Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

A Distributed and Integrated Method of Moments for High-Dimensional Correlated Data Analysis

Emily C. Hector & Peter X.-K. Song

To cite this article: Emily C. Hector & Peter X.-K. Song (2020): A Distributed and Integrated Method of Moments for High-Dimensional Correlated Data Analysis, Journal of the American Statistical Association, DOI: 10.1080/01621459.2020.1736082

To link to this article: https://doi.org/10.1080/01621459.2020.1736082

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2020.1736082

A Distributed and Integrated Method of Moments for High-Dimensional Correlated Data Analysis

Emily C. Hector and Peter X.-K. Song

Department of Biostatistics, University of Michigan, Ann Arbor, MI

ABSTRACT

This article is motivated by a regression analysis of electroencephalography (EEG) neuroimaging data with high-dimensional correlated responses with multilevel nested correlations. We develop a divide-and-conquer procedure implemented in a fully distributed and parallelized computational scheme for statistical estimation and inference of regression parameters. Despite significant efforts in the literature, the computational bottleneck associated with high-dimensional likelihoods prevents the scalability of existing methods. The proposed method addresses this challenge by dividing responses into subvectors to be analyzed separately and in parallel on a distributed platform using pairwise composite likelihood. Theoretical challenges related to combining results from dependent data are overcome in a statistically efficient way using a meta-estimator derived from Hansen's generalized method of moments. We provide a rigorous theoretical framework for efficient estimation, inference, and goodness-of-fit tests. We develop an R package for ease of implementation. We illustrate our method's performance with simulations and the analysis of the EEG data, and find that iron deficiency is significantly associated with two auditory recognition memory related potentials in the left parietal-occipital region of the brain. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2018
Accepted February 2020

KEYWORDS
Composite likelihood; Divide-and-conquer; Generalized method of moments; Parallel computing; Scalable computing.

1. Introduction

This article focuses on developing a systematic divide-and-conquer procedure, readily implemented in a parallel and scalable computational scheme, for statistical estimation and inference. We consider a regression setting with high-dimensional correlated responses with multilevel nested correlations. The proposed distributed and integrated method of moments (DIMM) is flexible, fast, and statistically efficient, and reduces computing time in two ways: (i) in the distributed step, composite likelihood is executed in parallel at a number of distributed computing nodes, and (ii) at the integrated step, an efficient one-step meta-estimator is derived from Hansen's (1982) seminal generalized method of moments (GMM) with no need to load the entire data on a common server.

Let $\boldsymbol{Y}_i$ be the M-dimensional correlated response for subject i, $i = 1, \ldots, N$, and $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\beta})$ the mean response-covariate relationship of interest for some $M \times p$ dimensional matrix of covariates $\boldsymbol{X}_i$ and a p-dimensional parameter of interest $\boldsymbol{\beta}$. In this article, we consider the case where the dimension M of $\boldsymbol{Y}_i$ may diverge to infinity, while the dimension p of $\boldsymbol{\beta}$ is fixed. For convenience this is referred to as high-dimensional correlated response or, in short, high-dimensional response. We model $\boldsymbol{\mu}_i$ by a generalized linear model of the form $h(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$, where h is a known link function. The difficulties associated with current methods for high-dimensional correlated response modeling stem from computational burdens and modeling challenges associated with a high-dimensional likelihood. The generalized estimating equation (GEE) proposed by Liang and Zeger (1986), one of the widely used methods for the analysis of correlated response data, uses a quasilikelihood approach based on the first two moments of the response to avoid the specification of a parametric joint distribution. GEE is not well suited to high-dimensionality due to the potentially large number of nuisance parameters to estimate and the inversion of large matrices (see Banerjee et al. 2008; Cressie and Johannesson 2008). Additionally, common assumptions by GEE on the correlation structure of the response are too simple to capture multilevel nested correlations, resulting in a substantial loss of efficiency (see Fitzmaurice, Laird, and Rotnitzky 1993). Simple cases where the estimator of the nuisance parameter does not exist are also outlined in Crowder (1995). Mixed effects models are also popular in the literature to analyze correlated outcomes, and in the linear mixed-effects model regression parameters may be interpreted as population-average effects, similar to the interpretation given by the GEE approach. In the nonlinear case, the interpretation of the population-average effects is obstructed by the random effects. Unfortunately, mixed effects model estimation can be computationally expensive due to the inversion of large matrices and non-convexity of the objective function (Laird, Lange, and Stram 1987; Lindstrom and Bates 1988; Perry 2017). Additionally, when the correlation of the response is complex, computation may become prohibitive due to the large number of random effects required to estimate mean

CONTACT Emily C. Hector ehector@umich.edu Department of Biostatistics, University of Michigan School of Public Health, 1415 Washington Heights, Ann Arbor, MI 48109; Peter X.-K. Song pxsong@umich.edu Department of Biostatistics, University of Michigan School of Public Health, 1415 Washington Heights, Ann Arbor, MI 48109.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

parameters efficiently. The computational burden can increase significantly due to the evaluation of high-dimensional integrals with respect to the distributions of random effects in nonlinear models (Song 2007, chap. 4).

Composite likelihood (CL) was proposed by Lindsay (1988) as a method to perform inference on $\boldsymbol{\beta}$ by only considering low dimensional marginals of the joint distribution. Pairwise CL, in particular, constructs a pseudolikelihood by multiplying the likelihood objects of pairs of observations. In this way, CL is free of the computational burden of inverting high-dimensional correlation matrices and benefits from an objective function that facilitates model selection. Pairwise CL has been used with longitudinal (Kuk and Nott 2000; Kong, Wang, and Gray 2015), spatial (Heagerty and Lele 1998; Arbia 2014), spatiotemporal (Bai, Song, and Raghunathan 2012; Bevilacqua et al. 2012), and genetic (Larribe and Fearnhead 2011) data. A well-known bottleneck of CL is the high computational cost of evaluating a large number of low-dimensional likelihoods and their derivatives, a problem exacerbated by large M.

The use of CL relies on knowledge of low-dimensional dependencies among $\boldsymbol{Y}_i$ to specify pairwise CLs properly. Fortunately, in practice, observations within $\boldsymbol{Y}_i$ can often be partitioned into groups of sub-responses with simple correlation structures according to previous science: for example, genomic response data can be grouped by gene or genetic function, metabolomic data by pathway, spatial data by proximity, and brain imaging data by brain function regions. This substantive scientific knowledge may be used to strategically partition response variables to speed up computations. The method of divide-and-conquer is a state of the art approach to analyzing data that can be partitioned. In the current literature, this method proposes to randomly split subjects into independent groups of subjects in the "divide" step (or "Mapper") and combines results in the "conquer" step (or "Reducer"); see, for example, kernel ridge regression (Zhang, Duchi, and Wainwright 2015) and matrix factorization (Mackey, Talwalkar, and Jordan 2015). The independent groups can be analyzed in parallel, greatly reducing computation time. Chen and Xie (2014) and Battey et al. (2015) used this approach to analyze large datasets by combining information from independent sources. These methods are not well suited to our problem due to assumptions of independence. Chang et al. (2015) proposed a divide-and-conquer CL approach for high-dimensional spatial data, but their Bayesian hierarchical model relies on the Metropolis–Hastings algorithm for estimation, which is time-consuming. Indeed, their divide-and-conquer strategy is primarily adopted in model building rather than to reduce computational speed. Extending the divide-and-conquer approach to our problem, we propose to split the high-dimensional correlated response into subvectors to form correlated response groups according to substantive scientific knowledge. Each subvector is analyzed separately, then results from these analyses are combined. While this method is computationally appealing, our groups of data are correlated, leading to new methodological challenges. In particular, correlation between groups of data must be taken into account when combining results. To our knowledge, our method is among the few attempts, including Li (2017) and Chang et al. (2015), to establish a rigorous theoretical framework for combining results from correlated groups of data. The key technique to establish the related theoretical framework relies on an extended version of the confidence distribution (CD) based on pairwise CL to derive a GMM estimator of $\boldsymbol{\beta}$. For discussion on the CD and related work with independent cross-sectional data see Singh, Xie, and Strawderman 2005; Xie, Singh, and Strawderman 2011; Xie and Singh 2013; Liu, Liu, and Xie 2015; for CD approaches to meta-analysis of independent studies see Claggett, Xie, and Tian 2014; Yang et al. 2014; for a divide-and-conquer approach with independent scalar responses see Lin and Xi 2011. We invoke an optimal weighting matrix that nonparametrically accounts for between-group correlations to alleviate the computational and modeling challenges associated with existing methods.

We illustrate our method with a motivating cohort study to assess the association between iron deficiency and auditory recognition memory in infants. Electrical activity in the brain during a 2000 msec period was measured in 157 infants under two vocal stimuli using an electroencephalography (EEG) net consisting of 64-channel sensors on the scalp as visualized in Figure 1. For each sensor and each stimulus, three important event-related potentials (ERPs) related to auditory recognition memory were calculated; as shown in Figure 2, P2 averages electrical signal between 175 and 300 msec, P750 between 350 and 600 msec, and late slow wave (LSW) between 850 and 1100 msec. The investigator wanted to analyze the data in subregions, where 46 of the nodes belong to six brain function regions related to auditory recognition memory, as seen in Figure 1. The complex data-generating mechanism results in a response of dimension $M = 46(\text{nodes}) \times 3(\text{ERPs}) \times 2(\text{stimuli}) = 276$ that has a multilevel nested correlation structure that is difficult to model, including longitudinal correlations between the three ERP's, spatial correlations between the 46 nodes and within the six brain function regions, and correlations within each voice stimulus. Due to this complex correlation structure and the large number of response variables, traditional methods for correlated data analysis are greatly challenged. Zhou and Song (2016) developed a method to analyze the LSW outcome, but no existing method is suitable to analyze this dataset in its entirety. We develop DIMM, a fast and efficient method to analyze all 276 responses simultaneously by partitioning the response according to ERPs and brain function regions. DIMM also performs well with higher dimensional correlated outcomes, as seen in simulations.

Our proposed DIMM loses minimal estimation efficiency for two reasons: first, CL performs well on smaller groups of responses with simple but well-approximated local correlation structure; and second, we use an optimal weighting matrix in the GMM. More importantly, our method is computationally attractive for two reasons: first, pairwise CL only evaluates low-dimensional likelihoods and CL analyses can be run in parallel; and second, we provide a closed-form of the combined estimator that only depends on CL estimates and group-specific sufficient statistics. Finally, this article contributes to the existing literature with two key innovations: DIMM provides a rigorous theoretical framework for combining estimates from dependent groups of data, and is scalable to large M. In addition, the proposed DIMM is illustrated on a

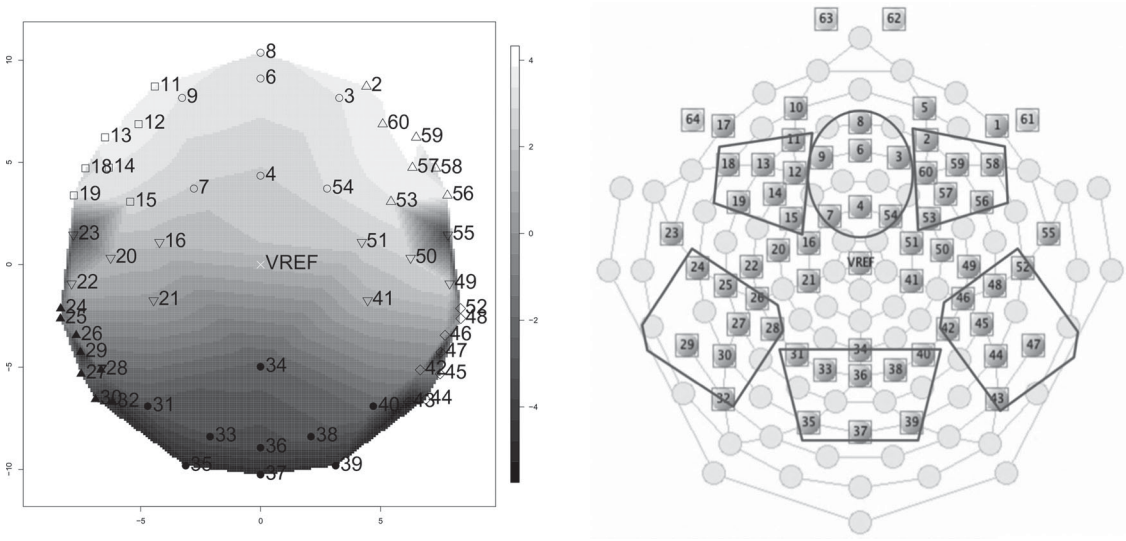

Figure 1. Left: Average P2 amplitude for iron sufficient children under stimulus of mother's voice. Color plot and additional plots in the supplementary materials. Right: Layout of the 64-channel sensor net with brain regions related to auditory recognition memory.

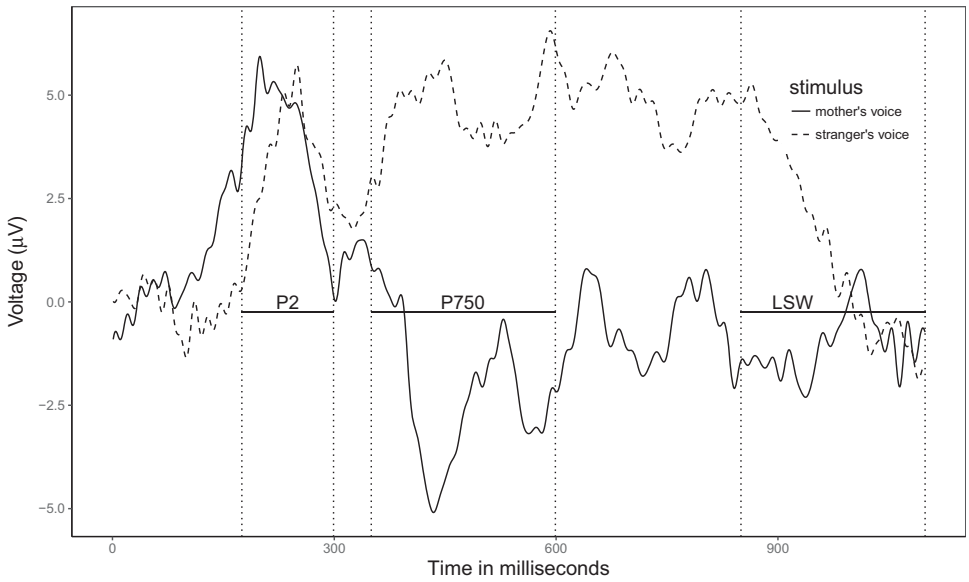

Figure 2. Plot of electrical potential for subject 1 at electrode 2 over time.

complex dataset that has previously not been analyzed in its entirety.

The rest of the article is organized as follows. Section 2 describes DIMM. Section 3 discusses large sample properties. Section 4 presents the closed form one-step meta-estimator, and its implementation in a parallel and scalable computational scheme. Section 5 illustrates DIMM's finite sample performance with simulations. Section 6 presents the EEG data analysis. Section 7 concludes with a discussion. Proofs of theorems and additional simulation and data analysis results are deferred to the Appendix and supplementary materials.

2. Formulation

Let $\{\boldsymbol{y}_i, \boldsymbol{X}_i\}_{i=1}^N$ be N independent observations, where the dimension M of $\boldsymbol{y}_i$ is so big and potentially diverging that a direct analysis of the data is computationally intensive or prohibitive. Let $f(\boldsymbol{Y}_i; \boldsymbol{\Gamma}_i, \boldsymbol{X}_i)$ be the M-variate joint distribution of $\boldsymbol{Y}_i|\boldsymbol{X}_i$, where $\boldsymbol{\Gamma}_i$ contains parameters of high-order dependencies that may be difficult to handle computationally. We aim to obtain a statistically efficient (small variance) and computationally fast estimator for the regression coefficient $\boldsymbol{\beta}$ given the challenges arising from the high-dimensionality and complex dependencies of the response. Our DIMM solution uses a divide-and-conquer approach based on pairwise CL methodology for locally homogeneous data blocks. We formulate an informal definition of homogeneous correlation: we say a vector of random variables is homogeneously correlated if their covariance (or second moments) can be parameterized with a small number of parameters. For example, responses with compound symmetric or AR(1) covariance structures are homogeneously correlated.

2.1. Division: Distributed Composite Likelihoods

For each $i \in \{1, \ldots, N\}$, we propose to split the M-dimensional response $\boldsymbol{y}_i$ and associated covariates into J blocks $\left\{\boldsymbol{y}_{i,j}, \boldsymbol{X}_{i,j}\right\}_{i=1}^{N}$ for $j = 1, \ldots, J$, J finite, as follows: $\boldsymbol{y}_i = (\ \boldsymbol{y}_{i,1}^T \ \ \ldots \ \ \boldsymbol{y}_{i,J}^T\ )^T$ and $\boldsymbol{X}_i = (\ \boldsymbol{X}_{i,1}^T \ \ \ldots \ \ \boldsymbol{X}_{i,J}^T\ )^T$. Within block j, let m_j be the dimension of subject i's response, $\sum_{j=1}^{J} m_j = M$, where $\boldsymbol{y}_{i,j} = \left(y_{i1,j}, \ldots, y_{im_j,j}\right)^T \in \mathbb{R}^{m_j}$ is subject i's jth sub-response vector and $\boldsymbol{X}_{i,j} \in \mathbb{R}^{m_j \times p}$ is the associated covariate matrix, and p is finite. For each j, $\left\{\boldsymbol{y}_{i,j}\right\}_{i=1}^{N}$ are independent realizations of the random variables $\boldsymbol{Y}_{i,j}|\boldsymbol{X}_{i,j}$ whose m_j-variate distributions conditional on $\boldsymbol{X}_{i,j}$ are denoted by $f(\boldsymbol{y}_{i,j}; \boldsymbol{\Gamma}_{i,j}, \boldsymbol{X}_{i,j})$. Parameter $\boldsymbol{\Gamma}_{i,j}$ encodes information on the marginal moments of $\boldsymbol{Y}_{i,j}$. This yields J regression models $h_j(\boldsymbol{\mu}_{i,j}) = \boldsymbol{X}_{i,j}\boldsymbol{\beta}_j$, where $\boldsymbol{\mu}_{i,j} = E(\boldsymbol{Y}_{i,j}|\boldsymbol{X}_{i,j}, \boldsymbol{\beta}_j)$ is the marginal mean of $\boldsymbol{Y}_{i,j}$, $j = 1, \ldots, J$. For simplification of the technical presentation, we assume homogeneity of the link function h_j and the regression parameter $\boldsymbol{\beta}_j$ holds such that $h_j \equiv h$ and $\boldsymbol{\beta}_j \equiv \boldsymbol{\beta}$ for $j = 1, \ldots, J$; we drop the subscript j by using $\boldsymbol{\beta}$ and h to denote $\boldsymbol{\beta}_j$ and h_j. On some occasions, homogeneity may not hold, for example when each sub-response $\boldsymbol{Y}_{i,j}$ corresponds to continuous, count, or dichotomous outcomes. In this case, we propose to perform a subgroup analysis by combining regression parameter estimates over the blocks where homogeneity in h_j and $\boldsymbol{\beta}_j$ holds; this approach will be illustrated in Section 6. Additionally, we propose a formal test of the homogeneity assumption in Section 3. To create blocks, we suggest splitting the response data according to substantive scientific knowledge, resulting in homogeneous correlations within each response subvector that are suitable for simplifications in structure. If such knowledge is lacking, data preprocessing may help to learn structural features of dependencies. As long as appropriate conditions are satisfied, estimation remains consistent, but may not be efficient, when the data split is not aligned with the true dependence structure.

We can obtain an estimate of $\boldsymbol{\beta}$ for each of the J blocks of data using pairwise CL methods. The above partition enables us to reduce the challenge of modeling M-order dependencies to that of modeling m_j-order dependencies of (approximately) local homogeneity. In addition, there may be tremendous computational burdens associated with the log-likelihood or its derivative, such as the computation of a high-dimensional inverse covariance matrix in the multivariate normal model. CL has been suggested by many researchers (see Varin, Reid, and Firth 2011 and references therein) to resolve this difficulty, and takes the following form:

$$\mathcal{L}_j(\boldsymbol{\beta}, \boldsymbol{\gamma}_j; \boldsymbol{y}_{i,j}) = \prod_{r=1}^{m_j-1} \prod_{t=r+1}^{m_j} f_j(y_{ir,j}, y_{it,j}; \boldsymbol{\beta}, \boldsymbol{\gamma}_j, \boldsymbol{X}_{i,j}), \qquad (1)$$

where $\boldsymbol{\gamma}_j$ only contains information on second-order moments of $\boldsymbol{Y}_{i,j}$. Let $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_{j0}$ the true values of $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\gamma}_j \in \mathbb{R}^{d_j}$, respectively, d_j finite, and denote $\boldsymbol{\gamma} = (\ \boldsymbol{\gamma}_1^T \ \ \ldots \ \ \boldsymbol{\gamma}_J^T\ )^T$, $\boldsymbol{\gamma}_0 = (\ \boldsymbol{\gamma}_{10}^T \ \ \ldots \ \ \boldsymbol{\gamma}_{J0}^T\ )^T$. The nature of the data partition gives rise to different dependence parameters $\boldsymbol{\gamma}_j$, allowing us to make simplifying assumptions on the high-order dependencies of $\boldsymbol{Y}_{i,j}$. Here, density f_j can be chosen according to the data type under investigation as bivariate margins of an m_j-variate joint distribution. For example, f_j can be bivariate Normal for continuous data, or, using bivariate dispersion models generated by Gaussian or vine copulas, can be bivariate Poisson or Bernoulli for count or dichotomous data; (see Song 2007, chap. 6; Joe 2014, chap. 3). We set f_j bivariate Normal for the EEG data. Within block j, the log-CL for the first and second moment parameters is

$$\begin{aligned} c\ell_j(\boldsymbol{\beta}, \boldsymbol{\gamma}_j; \boldsymbol{y}_j) &= \log \prod_{i=1}^{N} \mathcal{L}_j(\boldsymbol{\beta}, \boldsymbol{\gamma}_j; \boldsymbol{y}_{i,j}) \\ &= \sum_{i=1}^{N} \sum_{r=1}^{m_j-1} \sum_{t=r+1}^{m_j} \log f_j(y_{ir,j}, y_{it,j}; \boldsymbol{\beta}, \boldsymbol{\gamma}_j, \boldsymbol{X}_{i,j}). \end{aligned}$$

Define $\boldsymbol{\psi}_{j.\text{sub}}(\boldsymbol{\beta}; \boldsymbol{y}_{i,j}, \boldsymbol{\gamma}_j) = (1/m_j^2) \sum_{r=1}^{m_j-1} \sum_{t=r+1}^{m_j} \nabla_{\boldsymbol{\beta}} \log f_j(y_{ir,j}; y_{it,j}; \boldsymbol{\beta}, \boldsymbol{\gamma}_j, \boldsymbol{X}_{i,j}) \in \mathbb{R}^p$ and $\boldsymbol{g}_{j.\text{sub}}(\boldsymbol{\gamma}_j; \boldsymbol{y}_{i,j}, \boldsymbol{\beta}) = (1/m_j^2) \sum_{r=1}^{m_j-1} \sum_{t=r+1}^{m_j} \nabla_{\boldsymbol{\gamma}_j} \log f_j(y_{ir,j}; y_{it,j}; \boldsymbol{\beta}, \boldsymbol{\gamma}_j, \boldsymbol{X}_{i,j}) \in \mathbb{R}^{d_j}$. The pairwise CL estimating equations for the mean and covariance parameters are, respectively:

$$\boldsymbol{\Psi}_{j.\text{sub}}(\boldsymbol{\beta}; \boldsymbol{y}_j, \boldsymbol{\gamma}_j) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\psi}_{j.\text{sub}}(\boldsymbol{\beta}; \boldsymbol{y}_{i,j}, \boldsymbol{\gamma}_j) = \boldsymbol{0} \in \mathbb{R}^p, \qquad (2)$$

$$\boldsymbol{G}_{j.\text{sub}}(\boldsymbol{\gamma}_j; \boldsymbol{y}_j, \boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{g}_{j.\text{sub}}(\boldsymbol{\gamma}_j; \boldsymbol{y}_{i,j}, \boldsymbol{\beta}) = \boldsymbol{0} \in \mathbb{R}^{d_j}. \qquad (3)$$

Following Varin, Reid, and Firth (2011), the maximum composite likelihood estimators (MCLE) of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_j$ within block j, denoted, respectively, by $\widehat{\boldsymbol{\beta}}_j$ and $\widehat{\boldsymbol{\gamma}}_j$, are the joint solution to the system of unbiased estimating equations in (2) and (3). It is worth noting that the original CL proposed by Lindsay (1988) advocated for the use of weights in the log-CL function to improve estimation efficiency. This approach is shown to work well in Bevilacqua et al. (2012). Lindsay (1988) determined that the optimal weights that minimize the variance of the maximum composite likelihood estimator depend on higher order moments of the estimating function, and therefore can be demanding to compute. Again, we see the trade-off between computational and statistical efficiency.

Generally, $\boldsymbol{\gamma}_j$ is block-specific and unknown, and $\widehat{\boldsymbol{\beta}}_j$ depends on $\widehat{\boldsymbol{\gamma}}_j$. When $\boldsymbol{\gamma}_j$ is a function of $\boldsymbol{\beta}$ only, as in generalized linear models, finding $\widehat{\boldsymbol{\beta}}_j$ amounts to profile likelihood estimation. If $\boldsymbol{\gamma}_j$ is known or absent, then the above simplifies to finding $\widehat{\boldsymbol{\beta}}_j$ as the solution to $\boldsymbol{\Psi}_{j.\text{sub}}(\boldsymbol{\beta}; \boldsymbol{y}_j, \boldsymbol{\gamma}_j) = \boldsymbol{0}$. We denote $\widehat{\boldsymbol{\beta}}_{\text{MCLE}} = (\widehat{\boldsymbol{\beta}}_1^T, \ldots, \widehat{\boldsymbol{\beta}}_J^T)^T$ and $\widehat{\boldsymbol{\gamma}}_{\text{MCLE}} = (\widehat{\boldsymbol{\gamma}}_1^T, \ldots, \widehat{\boldsymbol{\gamma}}_J^T)^T$. In some practical studies where interest is in block-specific mean parameters and combined dependence parameters, we can treat $\boldsymbol{\beta}$ as a nuisance parameter and $\boldsymbol{\gamma}_j$ as the parameter of interest by switching the roles of $\boldsymbol{\Psi}_{j.\text{sub}}$ and $\boldsymbol{G}_{j.\text{sub}}$. In the case where both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_j$ are of interest and believed to be homogeneous over all blocks, we replace $\boldsymbol{\Psi}_{j.\text{sub}}$ with $(\boldsymbol{\Psi}_{j.\text{sub}}^T, \boldsymbol{G}_{j.\text{sub}}^T)^T$. The description of DIMM in the rest of the article, including Section 4, holds with these minor changes.

2.2. Integration: The Generalized Method of Moments

Suppose that we have successfully obtained J estimates of $\boldsymbol{\beta}$ based on J estimating equations (2). In the integration step, we treat each estimating equation $\boldsymbol{\Psi}_{j.\text{sub}}(\boldsymbol{\beta}; \boldsymbol{y}_j, \boldsymbol{\gamma}_j) = 0$ as a moment condition on $\boldsymbol{\beta}$ coming from block j, $j = 1, \ldots, J$. We would like to derive an estimator $\widehat{\boldsymbol{\beta}}_c$ of $\boldsymbol{\beta}$ that satisfies all J moment conditions. Unfortunately, there is no unique solution to all J estimating equations because they over-identify our parameter; that is, the dimension of parameter $\boldsymbol{\beta}$ is less than Jp, the dimension of the equation restrictions on $\boldsymbol{\beta}$. To overcome this, we invoke Hansen's (1982) seminal GMM to combine the moment conditions that arise from each block. Stack the J estimating equations by letting $\boldsymbol{\psi}(\boldsymbol{\beta}; \boldsymbol{y}_i) = (\boldsymbol{\psi}^T_{1.\text{sub}}(\boldsymbol{\beta}; \boldsymbol{y}_{i,1}, \boldsymbol{\gamma}_{10}), \ldots, \boldsymbol{\psi}^T_{J.\text{sub}}(\boldsymbol{\beta}; \boldsymbol{y}_{i,J}, \boldsymbol{\gamma}_{J0}))^T \in \mathbb{R}^{Jp}$ for each subject i, and

$$\begin{aligned}\boldsymbol{\Psi}_N(\boldsymbol{\beta}; \boldsymbol{y}) &= \left(\boldsymbol{\Psi}^T_{1.\text{sub}}(\boldsymbol{\beta}; \boldsymbol{y}_1, \boldsymbol{\gamma}_{10}) \quad \ldots \quad \boldsymbol{\Psi}^T_{J.\text{sub}}(\boldsymbol{\beta}; \boldsymbol{y}_J, \boldsymbol{\gamma}_{J0}) \right)^T \\ &= \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\psi}(\boldsymbol{\beta}; \boldsymbol{y}_i) \in \mathbb{R}^{Jp}.\end{aligned}$$

Define the outer-product as $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$ for $\boldsymbol{a} \in \mathbb{R}^{Jp}$. Since $\boldsymbol{\Psi}_N(\boldsymbol{\beta}; \boldsymbol{y}) = \boldsymbol{0}$ has no unique solution, following Hansen's GMM we minimize a quadratic form of $\boldsymbol{\Psi}_N$ with weight matrix $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$, the $Jp \times Jp$ sample variance-covariance matrix of $\boldsymbol{\Psi}_N(\boldsymbol{\beta}; \boldsymbol{y})$ evaluated at the MCLE's:

$$\begin{aligned}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}} = \frac{1}{N} \sum_{i=1}^{N} \big(&\boldsymbol{\psi}^T_{1.\text{sub}}(\widehat{\boldsymbol{\beta}}_1; \boldsymbol{y}_{i,1}, \widehat{\boldsymbol{\gamma}}_1), \ldots, \\ &\boldsymbol{\psi}^T_{J.\text{sub}}(\widehat{\boldsymbol{\beta}}_J; \boldsymbol{y}_{i,J}, \widehat{\boldsymbol{\gamma}}_J) \big)^{T\otimes 2}. \end{aligned} \tag{4}$$

Then define the combined GMM estimator of $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}}_c = \arg\min_{\boldsymbol{\beta}} \left\{ N \boldsymbol{\Psi}^T_N(\boldsymbol{\beta}; \boldsymbol{y}) \widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}} \boldsymbol{\Psi}_N(\boldsymbol{\beta}; \boldsymbol{y}) \right\} = \arg\min_{\boldsymbol{\beta}} Q_N(\boldsymbol{\beta}). \tag{5}$$

To solve (5), we replace $\boldsymbol{\gamma}_{j0}$ by $\widehat{\boldsymbol{\gamma}}_j$ in the evaluation of $\boldsymbol{\Psi}_N(\boldsymbol{\beta}; \boldsymbol{y})$. The role of the $\boldsymbol{\gamma}_j$'s is two-fold: first, their specification parametrizes the second-order moment in the block bivariate distributions in addition to the regression model for first moments; second, they may improve estimation efficiency of $\boldsymbol{\beta}$. Note that using plug-in estimators $\widehat{\boldsymbol{\gamma}}_j$ may impact efficiency of $\widehat{\boldsymbol{\beta}}_c$, but it will generally not affect consistency. A finite sample improvement on the efficiency may be obtained by re-estimating $\boldsymbol{\gamma}_j$ in the integration step, but this could become computationally intensive since these parameters are block-specific and heterogeneous. We notice similarities of (5) to Qu, Lindsay, and Li (2000) but with a completely different way of constructing moment conditions, and to Wang, Wang, and Song (2012) but with a completely different way of partitioning data and the added generality of allowing between-block correlations. The uniqueness of DIMM stems from combining estimating equations $\boldsymbol{\Psi}_{j.\text{sub}}$ with GMM instead of combining $\widehat{\boldsymbol{\beta}}_j$ or data blocks $\left\{\boldsymbol{y}_{i,j}, \boldsymbol{X}_{i,j}\right\}_{i=1}^{N}$ directly. This new approach allows us to find a GMM estimator $\widehat{\boldsymbol{\beta}}_c$ that benefits from a wealth of established theoretical properties. The sample covariance $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ is not parameter dependent and can therefore accommodate any between-block covariance, including unstructured. By using the sample covariance $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ we not only account for between-block correlations but find the optimal GMM estimator in the sense that $\widehat{\boldsymbol{\beta}}_c$ has variance at least as small as any other estimator exploiting the same moment conditions, hereafter referred to as "Hansen optimal." The combined GMM estimator $\widehat{\boldsymbol{\beta}}_c$ will yield significant computational advantages when the dimension of $\boldsymbol{\Psi}_N$ is smaller than that of $\boldsymbol{Y}$ by reducing the computational burden associated with handling $\boldsymbol{Y}$ directly. This is often the case in applications where M is very large, J is between M and p, and the number of covariates p is small enough that $p \ll M/J$.

To better understand our estimator, we can show that $\widehat{\boldsymbol{\beta}}_c$ maximizes a density in a manner similar to the classic maximum likelihood estimator (MLE) by deriving the quadratic form in (5) using an extended version of the confidence distribution (CD) (or density) (Fisher 1930). For more discussion on CD and applications to MLE with independent cross-sectional data, refer to Xie and Singh (2013), Singh, Xie, and Strawderman (2005), and Liu, Liu, and Xie (2015). So far, little work has been done on the development of CD for correlated data. Of note, a dissertation by Li (2017) considered a sequential split-and-conquer copula approach to extend the CD to combine information from correlated datasets. The proposed copula method assumes a known joint distribution or a known correlation matrix, which is mostly for theoretical convenience, and takes advantage of the structure of the spatial Gaussian process model to sequentially transform the dependent datasets into independent datasets. Li (2017) considered the case $N = 1$ and $M \to \infty$ for applications in spatial data modeling. Additional work on deriving a consistent estimator of the correlation matrix is required to make this method practically useful. $\boldsymbol{\Psi}_{j.\text{sub}}$ are sufficient statistics for $\boldsymbol{\beta}$ within each block and are asymptotically Normally distributed under mild assumptions by the central limit theorem (CLT). Their joint distribution is the distribution of $\boldsymbol{\Psi}_N$, which is also asymptotically Normal under the same mild assumptions of the CLT. Then if $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ is a consistent estimator of the variance of $\boldsymbol{\Psi}_N$, $\sqrt{N}\widehat{\boldsymbol{V}}^{-1/2}_{N,\boldsymbol{\psi}} \boldsymbol{\Psi}_N(\boldsymbol{\beta}_0; \boldsymbol{y})$ asymptotically follows a standard normal distribution. By maximizing the distribution of $\boldsymbol{\Psi}_N$ as a function of $\boldsymbol{\beta}$, we can find an estimator that accounts for correlation between sufficient statistics and is the most likely value to arise from the data. We define the confidence estimating function (CEF) as $F_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0) = \Phi\left(\sqrt{N}\widehat{\boldsymbol{V}}^{-1/2}_{N,\boldsymbol{\psi}} \boldsymbol{\Psi}_N(\boldsymbol{\beta}_0; \boldsymbol{y})\right)$, where $\Phi(\cdot)$ is the Jp-variate standard normal distribution function. Define the density of the CEF as

$$\begin{aligned} f_{\boldsymbol{\psi}}(\boldsymbol{\beta}) &= \phi\left(\sqrt{N}\widehat{\boldsymbol{V}}^{-1/2}_{N,\boldsymbol{\psi}} \boldsymbol{\Psi}_N(\boldsymbol{\beta}; \boldsymbol{y})\right) \\ &\propto \exp\left\{ -\frac{N}{2} \boldsymbol{\Psi}^T_N(\boldsymbol{\beta}; \boldsymbol{y}) \widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}} \boldsymbol{\Psi}_N(\boldsymbol{\beta}; \boldsymbol{y}) \right\}, \end{aligned} \tag{6}$$

where $\phi(\cdot)$ is the Jp-variate standard normal density. The CEF density has the advantage over the confidence density of not having a sandwich estimator for the variance, and thus not requiring the computation of a sensitivity matrix. It reflects the joint distribution of the J estimating equations (2). Maximizing $f_{\boldsymbol{\psi}}(\boldsymbol{\beta})$ in (6) yields the minimization defined in (5). The formulation in (6) is different from the aggregated estimating equation

approach proposed by Lin and Xi (2011) for independent scalar responses.

3. Asymptotic Properties

In this section, we study the asymptotic properties of $\widehat{\boldsymbol{\beta}}_c$ with J and p fixed, where we allow M to diverge, implying that m_j diverges for at least one sub-response dimension m_j. Due to the use of a simple correlation structure in each block, the dimension d_j of $\boldsymbol{\gamma}_j$ is fixed. It follows from (2) and (3) that $\boldsymbol{\Psi}_{j.\text{sub}}$ and $\boldsymbol{G}_{j.\text{sub}}$ are expressed as sums of two-dimensional marginal likelihoods as $m_j \to \infty$. Following Cox and Reid's (2004) study of the behavior of the CL when the dimension of the outcome grows with the sample size, we can similarly show the consistency of $(\widehat{\boldsymbol{\beta}}_j, \widehat{\boldsymbol{\gamma}}_j)$ with no conditions required on the divergence rate of M. This is formalized in the following proposition.

Proposition 1. Let $j \in \{1, \ldots, J\}$ such that $m_j \to \infty$. Suppose $\boldsymbol{\Psi}_{j.\text{sub}}$ and $\boldsymbol{G}_{j.\text{sub}}$ are unbiased at $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_{j0})$ and their expectations have a unique zero at $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_{j0})$. Then $(\widehat{\boldsymbol{\beta}}_j, \widehat{\boldsymbol{\gamma}}_j)$ are consistent estimators of $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_{j0})$ as $N \to \infty$.

The proof is given in the supplementary materials. Proposition 1 justifies why standard asymptotic theory is applicable even when $M \to \infty$. $\boldsymbol{\Psi}_{j.\text{sub}}$ and $\boldsymbol{G}_{j.\text{sub}}$ are unbiased if the bivariate marginals f_j are correctly specified. Existing model diagnostics can help detect ill-posed model structures on the f_j.

Let $\boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\beta}) = \lim_{M\to\infty} E_{\boldsymbol{\beta}}\left\{\boldsymbol{\psi}(\boldsymbol{\beta}; \boldsymbol{y}_i)\boldsymbol{\psi}^T(\boldsymbol{\beta}; \boldsymbol{y}_i)\right\} \in \mathbb{R}^{Jp\times Jp}$ and $\boldsymbol{s}_{\boldsymbol{\psi}}(\boldsymbol{\beta}) = \lim_{M\to\infty} -\nabla_{\boldsymbol{\beta}} E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\boldsymbol{\beta}; \boldsymbol{y}_i) \in \mathbb{R}^{Jp\times p}$ be, respectively, the positive definite variability matrix and the sensitivity matrix of $\boldsymbol{\Psi}_N$. Let $\left[\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta})\right]_{i,j}$ be the rows $(i-1)p+1$ to ip and columns $(j-1)p+1$ to jp of matrix $\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta})$. We assume throughout that $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ is nonsingular. Let $\|\cdot\|$ be the Euclidean norm. Let the variability and sensitivity matrices in block j, respectively, be

$$\begin{aligned}
\boldsymbol{v}_{j,\boldsymbol{\psi}_j}(\boldsymbol{\beta}) &= \lim_{M\to\infty} \text{var}_{\boldsymbol{\beta}}\left\{\sqrt{N}\boldsymbol{\Psi}_{j.\text{sub}}(\boldsymbol{\beta}; \boldsymbol{y}_j, \boldsymbol{\gamma}_{j0})\right\} \\
&= \lim_{M\to\infty} E_{\boldsymbol{\beta}}\left\{\boldsymbol{\psi}_{j.\text{sub}}^{\otimes 2}(\boldsymbol{\beta}; \boldsymbol{y}_{i,j}, \boldsymbol{\gamma}_{j0})\right\}, \\
\boldsymbol{s}_{j,\boldsymbol{\psi}_j}(\boldsymbol{\beta}) &= \lim_{M\to\infty} -\nabla_{\boldsymbol{\beta}} E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_{j.\text{sub}}(\boldsymbol{\beta}; \boldsymbol{y}_j, \boldsymbol{\gamma}_{j0})\right\} \\
&= \lim_{M\to\infty} -\nabla_{\boldsymbol{\beta}} E_{\boldsymbol{\beta}}\left\{\boldsymbol{\psi}_{j.\text{sub}}(\boldsymbol{\beta}; \boldsymbol{y}_{i,j}, \boldsymbol{\gamma}_{j0})\right\}.
\end{aligned}$$

As a GMM estimator, $\widehat{\boldsymbol{\beta}}_c$ enjoys several key asymptotic properties for valid statistical inference under mild regularity conditions C.1–C.3 listed in the Appendix, including consistency and asymptotic normality. We show in Lemma 1 that $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ in (4) converges to the true variability matrix of the estimating equations.

Lemma 1 (Hansen optimality). Under condition C.1, $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)$ as $N \to \infty$.

The proof of Lemma 1, given in the Appendix, is straightforward, and makes use of the consistency of the MCLE's and the CLT. Lemma 1 shows our GMM estimator is Hansen optimal because we use a weighting matrix that converges to the true variance of the estimating equations. Asymptotically, $\widehat{\boldsymbol{\beta}}_c$ has variance at least as small as any other estimator exploiting the same CL moment conditions. Since the pairwise CL is not a full likelihood, there are no general efficiency results about $\widehat{\boldsymbol{\beta}}_j$. In the linear setting with normally distributed responses, the mean and variance fully specify the joint distribution within each block, and therefore, if the first two moments are correctly specified, the MCLE loses minimal estimation efficiency. The MCLE in the nonlinear setting will inevitably lose some efficiency because higher order moments are not modeled. Extensive simulations were performed in the dissertation of Jin (2011) for linear and binary correlated data that show that the CL approach performs quite well, and generally shows little loss of efficiency in comparison to the full likelihood approach in the cases of compound symmetry, AR(1), and unstructured correlation structures. This means DIMM generally performs well. In Theorems 1 and 2, we show that $\widehat{\boldsymbol{\beta}}_c$ is consistent and asymptotically normal under mild moment conditions.

Theorem 1 (Consistency of $\widehat{\boldsymbol{\beta}}_c$). Given conditions C.1 and C.2, $\widehat{\boldsymbol{\beta}}_c \xrightarrow{p} \boldsymbol{\beta}_0$ as $N \to \infty$.

Theorem 2 (Asymptotic normality of $\widehat{\boldsymbol{\beta}}_c$). Given conditions C.1–C.3, $\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{j}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)\right)$ as $N \to \infty$, where the Godambe information of $\boldsymbol{\Psi}_N(\boldsymbol{\beta}; \boldsymbol{y})$ can be rewritten as $\boldsymbol{j}_{\boldsymbol{\psi}}(\boldsymbol{\beta}) = \boldsymbol{s}_{\boldsymbol{\psi}}^T(\boldsymbol{\beta})\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta})\boldsymbol{s}_{\boldsymbol{\psi}}(\boldsymbol{\beta}) = \sum_{i,j=1}^{J} \boldsymbol{s}_{i,\boldsymbol{\psi}_i}^T(\boldsymbol{\beta})\left[\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta})\right]_{i,j}\boldsymbol{s}_{j,\boldsymbol{\psi}_j}(\boldsymbol{\beta})$.

The proof of Theorem 1, given in the Appendix, derives from the consistency of the GMM estimator due to Hansen (1982) and, more generally, to Newey and McFadden (1994). The proof of Theorem 2 follows from Theorem 7.2 in Newey and McFadden (1994) and Theorem 1. Theorems 1 and 2 do not require the differentiability of $\boldsymbol{\Psi}_{j.\text{sub}}$ and Q_N. Instead, they require the differentiability of their population versions, and that $\boldsymbol{\Psi}_N$ behave "nicely" in a neighborhood of $\boldsymbol{\beta}_0$. These theoretical results provide a framework for constructing asymptotic confidence intervals and conducting Wald tests, so that we can perform inference for $\boldsymbol{\beta}$ when M diverges. Using an optimal weight matrix improves statistical power so DIMM can detect signals other methods may miss.

So far, we have been vague about how the data partition should be done, only suggesting it be done according to established scientific knowledge. There may be some uncertainty about how to partition data, which we discuss in Section 7. A formal approach to testing if the data split was done appropriately can be interpreted as a test of the over-identifying restrictions: if the blocks are improperly specified (in number, size, etc.), the estimating equation $\boldsymbol{\Psi}_N$ will have mismatched moment restrictions on $\boldsymbol{\beta}$. Formally, we can show that Q_N evaluated at $\widehat{\boldsymbol{\beta}}_c$ follows a chi-squared distribution with $(J-1)p$ degrees of freedom.

Theorem 3 (Test of over-identifying restrictions). Let $\widehat{\boldsymbol{\beta}}_c = \arg\min_{\boldsymbol{\beta}} Q_N(\boldsymbol{\beta})$. Given conditions C.1–C.3, $Q_N(\widehat{\boldsymbol{\beta}}_c) \xrightarrow{d} \chi^2_{(J-1)p}$ as $N \to \infty$.

The proof is given in the supplementary materials. Since the test statistic depends on $\widehat{\boldsymbol{\beta}}_c$, it should be performed after estimation of the model parameters to determine goodness of fit. It can be computed in a distributed fashion by computing $\boldsymbol{\psi}_{j.\text{sub}}(\widehat{\boldsymbol{\beta}}_c; \boldsymbol{y}_{i,j}, \widehat{\boldsymbol{\gamma}}_j)$ in parallel and plugging into the formula for Q_N. DIMM has the advantage of an objective function that allows for formal testing, whereas GEE model selection relies on information criteria that can be subjective. The test can also be thought of as a test of the homogeneity assumption on the mean parameter $\boldsymbol{\beta}$, since the model $h(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ translates into moment restrictions on $\boldsymbol{\beta}$. 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.

4. Implementation: The One-Step Estimator

In practice, searching for the integrated solution of the GMM equation (5) can be very slow or computationally prohibitive. Iterative methods must repeatedly evaluate $\boldsymbol{\Psi}_N(\boldsymbol{\beta}; \boldsymbol{y})$ at each candidate $\boldsymbol{\beta}$, which requires the computation of the pairwise CL from each block at every iteration. Additionally, it may not be the case that the dimension of $\boldsymbol{\Psi}_N$ is smaller than that of $\boldsymbol{Y}$. We propose a meta-estimator of $\boldsymbol{\beta}$ that delivers a one-step update via a linear function of MCLE's $\widehat{\boldsymbol{\beta}}_j$. Our derivation of the one-step estimator is rooted in asymptotic properties of the estimating equations $\boldsymbol{\Psi}_{j.\text{sub}}$ and $\boldsymbol{\Psi}_N$ in a neighborhood of $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_{j0})$, in a similar spirit to Newton–Raphson. Let $[\widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}}]_{i,j}$ be the rows $(i-1)p+1$ to ip and columns $(j-1)p+1$ to jp of matrix $\widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}}$. Let $\boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\boldsymbol{\beta}; \boldsymbol{y}_j)$ be a $\sqrt{N}$-consistent sample estimate of $\boldsymbol{s}_{j,\boldsymbol{\psi}_j}(\boldsymbol{\beta})$. We can obtain a one-step estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_{\text{DIMM}} = \left(\sum_{i,j=1}^{J} \boldsymbol{S}^T_{i,\boldsymbol{\psi}_i}(\widehat{\boldsymbol{\beta}}_i; \boldsymbol{y}_i) \left[\widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}}\right]_{i,j} \boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}}_j; \boldsymbol{y}_j) \right)^{-1} \sum_{i,j=1}^{J} \boldsymbol{S}^T_{i,\boldsymbol{\psi}_i}(\widehat{\boldsymbol{\beta}}_i; \boldsymbol{y}_i) \left[\widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}}\right]_{i,j} \boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}}_j; \boldsymbol{y}_j) \widehat{\boldsymbol{\beta}}_j. \quad (7)$$

With $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ in (7), DIMM can be implemented in a fully parallelized and scalable computational scheme following, for example, the MapReduce paradigm on the Hadoop platform, where only one pass through each block of data is required. These passes can be run on parallel CPUs, and return values of summary statistics $\{\widehat{\boldsymbol{\beta}}_j, \boldsymbol{\psi}_{j.\text{sub}}(\widehat{\boldsymbol{\beta}}_j; \boldsymbol{y}_{i,j}, \widehat{\boldsymbol{\gamma}}_j), \boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}}_j; \boldsymbol{y}_j)\}_{j=1}^J$. After computing $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ as a function of these summary statistics, computation of $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ in (7) can be done in one step. Big data stored on several servers never need be combined, meaning DIMM can be run on distributed correlated response data. $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ can also be used for subgroup analyses, as in Section 6, to combine estimates from specific subgroups of interest. In the following asymptotic theory, we assume J, p, and d_j are fixed; we allow M to diverge. We show in Theorem 4 that the one-step estimator $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ in (7) has the same asymptotic distribution as and is asymptotically equivalent to $\widehat{\boldsymbol{\beta}}_c$.

Theorem 4. Given conditions C.1–C.4, $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ and $\widehat{\boldsymbol{\beta}}_c$ have the same asymptotic distribution: $\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\text{DIMM}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{j}^{-1}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)\right)$ as $N \to \infty$. Moreover, $\widehat{\boldsymbol{\beta}}_c$ and $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ are asymptotically equivalent: $\sqrt{N}\left\|\widehat{\boldsymbol{\beta}}_{\text{DIMM}} - \widehat{\boldsymbol{\beta}}_c\right\| \xrightarrow{p} 0$ as $N \to \infty$.

The proof of this theorem is given in the supplementary materials. The additional conditions specify the convergence rate of the MCLE's $\widehat{\boldsymbol{\beta}}_j$ to ensure the proper convergence rate of $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$. These are necessary because the computation of the one-step estimator relies solely on the MCLE's. This theorem is the key result that allows us to use the one-step estimator, which is more computationally attractive than $\widehat{\boldsymbol{\beta}}_c$, without sacrificing any of the asymptotic properties enjoyed by $\widehat{\boldsymbol{\beta}}_c$, such as estimation efficiency.

Finally, it is clear from Theorem 4 and the form of the Godambe information $\boldsymbol{j}_{\boldsymbol{\psi}}(\boldsymbol{\beta}) = \sum_{i,j=1}^J \boldsymbol{s}^T_{i,\boldsymbol{\psi}_i}(\boldsymbol{\beta})[\boldsymbol{v}^{-1}_{\boldsymbol{\psi}}(\boldsymbol{\beta})]_{i,j} \boldsymbol{s}_{j,\boldsymbol{\psi}_j}(\boldsymbol{\beta})$ that under conditions C.1–C.4, a consistent estimator of the asymptotic covariance of $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ is $(N\sum_{i,j=1}^J \boldsymbol{S}^T_{i,\boldsymbol{\psi}_i}(\widehat{\boldsymbol{\beta}}_i; \boldsymbol{y}_i)[\widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}}]_{i,j}\boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}}_j; \boldsymbol{y}_j))^{-1}$. Equipped with $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ and an estimate of its asymptotic covariance, we can do Wald tests and construct confidence intervals for inference on $\boldsymbol{\beta}$. When conditions C.1–C.4 hold, it is clear that $Q_N(\widehat{\boldsymbol{\beta}}_{\text{DIMM}}) \xrightarrow{d} \chi^2_{(J-1)p}$ as $N \to \infty$, allowing us to test the goodness of fit of our model. For reasonably large Jp, say ≈5000, inversion of $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ can be numerically unstable, although we have never encountered such a situation. In this case, there are several options from the literature, such as linear shrinkage estimation (Han and Song 2011). Our preference is to use a regularized modified Cholesky decomposition of $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ following Pourahmadi (1999). Computation of a regularized estimate of $\widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}}$ requires the inversion of a diagonal matrix, which is fast to compute, and the selection of a tuning parameter by cross-validation. Details are available in the supplementary materials, and our R package allows for the implementation of a regularized weight matrix.

In summary, DIMM proceeds in three steps:

Step 1. Split the data according to established scientific knowledge to form J blocks of lower-dimensional response subvectors with homogeneous correlations.

Step 2. Analyze the J blocks in parallel using pairwise CL. MCLEs are obtained using the R function `optim`. We run 500 iterations of Nelder–Mead with initial values $\boldsymbol{\beta} = (1, \ldots, 1)^T$. End values of this optimization are used as starting values for the BFGS algorithm, which yields $\widehat{\boldsymbol{\beta}}_j$. We return $\{\widehat{\boldsymbol{\beta}}_j, \boldsymbol{\psi}_{j.\text{sub}}(\widehat{\boldsymbol{\beta}}_j; \boldsymbol{y}_{i,j}, \widehat{\boldsymbol{\gamma}}_j), \boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}}_j; \boldsymbol{y}_j)\}_{j=1}^J$.

Step 3. Compute $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ and then find $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ in (7).

An R package to implement DIMM is provided in the supplementary materials. We conclude this section with a brief discussion of the computational complexity of DIMM with general block-covariance structure. All methods depend on N

in the first order, which is therefore omitted from the discussion. Let $m_{\max} = \max_{j=1,\ldots,J} m_j$ and first consider the case where M is finite. In Step 2, inverting the two-dimensional covariance matrices is $O(2^{2+\epsilon})$ for some $\epsilon > 0$, and summing over all pairs of observations is $O(m_j^2)$. In Step 3, inverting $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ is $O((Jp)^{2+\epsilon})$. This yields a general computational complexity of $O((Jp)^{2+\epsilon} + m_{\max}^2)$ for DIMM. By contrast, GEE is generally $O(M^{2+\epsilon}) = O(J^{2+\epsilon} m_{\max}^{2+\epsilon})$ due to the inversion of the covariance matrix of the outcome. DIMM is computationally advantageous when $p^{2+\epsilon} \leq m_{\max}^{2+\epsilon} - m_{\max}^2/J^{2+\epsilon}$. As M diverges, $m_{\max}$ and M are of the same order since J is fixed, and $O(m_{\max}^{2+\epsilon} - m_{\max}^2/J^{2+\epsilon}) = O(M^{2+\epsilon} - M^2)$ so that DIMM becomes increasingly advantageous as M diverges. For computational complexity of mixed effects models see Perry (2017), which discusses various estimation procedures whose iterations are at best approximately $O(q^3)$, where q is the number of fixed and random effects. In the linear model, considering the simplest mixed model case with nested random effects for subjects and response groups, we can compare these two methods and find that DIMM is computationally advantageous when $(Jp)^{2+\epsilon} + m_{\max}^2 \leq (p + NJ)^3$ for fixed M. As M diverges, DIMM is $O(M^2)$ and its advantage depends on the relative rates of convergence of M and N.

5. Simulations

We examine through simulations the performance and finite sample properties in Theorem 4 of the one-step estimator $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ under the linear regression setting $\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta}$, where $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\beta})$, $\boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$. We consider two sets of simulations: the first illustrates DIMM for different dimensions M of $\boldsymbol{Y}$, $J = 5$ for all settings, with an intercept included in $\boldsymbol{X}_i$, and varying number of covariates; the second pushes DIMM to its extremes with very large M and J, and five covariates. In both settings, to mimic the infant EEG data, we let $\boldsymbol{\Sigma} = \boldsymbol{S} \otimes \boldsymbol{A}$ with nested correlation structure, where $\otimes$ denotes the Kronecker product, $\boldsymbol{A}$ an AR(1) covariance matrix, and $\boldsymbol{S}$ a $J \times J$ positive-definite matrix.

$\{\boldsymbol{Y}_i, \boldsymbol{X}_i\}_{i=1}^N$ can be partitioned into J blocks of data with local AR(1) covariance structure. Data within each block is modeled using the bivariate normal marginal distribution. We note that $\widehat{\boldsymbol{\beta}}_j$ has a closed-form solution following generalized least squares (GLS): estimating $\widehat{\boldsymbol{\beta}}_j$ can be done by iteratively updating $\widehat{\boldsymbol{\beta}}_j^{(k)} = (\boldsymbol{X}_j^T \widehat{\boldsymbol{\Sigma}}_j^{(k)} \boldsymbol{X}_j)^{-1} \boldsymbol{X}_j^T \{\widehat{\boldsymbol{\Sigma}}_j^{(k)}\}^{-1} \boldsymbol{y}_j$ and $\widehat{\boldsymbol{\Sigma}}_j^{(k)}$, where $\widehat{\boldsymbol{\Sigma}}_j^{(k)}$ has a known covariance structure, for $k = 1, 2, \ldots$ until convergence. We use GLS because it performs slightly faster, with the exception of Figure 4 where we use `optim` for computational reasons. True value of $\boldsymbol{\beta}$ is set to $\boldsymbol{\beta}_0 = (0.3, 0.6, 0.8, 1.2, 0.45, 1.6)^T$ in the case of five covariates, and subsets thereof for fewer covariates.

We discuss the first set of simulations. Let sample size be $N = 1000$ and the AR(1) covariance matrix $\boldsymbol{A}$ have standard deviation $\sigma = 2$ and correlation $\rho = 0.5$. CL estimation of $\widehat{\boldsymbol{\beta}}_j$ is done first by using the correct AR(1) block covariance structure (DIMM-AR(1)). To evaluate how our method performs under covariance misspecification, we estimate $\widehat{\boldsymbol{\beta}}_j$ using a compound symmetry (DIMM-CS) block covariance structure.

We compute $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ from (7) and its covariance, and report root mean squared error (RMSE), empirical standard error (ESE), mean asymptotic standard error (ASE), and mean bias (BIAS) with $M = 200$ and five scalar covariates (Table 1) and with $M = 1000$ and two vector covariates (Table 2). We compare DIMM to estimates of $\boldsymbol{\beta}$ obtained using GEE with a compound symmetry covariance structure (GEE-CS) and independence covariance structure (GEE-IND) using the

Table 1. Simulation results: RMSE, BIAS, ESE, ASE with five covariates, $N = 1000$, $M = 200$, $J = 5$, averaged over 500 simulations.

	Measure$\times 10^{-2}$	DIMM-AR(1)	DIMM-CS	GEE-CS	GEE-IND	LMM	GLS-oracle
β_0	RMSE/BIAS	4.34/−0.35	4.32/−0.32	4.88/−0.33	4.88/−0.33	4.85/−0.33	4.12/−0.36
	ESE/ASE	4.33/4.21	4.32/4.21	4.87/4.85	4.87/4.85	4.84/5.07	4.11/4.12
β_1	RMSE/BIAS	1.83/0.03	1.84/0.04	2.09/0.08	2.09/0.08	2.07/0.09	1.8/0.06
	ESE/ASE	1.83/1.78	1.84/1.78	2.09/2.05	2.09/2.05	2.07/2.14	1.8/1.74
β_2	RMSE/BIAS	3.41/−0.04	3.47/−0.07	3.75/0.08	3.75/0.08	3.69/0.09	3.24/−0.02
	ESE/ASE	3.41/3.23	3.47/3.23	3.76/3.72	3.76/3.72	3.7/3.89	3.25/3.17
β_3	RMSE/BIAS	1.51/0.14	1.51/0.14	1.67/0.09	1.67/0.09	1.66/0.1	1.45/0.13
	ESE/ASE	1.50/1.45	1.51/1.45	1.67/1.67	1.67/1.67	1.66/1.74	1.45/1.42
β_4	RMSE/BIAS	5.50/0.23	5.49/0.2	5.98/0.19	5.98/0.19	5.94/0.2	5.26/0.29
	ESE/ASE	5.50/5.15	5.49/5.15	5.98/5.92	5.98/5.92	5.94/6.19	5.25/5.04
β_5	RMSE/BIAS	3.53/−0.09	3.56/−0.07	3.99/−0.08	3.99/−0.08	3.97/−0.1	3.42/−0.04
	ESE/ASE	3.53/3.21	3.56/3.21	3.99/3.74	3.99/3.74	3.97/3.9	3.43/3.18

NOTE: Block sizes are $(m_1, m_2, m_3, m_4, m_5) = (45, 42, 50, 34, 29)$. $X_1 \sim \mathcal{N}(0, 1)$, $X_2 \sim$ Bernoulli(0.3), $X_3 \sim$ Categorical(0.1, 0.2, 0.4, 0.25, 0.05), $X_4 \sim$ Uniform(0, 1), and $X_5 = X_1 \times X_2$.

Table 2. Simulation results: RMSE, BIAS, ESE, ASE with two covariates, $N = 1000$, $M = 1000$, $J = 5$, averaged over 500 simulations.

	Measure$\times 10^{-2}$	DIMM-AR(1)	DIMM-CS	GEE-CS	GEE-IND	LMM	GLS-oracle
β_0	RMSE/BIAS	0.71/0.01	0.72/0.01	0.82/0.01	0.82/0.01	0.82/0.01	0.69/0.00
	ESE/ASE	0.71/0.72	0.72/0.72	0.82/0.82	0.82/0.82	0.82/0.85	0.69/0.7
β_1	RMSE/BIAS	0.15/0.00	0.19/0.00	0.21/0.00	0.21/0.00	0.15/0.00	0.13/0.00
	ESE/ASE	0.15/0.19	0.19/0.19	0.21/0.2	0.21/0.2	0.15/0.16	0.13/0.13
β_2	RMSE/BIAS	0.45/0.01	0.45/0.01	0.52/0.00	0.52/0.00	0.51/0.00	0.44/0.02
	ESE/ASE	0.45/0.46	0.46/0.46	0.52/0.52	0.52/0.52	0.51/0.52	0.44/0.45

NOTE: Block sizes are $(m_1, m_2, m_3, m_4, m_5) = (225, 209, 247, 170, 149)$. $X_1 \sim \text{Normal}_M(0, S)$, where S is a positive-definite $M \times M$ matrix, X_2 a vector of alternating 0's and 1's to imitate an exposure.

R package `geepack` (Højsgaard, Halekoh, and Yan 2006), using a linear mixed-effects (LMM) model with nested random intercepts for subject and block membership with AR(1) within-group correlation using the R package `nlme`, and using GLS with known covariance (GLS-oracle) (our code). The latter can be considered the "oracle setting," as we do not estimate the covariance of the response but use the true covariance to estimate $\boldsymbol{\beta}$. In the supplementary materials, we include simulations that show the statistical efficiency gain of using $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ to take into account the correlation between blocks. For these simulations, we compute an estimator derived by using a diagonal weighting matrix instead of $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ in equation (7), and compare the length of 95% confidence intervals. We examine Type I error of the test $H_0 : \beta_q = 0$ for $q = 1, \ldots, p$ for each simulation scenario, and the chi-squared distribution of test statistic $Q_N(\widehat{\boldsymbol{\beta}}_{\mathrm{DIMM}})$ with $M = 200, J = 3, 5$, with one and two covariates (see the supplementary materials). Simulations are conducted using R software on a standard Linux cluster with 16GB of random-access memory per CPU. CL evaluation is coded in C++ but minimization of the CL occurs in R. One simulation in each of the following settings failed to converge with LMM: one covariate with $M = 500$, five covariates with $M = 500$, one covariate with $M = 1000$. This is because of the numerical instability of LMM with high-dimensional outcomes.

In Table 1, $\widehat{\boldsymbol{\beta}}_{\mathrm{DIMM}}$ appears consistent since BIAS is close to zero. RMSE, ESE, and ASE are approximately equal, meaning DIMM is unbiased and has correct variance formula in Theorem 4. Moreover, DIMM mean variance is generally smaller than GEE and LMM mean variance. In data analyses, this results in increased statistical power and more signal detection. Finally, DIMM is close to attaining the estimation efficiency under the GLS-oracle case of known covariance, which is the best efficiency possible. In Table 2, we corroborate these observations for spatially/longitudinally varying vector covariates. Our method also still performs well when dimension is equal to sample size. Finally from Figure 3, we see that DIMM is computationally much faster than GEE and LMM and maintains

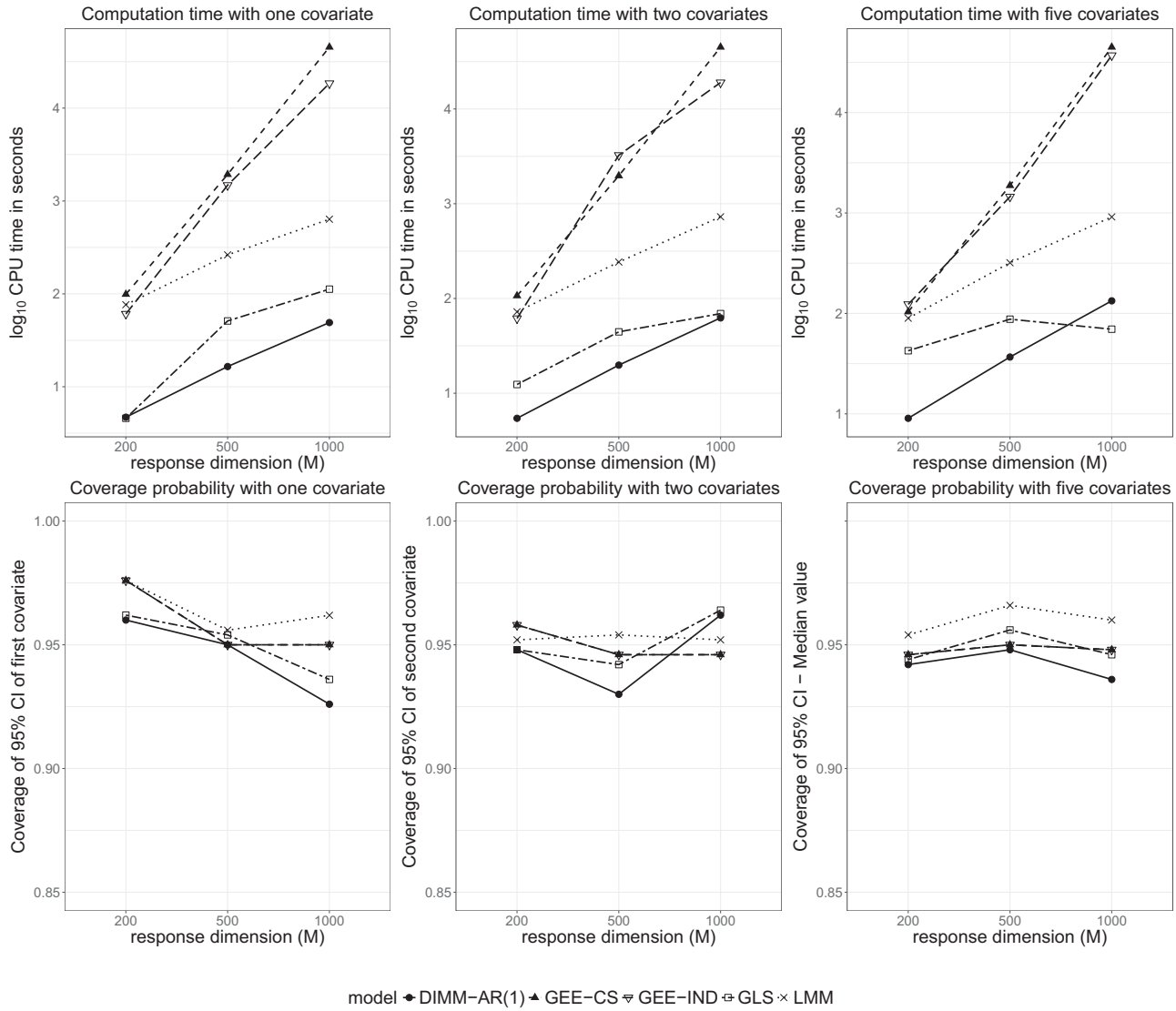

Figure 3. Upper panels: Comparison of computation time on $\log_{10}$ scale of five methods for varying dimension M based on 500 simulations. Lower panels: Comparison of 95% confidence interval coverage of five methods for varying dimension M based on 500 simulations. Left column has $X_1 \sim \mathcal{N}(0, 1)$; middle column has $X_1 \sim \mathcal{N}_M(0, S)$, where S is a positive-definite $M \times M$ matrix, and X_2 a vector of alternating 0's and 1's; right column has $X_1 \sim \mathcal{N}(0, 1)$, $X_2 \sim$ Bernoulli(0.3), $X_3 \sim$ Multinomial(0.1, 0.2, 0.4, 0.25, 0.05), $X_4 \sim$ Uniform(0, 1), and X_5 an interaction between X_1 and X_2.

appropriate confidence interval coverage, corroborating the theoretical asymptotic distribution in Theorem 4 for large sample size. For fixed m_j, DIMM is scalable, since the dimension of the response in each block does not increase. We remark that CPU time consists of time spent by the CPU on calculations and is generally shorter than elapsed time, especially for analyses that use the entire data such as GEE, LMM, and GLS-oracle. Elapsed time depends greatly on implementation and hardware, and is harder to compare between methods. For DIMM, CPU time is the sum of maximum CPU time over parallelized block analyses and CPU time spent on other computations, such as computing $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ and $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$.

We now discuss the second set of simulations. We let sample size $N = 1500$ and consider a very challenging linear regression problem with high-dimension $M = 10{,}000$, and $J = 12$ such that $(m_1, \ldots, m_{12}) = (917, 863, 988, 734, 906, 603, 756, 963, 915, 856, 641, 858)$. We let $\boldsymbol{X}_i$ be a matrix of five covariates and an intercept, and the AR(1) covariance matrix $\boldsymbol{A}$ with standard deviation $\sigma = 16$ and correlation $\rho = 0.8$. We compute $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ from (7) and its estimated covariance, and plot RMSE, ESE, ASE, and BIAS in Figure 4. We were unable to compare DIMM with existing competitors due to the tremendous computational burden associated with such high-dimensional M. As in the first set of simulations, $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ is consistent with ignorable BIAS. RMSE, ESE, and ASE are approximately equal, confirming the large-sample properties of DIMM in this numerical example. ASE slightly underestimates ESE for certain covariate types. This could be due to the high-dimensionality $Jp = 72$ of $\boldsymbol{\Psi}_N$, or the poorer performance of GMM in smaller samples (see Section 7). Beyond theoretical validation, the simulation results presented in this section highlight the applicability, flexibility, and computational power of DIMM. The empirical evidence from simulations is encouraging and advocates the ability of DIMM to deal with high-dimensional correlated response data with multilevel nested correlations.

6. Application to Infant EEG Data

We present the analysis of the infant EEG data introduced in Section 1. EEG data from 157 two-month-old infants under two stimuli at 46 nodes was used. Six brain regions were identified by the investigator as related to auditory recognition memory, with an additional reference node (VREF), as visualized in Figure 1: left frontal-central (11, 12, 13, 14, 15, 18, 19), middle frontal-central (3, 4, 6, 7, 8, 9, 54), right frontal-central (2, 53, 56, 57, 58, 59, 60), left parietal-occipital (24, 25, 26, 27, 28, 29, 30, 32), middle parietal-occipital (31, 33, 34, 35, 36, 37, 38, 39, 40), and right parietal-occipital (42, 43, 44, 45 46, 47, 48, 52).

The primary scientific objective of this study is to quantify the effect of iron deficiency on auditory recognition memory. From cord blood at birth, 50 infants were classified as iron deficient (sufficiency_status $= 1$) and 107 as iron sufficient based on serum ferritin and zinc protoporphyrin levels. Additional available covariates are age and type of stimulus (mother's voice coded with voice_stimulus $= 1$). The response for one infant has a complex nested correlation structure with response dimension $M = 276$; see Figure 5. This figure aligns with substantive scientific knowledge and suggests a partition of data into 18 blocks of response subvectors, one for each ERP and brain region. It also corroborates prior knowledge of high correlations within frontal-central regions, parietal-occipital regions, and between ERPs P2 and P750.

Let $\boldsymbol{Y}_{i,j}$ be the vector of EEG measurements in one brain region and ERP (block $j, j = 1, \ldots, 18$) for infant i, and consider the linear model with block-specific coefficients:

$$E\left(\boldsymbol{Y}_{i,j}\right) = \beta_{0,j} + \beta_{1,j}\text{age}_{i,j} + \beta_{2,j}\text{voice\_stimulus}_{i,j} + \beta_{3,j}\text{sufficiency\_status}_{i,j}. \quad (8)$$

Instead of assuming global homogeneous covariate effects, which is not biologically meaningful, we perform analyses based on certain locally homogeneous covariate-response relationships to identify specific regions affected or not by iron deficiency. Through individual block analyses (see the supplementary materials) and existing knowledge, we identify homogeneous covariate effects across frontal-central regions in each ERP ($M = 42$ for each ERP), the left parietal-occipital region in P2 and P750 ($M = 32$), the middle and right parietal-occipital regions from P2 ($M = 34$), the middle and right parietal-occipital regions from P750 ($M = 34$), and parietal-occipital regions from LSW ($M = 50$). As mentioned previously, DIMM's flexibility allows us to conduct subgroup analyses by combining blocks of homogeneous effects to improve statistical power.

We use an inverse normal transformation of the responses for each analysis. To estimate regression parameters using DIMM, we assume a compound symmetric covariance structure of the response within each brain region and each ERP; block analyses

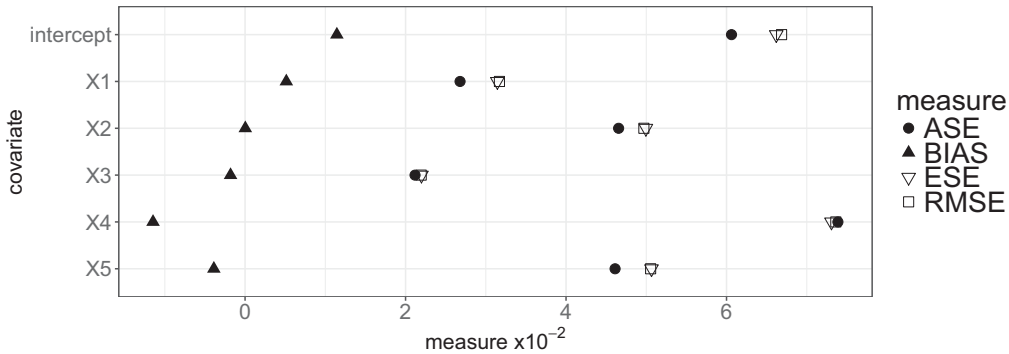

Figure 4. RMSE, BIAS, ESE, ASE based on 100 simulations with an intercept and five covariates, and $M = 10{,}000$. Covariates are simulated as in the right column of Figure 3.

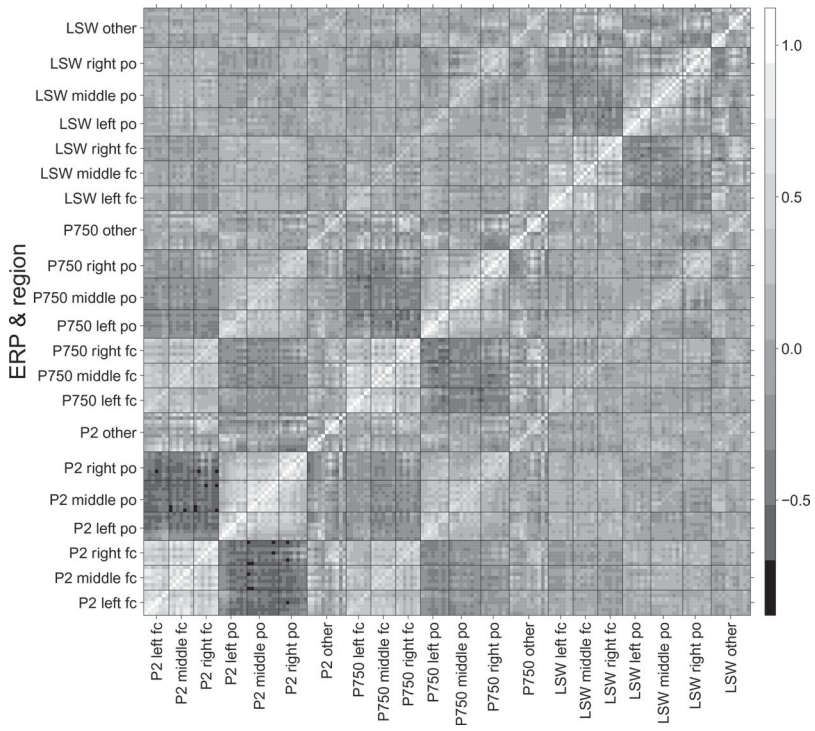

Figure 5. Correlation of electrical amplitude at three ERPs for iron sufficient children under stimulus of mother's voice (color plot and additional plots in the supplementary materials).

Table 3. Select EEG data analysis results: Iron sufficiency status effect estimates and statistics for each combination scheme.

Combine region, ERP	Method	Estimate ($SD \times 10^{-2}$)	*p*-value	CPU seconds	CPU time ratio*
Left, middle, and right fc, P2	GEE-CS	0.103 (12.0)	0.39	0.72	0.55
	LMM	0.103 (11.8)	0.38	1.97	1.49
	DIMM	0.087 (11.9)	0.47	1.32	1
Left po, P2 and P750	GEE-CS	−0.174 (8.3)	0.04	0.22	0.43
	LMM	−0.174 (8.3)	0.04	1.47	2.86
	DIMM	−0.226 (8.1)	0.005	0.51	1
Left, middle, and right po, LSW	GEE-CS	0.041 (8.7)	0.64	0.55	1.41
	LMM	0.041 (7.4)	0.58	3.53	9.07
	DIMM	0.087 (8.4)	0.30	0.39	1

NOTE: fc, frontal-central; po, parietal-occipital; SD, standard deviation.
*CPU time ratio is computed as CPU time of method divided by CPU time of DIMM.

are run in parallel; we compute the one-step estimator $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ for the set of homogeneous regions of interest. We compare DIMM to GEE-CS and LMM with nested random intercepts for subject, stimulus, ERP, and brain region with within-group compound symmetry correlation structure to reinforce gains in computation time and statistical power. Based on simulations mimicking our data setting (see the supplementary materials), we find that DIMM, GEE-CS, and LMM have adequate power. We present iron sufficiency status effect estimates for selected subgroup analyses in Table 3 (complete results available in the supplementary materials).

DIMM finds a more precise estimate than GEE for all analyses, and for a majority of analyses for LMM. This is because the covariance structures assumed by GEE and LMM over the entire response may not be close to the true covariance, resulting in a loss of efficiency. DIMM always performs faster than LMM, and for half the analyses DIMM also performs faster than GEE. This is because of the parallelization of DIMM. DIMM may be slower than GEE in the few analyses because of the limited sample size and small response dimensionality, limiting the improvements of DIMM over GEE. Nonetheless, in data simulations (see the supplementary materials), on average DIMM performs faster than GEE. Effect estimates from GEE, LMM, and DIMM tend to be in the same direction, increasing confidence in our results. The estimated effect for the left parietal-occipital region in P2 and P750 is significant: iron deficient infants had expected transformed left parietal-occipital P2 and P750 amplitude 0.226 units lower than iron sufficient infants of the same age and sex. We find more precise estimates faster than using GEE and LMM by making better model assumptions and running analyses in parallel. The proposed DIMM shows promise in simple data analyses, and has the theoretical justification to perform well in more complex scenarios.

7. Discussion

The proposed DIMM allows for the fast and efficient estimation of regression parameters with high-dimensional correlated

response. Simulations show the scalability of DIMM for fixed J and confirm key asymptotic properties of the DIMM estimator. The $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ estimator can be implemented using a fully parallelized computational scheme, for example using the MapReduce paradigm on the Hadoop platform. Investigators split data into blocks of responses with simple and homogeneous covariance structures. The data partition may be driven by some established scientific knowledge or certain data-driven approaches. Errors in prior knowledge can lead to misspecification of the data split, which may be checked via model diagnostics or goodness-of-fit tests. If sample size is large enough, investigators may consider imposing no or limited structure on $\boldsymbol{\gamma}_j$ to avoid misspecifying response blocks.

In the linear regression setting, the mean and variance of the composite likelihood approach fully specify the joint distribution of the subresponse $\boldsymbol{y}_{i,j}$, and minimal inferential efficiency is lost in the block analysis when the model is correctly specified. Empirical evidence from the simulations in Section 5 support this argument. In the nonlinear setting, inferential efficiency will inevitably be lost in the block analyses because the pairwise composite likelihood is a misspecified likelihood. This loss can be mitigated by using trivariate (or higher) marginal distributions to construct the block-specific estimating equations. By using the optimal weight matrix in the GMM, we avoid assumptions on the between-block covariance structure, and any further loss of efficiency. This may seem counter-intuitive given that divide-and-conquer approaches typically lead to a loss of efficiency. With DIMM, there is a trade-off between efficiency and homogeneity in the parameter $\boldsymbol{\beta}$. Indeed, the assumption of homogeneity in $\boldsymbol{\beta}$ can be restrictive but allows us to borrow information across blocks and use an efficient GMM, controlling the variance of $\boldsymbol{\beta}$ in the process.

In practice, potential trade-offs between number of blocks J and block size m_j should be evaluated when there is no strong substantive knowledge to guide the choice of partition. Our numerical experience has suggested that although large J leads to smaller m_j and therefore faster computation and less strict model assumptions, DIMM may yield inefficient results due to large dimensionality of the integrated CL score vector $\boldsymbol{\Psi}_N$. On the other hand, large m_j but small J will have the opposite effect of slower computation and stricter model assumptions within each block but better combination of results.

Finally, issues related to poor performance of GMM in small samples have been documented in the literature and must be considered when sample size is small (see Hansen, Heaton, and Yaron 1996 and others in the same issue). In this case, to reduce the dimensionality of the integrated CL score vector $\boldsymbol{\Psi}_N$, we suggest integrating analyses from a small number of blocks for more reliable results, as done in Section 6.

DIMM utilizes the full strength of GMM to combine information from multiple sources to achieve greater statistical power, an approach that has been shown to work well with longitudinal data (see, e.g., Wang, Wang, and Song 2012, 2016). DIMM has the potential to combine multimodal data, an important analytic task in biomedical data analysis for personalized medicine. Indeed, response data in each block can be modeled using any pairwise distribution f_j, where $\{f_j\}_{j=1}^{J}$ can be made compatible with $f(\boldsymbol{Y};\boldsymbol{\Gamma})$ using Fréchet classes (see Joe 1997, chap. 3). We anticipate numerous extensions to DIMM, including the addition of penalty terms to CL estimating equations, and allowing for spatially varying mean parameter $\boldsymbol{\beta}$ and prediction of neighboring response variables. Also of interest is the study of the asymptotic behavior of the DIMM estimator when J is allowed to grow with the sample size. Additional conditions to regularize the process of block (and dimension) growth, such as in Donald, Imbens, and Newey (2003), Newey (2004), and Qu, Lee, and Lindsay (2008), could be considered to study the GMM estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{L}}$, but much work remains to study the DIMM estimator $\widehat{\boldsymbol{\beta}}_{\text{DIMM}}$ since the dimensions of $\boldsymbol{\Psi}_N$ and $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ depend on J, introducing additional theoretical challenges. We anticipate that DIMM will be useful for many types of data, including genomic, epigenomic, and metabolomic, indicating the promising methodological potential of DIMM.

Appendix: Proofs of Asymptotic Properties

Let Θ be the compact parametric space of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. We list the regularity conditions required to establish large sample properties in the article.

C.1 Assume $E_{\boldsymbol{\beta}_0}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y})$ has a unique zero at $\boldsymbol{\beta}_0$, $E_{\boldsymbol{\gamma}_{j0}}\boldsymbol{G}_{j.\text{sub}}(\boldsymbol{\gamma}_j;\boldsymbol{y}_j,\boldsymbol{\beta}_0)$ has a unique zero at $\boldsymbol{\gamma}_{j0}$, $-\nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\boldsymbol{\beta};\boldsymbol{y}_i)$ is smooth in a neighborhood of $\boldsymbol{\beta}_0$ and positive definite, $\boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)$ is finite, positive-definite and nonsingular, and

$$\begin{aligned}&\left\|\boldsymbol{\psi}_{j.\text{sub}}(\boldsymbol{\beta}_1;\boldsymbol{y}_i,\boldsymbol{\gamma}_{j1})-\boldsymbol{\psi}_{j.\text{sub}}(\boldsymbol{\beta}_2;\boldsymbol{y}_i,\boldsymbol{\gamma}_{j2})\right\|\\&\quad\le C\left(\left\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\right\|+\left\|\boldsymbol{\gamma}_{j1}-\boldsymbol{\gamma}_{j2}\right\|\right)\end{aligned}$$

for all $\boldsymbol{\beta}_1,\boldsymbol{\gamma}_{j1},\boldsymbol{\beta}_2,\boldsymbol{\gamma}_{j2}$ in a neighborhood of $\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{j0}$ and some constant $C>0$.

C.2 Following Newey and McFadden (1994), assume $Q_0(\boldsymbol{\beta}) = E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{Y})\right\}\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)\,E_{\boldsymbol{\beta}}\{\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\}$ is twice-continuously differentiable in a neighborhood of $\boldsymbol{\beta}_0$.

C.3 Let $\widehat{\boldsymbol{\beta}}_c$ be as defined in (5), and $\boldsymbol{\beta}_0$ an interior point of Θ. Following Newey and McFadden (1994), assume $Q_N(\widehat{\boldsymbol{\beta}}_c) \le \inf_{\boldsymbol{\beta}\in\Theta} Q_N(\boldsymbol{\beta}) + o_p(1)$, and, for any $\delta_N \to 0$,

$$\begin{aligned}&\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le\delta_N}\frac{\sqrt{N}}{1+\sqrt{N}\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\right\|}\\&\quad\left\|\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y})-\boldsymbol{\Psi}_N(\boldsymbol{\beta}_0;\boldsymbol{y})-E_{\boldsymbol{\beta}}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right\|\xrightarrow{p}0.\end{aligned}$$

C.4 For each $j=1,\ldots,J$, assume $\widehat{\boldsymbol{\beta}}_j=\boldsymbol{\beta}_0+O_p(N^{-1/2})$ and $\widehat{\boldsymbol{\gamma}}_j=\boldsymbol{\gamma}_{j0}+O_p(N^{-1/2})$. Assume

$$\begin{aligned}&\sup\frac{\sqrt{N}}{1+\sqrt{N}\left\|(\boldsymbol{\beta},\boldsymbol{\gamma}_j)-(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{j0})\right\|}\\&\left\|\boldsymbol{\Psi}_{j.\text{sub}}(\boldsymbol{\beta};\boldsymbol{y}_j,\boldsymbol{\gamma}_j)-\boldsymbol{\Psi}_{j.\text{sub}}(\boldsymbol{\beta}_0;\boldsymbol{y}_j,\boldsymbol{\gamma}_{j0})\right.\\&\quad\left.-E_{\boldsymbol{\beta}}\boldsymbol{\psi}_{j.\text{sub}}(\boldsymbol{\beta};\boldsymbol{y}_{i,j},\boldsymbol{\gamma}_{j0})\right\|=O_p(N^{-1/2}).\end{aligned}$$

for any $\delta_N\to 0$, where the supremum is taken over the ball $\left\|(\boldsymbol{\beta},\boldsymbol{\gamma}_j)-(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{j0})\right\|\le\delta_N$.

Proof of Lemma 1. Denote $\boldsymbol{\psi}(\widehat{\boldsymbol{\beta}}_{\text{MCLE}};\boldsymbol{y}_i) = (\boldsymbol{\psi}^T_{1.\text{sub}}(\widehat{\boldsymbol{\beta}_1};\boldsymbol{y}_{i,1},\widehat{\boldsymbol{\gamma}_1}), \ldots, \boldsymbol{\psi}^T_{J.\text{sub}}(\widehat{\boldsymbol{\beta}_J};\boldsymbol{y}_{i,J},\widehat{\boldsymbol{\gamma}_J}))^T$. By consistency of the MCLE due to Proposition 1 and C.1, $\widehat{\boldsymbol{\beta}_j} - \boldsymbol{\beta}_0 = o_p(1)$ and $\widehat{\boldsymbol{\gamma}_j} - \boldsymbol{\gamma}_{j0} = o_p(1)$. Since J, p finite, $\|\widehat{\boldsymbol{\beta}}_{\text{MCLE}} - \boldsymbol{\beta}_0\| = o_p(1)$ and $\|\widehat{\boldsymbol{\gamma}}_{\text{MCLE}} - \boldsymbol{\gamma}_0\| = o_p(1)$. Then by C.1,

$$\begin{aligned}&\|\boldsymbol{\psi}(\widehat{\boldsymbol{\beta}}_{\text{MCLE}};\boldsymbol{y}_i) - \boldsymbol{\psi}(\boldsymbol{\beta}_0;\boldsymbol{y}_i)\| \\ &\quad\le C\left(\|\widehat{\boldsymbol{\beta}}_{\text{MCLE}} - \boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\gamma}}_{\text{MCLE}} - \boldsymbol{\gamma}_0\|\right) = o_p(1).\end{aligned}$$

Plugging into $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$, we have $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\psi}^{\otimes 2}(\widehat{\boldsymbol{\beta}}_{\text{MCLE}};\boldsymbol{y}_i) = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\psi}^{\otimes 2}(\boldsymbol{\beta}_0;\boldsymbol{y}_i) + o_p(1)$. Since $\frac{1}{N}\sum_{i=1}^N \boldsymbol{\psi}^{\otimes 2}(\boldsymbol{\beta}_0;\boldsymbol{y}_i) = \boldsymbol{\nu}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0) + o_p(1)$, then, $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}} = \boldsymbol{\nu}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0) + o_p(1)$. □

Proof of Theorem 1. It is sufficient to show that, by conditions C.1 and C.2, $\frac{1}{N}Q_N(\boldsymbol{\beta})$ converges uniformly in probability to $Q_0(\boldsymbol{\beta})$. Note that $\left\|\frac{1}{N}Q_N(\boldsymbol{\beta}) - Q_0(\boldsymbol{\beta})\right\|$ is equal to

$$\begin{aligned}&\Big\|\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{y})\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y}) \\ &\quad - 2E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{Y})\right\}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y}) \\ &\quad + 2E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{Y})\right\}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y}) \\ &\quad - 2E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{Y})\right\}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right\} \\ &\quad + 2E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{Y})\right\}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right\} \\ &\quad - E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{Y})\right\}\boldsymbol{\nu}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right\}\Big\| \\ &\le \Big\|\left[\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y}) - E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right\}\right]^T\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\left[\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y})\right. \\ &\quad \left. - E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right\}\right]\Big\| \\ &\quad + 2\left\|E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{Y})\right\}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\left[\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y}) - E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right\}\right]\right\| \\ &\quad + \left\|E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{Y})\right\}\left[\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1} - \boldsymbol{\nu}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)\right]E_{\boldsymbol{\beta}}\left\{\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right\}\right\| \\ &\le O_p(N^{-1/2}) + o_p(1).\end{aligned}$$

It follows that $\sup_{\boldsymbol{\beta}\in\Theta}\left\|\frac{1}{N}Q_N(\boldsymbol{\beta}) - Q_0(\boldsymbol{\beta})\right\| \xrightarrow{p} 0$ as $N\to\infty$. By Theorem 2.1 in Newey and McFadden (1994), the combined GMM estimator satisfies $\widehat{\boldsymbol{\beta}}_c \xrightarrow{p} \boldsymbol{\beta}_0$ as $N\to\infty$. □

Supplementary Materials

Additional technical details, proofs of theorems, simulations and data analysis results are in the supplementary materials, along with an R package.

Acknowledgments

The authors are grateful for the constructive comments given by the associate editor and the anonymous reviewers that led to a significant improvement of the article.

Funding

This research was funded by grants NSF DMS1811734, NIH R01ES024732, and NIH P01ES022844.

References

Arbia, G. (2014), "Pairwise Likelihood Inference for Spatial Regressions Estimated on Very Large Datasets," *Spatial Statistics*, 7, 21–39. [2]

Bai, Y., Song, P. X. K., and Raghunathan, T. E. (2012), "Joint Composite Estimating Functions in Spatiotemporal Models: Joint Composite Estimating Functions," *Journal of the Royal Statistical Society*, Series B, 74, 799–824. [2]

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian Predictive Process Models for Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 70, 825–848. [1]

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2015), "Distributed Estimation and Inference With Statistical Guarantees," arXiv no. 1509.05457. [2]

Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2012), "Estimating Space and Space-Time Covariance Functions for Large Data Sets: A Weighted Composite Likelihood Approach," *Journal of the American Statistical Association*, 107, 268–280. [2,4]

Chang, W., Haran, M., Olson, R., and Keller, K. (2015), "A Composite Likelihood Approach to Computer Model Calibration With High-Dimensional Spatial Data," *Statistica Sinica*, 25, 243–259. [2]

Chen, X., and Xie, M. (2014), "A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data," *Statistica Sinica*, 24, 1655–1684. [2]

Claggett, B., Xie, M., and Tian, L. (2014), "Meta-Analysis With Fixed, Unknown, Study-Specific Parameters," *Journal of the American Statistical Association*, 109, 1660–1671. [2]

Cox, D. R., and Reid, N. (2004), "A Note on Pseudolikelihood Constructed From Marginal Densities," *Biometrika*, 91, 729–737. [6]

Cressie, N., and Johannesson, G. (2008), "Fixed Rank Kriging for Very Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 70, 209–226. [1]

Crowder, M. (1995), "On the Use of a Working Correlation Matrix in Using Generalised Linear Models for Repeated Measures," *Biometrika*, 82, 407–410. [1]

Donald, S. G., Imbens, G. W., and Newey, W. K. (2003), "Empirical Likelihood Estimation and Consistent Tests With Conditional Moment Restrictions," *Journal of Econometrics*, 117, 55–93. [12]

Fisher, R. A. (1930), "Inverse Probability," in *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 26), Cambridge University Press, pp. 528–535. [5]

Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. (1993), "Regression Models for Discrete Longitudinal Responses," *Statistical Science*, 8, 284–309. [1]

Han, P., and Song, P. X.-K. (2011), "A Note on Improving Quadratic Inference Functions Using a Linear Shrinkage Approach," *Statistics and Probability Letters*, 81, 438–445. [7]

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054. [1,5,6]

Hansen, L. P., Heaton, J., and Yaron, A. (1996), "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Statistics*, 14, 262–280. [12]

Heagerty, P. J., and Lele, S. R. (1998), "A Composite Likelihood Approach to Binary Spatial Data," *Journal of the American Statistical Association*, 93, 1099–1111. [2]

Højsgaard, S., Halekoh, U., and Yan, J. (2006), "The R Package Geepack for Generalized Estimating Equations," *Journal of Statistical Software*, 15, 1–11. [9]

Jin, Z. (2011), "Aspects of Composite Likelihood Inference," Unpublished Ph.D. thesis, University of Toronto. [6]

Joe, H. (1997), *Multivariate Models and Dependence Concepts* (1st ed.), London: Chapman & Hall. [12]

——— (2014), *Dependence Modeling With Copulas* (1st ed.), London: Chapman & Hall. [4]

Kong, X., Wang, M.-C., and Gray, R. (2015), "Analysis of Longitudinal Multivariate Outcome Data From Couples Cohort Studies: Application to HPV Transmission Dynamics," *Journal of the American Statistical Association*, 110, 472–485. [2]

Kuk, A. Y., and Nott, D. J. (2000), "A Pairwise Likelihood Approach to Analyzing Correlated Binary Data," *Statistics and Probability Letters*, 47, 329–335. [2]

Laird, N. M., Lange, N., and Stram, D. (1987), "Maximum Likelihood Computations With Repeated Measures: Applications of the EM Algorithm," *Journal of the American Statistical Association*, 82, 97–105. [1]

Larribe, F., and Fearnhead, P. (2011), "On Composite Likelihoods in Statistical Genetics," *Statistica Sinica*, 21, 43–69. [2]

Li, C. (2017), "Fusion Learning of Dependent Studies by Confidence Distribution (CD): Theory and Applications," Unpublished Ph.D. thesis, Rutgers University. [2,5]

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22. [1]

Lin, N., and Xi, R. (2011), "Aggregated Estimating Equation Estimation," *Statistics and Its Interface*, 4, 73–83. [2,6]

Lindsay, B. G. (1988), "Composite Likelihood Methods," *Contemporary Mathematics*, 80, 220–239. [2,4]

Lindstrom, M. J., and Bates, D. M. (1988), "Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, 83, 1014–1022. [1]

Liu, D., Liu, R. Y., and Xie, M. (2015), "Multivariate Meta-Analysis of Heterogeneous Studies Using Only Summary Statistics: Efficiency and Robustness," *Journal of the American Statistical Association*, 110, 326–340. [2,5]

Mackey, L., Talwalkar, A., and Jordan, M. I. (2015), "Distributed Matrix Completion and Robust Factorization," *Journal of Machine Learning Research*, 16, 913–960. [2]

Newey, W. K. (2004), "Efficient Semiparametric Estimation via Moment Restrictions," *Econometrica*, 72, 1877–1897. [12]

Newey, W. K., and McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, 4, 2111–2245. [6,12,13]

Perry, P. O. (2017), "Fast Moment-Based Estimation for Hierarchical Models," *Journal of the Royal Statistical Society*, Series B, 79, 267–291. [1,8]

Pourahmadi, M. (1999), "Joint Mean-Covariance Models With Applications to Longitudinal Data: Unconstrained Parametrisation," *Biometrika*, 86, 677–690. [7]

Qu, A., Lee, J. J., and Lindsay, B. G. (2008), "Model Diagnostic Tests for Selecting Informative Correlation Structure in Correlated Data," *Biometrika*, 95, 891–905. [12]

Qu, A., Lindsay, B. G., and Li, B. (2000), "Improving Generalised Estimating Equations Using Quadratic Inference Functions," *Biometrika*, 87, 823–836. [5]

Singh, K., Xie, M., and Strawderman, W. E. (2005), "Combining Information From Independent Sources Through Confidence Distributions," *The Annals of Statistics*, 33, 159–183. [2,5]

Song, P. X.-K. (2007), *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer Series in Statistics, New York: Springer-Verlag. [2,4]

Varin, C., Reid, N., and Firth, D. (2011), "An Overview of Composite Likelihood Methods," *Statistica Sinica*, 21, 5–42. [4]

Wang, F., Wang, L., and Song, P. X.-K. (2012), "Quadratic Inference Function Approach to Merging Longitudinal Studies: Validation and Joint Estimation," *Biometrika*, 99, 755–762. [5,12]

——— (2016), "Fused Lasso With the Adaptation of Parameter Ordering in Combining Multiple Studies With Repeated Measurements," *Biometrics*, 72, 1184–1193. [12]

Xie, M., and Singh, K. (2013), "Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review," *International Statistical Review*, 81, 3–39. [2,5]

Xie, M., Singh, K., and Strawderman, W. E. (2011), "Confidence Distributions and a Unifying Framework for Meta-Analysis," *Journal of the American Statistical Association*, 106, 320–333. [2]

Yang, G., Liu, D., Liu, R. Y., Xie, M., and Hoaglin, D. C. (2014), "Efficient Network Meta-Analysis: A Confidence Distribution Approach," *Statistical Methodology*, 20, 105–125. [2]

Zhang, Y., Duchi, J., and Wainwright, M. (2015), "Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm With Minimax Optimal Rates," *Journal of Machine Learning Research*, 16, 3299–3340. [2]

Zhou, Y., and Song, P. X.-K. (2016), "Regression Analysis of Networked Data," *Biometrika*, 103, 287–301. [2]