diverges. This includes theoretical justifications for distributed inference when the dimension of the response and the number of response

divisions diverges, which allows the analysis of highly dense outcome data.
The rest of the paper is organized as follows. Section 2 describes the DDIMM, with examples introduced in Section 3. Section 4 discusses large sample properties of the proposed DDIMM. Section 5 presents the main contribution of the paper, a closed-form meta-estimator and its implementation in a parallel and scalable computational scheme. Section 6 illustrates the DDIMM's finite sample performance with simulations. Section 7 concludes with a discussion. Additional proofs and simulation results are deferred to the Appendices and Supplemental Material. An R package is available in the Supplemental Material.

2. Formulation

We begin with some notation. Let $\|\cdot\|$ be the ℓ_2-norm for a D-dimensional vector $\boldsymbol{a}$ and a $D_1 \times D_2$-dimensional matrix $\boldsymbol{A}$ defined by, respectively:

$$\begin{aligned} \|\boldsymbol{a}\| &= \left(\sum_{d=1}^{D} a_d^2\right)^{1/2} && \text{for} \quad \boldsymbol{a} = [a_d]_{d=1}^{D} \in \mathbb{R}^D, \\ \|\boldsymbol{A}\| &= \left(\sum_{d_1=1}^{D_1}\sum_{d_2=1}^{D_2} A_{d_1 d_2}^2\right)^{1/2} && \text{for} \quad \boldsymbol{A} = [A_{d_1 d_2}]_{d_1,d_2=1}^{D_1,D_2} \in \mathbb{R}^{D_1 \times D_2}. \end{aligned}$$

We define the stacking operator $\mathbb{S}(\cdot)$ for matrices $\{\boldsymbol{A}_{jk}\}_{j=1,k=1}^{J,K}$, $\boldsymbol{A}_{jk} \in \mathbb{R}^{D_1^{jk} \times D_2}$, as

$$\begin{aligned} \mathbb{S}\left(\boldsymbol{A}_{jk}, \boldsymbol{A}_{j'k'}\right) &= \begin{pmatrix} \boldsymbol{A}_{jk}^T & \boldsymbol{A}_{j'k'}^T \end{pmatrix}^T \in \mathbb{R}^{(D_1^{jk}+D_1^{j'k'}) \times D_2}, \\ \mathbb{S}^J\left(\boldsymbol{A}_{jk}\right) &= \begin{pmatrix} \boldsymbol{A}_{1k}^T & \dots & \boldsymbol{A}_{Jk}^T \end{pmatrix}^T \in \mathbb{R}^{D_1^k \times D_2}, \\ \mathbb{S}^{JK}\left(\boldsymbol{A}_{jk}\right) &= \begin{pmatrix} \boldsymbol{A}_{11}^T & \dots & \boldsymbol{A}_{J1}^T & \dots & \boldsymbol{A}_{1K}^T & \dots & \boldsymbol{A}_{JK}^T \end{pmatrix}^T \in \mathbb{R}^{D_1 \times D_2}, \end{aligned}$$

where $D_1^k = \sum_{j=1}^{J} D_1^{jk}$, $D_1 = \sum_{k=1}^{K} D_1^k$. Consider the collection of samples $\{\boldsymbol{y}_i, \boldsymbol{X}_i\}_{i=1}^{N}$, where $\boldsymbol{X}_i \in \mathbb{R}^{M \times q}$ is fixed, $\boldsymbol{Y}_i \in \mathbb{R}^M$, $q, M \in \mathbb{N}$. The number of covariates q is considered fixed in this paper. Let $\boldsymbol{\theta}, \boldsymbol{\zeta}$ take values in parameter spaces $\Theta \subseteq \mathbb{R}^p$, $\Xi \subseteq \mathbb{R}^d$, both compact subsets of p- and d-dimensional Euclidean space respectively. Let $p, d \in \mathbb{N}$, and consider $\boldsymbol{\theta}$ to be the parameter of interest, and $\boldsymbol{\zeta}$ to be a potentially large vector of parameters of secondary interest. Let $\boldsymbol{\theta}_0 \in \Theta, \boldsymbol{\zeta}_0 \in \Xi$ be the true values of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ respectively. Consider a class $\mathcal{P} = \{P_{\boldsymbol{\theta},\boldsymbol{\zeta}}\}$ of parametric models with associated estimating functions $\boldsymbol{\Psi}$ of parameter $\boldsymbol{\theta}$ (e.g. $\boldsymbol{\Psi}$ can be the derivative of some objective function). Suppose we want to learn the parameter $\boldsymbol{\theta}$ by finding the root of $\boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\zeta}) = \mathbf{0}$, which is computationally intensive or even prohibitive due to the large dimension M of $\boldsymbol{y}$, the large sample size N, or the large dimension d of $\boldsymbol{\zeta}$. We focus on a divide-and-combine approach utilizing modern distributed computing platforms to alleviate the computational and modeling challenges posed by analyzing the whole data.

2.1. Double Data Split Procedure

First, for each subject i, DDIMM divides the M-dimensional response $\boldsymbol{y}_i$ and its associated covariates into J blocks, denoted by:

$$\boldsymbol{y}_i = \begin{pmatrix} \boldsymbol{y}_{i,1}^T & \dots & \boldsymbol{y}_{i,J}^T \end{pmatrix}^T \text{ and } \boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{X}_{i,1}^T & \dots & \boldsymbol{X}_{i,J}^T \end{pmatrix}^T, \; i = 1, \dots, N.$$

Division into blocks is not restricted to the order of data entry: responses may be grouped according to pre-specified block memberships, according to, say, substantive scientific knowledge, such as functional regions of the brain. In this paper, with no loss of generality, we use the order of data entry in the data division procedure. Further, DDIMM randomly splits the N independent subjects to form K disjoint subject groups $\left\{\boldsymbol{y}_{i,jk}, \boldsymbol{X}_{i,jk}\right\}_{i=1}^{n_k}$. Then each group has sample size n_k, $k = 1, \ldots, K$, with $\sum_{k=1}^{K} n_k = N$. Refer to Table 1 for notation detail. For ease of exposition, we henceforth use the term "group" to refer to the division along subjects, and "block" to refer to the division along responses. We also use the term "block" to refer to the division along both responses and subjects.

We call $\left\{\boldsymbol{y}_{i,jk}, \boldsymbol{X}_{i,jk}\right\}_{i=1}^{n_k}$ block (j,k), $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Within block (j,k), let m_j be the dimension of the sub-response, $\boldsymbol{y}_{i,jk} = (y_{i1,jk}, \ldots, y_{im_j,jk})^T \in \mathbb{R}^{m_j}$, and $\boldsymbol{X}_{i,jk} \in \mathbb{R}^{m_j \times q}$ the associated covariate matrix, with $\sum_{j=1}^{J} m_j = M$. For each block $j \in \{1, \ldots, J\}$, we have K independent subject groups $\left\{\boldsymbol{y}_{i,jk}\right\}_{i=1,k=1}^{n_k,K}$. In contrast, each group $k \in \{1, \ldots, K\}$ has n_k subjects and for each subject $i \in \{1, \ldots, n_k\}$, the J response blocks $\left\{\boldsymbol{y}_{i,jk}\right\}_{j=1}^{m_j}$ are dependent.

The primary task is to solve $\boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\zeta}) = \mathbf{0}$ to learn parameter $\boldsymbol{\theta}$ ina supervised way over the entire data. Given the above double data split scheme, this task becomes a divide-and-combine procedure: the first step is to solve the following system of block-specific estimating equations: for $j \in \{1, \ldots, J\}$, $k \in \{1, \ldots, K\}$,

$$\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}; \boldsymbol{y}_{jk}, \boldsymbol{\zeta}_{jk}) = \mathbf{0}, \tag{1}$$

$$\boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk}; \boldsymbol{y}_{jk}, \boldsymbol{\theta}) = \mathbf{0}, \tag{2}$$

where $\boldsymbol{G}_{jk}$ is an estimating function used to learn parameters $\boldsymbol{\zeta}_{jk}$ (e.g. correlation parameters) that are allowed to be heterogeneous across blocks such that $\boldsymbol{\zeta} = \mathbb{S}^{JK}\left(\boldsymbol{\zeta}_{jk}\right)$. The true values $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0})$ of $(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})$ are the values such that $E_{\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}} \mathbb{S}(\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}_0; \boldsymbol{y}_{jk}, \boldsymbol{\zeta}_{jk0}), \boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk0}; \boldsymbol{y}_{jk}, \boldsymbol{\theta}_0)) = \mathbf{0}$. Parameters $\boldsymbol{\zeta}_{jk0}$ take values in parameter space $\Xi_{jk} \subset \mathbb{R}^{d_{jk}}$ for some $d_{jk} > 0$ such that $\boldsymbol{\zeta}_0 = \mathbb{S}^{JK}\left(\boldsymbol{\zeta}_{jk0}\right)$, $\Xi = \times_{j=1,k=1}^{J,K} \Xi_{jk}$, $d = \sum_{k=1}^{K} \sum_{j=1}^{J} d_{jk}$. Let $\boldsymbol{\zeta}_{k0} = \mathbb{S}^{J}\left(\boldsymbol{\zeta}_{jk0}\right)$ and $\boldsymbol{\zeta}_k = \mathbb{S}^{J}\left(\boldsymbol{\zeta}_{jk}\right)$. This is a similar approach to GEE2, proposed by Zhao and Prentice (1990), with details also in Liang et al. (1992), where unbiased estimating equations for the nuisance parameters are added in order to guarantee consistency. In this way, we impose homogeneity of the parameter of interest $\boldsymbol{\theta}$ across blocks but allow heterogeneity of the parameters of secondary interest. We assume that the class of parametric models $\mathcal{P}$ yields block-specific estimating functions satisfying the following regularity assumptions:

(A.1) (i) $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ are unbiased; that is, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{\zeta}_{jk} \in \Xi_{jk}$, $E_{\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}} \mathbb{S}(\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}; \boldsymbol{Y}_{jk}, \boldsymbol{\zeta}_{jk}), \boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk}; \boldsymbol{Y}_{jk}, \boldsymbol{\theta})) = \mathbf{0}$.

(ii) $E_{\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}} \mathbb{S}\left(\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}; \boldsymbol{Y}_{jk}, \boldsymbol{\zeta}_{jk}), \boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk}; \boldsymbol{y}_{jk}, \boldsymbol{\theta})\right)$ has a unique zero at $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0})$.

(iii) $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ are additive: for some kernel inference functions $\boldsymbol{\psi}_{jk}$ and $\boldsymbol{g}_{jk}$, they take the form

$$\begin{pmatrix} \boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}; \boldsymbol{y}_{jk}, \boldsymbol{\zeta}_{jk}) \\ \boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk}; \boldsymbol{y}_{jk}, \boldsymbol{\theta}) \end{pmatrix} = \frac{1}{n_k} \sum_{i=1}^{n_k} \begin{pmatrix} \boldsymbol{\psi}_{jk}(\boldsymbol{\theta}; \boldsymbol{y}_{i,jk}, \boldsymbol{\zeta}_{jk}) \\ \boldsymbol{g}_{jk}(\boldsymbol{\zeta}_{jk}; \boldsymbol{y}_{i,jk}, \boldsymbol{\theta}) \end{pmatrix}.$$

We define $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ as being "weakly regular" based on the above conditions (A.1) (i)-(iii) in which the defining properties of a regular inference function are applied to its mean; see Song (2007) Chapter 3.5 for a definition of regular inference functions. Additional conditions on the class $\mathcal{P}$ will be described throughout the paper where appropriate. Within block (j,k), denote by $\widehat{\boldsymbol{\theta}}_{jk}$ and $\widehat{\boldsymbol{\zeta}}_{jk}$ the joint solution to Equations 1 and 2, estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}_{jk}$ respectively. For notation purposes, let $\widehat{\boldsymbol{\theta}}_{list} = \mathbb{S}^{JK}(\widehat{\boldsymbol{\theta}}_{jk})$, $\widehat{\boldsymbol{\zeta}}_{k} = \mathbb{S}^{J}(\widehat{\boldsymbol{\zeta}}_{jk})$, and $\widehat{\boldsymbol{\zeta}}_{list} = \mathbb{S}^{JK}(\widehat{\boldsymbol{\zeta}}_{jk})$. Due to the homogeneity of $\boldsymbol{\theta}$, the next step is integration of the block-specific estimators $\widehat{\boldsymbol{\theta}}_{jk}$. By contrast, $\widehat{\boldsymbol{\zeta}}_{jk}$ remain heterogeneous and potentially high-dimensional. In the rest of the paper, for convenience of notation, we suppress the dependence of $\boldsymbol{\Psi}_{jk}$, $\boldsymbol{G}_{jk}$, $\boldsymbol{\psi}_{jk}$ and $\boldsymbol{g}_{jk}$ on $\boldsymbol{y}_{jk}$ and $\boldsymbol{y}_{i,jk}$:

$$\begin{aligned}\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta};\boldsymbol{\zeta}_{jk}) &= \boldsymbol{\Psi}_{jk}(\boldsymbol{\theta};\boldsymbol{y}_{jk},\boldsymbol{\zeta}_{jk}), \quad \boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk};\boldsymbol{\theta}) = \boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk};\boldsymbol{y}_{jk},\boldsymbol{\theta}),\\ \boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta};\boldsymbol{\zeta}_{jk}) &= \boldsymbol{\psi}_{jk}(\boldsymbol{\theta};\boldsymbol{y}_{i,jk},\boldsymbol{\zeta}_{jk}), \quad \boldsymbol{g}_{i,jk}(\boldsymbol{\zeta}_{jk};\boldsymbol{\theta}) = \boldsymbol{g}_{jk}(\boldsymbol{\zeta}_{jk};\boldsymbol{y}_{i,jk},\boldsymbol{\theta}).\end{aligned}$$

2.2. Integration

Integrating block estimates $\widehat{\boldsymbol{\theta}}_{jk}$ into an estimator of $\boldsymbol{\theta}$, denoted by $\widehat{\boldsymbol{\theta}}_{c}$, will yield a more efficient estimate of $\boldsymbol{\theta}$. In the integration step, our intuition is to treat each system of equations $\mathbb{S}\left(\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta};\boldsymbol{\zeta}_{jk}), \boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk};\boldsymbol{\theta})\right) = \mathbf{0}$ as a "moment condition" on $\boldsymbol{\theta}$ contributed by block (j,k), $j = 1,\ldots,J$, $k = 1,\ldots,K$. Technically, we want to derive an estimator $\widehat{\boldsymbol{\theta}}_{c}$ of $\boldsymbol{\theta}$ that satisfies all JK moment conditions that effectively makes use of the JK estimates of $\boldsymbol{\theta}$ obtained from Equations 1 and 2. To address the issue that $\boldsymbol{\theta}$ is over-identified by the JK moment conditions, we invoke Hansen (1982)'s seminal generalized method of moments (GMM) to combine the moment conditions that arise from each block. Another significant advantage of GMM is that it allows us to incorporate between-block dependencies, which cannot be easily done in classical meta-estimation. To this end, define the subject group indicator $\delta_i(k) = \mathbb{1}(\text{subject } i \text{ is in blocks } (j,k) \text{ for some } k \in \{1,\ldots,K\} \text{ and for all } j = 1,\ldots,J)$ for $i = 1,\ldots,N$, $k = 1,\ldots,K$. For subject i, let

$$\boldsymbol{\psi}_i(\boldsymbol{\theta};\boldsymbol{\zeta}) = \mathbb{S}^{JK}\left(\delta_i(k)\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta};\boldsymbol{\zeta}_{jk})\right), \quad \boldsymbol{g}_i(\boldsymbol{\zeta};\boldsymbol{\theta}) = \mathbb{S}^{JK}\left(\delta_i(k)\boldsymbol{g}_{i,jk}(\boldsymbol{\zeta}_{jk};\boldsymbol{\theta})\right),$$

where clearly only one $\mathbb{S}^{J}\left(\delta_i(k)\boldsymbol{\psi}_{i,jk}^{T}(\boldsymbol{\theta};\boldsymbol{\zeta}_{jk})\right)$ is non-zero. Let $\boldsymbol{a}^{\otimes 2}$ denote the outer product of a vector $\boldsymbol{a}$ with itself, namely $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{T}$. Then we can define $\boldsymbol{\Psi}_N(\boldsymbol{\theta};\boldsymbol{\zeta}) = (1/N)\sum_{i=1}^{N}\boldsymbol{\psi}_i(\boldsymbol{\theta};\boldsymbol{\zeta})$. It is easy to show that

$$\boldsymbol{\Psi}_N(\boldsymbol{\theta};\boldsymbol{\zeta}) = \frac{1}{N}\mathbb{S}^{JK}\left(\sum_{i=1}^{n_k}\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta};\boldsymbol{\zeta}_{jk})\right) = \frac{1}{N}\mathbb{S}^{JK}\left(n_k\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta};\boldsymbol{\zeta}_{jk})\right).$$

Similarly, define $\boldsymbol{G}_N(\boldsymbol{\zeta};\boldsymbol{\theta}) = (1/N)\sum_{i=1}^{N}\boldsymbol{g}_i(\boldsymbol{\zeta};\boldsymbol{\theta}) = (1/N)\mathbb{S}^{JK}\left(n_k\boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk};\boldsymbol{\theta})\right)$. Since $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ satisfy assumptions (A.1) for each j and k, $\boldsymbol{\Psi}_N$ and $\boldsymbol{G}_N$ are additive, unbiased, and $E_{\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0}\mathbb{S}\left(\boldsymbol{\Psi}_N(\boldsymbol{\theta};\boldsymbol{\zeta}), \boldsymbol{G}_N(\boldsymbol{\zeta};\boldsymbol{\theta})\right)$ has a unique zero at $(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)$. For convenience, we denote

$$\boldsymbol{T}_N(\boldsymbol{\theta},\boldsymbol{\zeta}) = \begin{pmatrix}\boldsymbol{\Psi}_N(\boldsymbol{\theta};\boldsymbol{\zeta})\\ \boldsymbol{G}_N(\boldsymbol{\zeta};\boldsymbol{\theta})\end{pmatrix}, \quad \boldsymbol{\tau}_i(\boldsymbol{\theta},\boldsymbol{\zeta}) = \begin{pmatrix}\boldsymbol{\psi}_i(\boldsymbol{\theta};\boldsymbol{\zeta})\\ \boldsymbol{g}_i(\boldsymbol{\zeta};\boldsymbol{\theta})\end{pmatrix}. \tag{3}$$

We assume that the class $\mathcal{P}$ yields $\boldsymbol{\psi}$, $\boldsymbol{g}$ satisfying the following conditions:

(A.2) (i) Both $\boldsymbol{\psi}_{jk}$ and $\boldsymbol{g}_{jk}$ are Lipschitz continuous in $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, namely for $j \in \{1, \ldots, J\}$, $k \in \{1, \ldots, K\}$, and some constants $c_{jk}, b_{jk} > 0$, for all $\left(\boldsymbol{\theta}_1, \boldsymbol{\zeta}_{jk1}\right), \left(\boldsymbol{\theta}_2, \boldsymbol{\zeta}_{jk2}\right)$ in a neighbourhood of $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0})$,

$$\left\|\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_1; \boldsymbol{\zeta}_{jk1}) - \boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_2; \boldsymbol{\zeta}_{jk2})\right\| \leq c_{jk} \left\|(\boldsymbol{\theta}_1, \boldsymbol{\zeta}_{jk1}) - (\boldsymbol{\theta}_2, \boldsymbol{\zeta}_{jk2})\right\|,$$
$$\left\|\boldsymbol{g}_{i,jk}(\boldsymbol{\zeta}_{jk1}; \boldsymbol{\theta}_1) - \boldsymbol{g}_{i,jk}(\boldsymbol{\zeta}_{jk2}; \boldsymbol{\theta}_2)\right\| \leq b_{jk} \left\|(\boldsymbol{\theta}_1, \boldsymbol{\zeta}_{jk1}) - (\boldsymbol{\theta}_2, \boldsymbol{\zeta}_{jk2})\right\|.$$

(ii) The sensitivity matrix $-\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}} E_{\boldsymbol{\theta},\boldsymbol{\zeta}} \boldsymbol{\tau}_i(\boldsymbol{\theta}, \boldsymbol{\zeta})$ is continuous in a compact neighbourhood $\mathbb{N}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ of $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$, and positive definite;

(iii) The variability matrix $E_{\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0}\left(\boldsymbol{\tau}_i(\boldsymbol{\theta}, \boldsymbol{\zeta})^{\otimes 2}\right)$ is finite and positive-definite.

Note that $\boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \boldsymbol{0}$ has no unique solution because its dimension is bigger than the dimension of $\boldsymbol{\theta}$. To overcome this issue, we follow Hansen's GMM for over-identified parameters. Let $\boldsymbol{W}$ be the weight matrix in the GMM Equation 4. Classical GMM theory states that any positive semi-definite matrix $\boldsymbol{W}$ can be used to guarantee consistency and asymptotic normality of the resulting estimator, and that an optimal choice of $\boldsymbol{W}$, corresponding to the inverse covariance of the estimating function $\boldsymbol{T}_N$ in Equation 3, leads to an efficient GMM estimator. In our setting, a possible formulation for a GMM estimator of $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ is

$$(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c) = \arg\min_{\boldsymbol{\theta},\boldsymbol{\zeta}} Q_N(\boldsymbol{\theta}, \boldsymbol{\zeta} | \boldsymbol{W}), \text{ where} \tag{4}$$
$$Q_N(\boldsymbol{\theta}, \boldsymbol{\zeta} | \boldsymbol{W}) = \boldsymbol{T}_N^T(\boldsymbol{\theta}, \boldsymbol{\zeta}) \boldsymbol{W} \boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}).$$

In Equation 4, the weight matrix $\boldsymbol{W}$ is a positive semi-definite $(JKp + d) \times (JKp + d)$ matrix. The heterogeneity of $\boldsymbol{\zeta}$ allowed by the use of $\boldsymbol{G}_N$ can lead to theoretical and computational challenges due to the high-dimensionality of the parameter, a problem from which GEE2 also suffers. See Chan et al. (1998) and Carey et al. (1993) for a discussion on the computational burden of inverting large matrices in GEE2. Note that block-specific estimators $\widehat{\boldsymbol{\zeta}}_{list}$ are consistent; the only possible improvement from re-learning $\boldsymbol{\zeta}$ in an iterative procedure between $\widehat{\boldsymbol{\theta}}_c$ and $\widehat{\boldsymbol{\zeta}}_c$ is a gain in efficiency. This is not necessary since $\boldsymbol{\zeta}$ are parameters of secondary interest and their efficiency is in general not of interest. We will derive a closed-form meta-estimator of $\boldsymbol{\theta}$ that avoids re-learning of $\boldsymbol{\zeta}$ in Section 5.
Following the work of Hansen (1982), we define a particular instance of the estimator in Equation 4 by specifying $\boldsymbol{W}$ as the inverse sample covariance of $\boldsymbol{T}_N$. We will show in Section 4 that this choice of $\boldsymbol{W}$ is optimal for the efficiency of the resulting estimator. Let $\widehat{\boldsymbol{V}}_N$ be the sample covariance of $\boldsymbol{T}_N(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$:

$$\widehat{\boldsymbol{V}}_N = \frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{\tau}_i(\widehat{\boldsymbol{\theta}}_{list}, \widehat{\boldsymbol{\zeta}}_{list})\right)^{\otimes 2} = \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} \boldsymbol{\psi}_i(\widehat{\boldsymbol{\theta}}_{list}; \widehat{\boldsymbol{\zeta}}_{list}) \\ \boldsymbol{g}_i(\widehat{\boldsymbol{\zeta}}_{list}; \widehat{\boldsymbol{\theta}}_{list}) \end{pmatrix}^{\otimes 2}, \tag{5}$$

where $\boldsymbol{\psi}_i(\widehat{\boldsymbol{\theta}}_{list}; \widehat{\boldsymbol{\zeta}}_{list}) = \mathbb{S}^{JK}\left(\delta_i(k) \boldsymbol{\psi}_{i,jk}(\widehat{\boldsymbol{\theta}}_{jk}; \widehat{\boldsymbol{\zeta}}_{jk})\right)$. Letting $\boldsymbol{W} = \widehat{\boldsymbol{V}}_N^{-1}$ yields the following optimal GMM estimator:

$$(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt}) = \arg\min_{\boldsymbol{\theta},\boldsymbol{\zeta}} \boldsymbol{T}_N^T(\boldsymbol{\theta}, \boldsymbol{\zeta}) \widehat{\boldsymbol{V}}_N^{-1} \boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}). \tag{6}$$

We assume that $\boldsymbol{W}$ and $\widehat{\boldsymbol{V}}_N$ are nonsingular; see Han and Song (2011) for optimal weighting matrix with QIF when the sample covariance is ill-defined. Before presenting large-sample properties of $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ and $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$ in Section 4, we demonstrate in Section 3 the flexibility of our framework through several important supervised learning methods.

3. Examples

We now present five examples to illustrate the flexibility of the unifying framework considered in this paper.

3.1. Likelihood-Based Methods

Consider the multidimensional regression model $h(\boldsymbol{\mu}_{i,jk}) = \boldsymbol{X}_{i,jk}(\ \boldsymbol{\theta}^T \quad \boldsymbol{\beta}_{jk}^T\ )^T$, where $\boldsymbol{\mu}_{i,jk} = E(\boldsymbol{Y}_{i,jk}|\boldsymbol{X}_{i,jk}, \boldsymbol{\theta}, \boldsymbol{\beta}_{jk})$ is the mean vector of $\boldsymbol{Y}_{i,jk}$ given $\boldsymbol{X}_{i,jk}$, $\boldsymbol{\beta}_{jk}$, and the p-dimensional parameter $\boldsymbol{\theta}$ ($p \leq q$ the number of covariates, which may include an intercept), and h is a known component-wise link function. Let $\boldsymbol{\zeta}_{jk}$ be parameters of the second-order moments of $\boldsymbol{Y}_{i,jk}$, such as dispersion parameters, and parameters in $\boldsymbol{\beta}_{jk}$ (which may be empty). If the full likelihood of $\boldsymbol{Y}_{i,jk}$ is computationally tractable, $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ correspond to the score functions, and $\widehat{\boldsymbol{\theta}}_{jk}$ and $\widehat{\boldsymbol{\zeta}}_{jk}$ may be given by the maximum likelihood estimates (MLEs). DDIMM can be applied straightforwardly by following the procedure in Section 2.
If the full likelihood is computationally intractable or difficult to construct, one can instead use pseudo-likelihoods such as the pairwise composite likelihood. The pairwise composite likelihood, originally proposed by Lindsay (1988) and detailed in Varin et al. (2011), provides the following forms of the equations for Equations 1 and 2:

$$\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}; \boldsymbol{\zeta}_{jk}) = \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{r=1}^{m_j - 1} \sum_{t=r+1}^{m_j} \nabla_{\boldsymbol{\theta}} \log f_j(y_{ir,jk}; y_{it,jk}; \boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}, \boldsymbol{X}_{i,jk}),$$

$$\boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk}; \boldsymbol{\theta}) = \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{r=1}^{m_j - 1} \sum_{t=r+1}^{m_j} \nabla_{\boldsymbol{\zeta}_{jk}} \log f_j(y_{ir,jk}; y_{it,jk}; \boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}, \boldsymbol{X}_{i,jk}),$$

for some bivariate marginal f_j which can be chosen according to the nature of the response data. As long as the bivariate marginals f_j are correctly specified, the composite score functions $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ satisfy the regularity conditions in (A.1). Hence the DDIMM can be used to overcome the computational challenges related to the MLE and pairwise composite likelihood. We refer readers to Chapter 6 of Song (2007) and Chapter 3 of Joe (2014) for details on constructing multivariate distributions using Gaussian and vine copulas respectively, but note that direct computation of the MLE is computationally very challenging when $m_j \geq 4$. Examples of applications of Gaussian copulas can be found in Song et al. (2009), Bodnar et al. (2010), Bai et al. (2014), and in the importance sampling algorithm proposed in Masarotto and Varin (2012), among others.

3.2. Generalized Estimating Equations

More generally, Wedderburn (1974)'s quasi-likelihood is a popular alternative method of supervised learning that does not require a fully specified multidimensional likelihood; it

receives a full treatment in Heyde (1997). Consider Liang and Zeger (1986)'s marginal mean model $h(\boldsymbol{\mu}_{i,jk}) = \boldsymbol{X}_{i,jk}(\ \boldsymbol{\theta}^T \quad \boldsymbol{\beta}_{jk}^T\ )^T$ for the analysis of longitudinal data, where $\boldsymbol{\mu}_{i,jk} = E(\boldsymbol{Y}_{i,jk}|\boldsymbol{X}_{i,jk}, \boldsymbol{\theta}, \boldsymbol{\beta}_{jk})$ is the marginal mean vector of serially correlated outcomes $\boldsymbol{Y}_{i,jk}$ given $\boldsymbol{X}_{i,jk}$, $\boldsymbol{\beta}_{jk}$, and the p-dimensional parameter $\boldsymbol{\theta}$ ($p \leq q$), and h is a known component-wise link function. In this setting, $\boldsymbol{\zeta}_{jk}$ consists of parameters in $\boldsymbol{\beta}_{jk}$ (which may be empty), parameters for the variances of $Y_{it,jk}$, $t = 1, \ldots, m_j$, and a nuisance parameter $\boldsymbol{\alpha}_{jk}$ which fully characterizes a working correlation matrix $\boldsymbol{R}_{jk}(\boldsymbol{\alpha}_{jk})$. In the case where $\boldsymbol{\beta}_{jk}$ is empty, the generalized estimating equation (GEE) proposed by Liang and Zeger (1986) yields the the kernel inference function $\boldsymbol{\psi}_{jk}(\boldsymbol{\theta}; \boldsymbol{\zeta}_{jk}) = \boldsymbol{D}_{i,jk}^T \boldsymbol{\Sigma}_{i,jk}^{-1} \boldsymbol{r}_{i,jk}$ in (A.1) (iii), where $\boldsymbol{D}_{i,jk} = \nabla_{\boldsymbol{\theta}} \boldsymbol{\mu}_{i,jk}$, $\boldsymbol{r}_{i,jk} = \boldsymbol{y}_{i,jk} - \boldsymbol{\mu}_{i,jk}$, and $\boldsymbol{\Sigma}_{i,jk} = \boldsymbol{A}_{i,jk} \boldsymbol{R}_{jk}(\boldsymbol{\alpha}_{jk}) \boldsymbol{A}_{i,jk}$, where $\boldsymbol{A}_{i,jk} = \text{diag}\left\{(Var(Y_{it,jk}))^{1/2}\right\}_{t=1}^{m_j}$. In GEE2, $\boldsymbol{G}_{jk}$ in Equation 2 is specified as another unbiased inference function satisfying (A.1) and (A.2). DDIMM provides a procedure for the application of distributed methods to high-dimensional longitudinal/clustered data.

3.3. M-Estimation

DDIMM can be applied to many learning methods proposed in robust statistics. In the robust statistics literature due to Huber (1964) and, more generally, Huber (2009), an M-estimator is defined as the root of an implicit equation of the form $\boldsymbol{\Psi}_{jk}(\widehat{\boldsymbol{\theta}}_{jk}) = \sum_{i=1}^{n_k} \boldsymbol{\psi}_{jk}(\widehat{\boldsymbol{\theta}}_{jk}) = \boldsymbol{0}$, where $\boldsymbol{\psi}_{jk}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \rho(\boldsymbol{\theta})$, ρ is a suitable function that primarily aims to provide estimators robust to influential data points, and $\widehat{\boldsymbol{\theta}}_{jk} \in \mathbb{R}^p$, and $\boldsymbol{\zeta}_{jk}$ is empty or known; additional details are available in Huber (2009) for the case when $\boldsymbol{\zeta}_{jk}$ is unknown. In the context of longitudinal data, Wang et al. (2005) robustify the generalized estimating equations of Liang and Zeger (1986) by replacing the standardized residuals with Huber's M-residuals.

3.4. Joint Mean-Variance Modeling

Following Pan and Mackenzie (2003), one can jointly model the marginal means and covariances of the longitudinal responses with $h(\boldsymbol{\mu}_{i,jk}) = \boldsymbol{X}_{i,jk,1}\boldsymbol{\beta}$, $\log(\boldsymbol{\sigma}_{i,jk}^2) = \boldsymbol{X}_{i,jk,2}\boldsymbol{\lambda}$, and $\phi_{irt,jk} = \boldsymbol{X}_{irt,jk,3}\boldsymbol{\gamma}$ for $1 \leq t < r \leq m_j$, where h is a known component-wise link function, $\boldsymbol{\beta} \in \mathbb{R}^{q_1}$, $\boldsymbol{\lambda} \in \mathbb{R}^{q_2}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{q_3}$ are unconstrained parameters, $\boldsymbol{\mu}_{i,jk} = E(\boldsymbol{Y}_{i,jk}|\boldsymbol{X}_{i,jk,1}, \boldsymbol{\theta})$ and $\boldsymbol{X}_{i,jk,1} \in \mathbb{R}^{m_j \times q_1}$ a submatrix of $\boldsymbol{X}_{i,jk}$, $\boldsymbol{\sigma}_{i,jk}^2 = \mathbb{S}\left(Var(Y_{ir,jk})\right)_{r=1}^{m_j}$ and $\boldsymbol{X}_{i,jk,2} \in \mathbb{R}^{m_j \times q_2}$ a submatrix of $\boldsymbol{X}_{i,jk}$, and $\phi_{irt,jk}$ are specified in Zhang et al. (2015a). Estimating functions $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ in Equations 1 and 2 are given in detail in Zhang et al. (2015a). There is some choice depending on the problem considered as to whether $\boldsymbol{\theta} = \boldsymbol{\beta}$, $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\gamma})$, or $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma})$. In the first case, learning of variance parameters only helps improve estimation efficiency. This type of framework is widely applied in biomedical studies where the mean parameters are of primary interest. In the second case, learning of covariance parameters is of interest and $\boldsymbol{\beta}$ is treated as a nuisance parameter. This is the situation where prediction is of primary interest, such as in kriging in spatial data analysis. In the third case, $\boldsymbol{G}_{jk}$ is null, and learning of variance parameters is of interest to the investigator. This case occurs for example in the study of volatility for risk management in financial data analysis.

3.5. Marginal Quantile Regression for Correlated Data

Consider the marginal quantile regression model $Q_{Y_{it,jk}|\boldsymbol{X}_{it,jk}}(\tau) = \boldsymbol{X}_{it,jk}\boldsymbol{\theta}$, where $Q_{Y_{it,jk}|\boldsymbol{X}_{it,jk}}(\tau) = F^{-1}_{Y_{it,jk}|\boldsymbol{X}_{it,jk}}(\tau) = \inf\{y_{it,jk} : F_{Y_{it,jk}|\boldsymbol{X}_{it,jk}}(y_{it,jk}) \geq \tau\}$ is the τth quantile of $Y_{it,jk}|\boldsymbol{X}_{it,jk}$, $\tau \in (0,1)$, where $f_{Y_{it,jk}|\boldsymbol{X}_{it,jk}}(y_{it,jk})$ is the conditional distribution function of $Y_{it,jk}$ given $\boldsymbol{X}_{it,jk}$, $t = 1, \ldots, m_j$. Many estimating functions $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ for the learning of $\boldsymbol{\theta}$ and association parameters $\boldsymbol{\zeta}_{jk}$ of $\boldsymbol{Y}_{i,jk}$ have been proposed; see Jung (1996), Fu and Wang (2012), Lu and Fan (2015), and Yang et al. (2017) for examples.

Each of these five examples requires additional work to fully develop a divide-and-conquer strategy via DDIMM, including specific computational details. Here we only present the general framework with a high-level discussion that sheds light on DDIMM's promising generality and flexibility, and its coverage of a wide range of supervised learning methods. The theoretical results presented in Sections 4 and 5 are developed under a general unified framework of estimating functions that includes the above five examples as special cases.

4. Asymptotic Properties

In this section we assume that K and J are fixed; this assumption will be relaxed in Section 5. Let $n_{\min} = \min_{k=1,\ldots,K} n_k$ and $n_{\max} = \max_{k=1,\ldots,K} n_k$. Suppose $\boldsymbol{W} \xrightarrow{p} \boldsymbol{w}$ as $n_{\min} \to \infty$. In this section we study the asymptotic properties of the GMM estimator $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ proposed in Equation 4 and its optimal version proposed in Equation 6. We assume throughout that subjects are monotonically allocated to subject groups; that is, as $n_{\min} \to \infty$, a subject cannot be reallocated to another group once it has been assigned to a subject group. Define the variability matrix of $\boldsymbol{\tau}_i(\boldsymbol{\theta}, \boldsymbol{\zeta})$ in Equation 3 as

$$\boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = Var_{\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0}\left\{\boldsymbol{\tau}_i(\boldsymbol{\theta}, \boldsymbol{\zeta})\right\} = \begin{pmatrix} \boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) & \boldsymbol{v}_{\boldsymbol{\psi g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) \\ \boldsymbol{v}^T_{\boldsymbol{\psi g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) & \boldsymbol{v}_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) \end{pmatrix}$$

where $\boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = Var_{\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0}\left\{\boldsymbol{\psi}_i(\boldsymbol{\theta}; \boldsymbol{\zeta})\right\}$, $\boldsymbol{v}_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = Var_{\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0}\left\{\boldsymbol{g}_i(\boldsymbol{\zeta}; \boldsymbol{\theta})\right\}$, and $\boldsymbol{v}_{\boldsymbol{\psi g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = E_{\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0}\left\{\boldsymbol{\psi}_i(\boldsymbol{\theta}; \boldsymbol{\zeta})\boldsymbol{g}_i^T(\boldsymbol{\zeta}; \boldsymbol{\theta})\right\}$. Let the sensitivity matrix of $\boldsymbol{\tau}_i(\boldsymbol{\theta}, \boldsymbol{\zeta})$ be

$$\boldsymbol{s}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = -\nabla_{\boldsymbol{\theta}, \boldsymbol{\zeta}} E_{\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0} \boldsymbol{\tau}_i(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \begin{pmatrix} \boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) & \boldsymbol{s}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) \\ \boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) & \boldsymbol{s}^{\boldsymbol{\zeta}}_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) \end{pmatrix}, \text{ where} \tag{7}$$

$$\boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \mathbb{S}^{JK}\left(\tfrac{n_k}{N}\boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})\right), \quad \boldsymbol{s}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \operatorname{diag}\left\{\tfrac{n_k}{N}\boldsymbol{s}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})\right\}^{J,K}_{j=1,k=1},$$

$$\boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \mathbb{S}^{JK}\left(\tfrac{n_k}{N}\boldsymbol{s}^{\boldsymbol{\theta}}_{g_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})\right), \quad \boldsymbol{s}^{\boldsymbol{\zeta}}_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \operatorname{diag}\left\{\tfrac{n_k}{N}\boldsymbol{s}^{\boldsymbol{\zeta}}_{\boldsymbol{g}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})\right\}^{J,K}_{j=1,k=1}$$

$$\boldsymbol{s}_{jk}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) = \begin{pmatrix} \boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) & \boldsymbol{s}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) \\ \boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{g}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) & \boldsymbol{s}^{\boldsymbol{\zeta}}_{\boldsymbol{g}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) \end{pmatrix}.$$

Following Theorem 3.4 of Song (2007), block-specific estimates $\widehat{\boldsymbol{\theta}}_{jk}$ and $\widehat{\boldsymbol{\zeta}}_{jk}$ are consistent given assumptions (A.1). Consistency and asymptotic normality of the GMM estimator $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ in Equation 4 have been established by Hansen (1982) and, more generally, by Newey and McFadden (1994). To establish consistency and asymptotic normality for the combined estimator $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$, we consider the following additional regularity conditions:

(A.3) Following Newey and McFadden (1994), define

$$Q_0(\boldsymbol{\theta}, \boldsymbol{\zeta}|\boldsymbol{W}) = E_{\boldsymbol{\theta},\boldsymbol{\zeta}} \left\{ \boldsymbol{T}_N^T(\boldsymbol{\theta}, \boldsymbol{\zeta}) \right\} \boldsymbol{w} E_{\boldsymbol{\theta},\boldsymbol{\zeta}} \left\{ \boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}) \right\}.$$

Assume $Q_0(\boldsymbol{\theta}, \boldsymbol{\zeta}|\boldsymbol{W})$ is twice-continuously differentiable in a neighbourhood of $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$.

(A.4) Let $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c) = \arg\min_{\boldsymbol{\theta},\boldsymbol{\zeta}} Q_N(\boldsymbol{\theta}, \boldsymbol{\zeta}|\boldsymbol{W})$. Following Newey and McFadden (1994), assume $Q_N(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c|\boldsymbol{W}) \leq \inf_{\boldsymbol{\theta}\in\Theta,\boldsymbol{\zeta}\in\Xi} Q_N(\boldsymbol{\theta}, \boldsymbol{\zeta}|\boldsymbol{W}) + \epsilon_N$ with $\epsilon_N = o_p(1)$. In addition, assume that $\boldsymbol{\theta}_0$, $\boldsymbol{\zeta}_0$ are interior points of Θ and Ξ respectively, and that for any $\delta_N \to 0$,

$$\sup_{\|(\boldsymbol{\theta},\boldsymbol{\zeta})-(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\|\leq\delta_N} \frac{N^{1/2}}{1+N^{1/2}\|(\boldsymbol{\theta},\boldsymbol{\zeta})-(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\|} \left\| \boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}) - \boldsymbol{T}_N(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) - E_{\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0} \boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}) \right\| = o_p(1).$$

Assumption (A.4) characterizes types of "stochastic equicontinuity" of non-smooth objective functions. As outlined in Newey and McFadden (1994), it essentially requires the convergence of $\boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta})$ to $\boldsymbol{T}_N(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ be uniform over any (shrinking) neighbourhood of $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ in estimating procedures, which is the standard condition widely used in the literature to relax the assumption of differentiability of objective functions. For most of the examples in Section 3, since the estimating functions are continuously differentiable, (A.4) holds automatically. Assumption (A.4) is mostly used for generalizability beyond continuous differentiability, so that our framework allows a broader class of methods. Checking assumption (A.4) has been extensively discussed in the literature, including some primitive conditions given in Pollard (1985) and Andrews (1994). We refer the reader to Chapter 7 of Newey and McFadden (1994) for a complete discussion of the topic. We may empirically check (A.4) by considering $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\zeta}})$ a $N^{1/2}$-consistent estimator of $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ and choosing $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$ in a small ball $\mathcal{B}_{\delta_N}$ centered at $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\zeta}})$ of radius $\delta_N = N^{-1}$. When the analytic form of $E\boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta})$ is available, which is the case in most applications, we can monitor the form

$$\frac{N^{1/2} \left\| \boldsymbol{T}_N(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) - \boldsymbol{T}_N(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\zeta}}) - E\boldsymbol{T}_N(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) \right\|}{1 + N^{1/2} \left\| (\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) - (\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\zeta}}) \right\|}$$

for a sequence of increasingly large sub-samples of subjects to empirically check how this form evolves as N increases. This form can be computed in a distributed fashion since $\boldsymbol{T}_N$ corresponds to stacked block estimating equations. We can choose $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\zeta}})$ as the DDIMM estimator, or a computationally cheaper meta-estimator such as an average of block estimators. If an analytic form for $E\boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta})$ is not available, a bootstrap method may be used for its estimation.

Heuristically, Theorems 1 and 2 do not require the differentiability of $\boldsymbol{T}_N$ and Q_N. Instead, they require the differentiability of their population versions, and that $\boldsymbol{T}_N$ behaves "nicely" in a neighbourhood of $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$, in the sense that higher order terms are asymptotically ignorable. The following two theorems state the consistency and asymptotic normality of $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ given in Equation 4 under Newey and McFadden's mild moment conditions given in (A.3) and (A.4).

Theorem 1 (Consistency of $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$) *Suppose assumptions (A.1)-(A.3) hold with $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ defined in Equation 4. Then $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c) \xrightarrow{p} (\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ as $n_{\min} \to \infty$.*

The proof of Theorem 1 follows closely the steps given in Hansen (1982) and Newey and McFadden (1994), and thus is omitted.

Theorem 2 (Asymptotic normality of $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$) *Suppose assumptions (A.1)-(A.4) hold with $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ defined in Equation 4. Then as $n_{\min} \to \infty$,*

$$N^{1/2} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_c - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_c - \boldsymbol{\zeta}_0 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{j}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\boldsymbol{s}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\tilde{\boldsymbol{v}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\boldsymbol{s}^T(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\boldsymbol{j}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right),$$

where $\tilde{\boldsymbol{v}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \boldsymbol{w}\boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{\zeta})\boldsymbol{w}$, and where the Godambe information $\boldsymbol{j}(\boldsymbol{\theta}, \boldsymbol{\zeta})$ of $\boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta})$ takes the form $\boldsymbol{j}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \boldsymbol{s}(\boldsymbol{\theta}, \boldsymbol{\zeta})\boldsymbol{w}\boldsymbol{s}^T(\boldsymbol{\theta}, \boldsymbol{\zeta})$.

The proof of Theorem 2 follows easily from Theorem 7.2 in Newey and McFadden (1994) and Theorem 1 above. We note that requiring K to be finite implies that N and $n_{\min}$ are asymptotically of the same order. We will relax this assumption in Section 5. Conditions (A.3) and (A.4) allow us to consider non-differentiable kernel inference functions in the block (j, k) analysis, extending Hector and Song (2020)'s DIMM beyond CL kernel inference functions. We can now consider quantile regression, M-estimation, and more general estimation functions than the score or CL score equations.

A test of the over-identifying restrictions follows from Hansen (1982) and Hector and Song (2020). This test is useful for detecting invalid moment restrictions, which can inform our choice of data partition and model. Formally, we show in Theorem 3 that the objective function NQ_N evaluated at $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ follows a χ^2 distribution with $(JK-1)p$ degrees of freedom. Unfortunately, it may be difficult to tell if invalid moment restrictions stem from an inappropriate data split or incorrect model specification. Residual analysis for model diagnostics can remove doubt in the latter case.

Theorem 3 (Test of over-identifying restrictions) *Suppose assumptions (A.1)-(A.4) hold with $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ defined in Equation 4. Then as $n_{\min} \to \infty$, $NQ_N(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c | \boldsymbol{W}) \xrightarrow{d} \chi^2_{(JK-1)p}$.*

The proof of Theorem 3 can be carried out with some minor changes from that of Theorem 3 in Hector and Song (2020). The GMM provides an objective function with which to do model selection even when the block analyses do not, such as with GEE and M-estimation. In the following, Theorem 4 and Corollary 5 show our combined GMM estimator derived from Equation 6 is optimal in the sense defined by Hansen (1982): it has an asymptotic covariance matrix at least as small (in terms of the Loewner ordering) as any other estimator exploiting the same over-identifying restrictions. We refer to this property as "Hansen optimality".

Theorem 4 *Suppose assumptions (A.1)-(A.2) hold. Then as $n_{\min} \to \infty$, $\widehat{\boldsymbol{V}}_N \xrightarrow{p} \boldsymbol{v}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$, i.e. $\boldsymbol{w} = \boldsymbol{v}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$.*

Proof The proof uses the consistency of the block estimators and the Central Limit Theorem, and is given in the Supplemental Material. ■

Corollary 5 (Hansen optimality) *Suppose assumptions (A.1)-(A.4) hold with* $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ *defined in Equation 4. Let* $\boldsymbol{j}(\boldsymbol{\theta}, \boldsymbol{\zeta})$ *as given in Theorem 2. Then as* $n_{\min} \to \infty$,

$$N^{1/2} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{opt} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{opt} - \boldsymbol{\zeta}_0 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{j}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right).$$

The theoretical results given in Theorems 1-4 provide a framework for constructing asymptotic confidence intervals and conducting hypothesis tests, so that we can perform inference for $\boldsymbol{\theta}$ when M and/or N are very large. Using an optimal weight matrix improves statistical power so DDIMM may detect some signals that other methods may miss. Since we consider a broad class of models $\mathcal{P}$, there are no general efficiency results about the block-specific estimator $\widehat{\boldsymbol{\theta}}_{jk}$. When a learning method based on $\boldsymbol{\Psi}_{jk}$ has known efficiency results and performs well enough, DDIMM generally inherits "local" efficiency to achieve overall efficiency.

Remark 6 *We discuss efficiency for selected examples in Section 3.*

(i) In Example 3.1, when the score function exists and satisfies mild regularity conditions, its variance is given by Fisher's information, and is a lower bound on the variances of estimating functions for $\boldsymbol{\theta}$ *and* $\boldsymbol{\zeta}$*. This, coupled with Hansen's optimality, means that using the score function for* $\boldsymbol{\psi}_{jk}$ *and* $\boldsymbol{g}_{jk}$ *yields an efficient estimator of* $\boldsymbol{\theta}$ *and* $\boldsymbol{\zeta}$*. In an unpublished dissertation, Jin (2011) studied the efficiency of the pairwise composite likelihood under different correlation structures. Hector and Song (2020) showed empirically that the efficiency of the pairwise composite likelihood propagates to the combined estimator.*

(ii) In Example 3.2, it is known that the GEE estimator $\widehat{\boldsymbol{\theta}}_{jk}$ *in Example 3.2 is semiparametrically efficient when the correlation structure of the response* $\boldsymbol{y}_{i,jk}$ *is correctly specified. This, coupled with Hansen's optimality, means that using GEE's for* $\boldsymbol{\psi}_{jk}$ *with the correct correlation structure of the response* $\boldsymbol{y}_{i,jk}$ *yields an efficient estimator of* $\boldsymbol{\theta}$*.*

Remark 7 *The GMM estimator* $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$ *can be interpreted as maximizing an extension of the confidence distribution density, as discussed in Hector and Song (2020). The confidence distribution approach is used for independent data in Xie and Singh (2013). Briefly, we can define the confidence estimating function (CEF) as* $U(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \Phi(N^{1/2}\widehat{\boldsymbol{V}}_N^{-1/2}\boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}))$*, where* $\Phi(\cdot)$ *is the* $(JKp + d)$*-variate standard normal distribution function. Clearly,* $U(\boldsymbol{\theta}, \boldsymbol{\zeta})$ *is asymptotically standard uniform at* $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ *as long as* $\widehat{\boldsymbol{V}}_N$ *is a consistent estimator of the covariance of* $\boldsymbol{T}_N$*. Then we can define the density of the CEF as* $u(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \phi(N^{1/2}\widehat{\boldsymbol{V}}_N^{-1/2}\boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}))$*. Maximizing* $u(\boldsymbol{\theta}, \boldsymbol{\zeta})$ *with respect to* $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ *yields the minimization defined in Equation 6.*

By framing our estimator as a GMM estimator, the theoretical framework of DIMM established only for CL can be extended to include a data split at the subject level and a generalization of $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$. Adding moment conditions allows the proposed method to enjoy the power and versatility of the GMM, and the necessary theoretical results to support its use. This divide-and-conquer strategy benefits from handling low dimensional blocks of data and estimating equations, yielding tremendous computational gains.

5. Distributed Estimation and Inference

Despite the computational gains offered by the divide-and-combine procedure and the GMM estimator, iteratively finding the solution $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$ (or $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$) to Equation 6 can be slow due to the high-dimensionality of parameter $\boldsymbol{\zeta}$ and the need to repeatedly evaluate $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$. To overcome this computational bottleneck, we propose a meta-estimator derived from Equation 6 that delivers a closed-form estimator via a linear function of block estimates $(\widehat{\boldsymbol{\theta}}_{list}, \widehat{\boldsymbol{\zeta}}_{list})$. We define the DDIMM estimator for $(\boldsymbol{\theta}, \boldsymbol{\zeta})$:

$$\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{DDIMM} \\ \widehat{\boldsymbol{\zeta}}_{DDIMM} \end{pmatrix} = \left(\sum_{k=1}^{K} \sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i} \right)^{-1} \sum_{k=1}^{K} \sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik} \\ \widehat{\boldsymbol{\zeta}}_{list} \end{pmatrix}. \tag{8}$$

where $\widehat{\boldsymbol{C}}_{k,i}$ is a function of sample variability and sensitivity matrices and block-specific estimators $\widehat{\boldsymbol{\theta}}_{jk}$ and $\widehat{\boldsymbol{\zeta}}_{jk}$ defined in detail in Section 5.1. If we do not plan to conduct inference for $\boldsymbol{\zeta}$, which is treated as a nuisance parameter, taking $\left[\widehat{\boldsymbol{C}}^{-1}\right]_{p:}$ to be rows 1 to p of matrix $(\sum_{k=1}^{K} \sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i})^{-1}$ leads to the closed-form estimator of $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}}_{DDIMM} = \left[\widehat{\boldsymbol{C}}^{-1}\right]_{p:} \sum_{k=1}^{K} \sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik}^T & \widehat{\boldsymbol{\zeta}}_{list}^T \end{pmatrix}^T. \tag{9}$$

We briefly define sample sensitivity matrices that will appear in the main body of the paper. Let $\boldsymbol{S}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})$ be a $n_k^{1/2}$-consistent sample estimator of $\boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})$, and similarly define $\boldsymbol{S}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})$, $\boldsymbol{S}^{\boldsymbol{\theta}}_{\boldsymbol{g}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})$ and $\boldsymbol{S}^{\boldsymbol{\zeta}}_{\boldsymbol{g}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})$. Let

$$\boldsymbol{S}_{jk}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) = \begin{pmatrix} \boldsymbol{S}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) & \boldsymbol{S}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) \\ \boldsymbol{S}^{\boldsymbol{\theta}}_{\boldsymbol{g}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) & \boldsymbol{S}^{\boldsymbol{\zeta}}_{\boldsymbol{g}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) \end{pmatrix}.$$

Note that the uppercase $\boldsymbol{S}$ denotes the sample sensitivity matrix, and the lower-case $\boldsymbol{s}$ denotes the population sensitivity matrix. Let $\widehat{\boldsymbol{S}}_{jk} = \boldsymbol{S}_{jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk})$ and similarly define $\widehat{\boldsymbol{S}}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}$, $\widehat{\boldsymbol{S}}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}$, $\widehat{\boldsymbol{S}}^{\boldsymbol{\theta}}_{\boldsymbol{g}_{jk}}$ and $\widehat{\boldsymbol{S}}^{\boldsymbol{\zeta}}_{\boldsymbol{g}_{jk}}$. Sensitivity formulas are summarized in Table A.1 in Appendix A.1. The DDIMM estimator in Equation 9 can be implemented in a fully parallelized and scalable computational scheme, where only one pass through each block of data is required. The block analyses are run on parallel CPUs, and return the values of summary statistics $\{\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk}, \boldsymbol{\psi}_{i,jk}(\widehat{\boldsymbol{\theta}}_{jk}; \widehat{\boldsymbol{\zeta}}_{jk}), \boldsymbol{g}_{i,jk}(\widehat{\boldsymbol{\zeta}}_{jk}; \widehat{\boldsymbol{\theta}}_{jk}), \widehat{\boldsymbol{S}}_{jk}\}_{j,k=1}^{J,K}$ to the main computing node, which computes $\widehat{\boldsymbol{\theta}}_{DDIMM}$ in Equation 9 in one step.

5.1. Construction of $\widehat{\boldsymbol{C}}_{k,i}$

We give details on the construction of $\widehat{\boldsymbol{C}}_{k,i}$. Readers may wish to omit this section on a first reading, as these details are not necessary for an understanding of the main body of the paper. We consider the optimal case where the GMM weighting matrix takes the form:

$$\boldsymbol{W} = \widehat{\boldsymbol{V}}_N^{-1} = \begin{pmatrix} \widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}} & \widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi g}} \\ \widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi g}}^T & \widehat{\boldsymbol{V}}_{N,\boldsymbol{g}} \end{pmatrix}^{-1} = \begin{pmatrix} \widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi}} & \widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi g}} \\ \widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi g}\,T} & \widehat{\boldsymbol{V}}_N^{\boldsymbol{g}} \end{pmatrix}.$$

For convenience, we introduce a subsetting operation, with technical details available in Appendix A.2: we let $\left[\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi}}\right]_{ij:k}$ subset the rows for the parameters corresponding to block (i,k) and the columns for the parameters corresponding to block (j,k) of matrix $\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi}}$. Similarly define $\left[\widehat{\boldsymbol{V}}_N^{\boldsymbol{g}}\right]_{ij:k}$, and $\left[\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi g}}\right]_{ij:k}$. For $\boldsymbol{\eta} \in \{\boldsymbol{\theta}, \boldsymbol{\zeta}\}$, let

$$\widehat{\boldsymbol{A}}_{k,ij}^{\boldsymbol{\eta}} = \left(\widehat{\boldsymbol{S}}_{\boldsymbol{\psi}_{jk}}^{\boldsymbol{\theta}\ T}\left[\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi}}\right]_{ji:k} + \widehat{\boldsymbol{S}}_{\boldsymbol{g}_{jk}}^{\boldsymbol{\theta}\ T}\left[\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi g}\ T}\right]_{ji:k}\right)\widehat{\boldsymbol{S}}_{\boldsymbol{\psi}_{ik}}^{\boldsymbol{\eta}} + \left(\widehat{\boldsymbol{S}}_{\boldsymbol{\psi}_{jk}}^{\boldsymbol{\theta}\ T}\left[\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi g}}\right]_{ji:k} + \widehat{\boldsymbol{S}}_{\boldsymbol{g}_{jk}}^{\boldsymbol{\theta}\ T}\left[\widehat{\boldsymbol{V}}_N^{\boldsymbol{g}}\right]_{ji:k}\right)\widehat{\boldsymbol{S}}_{\boldsymbol{g}_{ik}}^{\boldsymbol{\eta}},$$
$$\widehat{\boldsymbol{B}}_{k,ij}^{\boldsymbol{\eta}} = \left(\widehat{\boldsymbol{S}}_{\boldsymbol{\psi}_{jk}}^{\boldsymbol{\zeta}\ T}\left[\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi}}\right]_{ji:k} + \widehat{\boldsymbol{S}}_{\boldsymbol{g}_{jk}}^{\boldsymbol{\zeta}\ T}\left[\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi g}\ T}\right]_{ji:k}\right)\widehat{\boldsymbol{S}}_{\boldsymbol{\psi}_{ik}}^{\boldsymbol{\eta}} + \left(\widehat{\boldsymbol{S}}_{\boldsymbol{\psi}_{jk}}^{\boldsymbol{\zeta}\ T}\left[\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi g}}\right]_{ji:k} + \widehat{\boldsymbol{S}}_{\boldsymbol{g}_{jk}}^{\boldsymbol{\zeta}\ T}\left[\widehat{\boldsymbol{V}}_N^{\boldsymbol{g}}\right]_{ji:k}\right)\widehat{\boldsymbol{S}}_{\boldsymbol{g}_{ik}}^{\boldsymbol{\eta}}.$$

Define D^{ik} as the sum of the dimensions of $\boldsymbol{\zeta}_{11}, \ldots, \boldsymbol{\zeta}_{i-1k}$, and D^k as the sum of the dimensions of $\boldsymbol{\zeta}_{11}, \ldots, \boldsymbol{\zeta}_{Jk-1}$, with technical details in Appendix A.3. Let $d_k = \sum_{j=1}^J d_{jk}$. Then we can define the following,

$$\widehat{\boldsymbol{C}}_{k,i} = \begin{pmatrix} \sum_{j=1}^J \widehat{\boldsymbol{A}}_{k,ij}^{\boldsymbol{\theta}} & \mathbf{0}_{p\times D^{ik}} & \sum_{j=1}^J \widehat{\boldsymbol{A}}_{k,ij}^{\boldsymbol{\zeta}} & \mathbf{0}_{p\times(d-d_{ik}-D^{ik})} \\ & & \mathbf{0}_{D^k\times(p+d)} & \\ \widehat{\boldsymbol{B}}_{k,i1}^{\boldsymbol{\theta}} & \mathbf{0}_{d_{1k}\times D^{ik}} & \widehat{\boldsymbol{B}}_{k,i1}^{\boldsymbol{\zeta}} & \mathbf{0}_{d_{1k}\times(d-d_{ik}-D^{ik})} \\ & & \vdots & \\ \widehat{\boldsymbol{B}}_{k,iJ}^{\boldsymbol{\theta}} & \mathbf{0}_{d_{Jk}\times D^{ik}} & \widehat{\boldsymbol{B}}_{k,iJ}^{\boldsymbol{\zeta}} & \mathbf{0}_{d_{Jk}\times(d-d_{ik}-D^{ik})} \\ & & \mathbf{0}_{(d-d_k-D^k)\times(p+d)} & \end{pmatrix}. \tag{10}$$

5.2. Asymptotic Results for K and J Fixed

In this section we assume that K and J are fixed, which will be relaxed in Sections 5.3 and 5.4. Recall that we assume subjects are monotonically allocated to subject groups: as $n_{\min} \to \infty$, a subject cannot be reallocated to another group once it has been assigned to a subject group. Consider the following condition, which is a specification on the rate of convergence of Newey and McFadden (1994)'s condition in (A.4):

(A.4*) For each $j = 1, \ldots, J$, $k = 1, \ldots, K$, $\widehat{\boldsymbol{\theta}}_{jk} = \boldsymbol{\theta}_0 + O_p(n_k^{-1/2})$ and $\widehat{\boldsymbol{\zeta}}_{jk} = \boldsymbol{\zeta}_{jk0} + O_p(n_k^{-1/2})$. For any $\delta_{n_k} \to \infty$,

$$\sup_{\|(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})-(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{jk0})\|\le\delta_{n_k}} \frac{n_k^{1/2}}{1+n_k^{1/2}\|(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})-(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{jk0})\|}\left\|\boldsymbol{T}_{jk}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk}) - \boldsymbol{T}_{jk}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{jk0}) - E_{\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{jk0}}\boldsymbol{T}_{jk}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})\right\| = O_p(n_k^{-1/2}).$$

A similar discussion to the one for (A.4) can be given for (A.4*). Assumption (A.4*) is satisfied when moment conditions are continuously differentiable. Procedures for checking (A.4*) for non-smooth objective functions may be developed similarly to those for checking (A.4). Some large-sample results can be established which are helpful in studying the asymptotic behaviour of $\widehat{\boldsymbol{\theta}}_{DDIMM}$.

Lemma 8 *Suppose assumptions (A.1), (A.2) and (A.4*) hold. Then we have consistent estimation of information matrices:*

$$\widehat{\boldsymbol{V}}_N = \boldsymbol{v}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p(N^{-1/2}),$$

$$\widehat{\boldsymbol{S}}_{jk} = \boldsymbol{s}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) + O_p(n_k^{-1/2}) \quad \textit{for each } j,k, \textit{ and}$$
$$\frac{1}{N^2}\sum_{k=1}^{K}\sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i} = \widehat{\boldsymbol{S}}^T \widehat{\boldsymbol{V}}_N^{-1} \widehat{\boldsymbol{S}} = \boldsymbol{j}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p(N^{-1/2}),$$
$$\textit{where} \quad \widehat{\boldsymbol{S}} = \begin{pmatrix} \mathbb{S}\left(\frac{n_k}{N}\widehat{\boldsymbol{S}}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}\right)_{j=1,k=1}^{J,K} & diag\left\{\frac{n_k}{N}\widehat{\boldsymbol{S}}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}\right\}_{j=1,k=1}^{J,K} \\ \mathbb{S}\left(\frac{n_k}{N}\widehat{\boldsymbol{S}}^{\boldsymbol{\theta}}_{g_{jk}}\right)_{j=1,k=1}^{J,K} & diag\left\{\frac{n_k}{N}\widehat{\boldsymbol{S}}^{\boldsymbol{\zeta}}_{g_{jk}}\right\}_{j=1,k=1}^{J,K} \end{pmatrix}.$$

Proof A detailed proof is given in the Supplemental Material. ■

We show in Theorem 9 that the proposed closed-form estimator $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$ in Equation 8 is consistent and asymptotically normally distributed.

Theorem 9 *Suppose assumptions (A.1), (A.2) and (A.4*) hold. Let $\boldsymbol{j}(\boldsymbol{\theta}, \boldsymbol{\zeta})$ as given in Theorem 2. As $n_{\min} \to \infty$,*

$$N^{1/2}\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{DDIMM} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{DDIMM} - \boldsymbol{\zeta}_0 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{j}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right).$$

Proof [Proof of Theorem 9:] Here we present major steps, with all necessary details available in Appendix B.1. First, we show that $\widehat{\boldsymbol{\theta}}_{DDIMM}$ and $\widehat{\boldsymbol{\zeta}}_{DDIMM}$ are consistent. Define

$$\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i} \begin{pmatrix} \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{ik} \\ \boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}_{list} \end{pmatrix}. \tag{11}$$

By definition, $\lambda(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM}) = \boldsymbol{0}$. As shown in Lemma B.1.1 in Appendix B.1, $\lambda(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) \xrightarrow{p} \boldsymbol{0}$ as $n_{\min} \to \infty$. Given that $\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}}\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})$ exists and is nonsingular, for some $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$ between $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$ and $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$, the first-order Taylor expansion leads to

$$\lambda(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM}) - \lambda(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) = \nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}}\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})|_{\boldsymbol{\theta}^*,\boldsymbol{\zeta}^*} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{DDIMM} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{DDIMM} - \boldsymbol{\zeta}_0 \end{pmatrix}, \tag{12}$$

which converges in probability to $\boldsymbol{0}$ as $n_{\min} \to \infty$. This implies that $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM}) \xrightarrow{p} (\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ as $n_{\min} \to \infty$.

Now we derive the distribution of $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$. With a slight abuse of notation, let $\widehat{\boldsymbol{\theta}}_{list} - \boldsymbol{\theta}_0 = \mathbb{S}^{JK}\left(\widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0\right)$. We show in Lemma B.1.2 in Appendix B.1 that

$$\begin{pmatrix} \boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}_0; \boldsymbol{\zeta}_{jk0}) \\ \boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk0}; \boldsymbol{\theta}_0) \end{pmatrix} = \widehat{\boldsymbol{S}}_{jk} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix} + O_p(n_k^{-1}). \tag{13}$$

Recall the form of $\boldsymbol{T}_N$ in Equation 3. By the Central Limit Theorem, $N^{1/2}\boldsymbol{T}_N(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{v}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0))$. Then with $\widehat{\boldsymbol{S}}$ defined in Lemma 8, it follows from Equation 13 that

$$N^{1/2}\widehat{\boldsymbol{S}}\begin{pmatrix} (\widehat{\boldsymbol{\theta}}_{list} - \boldsymbol{\theta}_0)^T & (\widehat{\boldsymbol{\zeta}}_{list} - \boldsymbol{\zeta}_0)^T \end{pmatrix}^T \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{v}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right).$$

Moreover, by Lemma 8 and Slutsky's theorem we have:

$$N^{1/2}\left(\begin{array}{cc}(\widehat{\boldsymbol{\theta}}_{list}-\boldsymbol{\theta}_0)^T & (\widehat{\boldsymbol{\zeta}}_{list}-\boldsymbol{\zeta}_0)^T\end{array}\right)^T \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{j}^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\right).$$

Using the fact that the sum of jointly (asymptotically) Normal variables is (asymptotically) normal, by Lemma 8 and Slutsky's theorem again, we have

$$N^{1/2}\left(\begin{array}{c}\widehat{\boldsymbol{\theta}}_{DDIMM}-\boldsymbol{\theta}_0\\ \widehat{\boldsymbol{\zeta}}_{DDIMM}-\boldsymbol{\zeta}_0\end{array}\right) = N^{1/2}\left(\sum_{k=1}^{K}\sum_{i=1}^{J} n_k^2\widehat{\boldsymbol{C}}_{k,i}\right)^{-1}\sum_{k=1}^{K}\sum_{i=1}^{J} n_k^2\widehat{\boldsymbol{C}}_{k,i}\left(\begin{array}{c}\widehat{\boldsymbol{\theta}}_{ik}-\boldsymbol{\theta}_0\\ \widehat{\boldsymbol{\zeta}}_{list}-\boldsymbol{\zeta}_0\end{array}\right)$$

is asymptotically distributed $\mathcal{N}(\mathbf{0}, \boldsymbol{j}^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0))$. ■

This key theorem allows us to use $\widehat{\boldsymbol{\theta}}_{DDIMM}$, which is more computationally attractive than $\widehat{\boldsymbol{\theta}}_{opt}$ defined in Equation 6, without sacrificing any of the nice asymptotic properties for inference. Additionally, it follows easily from Theorem 9 that, under suitable conditions, the closed-form estimator $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$ in Equation 8 has the same asymptotic distribution as and is asymptotically equivalent to the GMM estimator $\widehat{\boldsymbol{\theta}}_{opt}$ in Equation 6.

Corollary 10 *Suppose assumptions (A.1)-(A.4*) hold with $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$ defined in Equation 6. Then $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$ and $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$ are asymptotically equivalent: as $n_{\min}\to\infty$,*

$$N^{1/2}\left\|\left(\begin{array}{c}\widehat{\boldsymbol{\theta}}_{DDIMM}-\widehat{\boldsymbol{\theta}}_{opt}\\ \widehat{\boldsymbol{\zeta}}_{DDIMM}-\widehat{\boldsymbol{\zeta}}_{opt}\end{array}\right)\right\| \xrightarrow{p} 0.$$

Proof A detailed proof is given in the Supplemental Material. ■

The computation of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ in Equation 9 relies solely on block-specific estimators $(\widehat{\boldsymbol{\theta}}_{list}, \widehat{\boldsymbol{\zeta}}_{list})$ and values of summary statistics from each block. To guarantee the appropriate asymptotic distribution of $\widehat{\boldsymbol{\theta}}_{DDIMM}$, we assume in condition (A.4*) that these block-specific estimators are $N^{1/2}$ consistent estimators of the true values, which restricts the scope of possible block-specific inference methods. For inference methods not satisfying this $N^{1/2}$ consistency in condition (A.4*), it is still possible to use $\widehat{\boldsymbol{\theta}}_{opt}$ in Equation 6.

5.3. Asymptotic Results for Diverging K with J Fixed

We show in Theorem 11 that the asymptotic distribution of $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$ remains unchanged as the number of subject groups K grows with the sample size.

Theorem 11 *Suppose $N^{\delta-1/2}K$ is bounded as $n_{\min}\to\infty$ for a positive constant $\delta<\frac{1}{2}$, and assumptions (A.1), (A.2) and (A.4*) hold. Let $\boldsymbol{H}\in\mathbb{R}^{h\times(p+d)}$ a matrix of rank $r\in\mathbb{N}$, $h\in\mathbb{N}$, $r\le h$, with finite maximum singular value $\bar{\sigma}(\boldsymbol{H})<\infty$. Let $\boldsymbol{j}(\boldsymbol{\theta},\boldsymbol{\zeta})$ as given in Theorem 2. Then, as $n_{\min}\to\infty$, we show that the limiting value $\boldsymbol{j}_{\boldsymbol{H}}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)$ of $\boldsymbol{H}\boldsymbol{j}^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\boldsymbol{H}^T$ is a positive semi-definite and symmetric variance matrix, and that*

$$N^{1/2}\boldsymbol{H}\left(\begin{array}{c}\widehat{\boldsymbol{\theta}}_{DDIMM}-\boldsymbol{\theta}_0\\ \widehat{\boldsymbol{\zeta}}_{DDIMM}-\boldsymbol{\zeta}_0\end{array}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{j}_{\boldsymbol{H}}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\right).$$

Proof [Proof of Theorem 11] Here we present major steps, with all necessary details available in Appendix B.2. First, we know that $\|\boldsymbol{H}\| \leq r\bar{\sigma}(\boldsymbol{H})$. Let $\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})$ defined by Equation 11, such that $\lambda(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM}) = \mathbf{0}$. We show in Lemma B.2.1 in Appendix B.2 that $\|\lambda(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\| = O_p(N^{-1/2-\delta} n_{\max}^{1/2})$ and $\left\|\{\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}}\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})\}^{-1}\right\| = O_p\left(N^{1/2+\delta} n_{\max}^{-1}\right)$. From the first-order Taylor expansion in Equation 12, we have

$$\begin{aligned}\left\|\boldsymbol{H}\begin{pmatrix}\widehat{\boldsymbol{\theta}}_{DDIMM} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{DDIMM} - \boldsymbol{\zeta}_0\end{pmatrix}\right\| &\leq \|\boldsymbol{H}\| \left\|\left(\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}}\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})|_{\boldsymbol{\theta}^*,\boldsymbol{\zeta}^*}\right)^{-1}\right\| \|\lambda(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\| \\ &\leq r\bar{\sigma}(\boldsymbol{H}) O_p(n_{\max}^{-1/2}).\end{aligned}$$

Then $\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_{DDIMM}^T, \widehat{\boldsymbol{\zeta}}_{DDIMM}^T)^T - \boldsymbol{H}(\boldsymbol{\theta}_0^T, \boldsymbol{\zeta}_0^T)^T \xrightarrow{p} \mathbf{0}$ as $n_{\min} \to \infty$.
To derive the distribution of $\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_{DDIMM}^T, \widehat{\boldsymbol{\zeta}}_{DDIMM}^T)^T$, first consider an arbitrary $k \in \{1, \ldots, K\}$. For convenience, denote

$$\begin{aligned}\boldsymbol{T}_k(\boldsymbol{\theta}, \boldsymbol{\zeta}_k) &= \mathbb{S}\left(\mathbb{S}^J\left(\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}; \boldsymbol{\zeta}_{jk})\right), \mathbb{S}^J\left(\boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk}; \boldsymbol{\theta})\right)\right), \\ \boldsymbol{\tau}_{i,k}(\boldsymbol{\theta}, \boldsymbol{\zeta}_k) &= \mathbb{S}\left(\mathbb{S}^J\left(\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}; \boldsymbol{\zeta}_{jk})\right), \mathbb{S}^J\left(\boldsymbol{g}_{i,jk}(\boldsymbol{\zeta}_{jk}; \boldsymbol{\theta})\right)\right).\end{aligned}$$

By the Central Limit Theorem, $n_k^{1/2}\boldsymbol{T}_k(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{k0}) = n_k^{-1/2}\sum_{i=1}^{n_k}\boldsymbol{\tau}_{i,k}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{k0}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{v}_k(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{k0}))$ as $n_k \to \infty$, where $\boldsymbol{v}_k(\boldsymbol{\theta}, \boldsymbol{\zeta}_k) = Var_{\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0}}\{\boldsymbol{\tau}_{i,k}(\boldsymbol{\theta}, \boldsymbol{\zeta}_k)\}$. Define

$$\begin{aligned}\boldsymbol{s}_k(\boldsymbol{\theta}, \boldsymbol{\zeta}_k) &= \begin{pmatrix}\mathbb{S}^J\left(\boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})\right) & \operatorname{diag}\left\{\boldsymbol{s}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})\right\}_{j=1}^J \\ \mathbb{S}^J\left(\boldsymbol{s}^{\boldsymbol{\theta}}_{g_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})\right) & \operatorname{diag}\left\{\boldsymbol{s}^{\boldsymbol{\zeta}}_{\boldsymbol{g}_{jk}}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})\right\}_{j=1}^J\end{pmatrix}, \text{ and} \\ \boldsymbol{j}_k(\boldsymbol{\theta}, \boldsymbol{\zeta}_k) &= \boldsymbol{s}_k^T(\boldsymbol{\theta}, \boldsymbol{\zeta})\boldsymbol{v}_k^{-1}(\boldsymbol{\theta}, \boldsymbol{\zeta}_k)\boldsymbol{s}_k(\boldsymbol{\theta}, \boldsymbol{\zeta}_k).\end{aligned}$$

By the same arguments as in the proof of Theorem 9,

$$n_k^{1/2}\boldsymbol{j}_k(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{k0})\begin{pmatrix}\mathbb{S}\left(\widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0\right)_{j=1}^J \\ \widehat{\boldsymbol{\zeta}}_k - \boldsymbol{\zeta}_{k0}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{j}_k^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{k0})\right).$$

Note that the above vectors are independent for $k = 1, \ldots, K$. We establish in Lemma B.2.2 in Appendix B.2 that, for some affine transformation matrices $\boldsymbol{E}_k$, $k = 1, \ldots, K$, of $\mathbf{0}$'s and $\mathbf{1}$'s,

$$\begin{aligned}\frac{n_k^2}{N^2}\sum_{i=1}^J \widehat{\boldsymbol{C}}_{k,i}\begin{pmatrix}\widehat{\boldsymbol{\theta}}_{ik} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{list} - \boldsymbol{\zeta}_0\end{pmatrix} &= \frac{n_k}{N}\boldsymbol{E}_k\boldsymbol{Z}_k + O_p\left(N^{-1}\right), \\ \text{and} \quad \frac{n_k^2}{N^2}\sum_{i=1}^J \widehat{\boldsymbol{C}}_{k,i} &= \frac{n_k}{N}\boldsymbol{E}_k\boldsymbol{j}_k(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{k0})\boldsymbol{E}_k^T + O_p\left(n_k^{1/2}N^{-1}\right),\end{aligned}$$

where $n_k^{1/2}\boldsymbol{Z}_k \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{j}_k^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{k0})\right)$. It is clear that $\boldsymbol{j}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \sum_{k=1}^K (n_k/N)\boldsymbol{E}_k\boldsymbol{j}_k(\boldsymbol{\theta}, \boldsymbol{\zeta}_k)\boldsymbol{E}_k^T$. Since $\boldsymbol{E}_k$ has finitely many 1's, $\|\boldsymbol{E}_k\|$ is bounded. Since $\|\boldsymbol{j}_k(\boldsymbol{\theta}, \boldsymbol{\zeta}_k)\|$ is also bounded, $\|\boldsymbol{j}(\boldsymbol{\theta}, \boldsymbol{\zeta})\| = O(Kn_{\max}N^{-1}) = O(1)$. $\boldsymbol{j}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ is positive semi-definite and symmetric,

implying that $\boldsymbol{H}\boldsymbol{j}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\boldsymbol{H}^T$ is also positive semi-definite and symmetric. Following the monotone convergence theorem, we can write $\boldsymbol{H}\boldsymbol{j}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\boldsymbol{H}^T \to \boldsymbol{j}_{\boldsymbol{H}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$, where $\boldsymbol{j}_{\boldsymbol{H}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ exists and is a proper variance matrix.
Using the fact that $\lambda(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM}) = \mathbf{0}$ and $K = O(N^{1/2-\delta})$, we show in Lemma B.2.3 in Appendix B.2 that $N^{1/2}\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_{DDIMM} - \boldsymbol{\theta}_0, \widehat{\boldsymbol{\zeta}}_{DDIMM} - \boldsymbol{\zeta}_0)$ can be rewritten as

$$\boldsymbol{H}\left\{\sum_{k=1}^{K}\frac{n_k}{N}\boldsymbol{E}_k\boldsymbol{j}_k(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{k0})\boldsymbol{E}_k^T + O_p\left(n_{\max}^{1/2}N^{-1/2-\delta}\right)\right\}^{-1}$$
$$\left[\sum_{k=1}^{K}\left\{\left(\frac{n_k}{N}\right)^{1/2}\boldsymbol{E}_k n_k^{1/2}\boldsymbol{Z}_k\right\} + O_p\left(N^{-\delta}\right)\right].$$

Since $O_p(n_{\max}^{1/2}N^{-1/2-\delta}) = o_p(1)$ and $O_p(N^{-\delta}) = o_p(1)$, it follows as in the proof of Theorem 9 that as $n_{\min} \to \infty$,

$$N^{1/2}\boldsymbol{H}\begin{pmatrix}\widehat{\boldsymbol{\theta}}_{DDIMM} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{DDIMM} - \boldsymbol{\zeta}_0\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{j}_{\boldsymbol{H}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right). \qquad \blacksquare$$

Theorem 11 suggests that we can tune our choice of K and $n_{\min}$ to attain the desired trade-off between inference and computational speed: smaller K and larger $n_{\min}$ will slow computations but improve estimation and asymptotic normality, whereas larger K and smaller $n_{\min}$ will speed computations but worsen estimation and asymptotic normality. In practice, increasing K decreases the estimated variance. This can be understood intuitively by noting that, when $n_{\min}$ is large enough to yield $n_k^{1/2}$-consistent block estimators, we are averaging more estimators in the integration step for $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$, which decreases the estimated variance. This is illustrated numerically in Section 6. In practice, the choice of K may be driven by practical considerations such as the number of available CPU's for parallelization and the sample sizes N and $n_{\min}$, as well as modeling considerations for the heterogeneous structure $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_{jk})_{j,k=1}^{J,K}$. As long as $n_{\min}$ remains reasonably large enough to yield $n_k^{1/2}$-consistent block estimators, the choice of K should be driven by the modeling considerations, since gains in estimation efficiency are obtained when the local structures in the outcome are appropriately specified for each subject group. Some preliminary analyses are recommended in practice in order to appropriately specify local structures in the outcome.

5.4. Asymptotic Results for Diverging K and J

In general, asymptotics for diverging J become very complicated and even analytically intractable depending on how, and to what extent, the dependence structure evolves as the dimension M of $\boldsymbol{Y}$ goes to infinity ($M \to \infty$). Cox and Reid (2004) propose constructing a pseudolikelihood from marginal densities when the full joint distribution is difficult to construct, and discuss asymptotics for increasing response dimensionality. To make the problem of diverging M tractable, we consider the following regularity conditions:

(A.5) Stationarity: for each $M^* \in \mathbb{N}$ and each (M^*+1)-dimensional measurable set B a subset of the sample space of $\boldsymbol{Y}$, the distribution of $\boldsymbol{Y}_i$ satisfies $P\{(Y_{i,r}, \ldots, Y_{i,r+M^*}) \in B\} = P\{(Y_{i,0}, \ldots, Y_{i,M^*}) \in B\}$ for every $r \in \mathbb{N}$.

(A.6) Let $\boldsymbol{C}_{k,i}$ be the version of $\widehat{\boldsymbol{C}}_{k,i}$ in Equation 10 evaluated at the true values $\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}$. For $k = 1, \ldots, K$, $i = 1, \ldots, J$, $(\sum_{l=1}^{K}\sum_{j=1}^{J} n_l^2 \boldsymbol{C}_{l,j})^{-1} n_k^2 \boldsymbol{C}_{k,i} = O_p(N^{-\delta_1})$ for a constant $0 \leq \delta_1 \leq 1/2$. This can be thought of as a type of Lindeberg condition.

(A.7) Conditions required for asymptotically normal distribution and efficiency of the GMM estimator $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$; see Theorem 5.4 in Donald et al. (2003) and the spanning condition in Newey (2004). See Newey (2004) for related work on semiparametric efficiency of the GMM estimator as the number of moment conditions goes to infinity.

Remark 12 *Condition (A.5) is typical for consistency and asymptotic normality of the GMM estimator $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$, following Hansen (1982) and Newey (2004). It is a typical condition for the application of the central limit theorem to stochastic processes, i.e. to infinite dimensional random vectors. Additionally, in order to make statements about convergence in probability, (A.5) is required to ensure a valid joint probability distribution as the dimension M increases.*

Remark 13 *Condition (A.6) ensures the covariance of the outcome $\boldsymbol{Y}_i$ is appropriately controlled as $M \to \infty$. Alternative conditions may be considered, such as α-mixing (Bradley (1985)), ρ-mixing (Peligrad (1986)), or ϕ-mixing (Peligrad (1986)), but this is beyond the scope of this paper. Condition (A.6) can be simplified for the case where $n_k = n$ for all $k = 1, \ldots, K$. Then (A.6) becomes $(\sum_{l=1}^{K}\sum_{j=1}^{J} \boldsymbol{C}_{l,j})^{-1} \boldsymbol{C}_{k,i} = O_p(N^{-\delta_1})$.*

In Theorem 14 we show the consistency and asymptotic normality of the DDIMM estimator as K and J diverge to ∞.

Theorem 14 *Suppose $N^{-\delta_2} n_{\min}$ and $N^{\delta_3 - 1/2} KJ$ are bounded as $n_{\min} \to \infty$ for constants $0 \leq \delta_2 \leq 1$ and $0 < \delta_3 < 1/2$ such that $\delta_3 + \delta_1 + \delta_2/2 > 1$. Suppose assumptions (A.1), (A.2), and (A.4*)-(A.7) hold. Let $\boldsymbol{H} \in \mathbb{R}^{h \times (p+d)}$ a matrix of rank $r \in \mathbb{N}$, $h \in \mathbb{N}$, $r \leq h$, with finite maximum singular value $\bar{\sigma}(\boldsymbol{H}) < \infty$. Let $\boldsymbol{j}_{\boldsymbol{H}}(\boldsymbol{\theta}, \boldsymbol{\zeta})$ as given in Theorem 11. Then as $n_{\min} \to \infty$,*

$$N^{1/2} \boldsymbol{H} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{DDIMM} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{DDIMM} - \boldsymbol{\zeta}_0 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{j}_{\boldsymbol{H}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right).$$

Proof Write

$$\boldsymbol{H} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{DDIMM} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{DDIMM} - \boldsymbol{\zeta}_0 \end{pmatrix} = \boldsymbol{H} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{DDIMM} - \widehat{\boldsymbol{\theta}}_{opt} \\ \widehat{\boldsymbol{\zeta}}_{DDIMM} - \widehat{\boldsymbol{\zeta}}_{opt} \end{pmatrix} + \boldsymbol{H} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{opt} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{opt} - \boldsymbol{\zeta}_0 \end{pmatrix}.$$

To show the asymptotic distribution of the left-hand side, it is sufficient to show that $\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_{DDIMM}^T - \widehat{\boldsymbol{\theta}}_{opt}^T, \widehat{\boldsymbol{\zeta}}_{DDIMM}^T - \widehat{\boldsymbol{\zeta}}_{opt}^T)^T = o_p(N^{-1/2})$.
Given the assumptions of the theorem, we have the asymptotic distribution of $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt,ik})$ and $(\widehat{\boldsymbol{\theta}}_{ik}, \widehat{\boldsymbol{\zeta}}_{ik})$: both are consistent estimators of $\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{ik0}$ and asymptotically normally distributed with rates $N^{-1/2}$ and $n_k^{-1/2}$ respectively. Then for each $k \in \{1, \ldots, K\}$,

$$\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{opt} - \widehat{\boldsymbol{\theta}}_{ik} \\ \widehat{\boldsymbol{\zeta}}_{opt,ik} - \widehat{\boldsymbol{\zeta}}_{ik} \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{opt} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{opt,ik} - \boldsymbol{\zeta}_{ik0} \end{pmatrix} - \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{ik} - \boldsymbol{\zeta}_{ik0} \end{pmatrix} = O_p(n_k^{-1/2}).$$

Defining $\widehat{\boldsymbol{C}}^*_{k,i}$ a subset of $\widehat{\boldsymbol{C}}_{k,i}$ in Appendix A.4, we can rewrite $(\widehat{\boldsymbol{\theta}}^T_{DDIMM} - \widehat{\boldsymbol{\theta}}^T_{opt}, \widehat{\boldsymbol{\zeta}}^T_{DDIMM} - \widehat{\boldsymbol{\zeta}}^T_{opt})^T$ as follows:

$$\begin{aligned}
&\left(\sum_{k=1}^{K}\sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i}\right)^{-1} \left\{\sum_{k=1}^{K}\sum_{i=1}^{J}\left[n_k^2 \widehat{\boldsymbol{C}}_{k,i}\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik} - \widehat{\boldsymbol{\theta}}_{opt} \\ \widehat{\boldsymbol{\zeta}}_{list} - \widehat{\boldsymbol{\zeta}}_{opt}\end{pmatrix}\right]\right\} \\
=&\sum_{k=1}^{K}\sum_{i=1}^{J}\left[\left(\sum_{l=1}^{K}\sum_{j=1}^{J} n_l^2 \widehat{\boldsymbol{C}}_{l,j}\right)^{-1} n_k^2 \widehat{\boldsymbol{C}}^*_{k,i}\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik} - \widehat{\boldsymbol{\theta}}_{opt} \\ \widehat{\boldsymbol{\zeta}}_{ik} - \widehat{\boldsymbol{\zeta}}_{opt,ik}\end{pmatrix}\right] \\
=&\sum_{k=1}^{K}\sum_{i=1}^{J}\left[O_p(N^{-\delta_1})O_p(n_k^{-1/2})\right] = O_p(KJN^{-\delta_1}n_{\min}^{-1/2}) \\
=&O_p(N^{1/2-\delta_3}N^{-\delta_1}N^{-\delta_2/2}) = O_p(N^{1/2-\delta_3-\delta_1-\delta_2/2}) = o_p(N^{-1/2}).
\end{aligned}$$

■

Theorem 14 guarantees the desirable inferential properties of the estimator of interest $\widehat{\boldsymbol{\theta}}_{DDIMM}$ as J grows with dimension M. Our procedure only requires the specification of local structures for subsets of the data and aggregates these to form a full, possibly complex, model that can better approximate the true structure than a whole data approach without data splitting. When $\boldsymbol{\zeta}$ consists of second-order moment parameters, this better model of the true covariance structure leads to improved efficiency, as discussed in Fitzmaurice et al. (1993). The choice of J is typically fixed *a priori* given prior knowledge of local structures in the outcome, for example from substantive biological knowledge. If substantive knowledge is lacking, J should be chosen to allow for precise modeling of local structures, which may be learned using data-driven techniques in preliminary analyses to develop adaptive models of local structures.

6. Simulations

In this section we consider four sets of simulations to examine the performance of the closed-form estimator $\widehat{\boldsymbol{\theta}}_{DDIMM}$. Three sets consider the linear regression setting $\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\theta}$, where $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\theta})$ and $\boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\theta}, \boldsymbol{\Sigma})$. One set considers the logistic regression setting $\log\{\mu_{ir}/(1-\mu_{ir})\} = \boldsymbol{X}_{ir}\boldsymbol{\theta}$ with $\mu_{ir} = E(Y_{ir}|\boldsymbol{X}_{ir}, \boldsymbol{\theta})$, $r = 1, \ldots, M$, where $\boldsymbol{Y}_i$ is a M-variate correlated Bernoulli random vector. In all settings, covariates consist of an intercept and two independently simulated M-dimensional multivariate normal variables. Simulations are conducted using R software on a standard Linux cluster.

The first set of simulations illustrates the finite sample performance and properties in Theorem 9 of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ under the linear regression setting with fixed sample size N, varying number of subject groups K, varying dimensions M of $\boldsymbol{Y}$, and fixed number of response blocks J. We specify $\boldsymbol{\Sigma} = \boldsymbol{S} \otimes \boldsymbol{A}$ with nested correlation structure, where $\otimes$ denotes the Kronecker product, $\boldsymbol{A}$ is an AR(1) covariance matrix with standard deviation $\sigma = 4$ and correlation $\rho = 0.8$, and $\boldsymbol{S}$ is a randomly simulated $J \times J$ positive-definite matrix. The true value of $\boldsymbol{\theta}$ is set to $\boldsymbol{\theta}_0 = (0.3, 0.6, 0.8)^T$. We consider varying dimensions M of $\boldsymbol{Y}$ with fixed $J = 5$, and a fixed sample size $N = 5,000$ with varying $K = 1, 2, 5$. We consider two supervised learning procedures: the pairwise composite likelihood using our own package,

and the GEE using R package `geepack` and our own package (see Supplemental Material). With each procedure, we fit the model with an AR(1) working block correlation structure. Results for the GEE are in Figure 1; results for the pairwise composite likelihood (CL) are in the Supplemental Material. We see that the mean asymptotic standard error (ASE) of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ approximates the empirical standard error (ESE) for all models, with slight variations due to the type of covariates simulated. This means the covariance formula in Theorem 9 is correct. Additionally, $\widehat{\boldsymbol{\theta}}_{DDIMM}$ appears consistent since root mean squared error (RMSE), ASE and ESE are approximately equal. Moreover, we notice the ASE of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ decreases as the response dimension M increases. This makes intuitive sense, since an increase in M corresponds to an increase in overall number of observations, resulting in increased power. We also see a decrease in the ASE as the number of groups increases. This is due to the heterogeneity of block covariance parameters. Lastly, we observe from Table 2 that the mean CPU time is very fast for the GEE, and decreases substantially as the number of subject groups increases.

The second set of simulations investigates the performance and properties of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ under the linear regression setting with fixed sample size $N = 12,000$, response dimension $M = 500$ of $\boldsymbol{Y}_i$ and response blocks $J = 6$. To illustrate the effect of data splitting when the correlation of the outcome does not vary with K, we consider homogeneous outcome covariance, i.e. $\boldsymbol{\zeta}_{jk}$ is absent. We consider varying number of subjects groups $K = 1, \ldots, 6, 24, 30, 40, 60, 120$ with $n_k = n_{\min}$ for all $k = 1, \ldots, K$ to illustrate (i) the gain in efficiency from splitting data at the subject level when $n_{\min}$ is large enough (as discussed in Section 5.3), and (ii) the loss of desirable inferential performance when $n_{\min}$ becomes too small to yield good block estimators $\widehat{\boldsymbol{\theta}}_{jk}$. Responses are simulated from a Multivariate Normal distribution with AR(1) covariance structure, with standard deviation $\sigma = 4$ and correlation $\rho = 0.8$: there are no heterogeneous parameters, so that the gain in efficiency observed is due only to the data splitting procedure and not to additional variability in the outcome. The true value of $\boldsymbol{\theta}$ is set to $\boldsymbol{\theta}_0 = (1, -2, 3)^T$. We learn mean and covariance parameters using GEE with an AR(1) working block correlation structure. We see from Table 3 that ASE and ESE are approximately equal for small values of K, but that as K grows large and $n_{\min}$ grows small, ASE underestimates the true standard error of $\widehat{\boldsymbol{\theta}}_{DDIMM}$. This is because the standard error calculation assumes the block estimators are asymptotically normally distributed, an assumption that is violated as $n_{\min}$ becomes small, resulting in undercoverage of the 95% confidence interval (COV) and inflation of Type-I error (ERR). We also observe in Table 3 that mean CPU time (CPU) has a U shape, with shortest computing time for $K = 6$. This suggests that the computing burden is reduced when K and $n_{\min}$ are moderately sized, so that the GEE block analysis and inversion of covariance matrices $\widehat{\boldsymbol{V}}_N$ and $\widehat{\boldsymbol{C}}_{k,i}$ are relatively fast. In this particular simulation setting, a good strategy based on the number of covariates and block correlation modeling appears to be a choice of K such that $n_{\min} \approx 2,000$ to ensure good block estimators and shortest computing time.

The third set of simulations illustrates the performance and properties in Theorem 14 of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ under the linear regression setting with growing sample size N and response dimension M of $\boldsymbol{Y}_i$, and varying number of subjects groups K and response blocks J. We consider diverging sample size N and response dimension M, and diverging number of subject groups K and response blocks J. We consider two settings: in Setting I, we let the

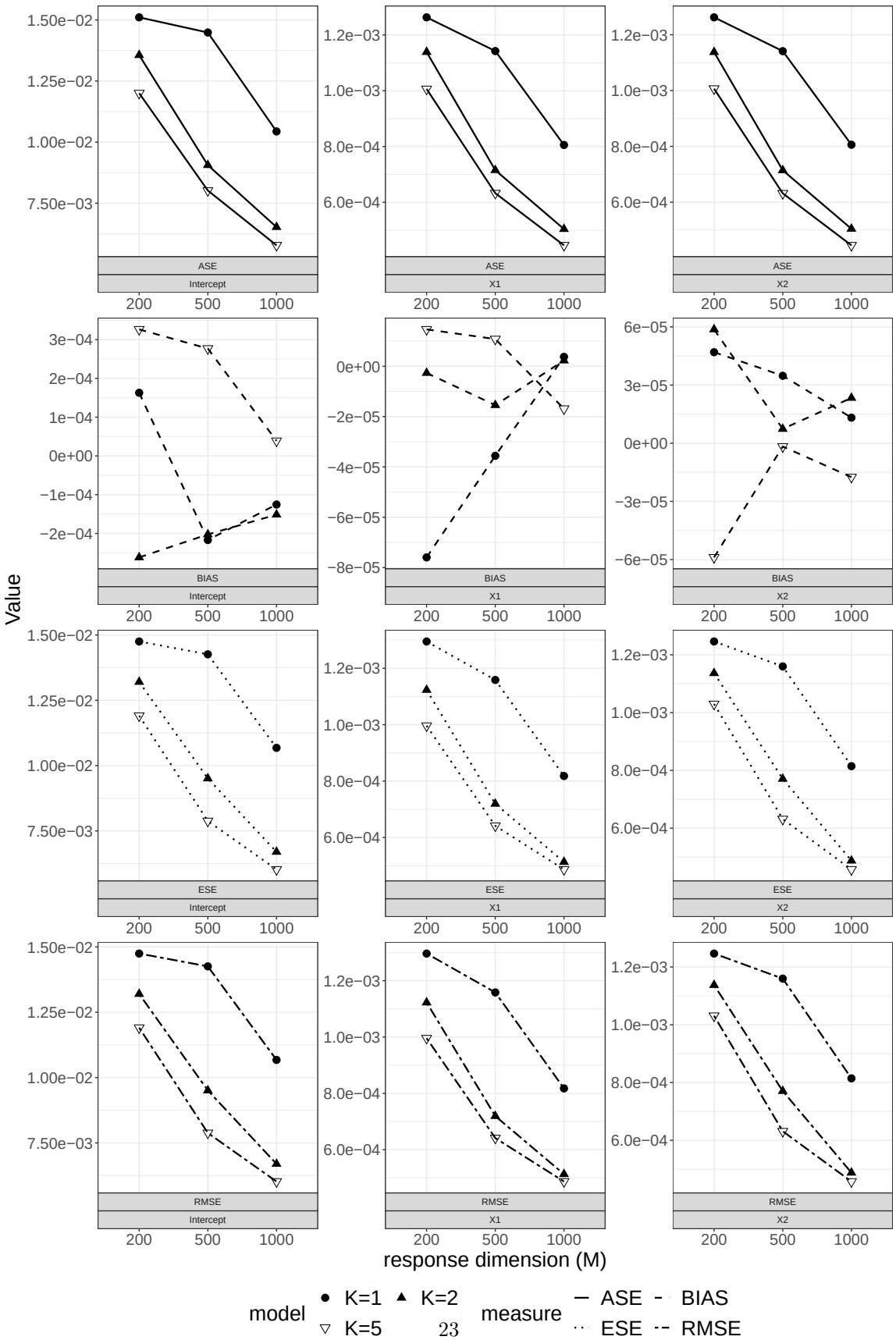

Figure 1: Plot of simulation metrics for first set of simulations with GEE, averaged over 1,000 simulations.

Response dimension	Number of subject groups		
	K=1	K=2	K=5
M=200	44	23	10
M=500	344	153	63
M=1,000	1680	876	364

Table 2: Mean CPU time in seconds for each setting in first set of simulations with the GEE block analysis, averaged over 1,000 simulations. Mean CPU time is computed as the maximum CPU time taken over parallelized block analyses added to the CPU time taken by the rest of the procedure.

K	n_k	ASE$\times 10^{-4}$	ESE$\times 10^{-4}$	COV	LEN$\times 10^{-4}$	ERR	CPU
1	12000	2.42	2.43	0.96	8.90	0.04	30
2	6000	2.42	2.44	0.96	8.90	0.04	27
3	4000	2.41	2.44	0.96	8.90	0.04	26
4	3000	2.41	2.46	0.96	8.90	0.04	26
5	2400	2.41	2.45	0.96	8.90	0.04	26
6	2000	2.40	2.45	0.96	8.90	0.04	25
12	1000	2.38	2.45	0.95	8.80	0.05	26
24	500	2.34	2.51	0.95	8.60	0.05	26
30	400	2.32	2.53	0.95	8.60	0.05	27
40	300	2.29	2.57	0.94	8.40	0.06	27
60	200	2.22	2.67	0.92	8.20	0.08	28
120	100	2.00	3.01	0.85	7.40	0.15	30

Table 3: ASE, ESE, mean 95% confidence interval coverage (COV), mean 95% confidence interval length (LEN), Type-I error (ERR) and mean CPU time (CPU) in minutes for second set of simulations with GEE with $N = 12,000$, $M = 500$ and $J = 6$ and homogeneous outcome covariance parameters, averaged over 500 simulations, taking the median over the intercept and two covariates. Mean CPU time is computed as the maximum CPU time taken over parallelized block analyses added to the CPU time taken by the rest of the procedure.

sample size $N = 5,000$ with number of response groups $K = 1$, and let response dimension $M = 4,500$ with number of response blocks $J = 6$; in Setting II, we let the sample size $N = 10,000$ with number of response groups $K = 2$, and let response dimension $M = 9,000$ with number of response blocks $J = 12$. Responses are simulated from a Multivariate Normal distribution with AR(1) covariance structure, with standard deviation $\sigma = 6$ and correlation $\rho = 0.8$. This means there are no heterogeneous block parameters, so we expect a slightly less efficient estimator since there is less variability in the outcome. The true value of $\boldsymbol{\theta}$ is set to $\boldsymbol{\theta}_0 = (0.3, 0.6, 0.8)^T$. We learn mean and covariance parameters using GEE with an AR(1) working block correlation structure. Mean bias (BIAS), RMSE, ESE and ASE of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ are in Table 4. We observe that RMSE, ESE and ASE are very close,

indicating appropriate estimation of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ and its covariance in Theorem 14. We also confirm DDIMM's ability to handle large sample size N and response dimension M.

Setting	Measure	Intercept	X_1	X_2
I: $K = 1$, $J = 6$	RMSE/BIAS	3.90/−1.80	0.64/0.09	0.60/−0.41
	ESE/ASE	3.90/3.78	0.64/0.59	0.60/0.59
II: $K = 2$, $J = 12$	RMSE/BIAS	1.86/−1.09	0.28/−0.03	0.28/−0.16
	ESE/ASE	1.86/1.70	0.28/0.27	0.28/0.27

Table 4: RMSE$\times 10^{-3}$, BIAS$\times 10^{-4}$, ESE$\times 10^{-3}$, ASE$\times 10^{-3}$ for each setting and each covariate in the third set of simulations, averaged over 500 simulations.

The fourth set of simulations illustrates the finite sample performance and properties in Theorem 14 of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ under the logistic regression setting with fixed sample size $N = 10,000$ and fixed number of subject groups $K = 2$, varying dimensions M of $\boldsymbol{Y}_i$ and varying number of response blocks J. We consider three settings: in Setting I, we let $M = 500$ with number of response blocks $J = 3$; in Setting II, we let $M = 1,000$ with $J = 6$; in Setting III, we let $M = 2,000$ with $J = 12$. $\boldsymbol{Y}_i$ is simulated using the `SimCorMultRes` R package (Touloumis, 2016) with block AR(1) correlation structures with varying variance and correlation parameters. The true value of $\boldsymbol{\theta}$ is set to $\boldsymbol{\theta}_0 = (-0.3, 0.4, 0.2)^T$. We learn mean and covariance parameters using GEE with an AR(1) working block correlation structure. BIAS, RMSE, ESE and ASE of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ are listed in Table 5. We again observe that RMSE, ESE and ASE are very close, indicating appropriate estimation of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ and the asymptotic variances in Theorem 14. ASE for the Intercept appears more variable than for X_1 and X_2, which is likely due to low variability in this predictor rendering parameter estimation more difficult. We also confirm DDIMM's ability to handle binomial distributed outcome data. Lastly, mean CPU times of 24, 21 and 22 minutes are observed for Settings I, II and III respectively. On a smaller scale, we observe a similar U shaped pattern to the computing time in Table 3, with smaller computing time for moderately sized values of J.

7. Discussion

We have presented the large sample theory as a theoretical guarantee for a Doubly Distributed and Integrated Method of Moments (DDIMM) that incorporates a broad class of supervised learning procedures into a doubly distributed and parallelizable computational scheme for the efficient analysis of large samples of high-dimensional correlated responses in the MapReduce framework. Theoretical challenges related to combining correlated estimators were addressed in the proofs, including the asymptotic properties of the proposed closed-form estimator with fixed and diverging numbers of subject groups and response blocks.

The GMM approach to deriving the combined estimator $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ proposed in Equation 4 requires only weak regularity of the estimating equations $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$. These assumptions are satisfied by a broad range of learning procedures. The closed-form estimator proposed in Equation 9, on the other hand, requires local $n_k^{1/2}$-consistent estimators in individual blocks

Setting	Measure	Intercept	X_1	X_2
I: $J = 3$, $M = 500$	RMSE/BIAS	1.53/−3.92	0.56/1.70	0.45/1.92
	ESE/ASE	1.53/1.55	0.56/0.56	0.45/0.44
II: $J = 6$, $M = 1000$	RMSE/BIAS	1.14/3.68	0.39/−1.14	0.32/−0.35
	ESE/ASE	1.14/1.19	0.39/0.41	0.32/0.32
III: $J = 12$, $M = 2000$	RMSE/BIAS	0.93/−0.57	0.30/0.90	0.23/1.14
	ESE/ASE	0.93/0.87	0.30/0.29	0.23/0.23

Table 5: RMSE$\times 10^{-3}$, BIAS$\times 10^{-5}$, ESE$\times 10^{-3}$, ASE$\times 10^{-3}$ for each setting and each covariate in the fourth set of simulations, averaged over 1,000 simulations.

of size n_k, which is easily satisfied if $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ are regular (see Song (2007) Chapter 3.5 for a definition of regular inference functions). This restricts the class of possible learning procedures, but still includes many analyses of interest.

A detailed discussion of the limitations and trade-offs of the single split DIMM with CL block analyses is featured in Hector and Song (2020). As mentioned in Section 5, the DDIMM introduces additional flexibility in trading off between computational speed and inference: the number of subject groups K and the smallest block size $n_{\min}$ can be chosen by the investigator to attain the desired speed and efficiency.

Particular applications of DDIMM to time series data are immediately obvious. Similarly, we envision potential application to nation-wide hospital daily visit numbers of, for example, asthma patients, over the course of the last decade. One could split the response (hospital daily intake/daily stock price) into J years and into K groups (of hospitals/stocks), analyze blocks separately and in parallel using GEE, and combine results using DDIMM. Finally, extensions of our work to stochastic process modeling are accessible, with more challenging work involving regularization of $\boldsymbol{\theta}$ also of interest.

Acknowledgments

The authors are grateful for the constructive comments given by the action editor and the anonymous reviewers that led to a significant improvement of the article. We would like to acknowledge support for this project from the National Science Foundation (NSF DMS1513595) and the National Institutes of Health (NIH R01ES024732).

Appendix A. Technical details

A.1. Summary of sensitivity matrix formulas

Sensitivity matrices are summarized in Table A.1.

A.2. Subsetting operation on variability matrices

Operation $\left[\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi}}\right]_{ij:k}$ extracts a submatrix of $\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi}}$ consisting of rows $\{(i-1)+(k-1)J\}\,p+1$ to $\{i+(k-1)J\}\,p$ and columns $\{j-1+(k-1)J\}\,p+1$ to $\{j+(k-1)J\}\,p$. Operation

sensitivity of	w.r.t.*	population	sample	plug-in sample
$\boldsymbol{\psi}_{i,jk}$	$\boldsymbol{\theta}$	$\boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})$	$\boldsymbol{S}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})$	$\widehat{\boldsymbol{S}}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}} = \boldsymbol{S}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}(\widehat{\boldsymbol{\theta}}_{jk},\widehat{\boldsymbol{\zeta}}_{jk})$
$\boldsymbol{\psi}_{i,jk}$	$\boldsymbol{\zeta}_{jk}$	$\boldsymbol{s}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})$	$\boldsymbol{S}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})$	$\widehat{\boldsymbol{S}}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}} = \boldsymbol{S}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}(\widehat{\boldsymbol{\theta}}_{jk},\widehat{\boldsymbol{\zeta}}_{jk})$
$\boldsymbol{g}_{i,jk}$	$\boldsymbol{\theta}$	$\boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{g}_{jk}}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})$	$\boldsymbol{S}^{\boldsymbol{\theta}}_{\boldsymbol{g}_{jk}}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})$	$\widehat{\boldsymbol{S}}^{\boldsymbol{\theta}}_{\boldsymbol{g}_{jk}} = \boldsymbol{S}^{\boldsymbol{\theta}}_{\boldsymbol{g}_{jk}}(\widehat{\boldsymbol{\theta}}_{jk},\widehat{\boldsymbol{\zeta}}_{jk})$
$\boldsymbol{g}_{i,jk}$	$\boldsymbol{\zeta}_{jk}$	$\boldsymbol{s}^{\boldsymbol{\zeta}}_{\boldsymbol{g}_{jk}}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})$	$\boldsymbol{S}^{\boldsymbol{\zeta}}_{\boldsymbol{g}_{jk}}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})$	$\widehat{\boldsymbol{S}}^{\boldsymbol{\zeta}}_{\boldsymbol{g}_{jk}} = \boldsymbol{S}^{\boldsymbol{\zeta}}_{\boldsymbol{g}_{jk}}(\widehat{\boldsymbol{\theta}}_{jk},\widehat{\boldsymbol{\zeta}}_{jk})$
$\mathbb{S}\left(\boldsymbol{\psi}_{i,jk},\boldsymbol{g}_{i,jk}\right)$	$(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})$	$\boldsymbol{s}_{jk}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})$	$\boldsymbol{S}_{jk}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})$	$\widehat{\boldsymbol{S}}_{jk} = \boldsymbol{S}_{jk}(\widehat{\boldsymbol{\theta}}_{jk},\widehat{\boldsymbol{\zeta}}_{jk})$

Table A.1: Summary of sensitivity formulas. *"w.r.t." shorthand for "with respect to". $\left[\widehat{\boldsymbol{V}}^{\boldsymbol{g}}_N\right]_{ij:k}$ extracts a submatrix of $\widehat{\boldsymbol{V}}^{\boldsymbol{g}}_N$ consisting of rows $1+D^{ik}$ to $d_{ik}+D^{ik}$ and columns $1+D^{jk}$ to $d_{jk}+D^{jk}$. Operation $\left[\widehat{\boldsymbol{V}}^{\boldsymbol{\psi g}}_N\right]_{ij:k}$ extracts a submatrix of $\widehat{\boldsymbol{V}}^{\boldsymbol{\psi g}}_N$ consisting of rows $\{(i-1)+(k-1)J\}\,p+1$ to $\{i+(k-1)J\}\,p$ and columns $1+D^{jk}$ to $d_{jk}+D^{jk}$, where d_{jk} is the dimension of $\boldsymbol{\zeta}_{jk}$ and D^{jk} is defined in Section 5.1.

A.3. Cumulative sum of dimensions of $\boldsymbol{\zeta}$

Recall that we define D^{ik} as the sum of the dimensions of $\boldsymbol{\zeta}_{11},\ldots,\boldsymbol{\zeta}_{i-1k}$, and D^k as the sum of the dimensions of $\boldsymbol{\zeta}_{11},\ldots,\boldsymbol{\zeta}_{Jk-1}$. Specifically, let $D^{ik}=\sum_{l=1}^{k-1}\sum_{j=1}^{J} d_{jl}+\sum_{j=1}^{i-1} d_{jk}$ for $i,k>1$, $D^{1k}=\sum_{l=1}^{k-1}\sum_{j=1}^{J} d_{jl}$ for $k>1$, and $D^{11}=0$. Let $D^k=\sum_{l=1}^{k-1} d_l$ for $k>1$ and $D^1=0$.

A.4. Definition of $\widehat{\boldsymbol{C}}^*_{k,i}$

Let $k\in\{1,\ldots,K\}$ and $i\in\{1,\ldots,J\}$. Recall the definitions of $\widehat{\boldsymbol{A}}^{\boldsymbol{\theta}}_{k,ij}$, $\widehat{\boldsymbol{A}}^{\boldsymbol{\zeta}}_{k,ij}$, $\widehat{\boldsymbol{B}}^{\boldsymbol{\theta}}_{k,ij}$ and $\widehat{\boldsymbol{B}}^{\boldsymbol{\zeta}}_{k,ij}$ in Section 5.1. Define

$$\widehat{\boldsymbol{C}}^*_{k,i}=\begin{pmatrix} \sum_{j=1}^{J}\widehat{\boldsymbol{A}}^{\boldsymbol{\theta}}_{k,ij} & \sum_{j=1}^{J}\widehat{\boldsymbol{A}}^{\boldsymbol{\zeta}}_{k,ij} \\ \multicolumn{2}{c}{\mathbf{0}_{D^{ik}\times(p+d)}} \\ \widehat{\boldsymbol{B}}^{\boldsymbol{\theta}}_{k,i1} & \widehat{\boldsymbol{B}}^{\boldsymbol{\zeta}}_{k,i1} \\ \multicolumn{2}{c}{\vdots} \\ \widehat{\boldsymbol{B}}^{\boldsymbol{\theta}}_{k,iJ} & \widehat{\boldsymbol{B}}^{\boldsymbol{\zeta}}_{k,iJ} \\ \multicolumn{2}{c}{\mathbf{0}_{(d-d_{ik}-D^{ik})\times(p+d)}} \end{pmatrix}.$$

Appendix B. Additional proofs

B.1. Proof of Theorem 9:

The following lemmas complete the proof of Theorem 9 given in the paper, under the assumed conditions.

Lemma B.1.1 *Define $\lambda(\boldsymbol{\theta},\boldsymbol{\zeta})$ as in Equation 11 in the proof of Theorem 9. Then $\lambda(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\xrightarrow{p}0$ as $n_{\min}\to\infty$.*

Proof Using Lemma 8,

$$\begin{aligned}\lambda(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) &= \frac{1}{N^2}\sum_{k=1}^{K}\sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i} \begin{pmatrix} \boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_{ik} \\ \boldsymbol{\zeta}_0 - \widehat{\boldsymbol{\zeta}}_{list} \end{pmatrix} \\ &= O_p\left(n_{\min}^{-1/2}\right)\left\{\boldsymbol{j}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p\left(N^{-1/2}\right)\right\} \\ &= O_p\left(n_{\min}^{-1/2}\right) + O_p\left(n_{\min}^{-1/2} N^{-1/2}\right) \xrightarrow{p} 0 \text{ as } n_{\min} \to \infty. \end{aligned}$$

■

Lemma B.1.2 *The following relationship holds:*

$$\begin{pmatrix} \boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}_0; \boldsymbol{\zeta}_{jk0}) \\ \boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk0}; \boldsymbol{\theta}_0) \end{pmatrix} = \widehat{\boldsymbol{S}}_{jk} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix} + O_p(n_k^{-1}).$$

Proof Let $j \in \{1, \ldots, J\}$, $k \in \{1, \ldots, K\}$ fixed. For convenience, denote

$$\boldsymbol{T}_{jk}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) = \begin{pmatrix} \boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}; \boldsymbol{\zeta}_{jk}) \\ \boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk}; \boldsymbol{\theta}) \end{pmatrix}, \quad \boldsymbol{\tau}_{i,jk}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) = \begin{pmatrix} \boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}; \boldsymbol{\zeta}_{jk}) \\ \boldsymbol{g}_{i,jk}(\boldsymbol{\zeta}_{jk}; \boldsymbol{\theta}) \end{pmatrix}.$$

By first-order Taylor expansion,

$$\begin{aligned} E_{\boldsymbol{\theta},\boldsymbol{\zeta}_{jk}}\left\{\boldsymbol{\tau}_{i,jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk})\right\} &= E_{\boldsymbol{\theta},\boldsymbol{\zeta}_{jk}}\left\{\boldsymbol{\tau}_{i,jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0})\right\} + \\ &\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta},\boldsymbol{\zeta}_{jk}}\left\{\boldsymbol{\tau}_{i,jk}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})\right\}|_{\boldsymbol{\theta}^*, \boldsymbol{\zeta}_{jk}^*} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix}, \end{aligned} \tag{14}$$

where $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}_{jk}^*)$ lies between $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0})$ and $(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk})$. By condition (A.4*),

$$\begin{aligned} &\boldsymbol{T}_{jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk}) - \boldsymbol{T}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) - E_{\boldsymbol{\theta},\boldsymbol{\zeta}_{jk}}\left\{\boldsymbol{\tau}_{i,jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk})\right\} \\ &\qquad = O_p(n_k^{-1/2}) \frac{1 + n_k^{1/2} O_p(n_k^{-1/2})}{n_k^{1/2}} = O_p(n_k^{-1}). \end{aligned} \tag{15}$$

In other words, the norm of the difference between $\boldsymbol{T}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0})$ and $\boldsymbol{T}_{jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk}) - E_{\boldsymbol{\theta},\boldsymbol{\zeta}_{jk}}\{\boldsymbol{\tau}_{i,jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk})\}$ goes to 0 at a rate faster than n_k^{-1}. Adding Equations 14 and 15, we have

$$\begin{aligned} -\boldsymbol{T}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) &= \boldsymbol{T}_{jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk}) - \boldsymbol{T}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) - E_{\boldsymbol{\theta},\boldsymbol{\zeta}_{jk}} \boldsymbol{\tau}_{i,jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) \\ &= \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta},\boldsymbol{\zeta}_{jk}} \boldsymbol{\tau}_{i,jk}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})|_{\boldsymbol{\theta}^*, \boldsymbol{\zeta}_{jk}^*} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix} + O_p(n_k^{-1}) \\ &= -\boldsymbol{s}_{jk}(\boldsymbol{\theta}^*, \boldsymbol{\zeta}_{jk}^*) \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix} + O_p(n_k^{-1}). \end{aligned}$$

Rearranging yields

$$\boldsymbol{T}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) = \boldsymbol{s}_{jk}(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*_{jk}) \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix} + O_p(n_k^{-1}). \tag{16}$$

Finally, note that $\widehat{\boldsymbol{S}}_{jk} = \boldsymbol{s}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) + O_p(n_k^{-1/2}) = \boldsymbol{s}_{jk}(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*_{jk}) + O_p(n_k^{-1/2})$. Then plugging this into Equation 16, we have:

$$\begin{aligned} \boldsymbol{T}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) &= \left(\widehat{\boldsymbol{S}}_{jk} + O_p(n_k^{-1/2})\right) \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix} + O_p(n_k^{-1}) \\ &= \widehat{\boldsymbol{S}}_{jk} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix} + O_p(n_k^{-1}). \end{aligned}$$

■

B.2. Proof of Theorem 11

The following lemmas complete the proof of Theorem 11 given in the paper, under the assumed conditions.

Lemma B.2.1 *Define $\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})$ as in Equation 11 in the proof of Theorem 9. Then $\|\lambda(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\| = O_p(N^{-1/2-\delta} n_{\max}^{1/2})$ and $\left\|\{\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}} \lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})\}^{-1}\right\| = O_p\left(N^{1/2+\delta} n_{\max}^{-1}\right)$.*

Proof Due to the independence between subject groups, $\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi}}$, $\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi g}}$ and $\widehat{\boldsymbol{V}}_N^{\boldsymbol{g}}$ are all block diagonal: $\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi}} = \operatorname{diag}\left\{\widehat{\boldsymbol{V}}_k^{\boldsymbol{\psi}}\right\}_{k=1}^K$, $\widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi g}} = \operatorname{diag}\left\{\widehat{\boldsymbol{V}}_k^{\boldsymbol{\psi g}}\right\}_{k=1}^K$, and $\widehat{\boldsymbol{V}}_N^{\boldsymbol{g}} = \operatorname{diag}\left\{\widehat{\boldsymbol{V}}_k^{\boldsymbol{g}}\right\}_{k=1}^K$. By the independence of subject groups, let

$$\begin{aligned} \boldsymbol{v}^{-1}(\boldsymbol{\theta}, \boldsymbol{\zeta}) &= \begin{pmatrix} \boldsymbol{v}^{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) & \boldsymbol{v}^{\boldsymbol{\psi g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) \\ \boldsymbol{v}^{\boldsymbol{\psi g}\ T}(\boldsymbol{\theta}, \boldsymbol{\zeta}) & \boldsymbol{v}^{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) \end{pmatrix} \\ &= \begin{pmatrix} \operatorname{diag}\left\{\frac{N}{n_k} \boldsymbol{v}_k^{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\zeta})\right\}_{k=1}^K & \operatorname{diag}\left\{\frac{N}{n_k} \boldsymbol{v}_k^{\boldsymbol{\psi g}}(\boldsymbol{\theta}, \boldsymbol{\zeta})\right\}_{k=1}^K \\ \operatorname{diag}\left\{\frac{N}{n_k} \boldsymbol{v}_k^{\boldsymbol{\psi g}\ T}(\boldsymbol{\theta}, \boldsymbol{\zeta})\right\}_{k=1}^K & \operatorname{diag}\left\{\frac{N}{n_k} \boldsymbol{v}_k^{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{\zeta})\right\}_{k=1}^K \end{pmatrix}. \end{aligned}$$

Similar to the proof of Lemma 8, it can easily be shown that for each $k = 1, \ldots, K$, $\widehat{\boldsymbol{V}}_k^{\boldsymbol{\psi}} = (N/n_k)\boldsymbol{v}_k^{\boldsymbol{\psi}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p(N^{-1/2})$, $\widehat{\boldsymbol{V}}_k^{\boldsymbol{\psi g}} = (N/n_k)\boldsymbol{v}_k^{\boldsymbol{\psi g}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p(N^{-1/2})$, and $\widehat{\boldsymbol{V}}_k^{\boldsymbol{g}} = (N/n_k)\boldsymbol{v}_k^{\boldsymbol{g}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p(N^{-1/2})$. Consider an arbitrary $k \in \{1, \ldots, K\}$. Let $(N/n_k)\left[\boldsymbol{v}_k^{\boldsymbol{\psi}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right]_{ji} = \left[\boldsymbol{v}^{\boldsymbol{\psi}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right]_{ji:k}$, and similarly define $\left[\boldsymbol{v}_k^{\boldsymbol{\psi g}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right]_{ji}$ and $\left[\boldsymbol{v}_k^{\boldsymbol{g}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right]_{ji}$. Then $\widehat{\boldsymbol{A}}_{k,ij}^{\boldsymbol{\theta}} = (N/n_k)\{\boldsymbol{a}_{k,ij}^{\boldsymbol{\theta}} + O_p(n_k^{-1/2})\}$, where $\boldsymbol{a}_{k,ij}^{\boldsymbol{\theta}}$ is defined as

$$\begin{aligned} &\left\{\boldsymbol{s}_{\boldsymbol{\psi}_{jk}}^{\boldsymbol{\theta}\ T}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) \left[\boldsymbol{v}_k^{\boldsymbol{\psi}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right]_{ji} + \boldsymbol{s}_{\boldsymbol{g}_{jk}}^{\boldsymbol{\theta}\ T}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) \left[\boldsymbol{v}_k^{\boldsymbol{\psi g}\ T}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right]_{ji}\right\} \boldsymbol{s}_{\boldsymbol{\psi}_{ik}}^{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + \\ &\left\{\boldsymbol{s}_{\boldsymbol{\psi}_{jk}}^{\boldsymbol{\theta}\ T}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) \left[\boldsymbol{v}_k^{\boldsymbol{\psi g}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right]_{ji} + \boldsymbol{s}_{\boldsymbol{g}_{jk}}^{\boldsymbol{\theta}\ T}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) \left[\boldsymbol{v}_k^{\boldsymbol{g}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right]_{ji}\right\} \boldsymbol{s}_{\boldsymbol{g}_{ik}}^{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0). \end{aligned}$$

We can show similar results for $\widehat{\boldsymbol{A}}^{\boldsymbol{\zeta}}_{k,ij}$, $\widehat{\boldsymbol{B}}^{\boldsymbol{\theta}}_{k,ij}$ and $\widehat{\boldsymbol{B}}^{\boldsymbol{\zeta}}_{k,ij}$. Then we can rewrite

$$\begin{aligned}
\|\lambda(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\| &\leq \sum_{k=1}^{K} O_p(n_k^{1/2}N^{-1}) = O_p(Kn_{\max}^{1/2}N^{-1}) = O_p(N^{-1/2-\delta}n_{\max}^{1/2}), \text{ and} \\
\|\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}}\lambda(\boldsymbol{\theta},\boldsymbol{\zeta})\| &\leq \frac{1}{N^2}\sum_{k=1}^{K}\sum_{i=1}^{J} n_k^2 \left\|\widehat{\boldsymbol{C}}_{k,i}\right\| \\
&\leq O_p\left(N^{-1/2-\delta}n_{\max}^{1/2}\right) + O\left(N^{-1/2-\delta}n_{\max}\right) = O_p\left(N^{-1/2-\delta}n_{\max}\right).
\end{aligned}$$

Since $\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}}\lambda(\boldsymbol{\theta},\boldsymbol{\zeta})$ is symmetric positive-definite, the above provides a bound on its eigenvalues. Therefore, $\left\|\{\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}}\lambda(\boldsymbol{\theta},\boldsymbol{\zeta})\}^{-1}\right\| = O_p\left(N^{1/2+\delta}n_{\max}^{-1}\right)$. ■

Lemma B.2.2 *For some matrices $\boldsymbol{E}_k$, $k = 1,\ldots,K$, of $\mathbf{0}$'s and $\mathbf{1}$'s, the following asymptotic properties hold:*

$$\begin{aligned}
\frac{n_k^2}{N^2}\sum_{i=1}^{J}\widehat{\boldsymbol{C}}_{k,i}\begin{pmatrix}\widehat{\boldsymbol{\theta}}_{ik}-\boldsymbol{\theta}_0\\ \widehat{\boldsymbol{\zeta}}_{list}-\boldsymbol{\zeta}_0\end{pmatrix} &= \frac{n_k}{N}\boldsymbol{E}_k\boldsymbol{Z}_k + O_p\left(N^{-1}\right), \\
\textit{and}\quad \frac{n_k^2}{N^2}\sum_{i=1}^{J}\widehat{\boldsymbol{C}}_{k,i} &= \frac{n_k}{N}\boldsymbol{E}_k\boldsymbol{j}_k(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0})\boldsymbol{E}_k^T + O_p\left(n_k^{1/2}N^{-1}\right),
\end{aligned}$$

where $n_k^{1/2}\boldsymbol{Z}_k \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{j}_k^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0})\right)$.

Proof Recall that $\widehat{\boldsymbol{C}}_{k,i}(\widehat{\boldsymbol{\theta}}_{ik}^T - \boldsymbol{\theta}_0^T, \widehat{\boldsymbol{\zeta}}_{list}^T - \boldsymbol{\zeta}_0^T)^T = \widehat{\boldsymbol{C}}^*_{k,i}(\widehat{\boldsymbol{\theta}}_{ik}^T - \boldsymbol{\theta}_0^T, \widehat{\boldsymbol{\zeta}}_{ik}^T - \boldsymbol{\zeta}_{ik0}^T)^T$. Let $\left[\boldsymbol{v}_k^{-1}(\boldsymbol{\theta},\boldsymbol{\zeta}_k)\right]_{ij}$ subset the rows for the parameters corresponding to block (i,k) and the columns for the parameters corresponding to block (j,k) of matrix $\boldsymbol{v}_k^{-1}(\boldsymbol{\theta},\boldsymbol{\zeta}_k)$. Define $\boldsymbol{j}_{jik}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk},\boldsymbol{\zeta}_{ik}) = \boldsymbol{s}_{jk}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})\left[\boldsymbol{v}_k^{-1}(\boldsymbol{\theta},\boldsymbol{\zeta}_k)\right]_{ji}\boldsymbol{s}_{ik}(\boldsymbol{\theta},\boldsymbol{\zeta}_{ik})$, and $\left[\boldsymbol{j}_k^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0})\right]_i$ the submatrix of $\boldsymbol{j}_k^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0})$ corresponding to parameters in block (i,k), such that

$$n_k^{1/2}\left\{\sum_{j=1}^{J}\boldsymbol{j}_{jik}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{jk0},\boldsymbol{\zeta}_{ik0})\right\}\begin{pmatrix}\widehat{\boldsymbol{\theta}}_{ik}-\boldsymbol{\theta}_0\\ \widehat{\boldsymbol{\zeta}}_{ik}-\boldsymbol{\zeta}_{ik0}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left[\boldsymbol{j}_k^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0})\right]_i\right).$$

Then using the results in the proof of Lemma B.2.1, let $\boldsymbol{E}_k$ and $\boldsymbol{E}_{k,i}$ matrices of $\mathbf{0}$'s and $\mathbf{1}$'s such that

$$\begin{aligned}
\frac{n_k^2}{N^2}\sum_{i=1}^{J}\widehat{\boldsymbol{C}}_{k,i} &= \frac{n_k}{N}\boldsymbol{E}_k\left\{\boldsymbol{j}_k(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0}) + O_p\left(n_k^{-1/2}\right)\right\}\boldsymbol{E}_k^T \\
&= \frac{n_k}{N}\boldsymbol{E}_k\boldsymbol{j}_k(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0})\boldsymbol{E}_k^T + O_p\left(n_k^{1/2}N^{-1}\right), \text{ and} \\
\frac{n_k^2}{N^2}\sum_{i=1}^{J}\widehat{\boldsymbol{C}}_{k,i}&\begin{pmatrix}\widehat{\boldsymbol{\theta}}_{ik}-\boldsymbol{\theta}_0\\ \widehat{\boldsymbol{\zeta}}_{list}-\boldsymbol{\zeta}_0\end{pmatrix}
\end{aligned}$$

$$= \frac{n_k}{N}\boldsymbol{E}_k \sum_{i=1}^{J} \boldsymbol{E}_{k,i} \left\{ \sum_{j=1}^{J} \boldsymbol{j}_{jik}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}, \boldsymbol{\zeta}_{ik0}) + O_p\left(n_k^{-1/2}\right) \right\} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{ik} - \boldsymbol{\zeta}_{ik0} \end{pmatrix}$$
$$= \frac{n_k}{N}\boldsymbol{E}_k \sum_{i=1}^{J} \boldsymbol{E}_{k,i} \sum_{j=1}^{J} \boldsymbol{j}_{jik}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}, \boldsymbol{\zeta}_{ik0}) \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{ik} - \boldsymbol{\zeta}_{ik0} \end{pmatrix} + O_p\left(N^{-1}\right).$$

To obtain the desired result, define

$$\boldsymbol{Z}_k = \sum_{i=1}^{J} \boldsymbol{E}_{k,i} \sum_{j=1}^{J} \boldsymbol{j}_{jik}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}, \boldsymbol{\zeta}_{ik0}) \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{ik} - \boldsymbol{\zeta}_{ik0} \end{pmatrix}. \qquad \blacksquare$$

Lemma B.2.3 *$N^{1/2}\boldsymbol{H}\left(\widehat{\boldsymbol{\theta}}_{DDIMM}^T - \boldsymbol{\theta}_0^T, \widehat{\boldsymbol{\zeta}}_{DDIMM}^T - \boldsymbol{\zeta}_0^T\right)$ can be rewritten as*

$$\boldsymbol{H}\left\{\sum_{k=1}^{K} \frac{n_k}{N}\boldsymbol{E}_k \boldsymbol{j}_k(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{k0})\boldsymbol{E}_k^T + O_p\left(n_{\max}^{1/2} N^{-1/2-\delta}\right)\right\}^{-1} \left[\sum_{k=1}^{K}\left\{\left(\frac{n_k}{N}\right)^{1/2}\boldsymbol{E}_k n_k^{1/2}\boldsymbol{Z}_k\right\} + O_p\left(N^{-\delta}\right)\right].$$

Proof

$$N^{1/2}\boldsymbol{H}\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{DDIMM} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{DDIMM} - \boldsymbol{\zeta}_0 \end{pmatrix}$$
$$= N^{1/2}\boldsymbol{H}\left(\sum_{k=1}^{K}\sum_{i=1}^{J} \frac{n_k^2}{N^2}\widehat{\boldsymbol{C}}_{k,i}\right)^{-1} \sum_{k=1}^{K}\sum_{i=1}^{J} \frac{n_k^2}{N^2}\widehat{\boldsymbol{C}}_{k,i}\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{list} - \boldsymbol{\zeta}_0 \end{pmatrix}$$
$$= \boldsymbol{H}\left[\sum_{k=1}^{K}\left\{\frac{n_k}{N}\boldsymbol{E}_k \boldsymbol{j}_k(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{k0})\boldsymbol{E}_k^T + O_p(n_k^{1/2}N^{-1})\right\}\right]^{-1} \cdot$$
$$\sum_{k=1}^{K}\left\{\frac{n_k}{N^{1/2}}\boldsymbol{E}_k \sum_{i=1}^{J}\boldsymbol{E}_{k,i}\boldsymbol{j}_{ik}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}, \boldsymbol{\zeta}_{ik0})\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{ik} - \boldsymbol{\zeta}_{ik0} \end{pmatrix} + O_p(N^{-1/2})\right\}$$
$$= \boldsymbol{H}\left\{\sum_{k=1}^{K}\frac{n_k}{N}\boldsymbol{E}_k \boldsymbol{j}_k(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{k0})\boldsymbol{E}_k^T + O_p\left(K n_{\max}^{1/2} N^{-1}\right)\right\}^{-1} \cdot$$
$$\left[\sum_{k=1}^{K}\left\{\frac{n_k}{N^{1/2}}\boldsymbol{E}_k \sum_{i=1}^{J}\boldsymbol{E}_{k,i}\boldsymbol{j}_{ik}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}, \boldsymbol{\zeta}_{ik0})\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{ik} - \boldsymbol{\zeta}_{ik0} \end{pmatrix}\right\} + O_p\left(K N^{-1/2}\right)\right]$$
$$= \boldsymbol{H}\left\{\sum_{k=1}^{K}\frac{n_k}{N}\boldsymbol{E}_k \boldsymbol{j}_k(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{k0})\boldsymbol{E}_k^T + O_p\left(n_{\max}^{1/2} N^{-1/2-\delta}\right)\right\}^{-1} \cdot$$
$$\left[\sum_{k=1}^{K}\left\{\left(\frac{n_k}{N}\right)^{1/2}\boldsymbol{E}_k n_k^{1/2}\boldsymbol{Z}_k\right\} + O_p\left(N^{-\delta}\right)\right]. \qquad \blacksquare$$

References

Donald W.K. Andrews. Empirical process methods in econometrics. *Handbook of Econometrics*, 4:2247 – 2294, 1994.

Yun Bai, Jian Kang, and Peter X.-K. Song. Efficient pairwise composite likelihood estimation for spatial-clustered data. *Biometrics*, 70(3):661–670, 2014.

Olha Bodnar, Taras Bodnar, and Arjun K. Gupta. Estimation and inference for dependence in multivariate data. *Journal of Multivariate Analysis*, 101(4):869–881, 2010.

Richard C. Bradley. On the central limit question under absolute regularity. *The Annals of Probability*, 13(4):1314–1325, 1985.

Vincent Carey, Scott L. Zeger, and Peter Diggle. Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80(3):517–526, 1993.

Jennifer S.K. Chan, Anthony Y.C. Kuk, James Bell, and Charles McGilchrist. The analysis of methadone clinic data using marginal and conditional logistic models with mixture of random effects. *Australian and New Zealand Journal of Statistics*, 40(1):1–10, 1998.

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

David R. Cox and Nancy Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

Stephen G. Donald, Guido W. Imbens, and Whitney K. Newey. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117(1):55–93, 2003.

Garrett M. Fitzmaurice, Nan M. Laird, and Andrea G. Rotnitzky. Regression models for discrete longitudinal responses. *Statistical Science*, 8(3):284–309, 1993.

Liya Fu and You-Gan Wang. Quantile regression for longitudinal data with a working correlation model. *Computational Statistics and Data Analysis*, 56(8):2526–2538, 2012.

Peisong Han and Peter X.-K. Song. A note on improving quadratic inference functions using a linear shrinkage approach. *Statistics and Probability Letters*, 81(3):438–445, 2011.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

Emily C. Hector and Peter X.-K. Song. A distributed and integrated method of moments for high-dimensional correlated data analysis. *Journal of the American Statistical Association*, pages 1–14, 2020. doi: 10.1080/01621459.2020.1736082.

Christopher C. Heyde. *Quasi-Likelihood and its Application: a General Approach to Optimal Parameter Estimation*. Springer Series in Statistics, 1997.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Peter J. Huber. *Robust Statistics*. Wiley Series in Probability and Statistics, 2nd edition, 2009.

Zi Jin. *Aspects of Composite Likelihood Inference*. PhD thesis, University of Toronto, 2011.

Harry Joe. *Dependence Modeling with Copulas*. Chapman & Hall, first edition, 2014.

Iain M. Johnstone and D. Michael Titterington. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4237–4253, 2009.

Sin-Ho Jung. Quasi-likelihood for median regression models. *Journal of the American Statistical Association*, 91(433):251–257, 1996.

Seyed Nima Khezr and Nima Jafari Navimipour. MapReduce and its applications, challenges and architecture: a comprehensive review and directions for future research. *Journal of Grid Computing*, 15(3):295–321, 2017.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Kung-Yee Liang, Scott L. Zeger, and Bahjat Qaqish. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, 54(1):3–40, 1992.

Dan-Yu Lin and Daniel Zeng. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):321–332, 2010.

Nan Lin and Ruibin Xi. Aggregated estimating equation estimation. *Statistics and its Interface*, 4(1):73–83, 2011.

Bruce G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:220–239, 1988.

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.

Xiaoming Lu and Zhaozhi Fan. Weighted quantile regression for longitudinal data. *Computational Statistics*, 30(2):569–592, 2015.

Lester Mackey, Ameet Talwalkar, and Michael I. Jordan. Divide-and-conquer matrix factorization. In *Advances in Neural Information Processing Systems 24*, pages 1134–1142, 2011.

Guido Masarotto and Cristiano Varin. Gaussian copula marginal regression. *Electronic Journal of Statistics*, 6:1517–1549, 2012.

Whitney K. Newey. Efficient semiparametric estimation via moment restrictions. *Econometrica*, 72(6):1877–1897, 2004.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

Jianxin Pan and Gilbert Mackenzie. On modelling mean-covariance structures in longitudinal studies. *Biometrika*, 90(1):239–244, 2003.

Magda Peligrad. Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables (a survey). In Ernst Eberlein and Murad S. Taqqu, editors, *Dependence in Probability and Statistics. Progress in Probability and Statistics*, volume 11. Birkhäuser, Boston, MA, 1986.

David Pollard. New ways to prove central limit theorems. *Econometric Theory*, 1(3): 295–313, 1985.

Piercesare Secchi. On the role of statistics in the era of big data: a call for a debate. *Statistics and probability letters*, 136:10–14, 2018.

Kesar Singh, Minge Xie, and William E. Strawderman. Combining information from independent sources through confidence distributions. *The Annals of Statistics*, 33(1): 159–183, 2005.

Peter X.-K. Song. *Correlated Data Analysis: Modeling, Analytics, and Applications.* Springer Series in Statistics, 2007.

Peter X.-K. Song, Mingyao Li, and Ying Yuan. Joint regression analysis of correlated data using gaussian copulas. *Biometrics*, 65(1):60–68, 2009.

Anestis Touloumis. Simulating correlated binary and multinomial responses under marginal model specification: The simcormultres package. *The R Journal*, 8(2):79–91, 2016.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

You-Gan Wang, Xu Lin, and Min Zhu. Robust estimating functions and bias correction for longitudinal data analysis. *Biometrics*, 61(3):684–691, 2005.

Robert W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3):439–447, 1974.

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

Chi-Chuan Yang, Yi-Hau Chen, and Hsing-Yi Chang. Joint regression analysis of marginal quantile and quantile association: application to longitudinal body mass index in adolescents. *Journal of the Royal Statistical Society, Series C*, 66(5):1075–1090, 2017.

Weiping Zhang, Chenlei Leng, and Cheng Yong Tang. A joint modelling approach for longitudinal studies. *Journal of the Royal Statistical Society, Series B*, 77(1):219–238, 2015a.

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*, 16:3299–3340, 2015b.

Lue Ping Zhao and Ross L. Prentice. Correlated binary regression using a quadratic exponential model. *Biometrika*, 77(3):642–648, 1990.