Distributed Estimation and Inference for the Analysis of Big Biomedical Data

by

Emily C. Hector

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2020

Doctoral Committee:

Professor Peter X.-K. Song, Chair
Professor Veerabhadran Baladandayuthapani
Professor Xuming He
Associate Professor Jian Kang

Emily C. Hector
ehector@umich.edu
ORCID iD: 0000-0003-1488-3150

To my home

ACKNOWLEDGEMENTS

It is a challenge to know where to begin. During my graduate studies in the Department of Biostatistics at the University of Michigan I have had the great privilege of working with generous, passionate, kind and brilliant colleagues and friends, without whom the journey and its final product would have been impossible. I will try to do them justice.

I am grateful to my parents and sister for their love, encouragement and support. They lead by example, living with balance, purpose, empathy and curiosity. I strive to be better because of them.

To my many friends and colleagues, especially Andrew Whiteman, thank you for your love, support and friendship. There have been times of challenge and of great happiness, and every moment was enriched by your presence. In alphabetical order, particular thanks go to Margaret Banker, Marco Benedetti, Mathieu Bray, Wei Hao, Holly Hartman, Lan Luo, Adam Peterson, Kelly Speth, Lu Tang, Lili Wang, Lu Xia, Xianyong Yin, Ling Zhou and Yiwang Zhou. I would also like to thank the Biostatistics department staff, and especially Nicole Fenech, for their tireless work.

I extend my sincere thanks to my many collaborators in the ELEMENT team, including Jackie Goodrich, Erica Jansen, Karen Peterson and Wei Perng, and to Tianwei Yu and Laura Scott, for their collaboration and guidance. I am grateful to Markku Laakso for generously letting me use his data in my third project. Special thanks go to Michael Boehnke for his generosity, support and accumen, and to the chair of the Biostatistics department, Bhramar Mukherjee, for her time, counsel and example. They foster a rich, diverse and stimulating environment in our department, and it is difficult to imagine Michigan without them.

I am deeply indebted to my committee for their time, advice and contributions, and especially: to Xuming He for his encouragement to do difficult theory that led to a new and deeper understanding and great improvement of my work; to Veera Baladandayuthapani for his insights, wisdom and kindness.

I am extremely grateful to Jian Kang, a mentor, motivator and supporter, for his invaluable encouragement and guidance over the past six years that have greatly enriched my graduate experience. I count myself lucky to have him as a mentor. Thank you for your brilliance.

Most of all, I cannot begin to express my deepest gratitude to my advisor, mentor and friend, Peter Song. I admire his approach of looking for elegant solutions to complex problems. His mentorship has put me on the path to becoming the statistician I aspire to be, and his deep insights have helped me hone my intuition. His enthusiasm and passion for research and data are infectious. He does so much for his students and I want to thank him for that. I hope to someday inspire my students the way he has inspired me.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF APPENDICES

Appendix

ABSTRACT

This thesis focuses on developing and implementing new statistical methods to address some of the current difficulties encountered in the analysis of high-dimensional correlated biomedical data. Following the divide-and-conquer paradigm, I develop a theoretically sound and computationally tractable class of distributed statistical methods that are made accessible to practitioners through R statistical software.

This thesis aims to establish a class of distributed statistical methods for regression analyses with very large outcome variables arising in many biomedical fields, such as in metabolomic or imaging research. The general distributed procedure divides data into blocks that are analyzed on a parallelized computational platform and combines these separate results via Hansen's (1982) generalized method of moments. These new methods provide distributed and efficient statistical inference in many different regression settings. Computational efficiency is achieved by leveraging recent developments in large scale computing, such as the MapReduce paradigm on the Hadoop platform.

In the first project presented in Chapter III, I develop a divide-and-conquer procedure implemented in a parallelized computational scheme for statistical estimation and inference of regression parameters with high-dimensional correlated responses. This project is motivated by an electroencephalography study whose goal is to determine the effect of iron deficiency on infant auditory recognition

memory. The proposed method (published as Hector and Song (2020a)), the Distributed and Integrated Method of Moments (DIMM), divides responses into subvectors to be analyzed in parallel using pairwise composite likelihood, and combines results using an optimal one-step meta-estimator.

In the second project presented in Chapter IV, I develop an extended theoretical framework of distributed estimation and inference to incorporate a broad range of classical statistical models and biomedical data types. To reduce computational speed and meet data privacy demands, I propose to divide data by outcomes and subjects, leading to a doubly divide-and-conquer paradigm. I also address parameter heterogeneity explicitly for added flexibility. I establish a new theoretical framework for the analysis of a broad class of big data problems to facilitate valid statistical inference for biomedical researchers. Possible applications include genomic data, metabolomic data, longitudinal and spatial data, and many more.

In the third project presented in Chapter V, I propose a distributed quadratic inference function framework to jointly estimate regression parameters from multiple potentially heterogeneous data sources with correlated vector outcomes. This project is motivated by the analysis of the association between smoking and metabolites in a large cohort study. The primary goal of this joint integrative analysis is to estimate covariate effects on all outcomes through a marginal regression model in a statistically and computationally efficient way. To overcome computational and modeling challenges arising from the high-dimensional likelihood of the correlated vector outcomes, I propose to analyze each data source using Qu et al.'s quadratic inference funtions, and then to jointly reestimate parameters from each data source by accounting for correlation between data sources.

CHAPTER I

Introduction

1.1 Motivation

Recent technological and computational advances have greatly reduced the cost of data generation and storage, leading to a new era of “big data”: data that is massive in volume, velocity, variety and complexity (Secchi, 2018). The wealth of information available presents an opportunity to gain unique insights in biomedical research. In particular, these developments have paved the way for new, exciting and meaningful scientific research in fields such as neuroscience, genomics, personalized medicine, and many more. Statisticians and applied researchers tend to formulate a hypothesis about the data generated by a scientific study and test its validity, with accompanying measures of uncertainty, to gain insights into the data. Several difficulties arise when applying this approach to high-dimensional data. With increasingly complex data, it becomes increasingly difficult to ask the right questions of the data, and obtain a meaningful and nuanced answer. Moreover, high dimensionality can lead to incorrect statistical inference and scientific conclusions due to noise accumulation, spurious correlations, and incidental endogeneity (Fan et al., 2014). Finally, classical statistical methods are burdened with tremendous, and oftentimes prohibitive computational costs when applied to high-dimensional datasets.

In this dissertation, I focus on developing divide-and-conquer solutions to the problem of analyzing high-dimensional response vectors with complex correlation structure. I describe the key computational and statistical challenges posed by this problem below.

1.2 Big Data Challenges

1.2.1 Modelling Challenges

When Big Data consists of a large number of correlated random variables, as is frequently the case for example in brain imaging, modelling their joint distribution can be challenging for many reasons. It can be difficult to model the full distribution of the data or high-order moments, especially as the number of moments increases beyond the sample size, because of a lack of information on them. Additionally, it can be challenging to capture the variety and heterogeneity of the data without using a large number of parameters, or to determine homogeneity/heterogeneity of these parameters. To address some of these difficulties, Qu et al. (2000) propose the Quadratic Inference Function (QIF) for generalized linear models with correlated outcomes. While this function does not model the correlation parameters, it imposes a correlation structure on the entire high-dimensional correlation structure, which can be statistically inefficient. In practice, when dealing with complex multi-level dependent data, it is ideal to begin by modelling local correlation structures and aggregate them into a global correlation specification. Indeed it is relatively easy to use a simple correlation structure, such as compound symmetry or AR(1), to appropriately capture local correlation. I will propose methods to estimate and carry out inference in a computationally efficient distributed fashion for a set of parameters of interest without modelling higher-order moments. The flexibility of these methods allows the high-dimensional response to have a complex multi-level correlation structure, minimizing loss of statistical efficiency.

1.2.2 Computational Challenges

One of the key computational challenges with correlated big data stem from inverting large matrices and optimizing over a large number of parameters (Cressie and Johannesson, 2008; Banerjee et al., 2008). Furthermore, iterative algorithms need to repeatedly evaluate an objective function over a very large dataset, which can be time consuming. Modern computing platforms use distributed systems to store data on different servers, and recent computing and algorithmic advances allow statistical methods to be run in a distributed fashion when the data on different servers are independent; computing platforms include the MapReduce paradigm on the Hadoop platform (Khezr and Navimipour, 2017) and Apache Spark (Zaharia et al., 2010); recent algorithmic advances include kernel ridge regression (Zhang et al., 2015b) and matrix factorization (Mackey et al., 2015). It is unclear how to proceed, however, when data on different servers are dependent. Moreover, these platforms are not accessible to applied researchers working in the biomedical field, who tend to work with R or SAS. Finally, some of these platforms, such as Apache Spark, still have an iterative component to them that is computationally challenging. I will provide distributed estimation and inference solutions for correlated distributed data problems that are of interest to applied researchers, with an R package for ease of implementation.

1.2.3 Theoretical Challenges

Theoretical challenges related to a large number of covariates p with a small sample size n are discussed in Johnstone and Titterington (2009). More frequently, biomedical big data includes a large number of observations on a large number of subjects. These massive datasets are often created by combining various datasets from different sources, such as multi-center cohort studies or consortia, which can lead to data heterogeneity and modelling challenges, as discussed above. More importantly, this data aggregation relies on a crucial independence assumption that is often not met. While the literature on combining information from independent sources is extensive (Singh et al., 2005; Xie et al., 2011; Lin and Xi, 2011; Xie and Singh, 2013; Chen and Xie, 2014; Claggett et al., 2014; Yang et al., 2014a; Battey et al., 2015; Liu et al., 2015; Tang and Song, 2016), to my knowledge no method has been proposed to combine information from dependent sources that provides a thorough description of the accompanying theory. In this dissertation, I establish needed methodology and asymptotic results for combining information from dependent sources.

1.3 Objectives

In this thesis, I focus on developing and implementing new statistical methods to address some of the current difficulties encountered in the analysis of high-dimensional correlated biomedical data. Following the divide-and-conquer paradigm, I develop a theoretically sound and computationally tractable class of distributed statistical methods that are made accessible to practitioners through R statistical software.

This thesis aims to establish a class of distributed statistical methods for regression analyses with very large outcome variables arising in many biomedical fields, such as in genetic or imaging research. The general distributed procedure divides data into blocks that are analyzed on a parallelized computational platform and combines these separate results via Hansen (1982)'s generalized method of moments. These new methods provide distributed and efficient statistical inference in many different regression settings. Computational efficiency is achieved by leveraging recent developments in large scale computing, such as the MapReduce paradigm on the Hadoop platform.

In Chapter III, I aim to address the modelling, computational, and theoretical challenges related to estimation and inference for regression parameters with high-dimensional responses with multi-level nested correlation structure. This project is motivated by an electroencephalography study whose goal is to determine the effect of iron deficiency on infant auditory recognition memory. I develop the Distributed and Integrated Method of Moments (DIMM) (Hector and Song, 2020a), a divide-and-conquer procedure implemented in a parallelized computational scheme. The DIMM divides responses into subvectors to be analyzed in parallel using pairwise composite likelihood, and combines results using an optimal one-step meta-estimator.

In Chapter IV, I aim to generalize the DIMM to types of analyses other than regression, and to further reduce the computational burden associated with high-dimensional correlated data. I also aim to establishing a clear theoretical foundation for this generalized DIMM. I develop an extended theoretical framework of distributed estimation and inference to incorporate a broad range of classical statistical models and biomedical data types. To reduce computational speed and meet data privacy demands, I propose to divide data by outcomes and subjects, leading to a doubly divide-and-conquer paradigm. I also address parameter heterogeneity explicitly for added flexibility. I establish a new theoretical framework for the analysis of a broad class of big data problems to facilitate valid statistical inference for biomedical researchers. Possible applications include genomic data, metabolomic data, longitudinal and spatial data, and many more.

In Chapter V, I propose a distributed quadratic inference function framework to jointly estimate regression parameters from multiple potentially heterogeneous data sources with correlated vector outcomes. This project is motivated by the analysis of the association between smoking and metabolites in a large cohort study. The primary goal of this joint integrative analysis is to estimate covariate effects on all outcomes through a marginal regression model in a statistically and computationally efficient way. To overcome computational and modeling challenges arising from the high-dimensional likelihood of the correlated vector outcomes, I propose to analyze each data source using Qu et al. (2000)'s QIF, and then to jointly reestimate parameters from each data source by accounting for correlation between data sources.

CHAPTER II

Modelling Correlated Data: a Framework

2.1 Introduction

The first part of this chapter is devoted to describing general approaches to modelling correlated data, and the second part to the general framework proposed in this thesis. I consider inference for an M-dimensional vector of correlated responses $\boldsymbol{y}_i$ with associated covariate $\boldsymbol{X}_i$, for $i = 1, \ldots, N$. Denote

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} & \ldots & y_{iM} \end{pmatrix}^T, \ \boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_{i1} & \ldots & \boldsymbol{x}_{iM} \end{pmatrix}.$$

Covariates $\boldsymbol{x}_{ij}$, $j = 1, \ldots, M$, are q-dimensional column vectors and may include an intercept. In a parametric or semi-parametric framework, the goal is to efficiently estimate and carry out inference for a parameter of interest $\boldsymbol{\theta}$, where $\boldsymbol{y}_i$ are independent realizations of $\boldsymbol{Y}_i$ which depend on $\boldsymbol{\theta}$ through their distribution:

$$\boldsymbol{Y}_i | \boldsymbol{X}_i \overset{ind.}{\sim} f(\boldsymbol{y} | \boldsymbol{X} = \boldsymbol{X}_i; \boldsymbol{\theta}, \boldsymbol{\Gamma}_i), \ i = 1, \ldots, N.$$

$\boldsymbol{\Gamma}_i$ represents other parameters required for the specification of the distribution of $\boldsymbol{y}_i$. Denote by Θ the parameter space of $\boldsymbol{\theta}$. Two main models are of interest and detailed below:

(a) (Marginal Dispersion Model) One can assume marginal densities of $\boldsymbol{Y}_i$ belong to the dispersion model family of distributions (Jørgensen, 1987):

$$f(y_{ij};\mu_{ij},\sigma_{ij}^2) = a(y_{ij};\sigma_{ij}^2)\exp\left\{-\frac{1}{2\sigma_{ij}^2}d(y_{ij};\mu_{ij})\right\},\ j = 1,\ldots,M,$$

with mean μ_{ij} and dispersion σ_{ij}^2. $d(y,\mu)$ is the deviance function, and $a(y;\sigma^2)$ is a normalizing term. Given a known link function h, the mean and dispersion parameters can be modelled as

$$h(\mu_{ij}) = \eta(\boldsymbol{x}_{ij};\boldsymbol{\beta}),\ \log(\sigma_{ij}^2) = \xi(\boldsymbol{x}_{ij};\boldsymbol{\zeta})\ j = 1,\ldots,M,$$

where η and ξ are systematic components. The parameter of interest $\boldsymbol{\theta}$ may be any subset of $(\boldsymbol{\beta},\boldsymbol{\zeta})$. $\boldsymbol{\theta}$ may take several forms such as a vector of regression coefficients in the Generalized Linear Model (GLM), a set of nonparametric regression functions as in the Generalized Additive Model (GAM), a nonparametric function and a vector of regression coefficients as in the semi-parametric model, and many more. See Chapter 4 of Song (2007) for a thorough discussion.

(b) (Marginal Quantile Regression) Quantile regression (Koenker and Bassett, 1978) models the quantiles of the response, rather than the mean, as a function of covariates. It provides a more comprehensive description of the relationship between response and covariates because it has ability to model any point in the distribution. To model the marginal quantiles of $\boldsymbol{y}_i$, following Lu and Fan (2015) let the τ^{th} quantile of y_{ij} given $\boldsymbol{x}_{ij}$ be

$$Q_\tau(y_{ij}|\boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\theta}_\tau. \tag{2.1}$$

In median regression, the parameter of interest becomes $\boldsymbol{\theta} = \boldsymbol{\theta}_2$ (where the 2 indicates the 2-quantile). Quantile regression also has the advantage of not specifying the error distribution, contrary to GLM. Thus, marginal quantile regression is useful when distributional assumptions of GLM fail or when trying to achieve an analysis robust to outliers in the data. Many modifications and extensions to the simple quantile regression model in (2.1) have been proposed, including Yang and He (2015) for spatially correlated data.

2.2 Estimation

2.2.1 Joint Modelling Approaches

Joint modelling approaches reconstruct the full distribution of $\boldsymbol{Y}_i$ to estimate $\boldsymbol{\theta}$. The Maximum Likelihood Estimator (MLE) maximizes the likelihood of the data as a function of the parameter of interest. When the elements of $\boldsymbol{Y}_i$ are independent, the joint likelihood can be constructed by multiplying the marginal likelihoods. When data are correlated, however, this construction is much more difficult due to the presence of high-order moments.

Specific examples of low-dimensional joint distributions exist in the literature. For example, for correlated binary data the log-linear model (Bishop et al., 1974) and the Bahadur representation (Bahadur, 1961) model the joint distribution of correlated binary random variables. For the former, interpretation of association parameters as conditional odds ratios is restrictive (Song, 2007). For the latter, maximum likelihood can fail to converge when the number of repeated observations M is small, such as $M = 10$ (Lipsitz et al., 1995).

In low-dimensional settings, Song (2000) studies a unified framework for dispersion models generated by Gaussian copulas. See also Chapter 3 of Joe (1997) and Chapter 3 of Joe (2014) for details on building joint distributions using Fréchet classes and vine copulas respectively.

2.2.2 Likelihood-Derived Approaches

Composite Likelihoods

Composite likelihoods (Lindsay, 1988) provide a principled approach to constructing a pseudo-likelihood by making assumptions on the functional form of low-dimensional marginal or conditional likelihoods of the data without specifying the full joint distribution; see Varin et al. (2011) for a comprehensive review. Generally, given nonnegative weights w_j and a set of likelihoods $\mathcal{L}_j(\boldsymbol{\theta})$, the composite likelihood is constructed following:

$$\mathcal{L}_{CL}(\boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{L}_j(\boldsymbol{\theta})^{w_j}.$$

The simplest example derives from assuming working independence and multiplying univariate marginals: $\mathcal{L}_{ICL}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{r=1}^{M} f(y_{ir}|\boldsymbol{X}_i, \boldsymbol{\theta})$. Perhaps of more interest in a setting with correlated data is the pairwise composite likelihood (Cox and Reid (2004), Varin (2008)):

$$\mathcal{L}_{PCL}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{r=1}^{M-1} \prod_{t=r+1}^{M} f(y_{ir}, y_{it}|\boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\gamma})$$

The composite likelihood inherits many desirable properties from the marginal likelihoods under suitable regularity conditions, such as unbiasedness, but computation time suffers greatly as dimension M of the response increases. Indeed, since the Bartlett identity does not hold, composite likelihood methods require the computation of the sensitivity matrix, which can be time consuming. Additionally, the pairwise composite likelihood requires the evaluation of a large number of bivariate marginals at every iteration of the optimization algorithm.

Wedderburn's Quasi-Likelihood

Wedderburn (1974) observed that only a specification of the mean and covariance of the response was necessary to compute the MLE of regression parameters in a

GLM, thus avoiding the need to fully specify the multivariate distribution of the data. He replaced assumptions on the likelihood with assumptions on the mean and covariance by defining a function, termed quasi-likelihood, which only specified the mean-covariance relationship and had similar properties to the likelihood function. Take for example the linear regression model $\boldsymbol{Y}_i = \boldsymbol{X}_i^T\boldsymbol{\theta} + \boldsymbol{\epsilon}_i$, with $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$ and $\boldsymbol{V} = E(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T)$. Define the quasi-likelihood function q as the weighted sum of squared residuals:

$$q(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i^T\boldsymbol{\theta}\right)^T \boldsymbol{V}^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i^T\boldsymbol{\theta}\right).$$

Its derivative takes the form

$$\frac{\partial q}{\partial \boldsymbol{\theta}} = \boldsymbol{Q}(\boldsymbol{\theta}) = \sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{V}^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i^T\boldsymbol{\theta}). \tag{2.2}$$

$\boldsymbol{Q}(\boldsymbol{\theta})$ behaves like a score function and is called the quasi-score function, since $E\{\boldsymbol{Q}(\boldsymbol{\theta})\} = \mathbf{0}$ and $E\{-\partial\boldsymbol{Q}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\} = \sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{V}^{-1}\boldsymbol{X}_i^T = \text{Var}\{\boldsymbol{Q}(\boldsymbol{\theta})\}$. Solving $\boldsymbol{Q}(\boldsymbol{\theta}) = \mathbf{0}$ for $\boldsymbol{\theta}$ yields a consistent estimator for $\boldsymbol{\theta}$. There are no assumptions on the functional form of the distribution of the error term $\boldsymbol{\epsilon}_i$ (or $\boldsymbol{Y}_i$); the quasi-likelihood approach relies only on the existence of the first two moments of the response.

The quasi-likelihood approach focuses on estimating the mean parameters while treating second-order moments as nuisance parameters. When $\boldsymbol{\Sigma}_i = \text{Var}(\boldsymbol{Y}_i|\boldsymbol{X}_i)$ is unknown, a plug-in estimate is used to estimate $\boldsymbol{\theta}$. Estimation efficiency relies on choosing $\boldsymbol{\Sigma}_i$ as close to the true covariance structure as possible (Fitzmaurice et al., 1993). As the correlation structure of the response becomes more complex, more nuisance parameters are needed to capture the underlying structure of the data, which can be computationally intensive. Additionally, simple cases where this approach fails are highlighted in Crowder (1987).

2.2.3 Estimating Equation Approaches

Generalized Estimating Equations

Perhaps the most famous approach to modelling correlated data is the generalized estimating equation proposed by Liang and Zeger (1986). Closely related to Wedderburn's quasi-likelihood, it estimates mean parameters in GLMs, forgoes the specification of a joint distribution and treats second moments as nuisance parameters. It goes one step further, however, by not providing an objective function from which the estimating equation is derived. Liang and Zeger generalize the quasi-score function (2.2) to non-normal data and replace $\boldsymbol{V}$ by a working correlation matrix that depends on nuisance parameters. They show that their estimator of $\boldsymbol{\theta}$ is semi-parametrically efficient when the correlation structure of the response is correctly specified, and that the estimator is still consistent even when the correlated structure is misspecified. This approach is available for discrete as well as continuous data, and has seen numerous extensions and applications. Limitations of the generalized estimating equations are well-known. Simple examples where estimation fails are outlined in Crowder (1995), Wang and Carey (2003), and Chaganty and Joe (2004). Computational issues related to inverting large matrices as M grows large and estimating a large number of nuisance parameters are covered in Cressie and Johannesson (2008) and Banerjee et al. (2008). Finally, model selection relies on subjective information criteria because there is no objective function to evaluate model fit.

Generalized Method of Moments

Hansen (1982) introduced the generalized method of moments to estimate a parameter that is over-identified; that is, a parameter that has more estimating equations than it has components. For example, if $\boldsymbol{y}_{ir}$ is Poisson(λ) distributed, the mean and variance parameter λ satisfies the moment conditions $E(\boldsymbol{y}_{ir}) = \lambda$ and $\mathrm{Var}(\boldsymbol{y}_{ir}) = \lambda$. Deriving the moment conditions for the mean and variance yields two estimating equations for λ, and solving these does not lead to a unique solution. To overcome this challenge, Hansen (1982) proposed solving a quadratic form of the estimating equations as follows:

$$\arg\min_{\boldsymbol{\theta}} Q_N(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \boldsymbol{\Psi}_N^T(\boldsymbol{\theta}) \boldsymbol{W} \boldsymbol{\Psi}_N(\boldsymbol{\theta}),$$

where $\boldsymbol{\Psi}_N(\boldsymbol{\theta})$ is the vector of over-identifying estimating equations for $\boldsymbol{\theta}$ and $\boldsymbol{W}$ is a positive semi-definite weight matrix. Under suitable regularity conditions defined in Hansen (1982) and, more generally, in Newey and McFadden (1994), the minimizer of $Q_N(\boldsymbol{\theta})$ is consistent and asymptotically normal. Moreover, Hansen showed that an optimal choice of $\boldsymbol{W}$, corresponding to the inverse sample covariance of $\boldsymbol{\Psi}_N$, leads to an estimator that has asymptotic covariance at least as small as any other estimator derived from the same estimating equation. I hereafter refer to this property as Hansen optimality. Finally, the generalized method of moments also provides a goodness-of-fit test derived from a χ^2 statistic to facilitate model fit evaluation. The generalized method of moments receives a thorough treatment in Hall (2004).

Quadratic Inference Functions

For longitudinal data, Qu et al. (2000) propose Quadratic Inference Functions (QIF) to estimate mean regression parameters in a generalized linear model setting. They model the inverse working correlation matrix of the response by a linear combination of known basis matrices. This approach allows them to build a vector of over-identified moment restrictions on the mean regression parameters, leading to a modified generalized method of moments equation where correlation parameters do not need to be estimated and the weight matrix depends on the parameter of interest. The quadratic inference function estimator minimizes this quadratic form, and is shown to be semi-parametrically efficient when the true correlation structure of the response belongs to the class of linear combinations used to model the inverse working correlation. Additionally, the estimator is still Hansen optimal when the true correlation does not belong to this class. A list of the advantages of quadratic inference functions over generalized estimating equations, such as model selection, robustness, and treatment of nuisance parameters, is given in Hu and Song (2012). Quadratic inference functions suffer computationally from the iterative optimization procedure and inversion of large matrices.

2.3 A Unifying Framework: Estimating Function Theory

With the exception of the generalized method of moments and quadratic inference functions, each of these approaches leads to an estimator of $\boldsymbol{\theta}$ that is the root of some estimating function

$$\boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{X}) = \mathbf{0}.$$

In the joint modelling framework, this function is the score function derived from the likelihood, but the estimating function is the only necessary part of the estimation process. In the case of the generalized method of moments and quadratic inference functions, finding the root of the estimating function can be generalized to finding its minimum. Alternatively, taking first derivatives of the quadratic form leads to estimating functions, and finding their root leads to an estimator of $\boldsymbol{\theta}$.

The estimation approaches described in section 2.2 can be unified under estimating function theory, which justifies why the quasi-likelihood, generalized estimating equations, generalized method of moments and quadratic inference functions are able to estimate $\boldsymbol{\theta}$ without a full specification of the distribution of $\boldsymbol{Y}$. Let us start with some definitions from Godambe (1960) and Song (2007).

Definition 1 (Estimating function)**.** Let $\mathcal{X}$ be the sample space. A function $\boldsymbol{\Psi} : \Theta \times \mathcal{X} \to \mathbb{R}^p$ is called an estimating (or inference) function if $\boldsymbol{\Psi}(\boldsymbol{\theta}; \cdot)$ is measurable for any $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\Psi}(\cdot; \boldsymbol{x})$ is continuous in a compact subspace of Θ containing the true parameter $\boldsymbol{\theta}_0$ for any sample $\boldsymbol{x} \in \mathcal{X}$.

Definition 2 (Additive estimating function)**.** An estimating function $\boldsymbol{\Psi}$ is additive if $\boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{i=1}^{N} \boldsymbol{\psi}(\boldsymbol{\theta}; \boldsymbol{X}_i)$ where $\boldsymbol{X}_i \in \mathcal{X}$. $\boldsymbol{\psi}$ is called the kernel estimating (or inference) function.

Definition 3 (Unbiased estimating function)**.** An estimating function $\boldsymbol{\Psi}$ is unbiased if $E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{X})\right) = \mathbf{0}$ for all $\boldsymbol{\theta} \in \Theta$.

Definition 4 (Regular inference function)**.** An estimating function $\boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{X})$ is regular if

(i) it is unbiased: $E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{X})\right) = \mathbf{0}$ for all $\boldsymbol{\theta} \in \Theta$.

(ii) $\nabla_{\boldsymbol{\theta}} \boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{X})$ exists for almost all $\boldsymbol{X} \in \mathcal{X}$ and for all $\boldsymbol{\theta} \in \Theta$.

(iii) For any bounded function $g(\boldsymbol{x})$ independent of $\boldsymbol{\theta}$, $\int_{\mathcal{X}} g(\boldsymbol{x}) \boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{x}) f(\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{x}$ is differentiable under the integral sign.

(iv) The variability matrix $\boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left\{\boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{X}) \boldsymbol{\Psi}^T(\boldsymbol{\theta}; \boldsymbol{X})\right\}$ exists and is positive-definite.

(v) The sensitivity matrix $\boldsymbol{s}_{\boldsymbol{\psi}}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left\{\nabla_{\boldsymbol{\theta}} \boldsymbol{\Psi}(\boldsymbol{\theta}; \boldsymbol{X})\right\}$ is non-singular.

Optimal estimating function theory was initially developed by Godambe (1960) and Durbin (1960), and summarized in Godambe and Heyde (1987). If an additive regular estimating function has a unique zero at the true value, then its root is a consistent estimator of $\boldsymbol{\theta}$. Additionally, if its second derivative is bounded in a neighbourhood of the true value, this estimator is asymptotically Normal. More details are available in McLeish and Small (1988), Godambe (1991) and Heyde (1997). The key ideas of this thesis derive from two observations.

First, an estimating function for the parameter $\boldsymbol{\theta}$ can be constructed from subsets of the whole data if $\boldsymbol{\theta}$ is homogeneous over the entire response $\boldsymbol{Y}$. Typically, statistical methods are concerned with using as much of the data as possible to achieve large sample results. With big data, using all of the data is computationally prohibitive, and subsets of the data typically provide adequate sample size. Using subsets of the data, however, raises concerns of biased sub-sampling and generalizability to the whole sample; additionally, subsetting $\boldsymbol{Y}$ yields results that only hold for that subset. The trick is to derive estimating functions for each data subset, and combine them in a computationally tractable and statistically efficient way.

Second, rather than combining data or estimators directly, one can combine estimating functions. As functions of the data and the parameter, estimating functions inherently take into account sampling uncertainty and behave like random variables. Whereas the joint distribution of the data or the estimators may be intractable, with suitable regularity conditions the estimating functions inherit asymptotic normality from the Central Limit Theorem and their joint distribution can be reconstructed with ease. Maximizing this distribution yields the same optimization problem as combining the estimating equations using the generalized method of moments.

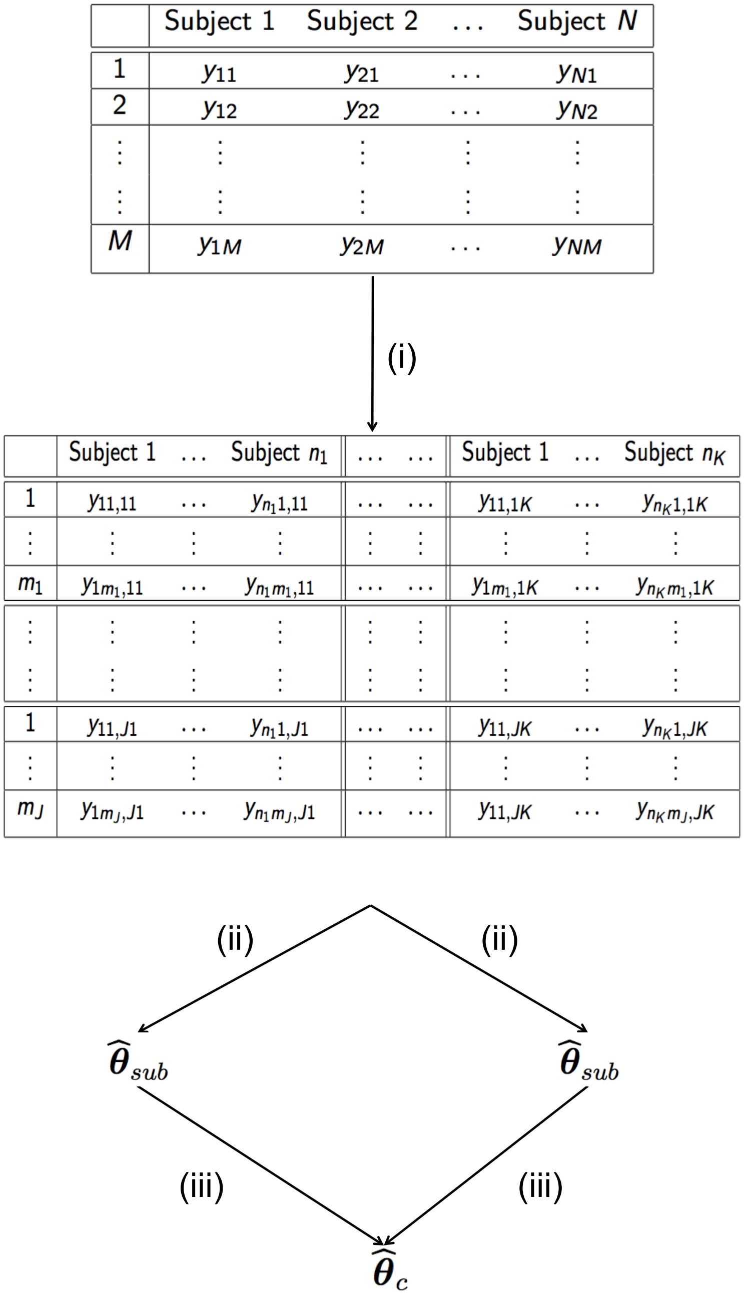
These two keys observations lead to a novel approach to high-dimensional correlation data analysis. The following informal steps describe the general framework proposed by this dissertation:

(i) Divide the data $\{\boldsymbol{y}_i, \boldsymbol{X}_i\}_{i=1}^N$ into blocks $\left\{\boldsymbol{y}_{i,sub}, \boldsymbol{X}_{i,sub}\right\}_{i=1}^N$ for $sub = 1, \ldots, S$.

(ii) Estimate the parameter of interest in blocks $\left\{\boldsymbol{y}_{i,sub}, \boldsymbol{X}_{i,sub}\right\}_{i=1}^N$, $sub = 1, \ldots, S$, separately and in parallel using additive estimating functions. Each block yields an estimator $\widehat{\boldsymbol{\theta}}_{sub}$ of $\boldsymbol{\theta}$.

(iii) Combine individual estimators $\widehat{\boldsymbol{\theta}}_{sub}$.

This can be visualized in Figure 2.1 for $S = JK$. The notation sub is used to denote blocks: for example, for $sub = 1$, $\boldsymbol{y}_{i,sub} = \{(y_{i1,11}, \ldots, y_{im_1,11})\}$ for $i = 1, \ldots, n_1$. In Chapter III, in step (i) the data is divided at the outcome level to form blocks of low-dimensional sub-responses. In step (ii), the blocks are analyzed using pairwise composite likelihood. In step (iii), the estimators are combined using a one-step meta-estimator derived from the optimal generalized method of moments equation. In Chapter IV, in step (i) the data is divided at the outcome level as in Chapter III and additionally at the subject level, as in Figure 2.1, to form blocks of low-dimensional sub-responses on a subset of the population. In step (ii), I outline a broad class of estimating functions that can be used to obtain $\widehat{\boldsymbol{\theta}}_{sub}$. In step (iii) I generalize the one-step meta-estimator from Chapter III to account for block-specific sample sizes. In Chapter V, in step (i) the data is divided at the outcome and subject level as in Chapter IV.



	Subject 1	Subject 2	…	Subject N
1	y_{11}	y_{21}	…	y_{N1}
2	y_{12}	y_{22}	…	y_{N2}
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
M	y_{1M}	y_{2M}	…	y_{NM}

	Subject 1	…	Subject n_1	…	…	Subject 1	…	Subject n_K
1	$y_{11,11}$	…	$y_{n_1 1,11}$	…	…	$y_{11,1K}$	…	$y_{n_K 1,1K}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
m_1	$y_{1m_1,11}$	…	$y_{n_1 m_1,11}$	…	…	$y_{1m_1,1K}$	…	$y_{n_K m_1,1K}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	$y_{11,J1}$	…	$y_{n_1 1,J1}$	…	…	$y_{11,JK}$	…	$y_{n_K 1,JK}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
m_J	$y_{1m_J,J1}$	…	$y_{n_1 m_J,J1}$	…	…	$y_{11,JK}$	…	$y_{n_K m_J,JK}$

(i)
(ii)
(ii)
$\widehat{\boldsymbol{\theta}}_{sub}$
$\widehat{\boldsymbol{\theta}}_{sub}$
(iii)
(iii)
$\widehat{\boldsymbol{\theta}}_c$

Figure 2.1: Schematic of general procedure.

CHAPTER III

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

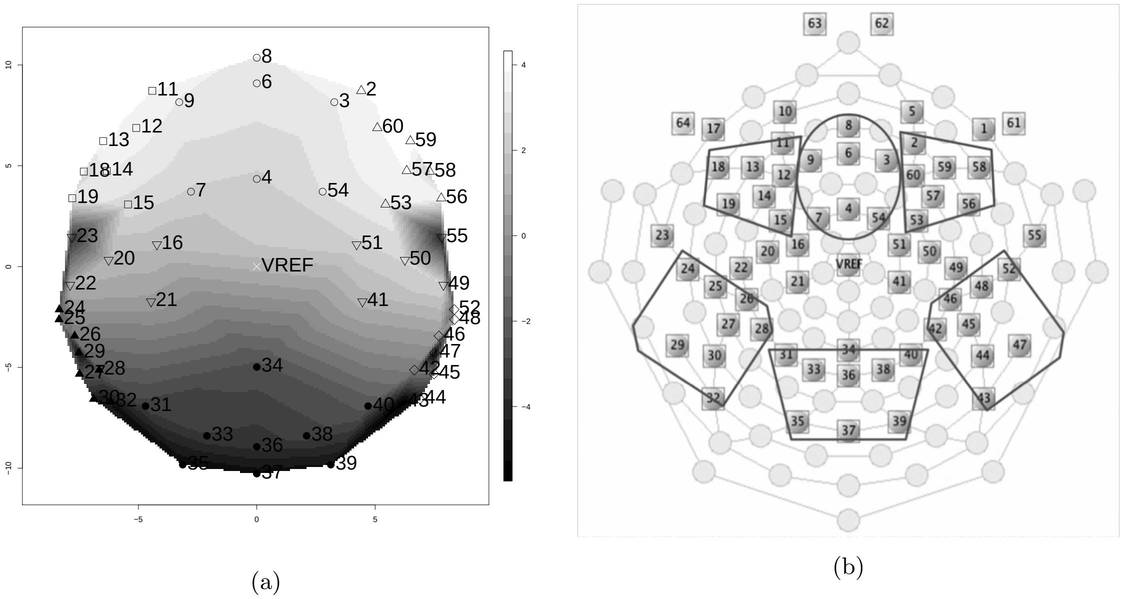
3.1 Introduction

This chapter 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 multi-level 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 (1982)'s 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 chapter 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 Cressie and Johannesson (2008) and Banerjee et al. (2008). Additionally, common assumptions by GEE on the correlation structure of the response are too simple to capture multi-level nested correlations, resulting in a substantial loss of efficiency; see Fitzmaurice et al. (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 et al. (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 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 (Chapter 4 of Song (2007)).

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 et al. (2015)), spatial (Heagerty and Lele (1998), Arbia (2014)), spatiotemporal (Bai et al. (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 $\boldsymbol{Y}_i$ in order 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 in order 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 et al. (2015b)) and matrix factorization (Mackey et al. (2015)). The independent groups can be analyzed in parallel, greatly reducing computation time. Chen and Xie (2014) and Battey et al. (2015) use 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) propose 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 et al. (2005), Xie et al. (2011), Xie and Singh (2013) and Liu et al. (2015); for CD approaches to meta-analysis of independent studies, see Claggett et al. (2014) and Yang et al. (2014b); for a



(a)

(b)

Figure 3.1: (a) Average P2 amplitude for iron sufficient children under stimulus of mother's voice. Color plot and additional plots in Appendix D. (b) Layout of the 64 channel sensor net with brain regions related to auditory recognition memory.

divide-and-conquer approach with independent scalar responses, see Lin and Xi (2011). We invoke an optimal weighting matrix that non-parametrically 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 milliseconds 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 3.1a. For each sensor and each stimulus, three important event-related potentials (ERPs) related to auditory recognition memory were calculated; as shown in Figure 3.2, P2 averages electrical signal between 175 and 300 milliseconds, P750 between 350 and 600 milliseconds, and late slow wave (LSW) between 850 and 1100 milliseconds. The investigator wanted to analyze the data in sub-regions, where 46 of the nodes belong to six brain function regions

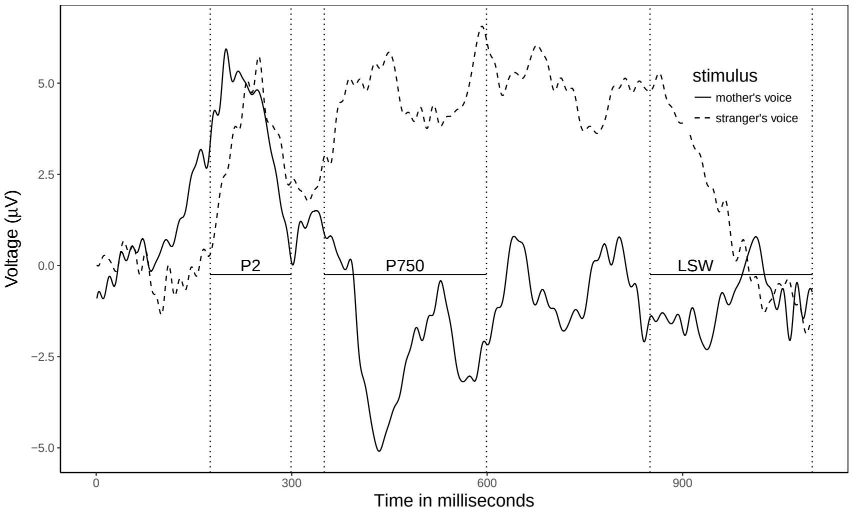


Figure 3.2: Plot of electrical potential for subject 1 at electrode 2 over time.

related to auditory recognition memory, as seen in Figure 3.1b. The complex data-generating mechanism results in a response of dimension $M = 46(nodes) \times 3(ERPs) \times 2(stimuli) = 276$ that has a multi-level 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 Distributed and Integrated Method of Moments (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 chapter 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 complex dataset that has previously not been analyzed in its entirety.

The rest of the chapter is organized as follows. Section 3.2 describes DIMM. Section 3.3 discusses large sample properties. Section 3.4 presents the closed form one-step meta-estimator, and its implementation in a parallel and scalable computational scheme. Section 3.5 illustrates DIMM's finite sample performance with simulations. Section 3.6 presents the EEG data analysis. Section 3.7 concludes with a discussion. Proofs of theorems and additional simulation and data analysis results are deferred to Appendices A-D.

3.2 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 parametrized with a small number of parameters. For example, responses with compound symmetric or AR(1) covariance structures are homogeneously correlated.

3.2.1 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: $\boldsymbol{y}_i = (\begin{array}{ccc} \boldsymbol{y}_{i,1}^T & \ldots & \boldsymbol{y}_{i,J}^T \end{array})^T$ and $\boldsymbol{X}_i = (\begin{array}{ccc} \boldsymbol{X}_{i,1}^T & \ldots & \boldsymbol{X}_{i,J}^T \end{array})^T$. 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 $\boldsymbol{\beta}_j$ hold 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 sub-group 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 3.6. Additionally, we propose a formal test of the homogeneity assumption in Section 3.3. 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 pre-processing 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 et al. (2011) and references therein) to resolve this difficulty, and takes the following form:

$$\mathcal{L}_j(\boldsymbol{\beta}, \boldsymbol{\gamma}_j; \boldsymbol{y}_{i,j}) = \prod_{r=1}^{m_j-1} \prod_{t=r+1}^{m_j} f_j(y_{ir,j}, y_{it,j}; \boldsymbol{\beta}, \boldsymbol{\gamma}_j, \boldsymbol{X}_{i,j}), \tag{3.1}$$

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 $\boldsymbol{\gamma} = (\begin{array}{ccc} \boldsymbol{\gamma}_1^T & \dots & \boldsymbol{\gamma}_J^T \end{array})^T$, $\boldsymbol{\gamma}_0 = (\begin{array}{ccc} \boldsymbol{\gamma}_{10}^T & \dots & \boldsymbol{\gamma}_{J0}^T \end{array})^T$. 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 Chapter 6 of Song (2007) and Chapter 3 of Joe (2014). We set f_j bivariate Normal for the EEG data. Within block j, the log-CL for the first and second moment parameters is

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

Define $\boldsymbol{\psi}_{j.sub}(\boldsymbol{\beta}; \boldsymbol{y}_{i,j}, \boldsymbol{\gamma}_j) = (1/m_j^2) \sum_{r=1}^{m_j-1} \sum_{t=r+1}^{m_j} \nabla_{\boldsymbol{\beta}} \log f_j(y_{ir,j}; y_{it,j}; \boldsymbol{\beta}, \boldsymbol{\gamma}_j, \boldsymbol{X}_{i,j}) \in \mathbb{R}^p$ and

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

$$\boldsymbol{\Psi}_{j.sub}(\boldsymbol{\beta}; \boldsymbol{y}_j, \boldsymbol{\gamma}_j) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\psi}_{j.sub}(\boldsymbol{\beta}; \boldsymbol{y}_{i,j}, \boldsymbol{\gamma}_j) = \mathbf{0} \in \mathbb{R}^p, \tag{3.2}$$

$$\boldsymbol{G}_{j.sub}(\boldsymbol{\gamma}_j; \boldsymbol{y}_j, \boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{g}_{j.sub}(\boldsymbol{\gamma}_j; \boldsymbol{y}_{i,j}, \boldsymbol{\beta}) = \mathbf{0} \in \mathbb{R}^{d_j}. \tag{3.3}$$

Following Varin et al. (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 (3.2) and (3.3). 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 $\boldsymbol{\Psi}_{j.sub}(\boldsymbol{\beta}; \boldsymbol{y}_j, \boldsymbol{\gamma}_j) = \mathbf{0}$. We denote $\widehat{\boldsymbol{\beta}}_{MCLE} = (\widehat{\boldsymbol{\beta}_1}^T, \dots, \widehat{\boldsymbol{\beta}_J}^T)^T$ and $\widehat{\boldsymbol{\gamma}}_{MCLE} = (\widehat{\boldsymbol{\gamma}_1}^T, \dots, \widehat{\boldsymbol{\gamma}_J}^T)^T$. 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 $\boldsymbol{\Psi}_{j.sub}$ and $\boldsymbol{G}_{j.sub}$. In the case where both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_j$ are of interest and believed to be homogeneous over all blocks, we replace $\boldsymbol{\Psi}_{j.sub}$ with $(\boldsymbol{\Psi}_{j.sub}^T, \boldsymbol{G}_{j.sub}^T)^T$. The description of DIMM in the rest of the chapter, including Section 3.4, holds with these minor changes.

3.2.2 Integration: the Generalized Method of Moments

Suppose that we have successfully obtained J estimates of $\boldsymbol{\beta}$ based on J estimating equations (3.2). In the integration step, we treat each estimating equation $\boldsymbol{\Psi}_{j.sub}(\boldsymbol{\beta}; \boldsymbol{y}_j, \boldsymbol{\gamma}_j) = 0$ as a moment condition on $\boldsymbol{\beta}$ coming from block j, $j = 1, \dots, J$. 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 (1982)'s seminal GMM to combine the moment conditions that arise from each block. Stack the J estimating equations by letting $\boldsymbol{\psi}(\boldsymbol{\beta}; \boldsymbol{y}_i) = (\boldsymbol{\psi}_{1.sub}^T(\boldsymbol{\beta}; \boldsymbol{y}_{i,1}, \boldsymbol{\gamma}_{10}), \dots, \boldsymbol{\psi}_{J.sub}^T(\boldsymbol{\beta}; \boldsymbol{y}_{i,J}, \boldsymbol{\gamma}_{J0}))^T \in \mathbb{R}^{Jp}$ for each subject i, and

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

Define the outer-product as $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$ for $\boldsymbol{a} \in \mathbb{R}^{Jp}$. Since $\boldsymbol{\Psi}_N(\boldsymbol{\beta}; \boldsymbol{y}) = \mathbf{0}$ 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:

$$\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}} = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{\psi}^T_{1.sub}(\widehat{\boldsymbol{\beta}_1};\boldsymbol{y}_{i,1},\widehat{\boldsymbol{\gamma}_1}),\ldots,\boldsymbol{\psi}^T_{J.sub}(\widehat{\boldsymbol{\beta}_J};\boldsymbol{y}_{i,J},\widehat{\boldsymbol{\gamma}_J})\right)^{T\ \otimes 2}, \tag{3.4}$$

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

$$\widehat{\boldsymbol{\beta}_c} = \arg\min_{\boldsymbol{\beta}}\left\{N\boldsymbol{\Psi}^T_N(\boldsymbol{\beta};\boldsymbol{y})\widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y})\right\} = \arg\min_{\boldsymbol{\beta}} Q_N(\boldsymbol{\beta}). \tag{3.5}$$

To solve (3.5), 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 (3.5) to Qu et al. (2000) but with a completely different way of constructing moment conditions, and to Wang et al. (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 $\boldsymbol{\Psi}_{j.sub}$ 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 (3.5) 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 et al. (2005), and Liu et al. (2015). So far, little work has been done on the development of CD for correlated data. Of note, a dissertation by Li (2017) considers 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) considers 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 in order to make this method practically useful. $\boldsymbol{\Psi}_{j.sub}$ 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 $\boldsymbol{\Psi}_N$, $\sqrt{N}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1/2}\boldsymbol{\Psi}_N(\boldsymbol{\beta}_0;\boldsymbol{y})$ 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 $F_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0) = \Phi\left(\sqrt{N}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1/2}\boldsymbol{\Psi}_N(\boldsymbol{\beta}_0;\boldsymbol{y})\right)$, where $\Phi(\cdot)$ is the Jp-variate standard normal distribution function. Define the density of the CEF as

$$f_{\boldsymbol{\psi}}(\boldsymbol{\beta}) = \phi\left(\sqrt{N}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1/2}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y})\right) \propto \exp\left\{-\frac{N}{2}\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{y})\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y})\right\}, \qquad (3.6)$$

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 (3.2). Maximizing $f_{\boldsymbol{\psi}}(\boldsymbol{\beta})$ in (3.6) yields the minimization defined in (3.5). The formulation in (3.6) is different from the aggregated estimating equation approach proposed by Lin and Xi (2011) for independent scalar responses.

3.3 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 (3.2) and (3.3) that $\boldsymbol{\Psi}_{j.sub}$ and $\boldsymbol{G}_{j.sub}$ are expressed as sums of 2-dimensional marginal likelihoods as $m_j \to \infty$. Following Cox and Reid (2004)'s 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 $j \in \{1, \dots, J\}$ such that $m_j \to \infty$. Suppose $\boldsymbol{\Psi}_{j.sub}$ and $\boldsymbol{G}_{j.sub}$ are unbiased at $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_{j0})$ and their expectations have a unique zero at $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_{j0})$. Then $(\widehat{\boldsymbol{\beta}_j}, \widehat{\gamma}_j)$ are consistent estimators of $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_{j0})$ as $N \to \infty$.*

The proof is given in Appendix A. Proposition 1 justifies why standard asymptotic theory is applicable even when $M \to \infty$. $\boldsymbol{\Psi}_{j.sub}$ and $\boldsymbol{G}_{j.sub}$ 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

$$\boldsymbol{v}_{j,\psi_j}(\boldsymbol{\beta}) = \lim_{M\to\infty} Var_{\boldsymbol{\beta}} \left\{ \sqrt{N} \boldsymbol{\Psi}_{j.sub}(\boldsymbol{\beta}; \boldsymbol{y}_j, \boldsymbol{\gamma}_{j0}) \right\} = \lim_{M\to\infty} E_{\boldsymbol{\beta}} \left\{ \boldsymbol{\psi}_{j.sub}^{\otimes 2}(\boldsymbol{\beta}; \boldsymbol{y}_{i,j}, \boldsymbol{\gamma}_{j0}) \right\},$$

$$\boldsymbol{s}_{j,\psi_j}(\boldsymbol{\beta}) = \lim_{M\to\infty} -\nabla_{\boldsymbol{\beta}} E_{\boldsymbol{\beta}} \left\{ \boldsymbol{\Psi}_{j.sub}(\boldsymbol{\beta}; \boldsymbol{y}_j, \boldsymbol{\gamma}_{j0}) \right\} = \lim_{M\to\infty} -\nabla_{\boldsymbol{\beta}} E_{\boldsymbol{\beta}} \left\{ \boldsymbol{\psi}_{j.sub}(\boldsymbol{\beta}; \boldsymbol{y}_{i,j}, \boldsymbol{\gamma}_{j0}) \right\}.$$

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 III.1 that $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ in (3.4) converges to the true variability matrix of the estimating equations.

Lemma III.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 III.1, given in Appendix A, is straightforward, and makes use of the consistency of the MCLE's and the Central Limit Theorem. Lemma III.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 III.1 and III.2, we show that $\widehat{\boldsymbol{\beta}_c}$ is consistent and asymptotically normal under mild moment conditions.

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

The proof of Theorem III.1, given in Appendix A, derives from the consistency of the GMM estimator due to Hansen (1982) and, more generally, to Newey and

McFadden (1994). The proof of Theorem III.2 follows from Theorem 7.2 in Newey and McFadden (1994) and Theorem III.1. Theorems III.1 and III.2 do not require the differentiability of $\mathbf{\Psi}_{j.sub}$ and Q_N. Instead, they require the differentiability of their population versions, and that $\mathbf{\Psi}_N$ behave "nicely" in a neighbourhood 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 3.7. 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 $\mathbf{\Psi}_N$ 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 III.3 (Test of over-identifying restrictions)**.** *Let* $\widehat{\boldsymbol{\beta}_c} = \arg\min_{\boldsymbol{\beta}} Q_N(\boldsymbol{\beta})$. *Given conditions C.1, C.2 and C.3,* $Q_N(\widehat{\boldsymbol{\beta}_c}) \overset{d}{\to} \chi^2_{(J-1)p}$ *as* $N \to \infty$.

The proof is given in Appendix A. 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 $\boldsymbol{\psi}_{j.sub}(\widehat{\boldsymbol{\beta}_c}; \boldsymbol{y}_{i,j}, \widehat{\gamma}_j)$ 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.

3.4 Implementation: the Parallelized One-Step Estimator

In practice, searching for the integrated solution of the GMM equation (3.5) 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 $\boldsymbol{\Psi}_{j.sub}$ and $\boldsymbol{\Psi}_N$ in a neighbourhood 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 $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_{DIMM} = \left(\sum_{i,j=1}^{J} \boldsymbol{S}^T_{i,\boldsymbol{\psi}_i}(\widehat{\boldsymbol{\beta}}_i; \boldsymbol{y}_i) \left[\widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}} \right]_{i,j} \boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}_j}; \boldsymbol{y}_j) \right)^{-1} \sum_{i,j=1}^{J} \boldsymbol{S}^T_{i,\boldsymbol{\psi}_i}(\widehat{\boldsymbol{\beta}}_i; \boldsymbol{y}_i) \left[\widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}} \right]_{i,j} \boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}_j}; \boldsymbol{y}_j) \widehat{\boldsymbol{\beta}_j}. \tag{3.7}$$

With $\widehat{\boldsymbol{\beta}}_{DIMM}$ in (3.7), 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 $\{\widehat{\boldsymbol{\beta}_j}, \boldsymbol{\psi}_{j.sub}(\widehat{\boldsymbol{\beta}_j}; \boldsymbol{y}_{i,j}, \widehat{\boldsymbol{\gamma}_j}), \boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}_j}; \boldsymbol{y}_j)\}_{j=1}^{J}$. After computing $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ as a function of these summary statistics, computation of $\widehat{\boldsymbol{\beta}}_{DIMM}$ in (3.7) 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. $\widehat{\boldsymbol{\beta}}_{DIMM}$ can also be used for sub-group analyses, as in Section 3.6, to combine estimates from specific sub-groups 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 III.4 that the one-step estimator $\widehat{\boldsymbol{\beta}}_{DIMM}$ in (3.7) has the same asymptotic distribution as and is asymptotically equivalent to $\widehat{\boldsymbol{\beta}_c}$.

Theorem III.4. *Given conditions C.1, C.2, C.3 and C.4,* $\widehat{\boldsymbol{\beta}}_{DIMM}$ *and* $\widehat{\boldsymbol{\beta}_c}$ *have the same asymptotic distribution:* $\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{DIMM} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{j}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)\right)$ *as* $N \to \infty$*. Moreover,* $\widehat{\boldsymbol{\beta}_c}$ *and* $\widehat{\boldsymbol{\beta}}_{DIMM}$ *are asymptotically equivalent:* $\sqrt{N}\left\|\widehat{\boldsymbol{\beta}}_{DIMM} - \widehat{\boldsymbol{\beta}_c}\right\| \xrightarrow{p} 0$ *as* $N \to \infty$*.*

The proof of this theorem is given in Appendix A. The additional conditions specify the convergence rate of the MCLE's $\widehat{\boldsymbol{\beta}_j}$ to ensure the proper convergence rate of $\widehat{\boldsymbol{\beta}}_{DIMM}$. 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 III.4 and the form of the Godambe information $\boldsymbol{j}_{\boldsymbol{\psi}}(\boldsymbol{\beta}) = \sum_{i,j=1}^{J} \boldsymbol{s}_{i,\boldsymbol{\psi}_i}^T(\boldsymbol{\beta})[\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta})]_{i,j}\boldsymbol{s}_{j,\boldsymbol{\psi}_j}(\boldsymbol{\beta})$ that under conditions C.1-C.4, a consistent estimator of the asymptotic covariance of $\widehat{\boldsymbol{\beta}}_{DIMM}$ is $(N\sum_{i,j=1}^{J} \boldsymbol{S}_{i,\boldsymbol{\psi}_i}^T(\widehat{\boldsymbol{\beta}}_i; \boldsymbol{y}_i)[\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}]_{i,j}\boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}_j}; \boldsymbol{y}_j))^{-1}$. Equipped with $\widehat{\boldsymbol{\beta}}_{DIMM}$ 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 $Q_N(\widehat{\boldsymbol{\beta}}_{DIMM}) \xrightarrow{d} \chi^2_{(J-1)p}$ 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 $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}$ 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 Appendix B, 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. MCLE's are obtained using the R function `optim`. We run 500 iterations of Nelder-Mead with initial values $\boldsymbol{\beta} = (1, \dots, 1)^T$. End values of this optimization are used as starting values for the BFGS algorithm, which yields $\widehat{\boldsymbol{\beta}_j}$. We return $\{\widehat{\boldsymbol{\beta}_j}, \boldsymbol{\psi}_{j.sub}(\widehat{\boldsymbol{\beta}_j}; \boldsymbol{y}_{i,j}, \widehat{\gamma}_j), \boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}_j}; \boldsymbol{y}_j)\}_{j=1}^J$.

Step 3 Compute $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ and then find $\widehat{\boldsymbol{\beta}}_{DIMM}$ in (3.7).

An R package to implement DIMM is available and will be submitted to the Comprehensive R Archive Network (CRAN) shortly. 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 $m_{\max} = \max_{j=1,\dots,J} m_j$ 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.

3.5 Simulations

We examine through simulations the performance and finite sample properties in Theorem III.4 of the one-step estimator $\widehat{\boldsymbol{\beta}}_{DIMM}$ 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 3.4 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 $N = 1,000$ 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 $\widehat{\boldsymbol{\beta}}_{DIMM}$ from (3.7) 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 3.1) and with $M = 1,000$ and two vector covariates (Table 3.2). 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 R package `geepack` (Højsgaard et al. (2006)), using a linear mixed-effects (LMM) model with nested random intercepts for subject and

Table 3.1: Simulation results: RMSE, BIAS, ESE, ASE with five covariates, $N = 1,000$, $M = 200$, $J = 5$, 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

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 $X_5 = X_1 \times X_2$.

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 Appendix C, 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 (3.7), 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 $Q_N(\widehat{\boldsymbol{\beta}}_{DIMM})$ with $M = 200$, $J = 3, 5$, with one and two covariates (see Appendix C). 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

Table 3.2: Simulation results: RMSE, BIAS, ESE, ASE with two covariates, $N = 1,000$, $M = 1,000$, $J = 5$, 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

Block sizes are $(m_1, m_2, m_3, m_4, m_5) = (225, 209, 247, 170, 149)$. $X_1 \sim Normal_M(0, S)$, 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.

covariate with $M = 500$, five covariates with $M = 500$, one covariate with $M = 1,000$. This is because of the numerical instability of LMM with high-dimensional outcomes.

In Table 3.1, $\widehat{\boldsymbol{\beta}}_{DIMM}$ 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 III.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 3.2, 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 3.3, we see that DIMM is computationally much faster than GEE and LMM and maintains appropriate confidence interval coverage, corroborating the theoretical asymptotic distribution in Theorem III.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 $\widehat{\boldsymbol{\beta}}_{DIMM}$.

We now discuss the second set of simulations. We let sample size $N = 1,500$ 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 $\widehat{\boldsymbol{\beta}}_{DIMM}$ from (3.7) and its estimated covariance, and plot RMSE, ESE, ASE, and BIAS in Figure 3.4. 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, $\widehat{\boldsymbol{\beta}}_{DIMM}$ 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

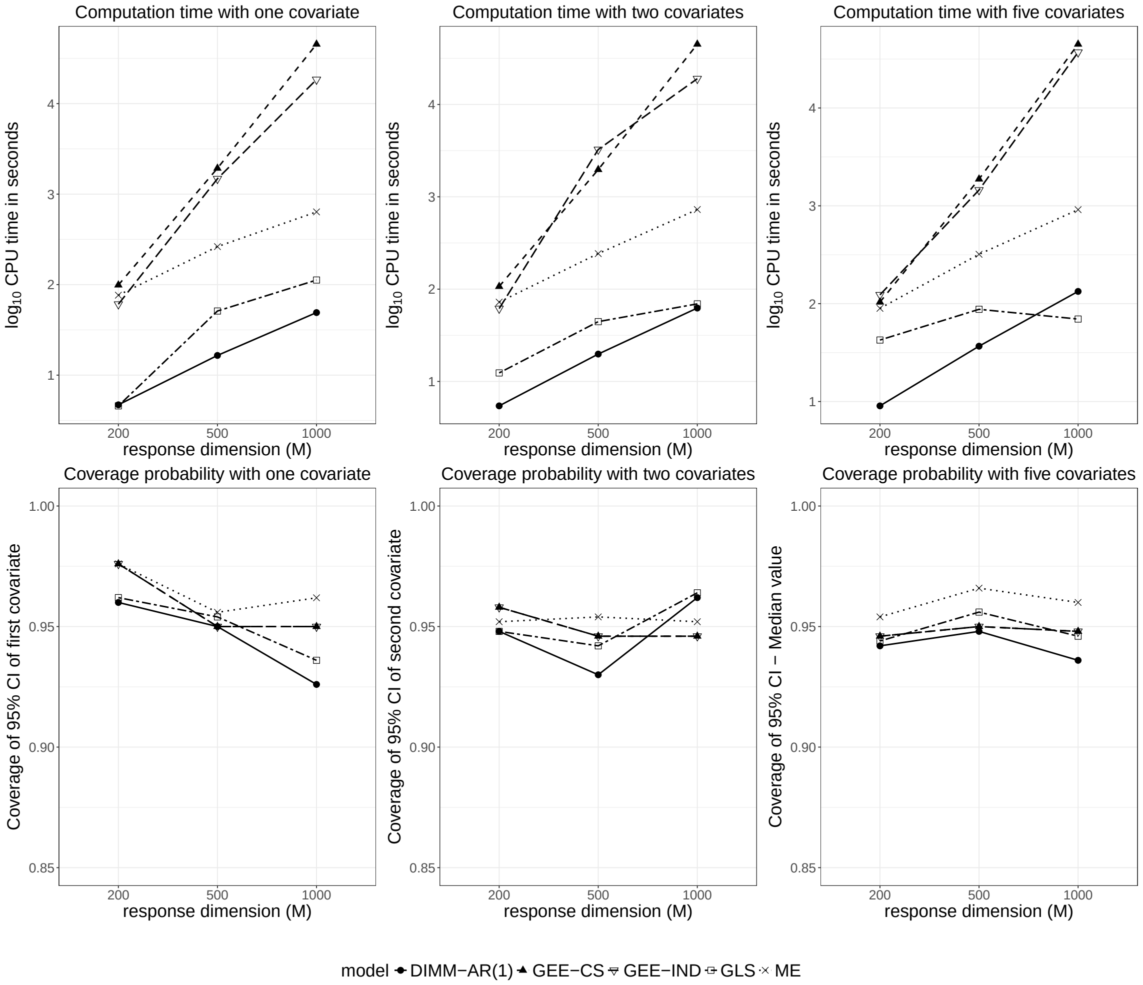


Figure 3.3: Upper panels: comparison of computation time on $\log_{10}$ scale of five methods for varying dimension M based on 500 simulations. Lower panels: comparison of 95% confidence interval coverage of four methods for varying dimension M based on 500 simulations. Left column has $X_1 \sim \mathcal{N}(0,1)$; 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.

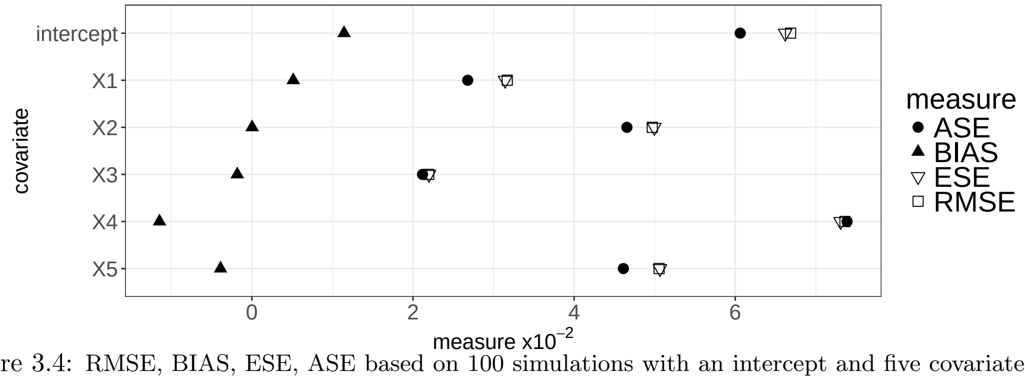


Figure 3.4: RMSE, BIAS, ESE, ASE based on 100 simulations with an intercept and five covariates, and $M = 10,000$. Covariates are simulated as in the right column of Figure 3.3.

$\boldsymbol{\Psi}_N$, or the poorer performance of GMM in smaller samples (see Section 3.7). 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 multi-level nested correlations.

3.6 Application to Infant EEG Data

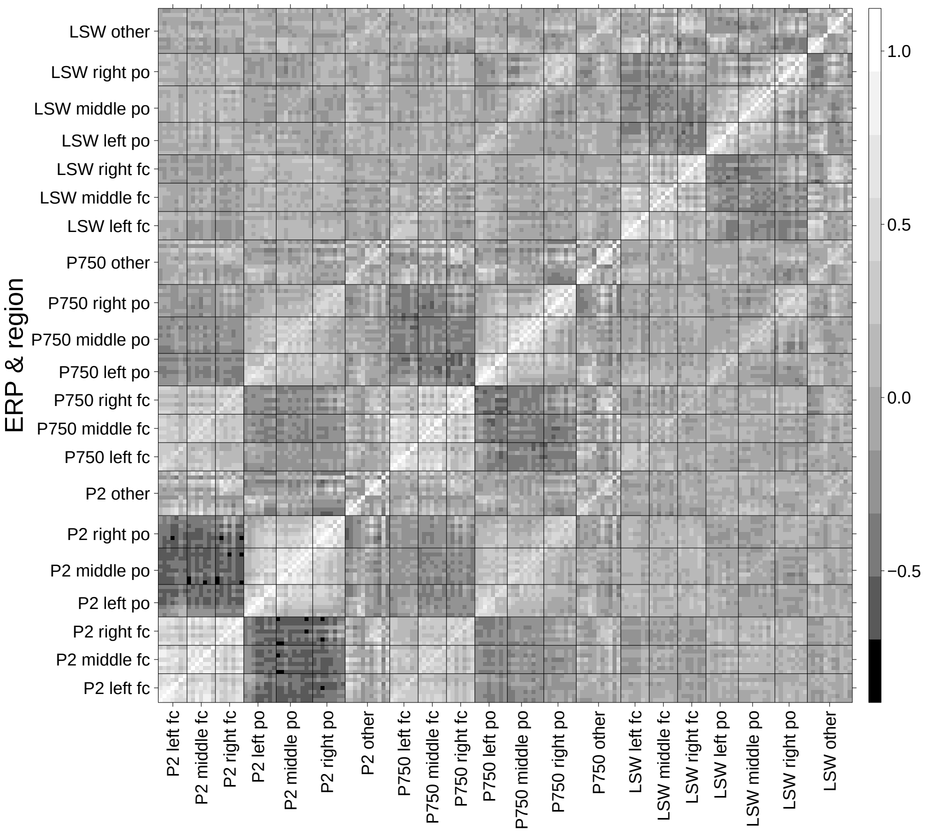


Figure 3.5: Correlation of electrical amplitude at three ERP's for iron sufficient children under stimulus of mother's voice (color plot and additional plots in Appendix D).

We present the analysis of the infant EEG data introduced in Section 3.1. 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 3.1b:

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 3.5. 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:

$$E\left(\boldsymbol{Y}_{i,j}\right) = \beta_{0,j} + \beta_{1,j} age_{i,j} + \beta_{2,j} voice\_stimulus_{i,j} + \beta_{3,j} sufficiency\_status_{i,j}. \quad (3.8)$$

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 Appendix D) 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 sub-group 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 are run in parallel; we compute the one-step estimator $\widehat{\boldsymbol{\beta}}_{DIMM}$ 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 Appendix D), we find that DIMM, GEE-CS and LMM have adequate power. We present iron sufficiency status effect estimates for selected sub-group analyses in Table 3.3 (complete results available in Appendix D).

Table 3.3: Select EEG data analysis results: iron sufficiency status effect estimates and statistics for each combination scheme.

combine region, ERP	method	estimate (s.d.$\times 10^{-2}$)	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 & 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

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

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 Appendix D), 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 & P750 is significant: iron deficient infants had expected transformed left parietal-occipital P2 & 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.

3.7 Discussion

The proposed DIMM, published as Hector and Song (2020a), 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 $\widehat{\boldsymbol{\beta}}_{DIMM}$ 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 3.5 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 et al. (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 3.6.

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 for examples Wang et al. (2012) and Wang et al. (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 Chapter 3 of Joe (1997)). 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 neighbouring 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 et al. (2003), Newey (2004) and Qu et al. (2008), could be considered to study the GMM estimator $\widehat{\boldsymbol{\beta}_c}$, but much work remains to study the DIMM estimator $\widehat{\boldsymbol{\beta}}_{DIMM}$ 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.

CHAPTER IV

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

4.1 Introduction

Although the divide-and-conquer paradigm has been widely used in statistics and computer science, its application with correlated data has been little investigated in the literature. We provide a theoretical justification, with theoretical guarantees, for divide-and-conquer methods with correlated data through a general unified estimating function theory framework. In particular, in this chapter we focus on the large sample properties of a class of distributed and integrated estimators for supervised learning and inference with high-dimensional correlated outcomes. We consider N independent observations $\{\boldsymbol{y}_i, \boldsymbol{X}_i\}_{i=1}^N$ where both the sample size N and the dimension M of the response vector $\boldsymbol{y}_i$ may be so big that a direct analysis of the data using conventional methodology is computationally intensive, or even prohibitive. Such data may arise, for example, from imaging measurements of brain activity or from genomic data. Denote by $f(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\Gamma}_i)$ the M-variate joint parametric distribution of $\boldsymbol{Y}_i$ conditioned on $\boldsymbol{X}_i$, where $\boldsymbol{\theta}$ is the parameter of interest and $\boldsymbol{\Gamma}_i$ contains parameters, such as for high-order dependencies, that may be difficult to model or handle computationally. Statistical inference with big data can be extremely challenging due to the high volume and high variety of these data, as noted recently by Secchi (2018). In the statistics literature, methodological efforts to date have primarily focused on high-dimensional covariates (i.e. high-dimensional $\boldsymbol{X}_i$) with univariate responses (corresponding to $M = 1$); see Johnstone and Titterington (2009) for an overview of the difficulties and methods in linear regression, and the citations therein for references to the extensive publications in this field. By contrast, little work has focused on high-dimensional correlated outcomes (corresponding to large M), which pose an entirely new and different set of methodological challenges stemming from a high-dimensional likelihood. The divide-and-combine paradigm holds promise in overcoming these challenges; see Mackey et al. (2015) and Zhang et al. (2015b) for early examples of the power of divide-and-combine algorithms. Some recent divide-and-combine methods for independent outcomes can be found in Singh et al. (2005), Lin and Zeng (2010), Lin and Xi (2011), Chen and Xie (2014), and Liu et al. (2015), among others.

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

This chapter proposes a new doubly divided procedure to learn and perform inference for a homogeneous parameter of interest in the presence of heterogeneous parameters with a general class of supervised learning procedures. The double division at the response and subject levels further speeds up computations in comparison to DIMM and results in a double division of the data, visualized in Table 4.1: a division of the response $\boldsymbol{Y}$, and a random division of subjects into independent subject groups, resulting in blocks of data with a smaller sample of low-dimensional response subvectors. We consider a general class of supervised learning procedures to analyze these blocks separately and in parallel. Then we establish a GMM-type combination procedure that yields a meta-estimator of the parameter of interest. This proposed estimator is more general than the DIMM estimator in Hector and Song (2020a), and thus appealing in many practical settings where analyzing data with both large M and N is challenging. We achieve a doubly divided learning and inference procedure implemented in a distributed and parallelized computational scheme. The proposed class of supervised learning procedures is very general, including many important estimation methods as special cases, such as Fisher's maximum likelihood, Wedderburn (1974)'s quasi-likelihood, Liang and Zeger (1986)'s generalized estimating equations, Huber (1964)'s M-estimation for robust inference, with possible extensions to semi-parametric and non-parametric models.

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

Table 4.1: Double division of outcome data on both the dimension of responses (into blocks) and sample size (into groups).

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

The rest of the chapter is organized as follows. Section 4.2 describes the DDIMM, with examples introduced in Section 4.3. Section 4.4 discusses large sample properties of the proposed DDIMM. Section 4.5 presents the main contribution of the chapter, a closed-form meta-estimator and its implementation in a parallel and scalable computational scheme. Section 4.6 illustrates the DDIMM's finite sample performance with simulations. Section 4.7 concludes with a discussion. Additional proofs and simulation results are deferred to Appendices E-G. An R package is also available.

4.2 Formulation

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

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

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

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

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

4.2.1 Double Data Split Procedure

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

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

Division into blocks is not restricted to the order of data entry: responses may be grouped according to pre-specified block memberships, according to, say, substantive scientific knowledge, such as functional regions of the brain. In this chapter, with no loss of generality, we use the order of data entry in the data division procedure. Further, DDIMM randomly splits the N independent subjects to form K disjoint subject groups $\{\boldsymbol{y}_{i,jk}, \boldsymbol{X}_{i,jk}\}_{i=1}^{n_k}$. Then each group has sample size n_k, $k = 1, \dots, K$, with $\sum_{k=1}^{K} n_k = N$. Refer to Table 4.1 for notation detail. For ease of exposition, we henceforth use the term "group" to refer to the division along subjects, and "block" to refer to the division along responses. We also use the term "block" to refer to the division along both responses and subjects.

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

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

$$\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta};\boldsymbol{y}_{jk},\boldsymbol{\zeta}_{jk}) = \mathbf{0}, \tag{4.1}$$

$$\boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk};\boldsymbol{y}_{jk},\boldsymbol{\theta}) = \mathbf{0}, \tag{4.2}$$

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

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

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

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

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

We define $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ as being "weakly regular" based on the above conditions (A.1) (i)-(iii) in which the defining properties of a regular inference function are applied to its mean; see Song (2007) Chapter 3.5 for a definition of regular inference functions. Additional conditions on the class $\mathcal{P}$ will be described throughout the chapter where appropriate. Within block (j,k), denote by $\widehat{\boldsymbol{\theta}}_{jk}$ and $\widehat{\boldsymbol{\zeta}}_{jk}$ the joint solution to (4.1) and (4.2), estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}_{jk}$ respectively. For notation purposes, let $\widehat{\boldsymbol{\theta}}_{list} = \mathbb{S}^{JK}(\widehat{\boldsymbol{\theta}}_{jk})$, $\widehat{\boldsymbol{\zeta}}_k = \mathbb{S}^{J}(\widehat{\boldsymbol{\zeta}}_{jk})$, and $\widehat{\boldsymbol{\zeta}}_{list} = \mathbb{S}^{JK}(\widehat{\boldsymbol{\zeta}}_{jk})$. Due to the homogeneity of $\boldsymbol{\theta}$, the next step is integration of the block-specific estimators $\widehat{\boldsymbol{\theta}}_{jk}$. By contrast, $\widehat{\boldsymbol{\zeta}}_{jk}$ remain heterogeneous and potentially high-dimensional. In the rest of the chapter, for convenience of notation, we suppress the dependence of

$\boldsymbol{\Psi}_{jk}$, $\boldsymbol{G}_{jk}$, $\boldsymbol{\psi}_{jk}$ and $\boldsymbol{g}_{jk}$ on $\boldsymbol{y}_{jk}$ and $\boldsymbol{y}_{i,jk}$:

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

4.2.2 Integration

Integrating block estimates $\widehat{\boldsymbol{\theta}}_{jk}$ into an estimator of $\boldsymbol{\theta}$, denoted by $\widehat{\boldsymbol{\theta}}_c$, will yield a more efficient estimate of $\boldsymbol{\theta}$. In the integration step, our intuition is to treat each system of equations $\mathbb{S}\left(\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta};\boldsymbol{\zeta}_{jk}),\boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk};\boldsymbol{\theta})\right) = \mathbf{0}$ as a "moment condition" on $\boldsymbol{\theta}$ contributed by block (j,k), $j = 1,\ldots,J$, $k = 1,\ldots,K$. Technically, we want to derive an estimator $\widehat{\boldsymbol{\theta}}_c$ of $\boldsymbol{\theta}$ that satisfies all JK moment conditions that effectively makes use of the JK estimates of $\boldsymbol{\theta}$ obtained from equations (4.1) and (4.2). To address the issue that $\boldsymbol{\theta}$ is over-identified by the JK moment conditions, we invoke Hansen (1982)'s seminal generalized method of moments (GMM) to combine the moment conditions that arise from each block. Another significant advantage of GMM is that it allows us to incorporate between-block dependencies, which cannot be easily done in classical meta-estimation. To this end, define the subject group indicator $\delta_i(k) = \mathbb{1}($subject i is in blocks (j,k) for some $k \in \{1,\ldots,K\}$ and for all $j = 1,\ldots,$ $J)$ for $i = 1,\ldots,N$, $k = 1,\ldots,K$. For subject i, let

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

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

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

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

$$\boldsymbol{T}_N(\boldsymbol{\theta},\boldsymbol{\zeta}) = \begin{pmatrix} \boldsymbol{\Psi}_N(\boldsymbol{\theta};\boldsymbol{\zeta}) \\ \boldsymbol{G}_N(\boldsymbol{\zeta};\boldsymbol{\theta}) \end{pmatrix}, \; \boldsymbol{\tau}_i(\boldsymbol{\theta},\boldsymbol{\zeta}) = \begin{pmatrix} \boldsymbol{\psi}_i(\boldsymbol{\theta};\boldsymbol{\zeta}) \\ \boldsymbol{g}_i(\boldsymbol{\zeta};\boldsymbol{\theta}) \end{pmatrix}. \tag{4.3}$$

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

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

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

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

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

Note that $\boldsymbol{T}_N(\boldsymbol{\theta},\boldsymbol{\zeta}) = \boldsymbol{0}$ has no unique solution because its dimension is bigger than the dimension of $\boldsymbol{\theta}$. To overcome this issue, we follow Hansen's GMM for over-identified parameters. Let $\boldsymbol{W}$ be the weight matrix in the GMM equation (4.4). Classical GMM theory states that any positive semi-definite matrix $\boldsymbol{W}$ can be used to guarantee consistency and asymptotic normality of the resulting estimator, and that an optimal choice of $\boldsymbol{W}$, corresponding to the inverse covariance of the estimating function $\boldsymbol{T}_N$ in (4.3), leads to an efficient GMM estimator. In our setting, a possible formulation for a GMM estimator of $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ is

$$(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c) = \arg\min_{\boldsymbol{\theta},\boldsymbol{\zeta}} Q_N(\boldsymbol{\theta}, \boldsymbol{\zeta}|\boldsymbol{W}), \text{ where} \tag{4.4}$$

$$Q_N(\boldsymbol{\theta}, \boldsymbol{\zeta}|\boldsymbol{W}) = \boldsymbol{T}_N^T(\boldsymbol{\theta}, \boldsymbol{\zeta})\boldsymbol{W}\boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}).$$

In (4.4), the weight matrix $\boldsymbol{W}$ is a positive semi-definite $(JKp + d) \times (JKp + d)$ matrix. The heterogeneity of $\boldsymbol{\zeta}$ allowed by the use of $\boldsymbol{G}_N$ can lead to theoretical and computational challenges due to the high-dimensionality of the parameter, a problem from which GEE2 also suffers. See Chan et al. (1998) and Carey et al. (1993) for a discussion on the computational burden of inverting large matrices in GEE2. Note that block-specific estimators $\widehat{\boldsymbol{\zeta}}_{list}$ are consistent; the only possible improvement from re-learning $\boldsymbol{\zeta}$ in an iterative procedure between $\widehat{\boldsymbol{\theta}}_c$ and $\widehat{\boldsymbol{\zeta}}_c$ is a gain in efficiency. This is not necessary since $\boldsymbol{\zeta}$ are parameters of secondary interest and their efficiency is in general not of interest. We will derive a closed-form meta-estimator of $\boldsymbol{\theta}$ that avoids re-learning of $\boldsymbol{\zeta}$ in Section 4.5.

Following the work of Hansen (1982), we define a particular instance of the estimator in (4.4) by specifying $\boldsymbol{W}$ as the inverse sample covariance of $\boldsymbol{T}_N$. We will show in Section 4.4 that this choice of $\boldsymbol{W}$ is optimal for the efficiency of the resulting estimator. Let $\widehat{\boldsymbol{V}}_N$ be the sample covariance of $\boldsymbol{T}_N(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$:

$$\widehat{\boldsymbol{V}}_N = \frac{1}{N}\sum_{i=1}^N \left(\boldsymbol{\tau}_i(\widehat{\boldsymbol{\theta}}_{list}, \widehat{\boldsymbol{\zeta}}_{list})\right)^{\otimes 2} = \frac{1}{N}\sum_{i=1}^N \begin{pmatrix} \boldsymbol{\psi}_i(\widehat{\boldsymbol{\theta}}_{list}; \widehat{\boldsymbol{\zeta}}_{list}) \\ \boldsymbol{g}_i(\widehat{\boldsymbol{\zeta}}_{list}; \widehat{\boldsymbol{\theta}}_{list}) \end{pmatrix}^{\otimes 2}, \tag{4.5}$$

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

$$(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt}) = \arg\min_{\boldsymbol{\theta},\boldsymbol{\zeta}} \boldsymbol{T}_N^T(\boldsymbol{\theta}, \boldsymbol{\zeta})\widehat{\boldsymbol{V}}_N^{-1}\boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}). \tag{4.6}$$

We assume that $\boldsymbol{W}$ and $\widehat{\boldsymbol{V}}_N$ are nonsingular; see Han and Song (2011) for optimal weighting matrix with QIF when the sample covariance is ill-defined. Before presenting large-sample properties of $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ and $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$ in Section 4.4, we demonstrate in Section 4.3 the flexibility of our framework through several important supervised learning methods.

4.3 Examples

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

4.3.1 Likelihood-based methods

Consider the multidimensional regression model $h(\boldsymbol{\mu}_{i,jk}) = \boldsymbol{X}_{i,jk}(\; \boldsymbol{\theta}^T \quad \boldsymbol{\beta}_{jk}^T \;)^T$, where $\boldsymbol{\mu}_{i,jk} = E(\boldsymbol{Y}_{i,jk} | \boldsymbol{X}_{i,jk}, \boldsymbol{\theta}, \boldsymbol{\beta}_{jk})$ is the mean vector of $\boldsymbol{Y}_{i,jk}$ given $\boldsymbol{X}_{i,jk}$, $\boldsymbol{\beta}_{jk}$, and the p-dimensional parameter $\boldsymbol{\theta}$ ($p \leq q$ the number of covariates, which may include an intercept), and h is a known component-wise link function. Let $\boldsymbol{\zeta}_{jk}$ be parameters of the second-order moments of $\boldsymbol{Y}_{i,jk}$, such as dispersion parameters, and parameters in $\boldsymbol{\beta}_{jk}$ (which may be empty). If the full likelihood of $\boldsymbol{Y}_{i,jk}$ is computationally tractable, $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ correspond to the score functions, and $\widehat{\boldsymbol{\theta}}_{jk}$ and $\widehat{\boldsymbol{\zeta}}_{jk}$ may be given by the maximum likelihood estimates (MLEs). DDIMM can be applied straightforwardly by following the procedure in Section 4.2.

If the full likelihood is computationally intractable or difficult to construct, one can instead use pseudo-likelihoods such as the pairwise composite likelihood. The pairwise composite likelihood, originally proposed by Lindsay (1988) and detailed in Varin et al. (2011), provides the following forms of the equations for (4.1) and (4.2):

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

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

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

4.3.2 Generalized estimating equations

More generally, Wedderburn (1974)'s quasi-likelihood is a popular alternative method of supervised learning that does not require a fully specified multidimensional likelihood; it receives a full treatment in Heyde (1997). Consider Liang and Zeger (1986)'s marginal mean model $h(\boldsymbol{\mu}_{i,jk}) = \boldsymbol{X}_{i,jk}(\; \boldsymbol{\theta}^T \quad \boldsymbol{\beta}_{jk}^T \;)^T$ for the analysis of longitudinal data, where $\boldsymbol{\mu}_{i,jk} = E(\boldsymbol{Y}_{i,jk} | \boldsymbol{X}_{i,jk}, \boldsymbol{\theta}, \boldsymbol{\beta}_{jk})$ is the marginal mean vector of serially correlated outcomes $\boldsymbol{Y}_{i,jk}$ given $\boldsymbol{X}_{i,jk}$, $\boldsymbol{\beta}_{jk}$, and the p-dimensional parameter $\boldsymbol{\theta}$ ($p \leq q$), and h is a known component-wise link function. In this setting, $\boldsymbol{\zeta}_{jk}$ consists of parameters in $\boldsymbol{\beta}_{jk}$ (which may be empty), parameters for the variances of $Y_{it,jk}$, $t = 1, \ldots, m_j$, and a nuisance parameter $\boldsymbol{\alpha}_{jk}$ which fully characterizes a working correlation matrix $\boldsymbol{R}_{jk}(\boldsymbol{\alpha}_{jk})$. In the case where $\boldsymbol{\beta}_{jk}$ is empty, the generalized estimating equation (GEE) proposed by Liang and Zeger (1986) yields the the kernel inference function $\boldsymbol{\psi}_{jk}(\boldsymbol{\theta}; \boldsymbol{\zeta}_{jk}) = \boldsymbol{D}_{i,jk}^T \boldsymbol{\Sigma}_{i,jk}^{-1} \boldsymbol{r}_{i,jk}$ in (A.1) (iii), where $\boldsymbol{D}_{i,jk} = \nabla_{\boldsymbol{\theta}} \boldsymbol{\mu}_{i,jk}$, $\boldsymbol{r}_{i,jk} = \boldsymbol{y}_{i,jk} - \boldsymbol{\mu}_{i,jk}$, and $\boldsymbol{\Sigma}_{i,jk} = \boldsymbol{A}_{i,jk} \boldsymbol{R}_{jk}(\boldsymbol{\alpha}_{jk}) \boldsymbol{A}_{i,jk}$, where $\boldsymbol{A}_{i,jk} = \operatorname{diag}\{(Var(Y_{it,jk}))^{1/2}\}_{t=1}^{m_j}$. In GEE2, $\boldsymbol{G}_{jk}$ in (4.2) is specified as another unbiased inference function satisfying (A.1) and (A.2). DDIMM provides a procedure for the application of distributed methods to high-dimensional longitudinal/clustered data.

4.3.3 M-estimation

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

4.3.4 Joint mean-variance modelling

Following Pan and Mackenzie (2003), one can jointly model the marginal means and covariances of the longitudinal responses with $h(\boldsymbol{\mu}_{i,jk}) = \boldsymbol{X}_{i,jk,1}\boldsymbol{\beta}$, $\log(\boldsymbol{\sigma}^2_{i,jk}) = \boldsymbol{X}_{i,jk,2}\boldsymbol{\lambda}$, and $\phi_{irt,jk} = \boldsymbol{X}_{irt,jk,3}\boldsymbol{\gamma}$ for $1 \le t < r \le m_j$, where h is a known component-wise link function, $\boldsymbol{\beta} \in \mathbb{R}^{q_1}$, $\boldsymbol{\lambda} \in \mathbb{R}^{q_2}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{q_3}$ are unconstrained parameters, $\boldsymbol{\mu}_{i,jk} = E(\boldsymbol{Y}_{i,jk}|\boldsymbol{X}_{i,jk,1}, \boldsymbol{\theta})$ and $\boldsymbol{X}_{i,jk,1} \in \mathbb{R}^{m_j \times q_1}$ a submatrix of $\boldsymbol{X}_{i,jk}$, $\boldsymbol{\sigma}^2_{i,jk} = \mathbb{S}\left(Var(Y_{ir,jk})\right)_{r=1}^{m_j}$ and $\boldsymbol{X}_{i,jk,2} \in \mathbb{R}^{m_j \times q_2}$ a submatrix of $\boldsymbol{X}_{i,jk}$, and $\phi_{irt,jk}$ are specified in Zhang et al. (2015a). Estimating functions $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ in (4.1) and (4.2) are given in detail in Zhang et al. (2015a). There is some choice depending on the problem considered as to whether $\boldsymbol{\theta} = \boldsymbol{\beta}$, $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\gamma})$, or $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma})$. In the first case, learning of variance parameters only helps improve estimation efficiency. This type of framework is widely applied in biomedical studies where the mean parameters are of primary interest. In the second case, learning of covariance parameters is of interest and $\boldsymbol{\beta}$ is treated as a nuisance parameter. This is the situation where prediction is of primary interest, such as in kriging in spatial data analysis. In the third case, $\boldsymbol{G}_{jk}$ is null, and learning of variance parameters is of interest to the investigator. This case occurs for example in the study of volatility for risk management in financial data analysis.

4.3.5 Marginal quantile regression for correlated data

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

Each of these five examples requires additional work to fully develop a divide-and-conquer strategy via DDIMM, including specific computational details. Here we only present the general framework with a high-level discussion that sheds light on DDIMM's promising generality and flexibility, and its coverage of a wide range of supervised learning methods. The theoretical results presented in Sections 4.4 and 4.5 are developed under a general unified framework of estimating functions that includes the above five examples as special cases.

4.4 Asymptotic Properties

In this section we assume that K and J are fixed; this assumption will be relaxed in Section 4.5. Let $n_{\min} = \min_{k=1,\ldots,K} n_k$ and $n_{\max} = \max_{k=1,\ldots,K} n_k$. Suppose $\boldsymbol{W} \overset{p}{\to} \boldsymbol{w}$ as $n_{\min} \to \infty$. In this section we study the asymptotic properties of the GMM estimator $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}_c})$ proposed in (4.4) and its optimal version proposed in (4.6). We assume throughout that subjects are monotonically allocated to subject groups; that is, as $n_{\min} \to \infty$, a subject cannot be reallocated to another group once it has been assigned to a subject group. Define the variability matrix of $\boldsymbol{\tau}_i(\boldsymbol{\theta}, \boldsymbol{\zeta})$ in (4.3) as

$$\boldsymbol{v}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = Var_{\boldsymbol{\theta}_0, \zeta_0} \left\{ \boldsymbol{\tau}_i(\boldsymbol{\theta}, \boldsymbol{\zeta}) \right\} = \begin{pmatrix} \boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) & \boldsymbol{v}_{\boldsymbol{\psi g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) \\ \boldsymbol{v}^T_{\boldsymbol{\psi g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) & \boldsymbol{v}_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) \end{pmatrix}$$

where $\boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = Var_{\boldsymbol{\theta}_0, \zeta_0} \left\{ \boldsymbol{\psi}_i(\boldsymbol{\theta}; \boldsymbol{\zeta}) \right\}$, $\boldsymbol{v}_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = Var_{\boldsymbol{\theta}_0, \zeta_0} \left\{ \boldsymbol{g}_i(\boldsymbol{\zeta}; \boldsymbol{\theta}) \right\}$, and $\boldsymbol{v}_{\boldsymbol{\psi g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = E_{\boldsymbol{\theta}_0, \zeta_0} \left\{ \boldsymbol{\psi}_i(\boldsymbol{\theta}; \boldsymbol{\zeta}) \boldsymbol{g}_i^T(\boldsymbol{\zeta}; \boldsymbol{\theta}) \right\}$. Let the sensitivity matrix of $\boldsymbol{\tau}_i(\boldsymbol{\theta}, \boldsymbol{\zeta})$ be

$$\boldsymbol{s}(\boldsymbol{\theta}, \boldsymbol{\zeta}) = -\nabla_{\boldsymbol{\theta}, \zeta} E_{\boldsymbol{\theta}_0, \zeta_0} \boldsymbol{\tau}_i(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \begin{pmatrix} \boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) & \boldsymbol{s}^{\zeta}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) \\ \boldsymbol{s}^{\boldsymbol{\theta}}_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) & \boldsymbol{s}^{\zeta}_{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) \end{pmatrix}, \text{ where} \tag{4.7}$$

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

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

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

Following Theorem 3.4 of Song (2007), block-specific estimates $\widehat{\boldsymbol{\theta}}_{jk}$ and $\widehat{\boldsymbol{\zeta}}_{jk}$ are consistent given assumptions (A.1). Consistency and asymptotic normality of the GMM estimator $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ in (4.4) have been established by Hansen (1982) and, more generally, by Newey and McFadden (1994). To establish consistency and asymptotic normality for the combined estimator $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$, we consider the following additional regularity conditions:

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

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

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

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

$$\sup_{\|(\boldsymbol{\theta},\boldsymbol{\zeta})-(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\| \le \delta_N} \frac{N^{1/2}}{1+N^{1/2}\|(\boldsymbol{\theta},\boldsymbol{\zeta})-(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\|} \left\| \boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}) - \boldsymbol{T}_N(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) - E_{\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0} \boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}) \right\| \xrightarrow{p} 0.$$

Theorems IV.1 and IV.2 do not require the differentiability of $\boldsymbol{T}_N$ and Q_N. Instead, they require the differentiability of their population versions, and that $\boldsymbol{T}_N$ behaves "nicely" in a neighbourhood of $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$, in the sense that higher order terms are asymptotically ignorable. The following two theorems state the consistency and asymptotic normality of $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ given in (4.4) under Newey and McFadden's mild moment conditions given in (A.3) and (A.4).

Theorem IV.1 (Consistency of $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$)**.** *Suppose assumptions (A.1)-(A.3) hold with* $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ *defined in* (4.4). *Then* $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c) \xrightarrow{p} (\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ *as* $n_{\min} \to \infty$.

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

Theorem IV.2 (Asymptotic normality of $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$**.** *Suppose assumptions (A.1)-(A.4) hold with* $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ *defined in* (4.4). *Then as* $n_{\min} \to \infty$,

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

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

The proof of Theorem IV.2 follows easily from Theorem 7.2 in Newey and McFadden (1994) and Theorem IV.1 above. We note that requiring K to be finite implies that N and $n_{\min}$ are asymptotically of the same order. We will relax this assumption in Section 4.5. Conditions (A.3) and (A.4) allow us to consider non-differentiable kernel inference functions in the block (j, k) analysis, extending Hector and Song (2020a)'s DIMM beyond CL kernel inference functions. We can now consider quantile regression, M-estimation, and more general estimation functions than the score or CL score equations.

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

Theorem IV.3 (Test of over-identifying restrictions)**.** *Suppose assumptions (A.1)-(A.4) hold with* $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ *defined in* (4.4)*. Then as* $n_{\min} \to \infty$*,* $NQ_N(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c | \boldsymbol{W}) \xrightarrow{d} \chi^2_{(JK-1)p}$*.*

The proof of Theorem IV.3 can be carried out with some minor changes from that of

Theorem 3 in Hector and Song (2020a). The GMM provides an objective function with which to do model selection even when the block analyses do not, such as with GEE and M-estimation. In the following, Theorem IV.4 and Corollary IV.1 show our combined GMM estimator derived from (4.6) is optimal in the sense defined by Hansen (1982): it has an asymptotic covariance matrix at least as small (in terms of the Loewner ordering) as any other estimator exploiting the same over-identifying restrictions. We refer to this property as "Hansen optimality".

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

Proof. The proof uses the consistency of the block estimators and the Central Limit Theorem, and is given in Appendix F. □

Corollary IV.1 (Hansen optimality)**.** *Suppose assumptions (A.1)-(A.4) hold with* $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ *defined in* (4.4)*. Let* $\boldsymbol{j}(\boldsymbol{\theta}, \boldsymbol{\zeta})$ *as given in Theorem IV.2. Then as* $n_{\min} \to \infty$,

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

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

Remark 1. We discuss efficiency for selected examples in Section 4.3.

(i) In Example 4.3.1, when the score function exists and satisfies mild regularity conditions, its variance is given by Fisher's information, and is a lower bound on the variances of estimating functions for $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$. This, coupled with Hansen's optimality, means that using the score function for $\boldsymbol{\psi}_{jk}$ and $\boldsymbol{g}_{jk}$ yields an efficient estimator of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$. In an unpublished dissertation, Jin (2011) studied the efficiency of the pairwise composite likelihood under different correlation structures. Hector and Song (2020a) showed empirically that the efficiency of the pairwise composite likelihood propagates to the combined estimator.

(ii) In Example 4.3.2, it is known that the GEE estimator $\widehat{\boldsymbol{\theta}}_{jk}$ in Example 4.3.2 is semi-parametrically efficient when the correlation structure of the response $\boldsymbol{y}_{i,jk}$ is correctly specified. This, coupled with Hansen's optimality, means that using GEE's for $\boldsymbol{\psi}_{jk}$ with the correct correlation structure of the response $\boldsymbol{y}_{i,jk}$ yields an efficient estimator of $\boldsymbol{\theta}$.

Remark 2. The GMM estimator $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$ can be interpreted as maximizing an extension of the confidence distribution density, as discussed in Hector and Song (2020a). The confidence distribution approach is used for independent data in Xie and Singh (2013). Briefly, we can define the confidence estimating function (CEF) as $U(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \Phi(N^{1/2} \widehat{\boldsymbol{V}}_N^{-1/2} \boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}))$, where $\Phi(\cdot)$ is the $(JKp + d)$-variate standard normal distribution function. Clearly, $U(\boldsymbol{\theta}, \boldsymbol{\zeta})$ is asymptotically standard uniform at $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ as long as $\widehat{\boldsymbol{V}}_N$ is a consistent estimator of the covariance of $\boldsymbol{T}_N$. Then we can define the density of the CEF as $u(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \phi(N^{1/2} \widehat{\boldsymbol{V}}_N^{-1/2} \boldsymbol{T}_N(\boldsymbol{\theta}, \boldsymbol{\zeta}))$. Maximizing $u(\boldsymbol{\theta}, \boldsymbol{\zeta})$ with respect to $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ yields the minimization defined in (4.6).

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

4.5 Distributed Estimation and Inference

Despite the computational gains offered by the divide-and-combine procedure and the GMM estimator, iteratively finding the solution $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$ (or $(\widehat{\boldsymbol{\theta}}_{c}, \widehat{\boldsymbol{\zeta}}_{c})$) to (4.6) can be slow due to the high-dimensionality of parameter $\boldsymbol{\zeta}$ and the need to repeatedly evaluate $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$. To overcome this computational bottleneck, we propose a meta-estimator derived from (4.6) that delivers a closed-form estimator via a linear function of block estimates $(\widehat{\boldsymbol{\theta}}_{list}, \widehat{\boldsymbol{\zeta}}_{list})$. We define the DDIMM estimator for $(\boldsymbol{\theta}, \boldsymbol{\zeta})$:

$$\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{DDIMM} \\ \widehat{\boldsymbol{\zeta}}_{DDIMM} \end{pmatrix} = \left(\sum_{k=1}^{K} \sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i} \right)^{-1} \sum_{k=1}^{K} \sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik} \\ \widehat{\boldsymbol{\zeta}}_{list} \end{pmatrix}. \tag{4.8}$$

where $\widehat{\boldsymbol{C}}_{k,i}$ is a function of sample variability and sensitivity matrices and block-specific estimators $\widehat{\boldsymbol{\theta}}_{jk}$ and $\widehat{\boldsymbol{\zeta}}_{jk}$ defined in detail in Section 4.5.1. If we do not plan to conduct inference for $\boldsymbol{\zeta}$, which is treated as a nuisance parameter, taking $\left[\widehat{\boldsymbol{C}}^{-1}\right]_{p:}$ to be rows 1 to p of matrix $(\sum_{k=1}^{K} \sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i})^{-1}$ leads to the closed-form estimator of $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}}_{DDIMM} = \left[\widehat{\boldsymbol{C}}^{-1}\right]_{p:} \sum_{k=1}^{K} \sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{ik}^T & \widehat{\boldsymbol{\zeta}}_{list}^T \end{pmatrix}^T. \tag{4.9}$$

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

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

Note that the uppercase $\boldsymbol{S}$ denotes the sample sensitivity matrix, and the lower-case $\boldsymbol{s}$ denotes the population sensitivity matrix. Let $\widehat{\boldsymbol{S}}_{jk} = \boldsymbol{S}_{jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk})$ and similarly define $\widehat{\boldsymbol{S}}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}$, $\widehat{\boldsymbol{S}}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}$, $\widehat{\boldsymbol{S}}^{\boldsymbol{\theta}}_{\boldsymbol{g}_{jk}}$ and $\widehat{\boldsymbol{S}}^{\boldsymbol{\zeta}}_{\boldsymbol{g}_{jk}}$. Sensitivity formulas are summarized in Appendix E.

The DDIMM estimator in (4.9) can be implemented in a fully parallelized and scalable computational scheme, where only one pass through each block of data is required. The block analyses are run on parallel CPUs, and return the values of summary statistics $\{\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk}, \boldsymbol{\psi}_{i,jk}(\widehat{\boldsymbol{\theta}}_{jk}; \widehat{\boldsymbol{\zeta}}_{jk}), \boldsymbol{g}_{i,jk}(\widehat{\boldsymbol{\zeta}}_{jk}; \widehat{\boldsymbol{\theta}}_{jk}), \widehat{\boldsymbol{S}}_{jk}\}_{j,k=1}^{J,K}$ to the main computing node, which computes $\widehat{\boldsymbol{\theta}}_{DDIMM}$ in (4.9) in one step.

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

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

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

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

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

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

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

$$\widehat{\boldsymbol{C}}_{k,i} = \begin{pmatrix} \sum\limits_{j=1}^{J} \widehat{\boldsymbol{A}}^{\boldsymbol{\theta}}_{k,ij} & \mathbf{0}_{p\times D^{ik}} & \sum\limits_{j=1}^{J} \widehat{\boldsymbol{A}}^{\boldsymbol{\zeta}}_{k,ij} & \mathbf{0}_{p\times(d-d_{ik}-D^{ik})} \\ \multicolumn{4}{c}{\mathbf{0}_{D^{k}\times(p+d)}} \\ \widehat{\boldsymbol{B}}^{\boldsymbol{\theta}}_{k,i1} & \mathbf{0}_{d_{1k}\times D^{ik}} & \widehat{\boldsymbol{B}}^{\boldsymbol{\zeta}}_{k,i1} & \mathbf{0}_{d_{1k}\times(d-d_{ik}-D^{ik})} \\ & & \vdots & \\ \widehat{\boldsymbol{B}}^{\boldsymbol{\theta}}_{k,iJ} & \mathbf{0}_{d_{Jk}\times D^{ik}} & \widehat{\boldsymbol{B}}^{\boldsymbol{\zeta}}_{k,iJ} & \mathbf{0}_{d_{Jk}\times(d-d_{ik}-D^{ik})} \\ \multicolumn{4}{c}{\mathbf{0}_{(d-d_{k}-D^{k})\times(p+d)}} \end{pmatrix}. \tag{4.10}$$

4.5.2 Asymptotic results for K and J fixed

In this section we assume that K and J are fixed, which will be relaxed in Sections 4.5.3 and 4.5.4. Recall that we assume subjects are monotonically allocated to subject groups: as $n_{\min} \to \infty$, a subject cannot be reallocated to another group once it has been assigned to a subject group. Consider the following condition:

(A.5) For each $j = 1, \ldots, J$, $k = 1, \ldots, K$, $\widehat{\boldsymbol{\theta}}_{jk} = \boldsymbol{\theta}_0 + O_p(n_k^{-1/2})$ and $\widehat{\boldsymbol{\zeta}}_{jk} = \boldsymbol{\zeta}_{jk0} + O_p(n_k^{-1/2})$.

For any $\delta_N \to 0$,

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

Consequently, some large-sample results can be established which are helpful in studying the asymptotic behaviour of $\widehat{\boldsymbol{\theta}}_{DDIMM}$.

Lemma IV.1. *Suppose assumptions (A.1), (A.2) and (A.5) hold. Then we have consistent estimation of information matrices:*

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

$$\widehat{\boldsymbol{S}}_{jk} = \boldsymbol{s}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) + O_p(n_k^{-1/2}) \ \ \textit{for each } j, k, \ \textit{and}$$

$$\frac{1}{N^2}\sum_{k=1}^{K}\sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i} = \widehat{\boldsymbol{S}}^T \widehat{\boldsymbol{V}}_N^{-1} \widehat{\boldsymbol{S}} = \boldsymbol{j}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p(N^{-1/2}),$$

$$\textit{where } \widehat{\boldsymbol{S}} = \begin{pmatrix} \mathbb{S}\left(\frac{n_k}{N}\widehat{\boldsymbol{S}}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}\right)_{j=1,k=1}^{J,K} & diag\left\{\frac{n_k}{N}\widehat{\boldsymbol{S}}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}\right\}_{j=1,k=1}^{J,K} \\ \mathbb{S}\left(\frac{n_k}{N}\widehat{\boldsymbol{S}}^{\boldsymbol{\theta}}_{g_{jk}}\right)_{j=1,k=1}^{J,K} & diag\left\{\frac{n_k}{N}\widehat{\boldsymbol{S}}^{\boldsymbol{\zeta}}_{g_{jk}}\right\}_{j=1,k=1}^{J,K} \end{pmatrix}.$$

Proof. A detailed proof is given in Appendix F. □

We show in Theorem IV.5 that the proposed closed-form estimator $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$ in (4.8) is consistent and asymptotically normally distributed.

Theorem IV.5. *Suppose assumptions (A.1), (A.2) and (A.5) hold. Let $\boldsymbol{j}(\boldsymbol{\theta}, \boldsymbol{\zeta})$ as given in Theorem IV.2. As $n_{\min} \to \infty$,*

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

Proof of Theorem IV.5: Here we present major steps, with all necessary details available in Appendix F. First, we show that $\widehat{\boldsymbol{\theta}}_{DDIMM}$ and $\widehat{\boldsymbol{\zeta}}_{DDIMM}$ are consistent. Define

$$\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{i=1}^{J} n_k^2 \widehat{\boldsymbol{C}}_{k,i}\begin{pmatrix} \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{ik} \\ \boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}_{list} \end{pmatrix}. \tag{4.11}$$

By definition, $\lambda(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM}) = \mathbf{0}$. As shown in Lemma F.0.0.1 in Appendix F, $\lambda(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) \xrightarrow{p} \mathbf{0}$ as $n_{\min} \to \infty$. Given that $\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}}\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})$ exists and is nonsingular, for some $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$ between $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$ and $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$, the first-order Taylor expansion leads to

$$\lambda(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM}) - \lambda(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) = \nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}} \lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})|_{\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{DDIMM} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{DDIMM} - \boldsymbol{\zeta}_0 \end{pmatrix}, \tag{4.12}$$

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

Now we derive the distribution of $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$. With a slight abuse of notation, let $\widehat{\boldsymbol{\theta}}_{list} - \boldsymbol{\theta}_0 = \mathbb{S}^{JK}\left(\widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0\right)$. We show in Lemma F.0.0.2 in Appendix F that

$$\begin{pmatrix} \boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}_0; \boldsymbol{\zeta}_{jk0}) \\ \boldsymbol{G}_{jk}(\boldsymbol{\zeta}_{jk0}; \boldsymbol{\theta}_0) \end{pmatrix} = \widehat{\boldsymbol{S}}_{jk} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix} + O_p(n_k^{-1}). \tag{4.13}$$

Recall the form of $\boldsymbol{T}_N$ in (4.3). By the Central Limit Theorem, $N^{1/2}\boldsymbol{T}_N(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{v}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0))$. Then with $\widehat{\boldsymbol{S}}$ defined in Lemma IV.1, it follows from equation (4.13) that

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

Moreover, by Lemma IV.1 and Slutsky's theorem we have:

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

Using the fact that the sum of jointly (asymptotically) Normal variables is (asymptotically) normal, by Lemma IV.1 and Slutsky's theorem again, we have

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

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

This key theorem allows us to use $\widehat{\boldsymbol{\theta}}_{DDIMM}$, which is more computationally attractive than $\widehat{\boldsymbol{\theta}}_{opt}$ defined in (4.6), without sacrificing any of the nice asymptotic properties for inference. Additionally, it follows easily from Theorem IV.5 that, under suitable conditions, the closed-form estimator $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$ in (4.8) has the same asymptotic distribution as and is asymptotically equivalent to the GMM estimator $\widehat{\boldsymbol{\theta}}_{opt}$ in (4.6).

Corollary IV.2. *Suppose assumptions (A.1)-(A.5) hold with $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$ defined in* (4.6)*. Then $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$ and $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$ are asymptotically equivalent: as $n_{\min} \to \infty$,*

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

Proof. A detailed proof is given in Appendix F. □

The computation of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ in (4.9) relies solely on block-specific estimators $(\widehat{\boldsymbol{\theta}}_{list}, \widehat{\boldsymbol{\zeta}}_{list})$ and values of summary statistics from each block. To guarantee the appropriate asymptotic distribution of $\widehat{\boldsymbol{\theta}}_{DDIMM}$, we assume in condition (A.5) that these block-specific estimators are $N^{1/2}$ consistent estimators of the true values, which restricts the scope of possible block-specific inference methods. For inference methods not satisfying this $N^{1/2}$ consistency in condition (A.5), it is still possible to use $\widehat{\boldsymbol{\theta}}_{opt}$ in (4.6).

4.5.3 Asymptotic results for diverging K with J fixed

We show in Theorem IV.6 that the asymptotic distribution of $(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM})$ remains unchanged as the number of subject groups K grows with the sample size.

Theorem IV.6. *Suppose $N^{\delta - 1/2}K$ is bounded as $n_{\min} \to \infty$ for a positive constant $\delta < \frac{1}{2}$, and assumptions (A.1), (A.2) and (A.5) hold. Let $\boldsymbol{H} \in \mathbb{R}^{h \times (p+d)}$ a matrix of rank $r \in \mathbb{N}$, $h \in \mathbb{N}$, $r \leq h$, with finite maximum singular value $\bar{\sigma}(\boldsymbol{H}) < \infty$. Let $\boldsymbol{j}(\boldsymbol{\theta}, \boldsymbol{\zeta})$ as given in Theorem IV.2. Then, as $n_{\min} \to \infty$, we show that the limiting value*

$\boldsymbol{j}_{\boldsymbol{H}}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)$ *of* $\boldsymbol{H}\boldsymbol{j}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\boldsymbol{H}^T$ *is a positive semi-definite and symmetric variance matrix, and that*

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

Proof [Proof of Theorem IV.6] Here we present major steps, with all necessary details available in Appendix F. First, we know that $\|\boldsymbol{H}\| \leq r\bar{\sigma}(\boldsymbol{H})$. Let $\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})$ defined by (4.11), such that $\lambda(\widehat{\boldsymbol{\theta}}_{DDIMM}, \widehat{\boldsymbol{\zeta}}_{DDIMM}) = \mathbf{0}$. We show in Lemma F.0.0.3 in Appendix F that $\|\lambda(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\| = O_p(N^{-1/2-\delta}n_{\max}^{1/2})$ and $\left\|\{\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}}\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})\}^{-1}\right\| = O_p\left(N^{1/2+\delta}n_{\max}^{-1}\right)$. From the first-order Taylor expansion in (4.12), we have

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

Then $\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_{DDIMM}^T, \widehat{\boldsymbol{\zeta}}_{DDIMM}^T)^T - \boldsymbol{H}(\boldsymbol{\theta}_0^T, \boldsymbol{\zeta}_0^T)^T \xrightarrow{p} \mathbf{0}$ as $n_{\min} \to \infty$.

To derive the distribution of $\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_{DDIMM}^T, \widehat{\boldsymbol{\zeta}}_{DDIMM}^T)^T$, first consider an arbitrary $k \in \{1, \dots, K\}$. For convenience, denote

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

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

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

$$\boldsymbol{j}_k(\boldsymbol{\theta}, \boldsymbol{\zeta}_k) = \boldsymbol{s}_k^T(\boldsymbol{\theta}, \boldsymbol{\zeta})\boldsymbol{v}_k^{-1}(\boldsymbol{\theta}, \boldsymbol{\zeta}_k)\boldsymbol{s}_k(\boldsymbol{\theta}, \boldsymbol{\zeta}_k).$$

By (4.13) in the proof of Theorem IV.5, Lemma IV.1, and Slutsky's theorem,

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

Note that the above vectors are independent for $k=1,\ldots,K$. We establish in Lemma F.0.0.4 in Appendix F that, for some affine transformation matrices $\boldsymbol{E}_k$, $k=1,\ldots,K$, of $\mathbf{0}$'s and $\mathbf{1}$'s,

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

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

where $n_k^{1/2}\boldsymbol{Z}_k\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\boldsymbol{j}_k^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0})\right)$. It is clear that $\boldsymbol{j}(\boldsymbol{\theta},\boldsymbol{\zeta})=\sum_{k=1}^K(n_k/N)$ $\boldsymbol{E}_k\boldsymbol{j}_k(\boldsymbol{\theta},\boldsymbol{\zeta}_k)\boldsymbol{E}_k^T$. Since $\boldsymbol{E}_k$ has finitely many 1's, $\|\boldsymbol{E}_k\|$ is bounded. Since $\|\boldsymbol{j}_k(\boldsymbol{\theta},\boldsymbol{\zeta}_k)\|$ is also bounded, $\|\boldsymbol{j}(\boldsymbol{\theta},\boldsymbol{\zeta})\|=O(Kn_{\max}N^{-1})=O(1)$. $\boldsymbol{j}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)$ is positive semi-definite and symmetric, implying that $\boldsymbol{H}\boldsymbol{j}^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\boldsymbol{H}^T$ is also positive semi-definite and symmetric. Following the monotone convergence theorem, we can write $\boldsymbol{H}\boldsymbol{j}^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)\boldsymbol{H}^T\to\boldsymbol{j}_{\boldsymbol{H}}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)$, where $\boldsymbol{j}_{\boldsymbol{H}}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_0)$ exists and is a proper variance matrix.

Using the fact that $\lambda(\widehat{\boldsymbol{\theta}}_{DDIMM},\widehat{\boldsymbol{\zeta}}_{DDIMM})=\mathbf{0}$ and $K=O(N^{1/2-\delta})$, we show in Lemma F.0.0.5 in Appendix F that $N^{1/2}\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_{DDIMM}-\boldsymbol{\theta}_0,\widehat{\boldsymbol{\zeta}}_{DDIMM}-\boldsymbol{\zeta}_0)$ can be rewritten as

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

Since $O_p(n_{\max}^{1/2}N^{-1/2-\delta})=o_p(1)$ and $O_p(N^{-\delta})=o_p(1)$, it follows as in the proof of Theorem IV.5 that as $n_{\min}\to\infty$,

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

In practice, Theorem IV.6 suggests that we can tune our choice of K and $n_{\min}$ to attain the desired trade-off between inference and computational speed: smaller K and larger $n_{\min}$ will slow computations but improve estimation and asymptotic normality, whereas larger K and smaller $n_{\min}$ will speed computations but worsen estimation and asymptotic normality.

4.5.4 Asymptotic results for diverging K and J

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

(A.6) Stationarity: for each $M^* \in \mathbb{N}$ and each $(M^* + 1)$-dimensional measurable set B a subset of the sample space of $\boldsymbol{Y}$, the distribution of $\boldsymbol{Y}_i$ satisfies $P\{(Y_{i,r}, \ldots, Y_{i,r+M^*}) \in B\} = P\{(Y_{i,0}, \ldots, Y_{i,M^*}) \in B\}$ for every $r \in \mathbb{N}$.

(A.7) Let $\boldsymbol{C}_{k,i}$ be the version of $\widehat{\boldsymbol{C}}_{k,i}$ in (4.10) evaluated at the true values $\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}$. For $k = 1, \ldots, K$, $i = 1, \ldots, J$, $(\sum_{l=1}^{K} \sum_{j=1}^{J} n_l^2 \boldsymbol{C}_{l,j})^{-1} n_k^2 \boldsymbol{C}_{k,i} = O_p(N^{-\delta_1})$ for a constant $0 \leq \delta_1 \leq 1/2$. This can be thought of as a type of Lindeberg condition.

(A.8) Conditions required for asymptotically normal distribution and efficiency of the GMM estimator $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$; see Theorem 5.4 in Donald et al. (2003) and the spanning condition in Newey (2004). See Newey (2004) for related work on semiparametric efficiency of the GMM estimator as the number of moment conditions goes to infinity.

Remark 3. Condition (A.6) is typical for consistency and asymptotic normality of the GMM estimator $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt})$, following Hansen (1982) and Newey (2004). It is a typical condition for the application of the central limit theorem to stochastic processes, i.e. to infinite dimensional random vectors. Additionally, in order to make statements about convergence in probability, (A.6) is required to ensure a valid joint probability distribution as the dimension M increases.

Remark 4. Condition (A.7) ensures the covariance of the outcome $\boldsymbol{Y}_i$ is appropriately controlled as $M \to \infty$. Alternative conditions may be considered, such as α-mixing (Bradley (1985)), ρ-mixing (Peligrad (1986)), or ϕ-mixing (Peligrad (1986)), but this is beyond the scope of this chapter. Condition (A.7) can be simplified for the case where $n_k = n$ for all $k = 1, \ldots, K$. Then (A.7) becomes $(\sum_{l=1}^{K} \sum_{j=1}^{J} \boldsymbol{C}_{l,j})^{-1} \boldsymbol{C}_{k,i} = O_p(N^{-\delta_1})$.

In Theorem IV.7 we show the consistency and asymptotic normality of the DDIMM estimator as K and J diverge to ∞.

Theorem IV.7. *Suppose $N^{-\delta_2} n_{\min}$ and $N^{\delta_3 - 1/2} KJ$ are bounded as $n_{\min} \to \infty$ for constants $0 \le \delta_2 \le 1$ and $0 < \delta_3 < 1/2$ such that $\delta_3 + \delta_1 + \delta_2/2 > 1$. Suppose assumptions (A.1), (A.2), and (A.5)-(A.8) hold. Let $\boldsymbol{H} \in \mathbb{R}^{h \times (p+d)}$ a matrix of rank $r \in \mathbb{N}$, $h \in \mathbb{N}$, $r \le h$, with finite maximum singular value $\bar{\sigma}(\boldsymbol{H}) < \infty$. Let $\boldsymbol{j}_{\boldsymbol{H}}(\boldsymbol{\theta}, \boldsymbol{\zeta})$ as given in Theorem IV.6. Then as $n_{\min} \to \infty$,*

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

Proof Write

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

To show the asymptotic distribution of the left-hand side, it is sufficient to show that $\boldsymbol{H}(\widehat{\boldsymbol{\theta}}^T_{DDIMM} - \widehat{\boldsymbol{\theta}}^T_{opt}, \widehat{\boldsymbol{\zeta}}^T_{DDIMM} - \widehat{\boldsymbol{\zeta}}^T_{opt})^T = o_p(N^{-1/2})$.

Given the assumptions of the theorem, we have the asymptotic distribution of $(\widehat{\boldsymbol{\theta}}_{opt}, \widehat{\boldsymbol{\zeta}}_{opt,ik})$ and $(\widehat{\boldsymbol{\theta}}_{ik}, \widehat{\boldsymbol{\zeta}}_{ik})$: both are consistent estimators of $\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{ik0}$ and asymptotically normally distributed with rates $N^{-1/2}$ and $n_k^{-1/2}$ respectively. Then for each $k \in \{1, \dots, K\}$,

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

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

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

4.6 Simulations

In this section we consider two sets of simulations to examine the performance of the closed-form estimator $\widehat{\boldsymbol{\theta}}_{DDIMM}$ under the linear regression setting $\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\theta}$, where $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\theta})$ and $\boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\theta}, \boldsymbol{\Sigma})$. The first set illustrates the finite sample performance and properties in Theorem IV.5 of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ with fixed sample size N, varying number of subject groups K, varying dimensions M of $\boldsymbol{Y}$, and fixed number of response blocks J. The second set of simulations illustrates the performance and properties in Theorem IV.7 of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ with growing sample size N and response dimension M of $\boldsymbol{Y}$, and varying number of subjects groups K and response blocks J. In both settings, covariates consist of an intercept and two independently simulated M-dimensional multivariate normal variables, and the true value of $\boldsymbol{\theta}$ is set to $\boldsymbol{\theta}_0 = (0.3, 0.6, 0.8)^T$. Simulations are conducted using R software on a standard Linux cluster.

We describe the first set of simulations. We specify $\boldsymbol{\Sigma} = \boldsymbol{S} \otimes \boldsymbol{A}$ with nested correlation structure, where $\otimes$ denotes the Kronecker product, $\boldsymbol{A}$ is an AR(1) covariance matrix with standard deviation $\sigma = 4$ and correlation $\rho = 0.8$, and $\boldsymbol{S}$ is a randomly simulated $J \times J$ positive-definite matrix. We consider varying dimensions M of $\boldsymbol{Y}$ with fixed $J = 5$, and a fixed sample size $N = 5,000$ with varying $K = 1, 2, 5$. We consider two supervised learning procedures: the pairwise composite likelihood using our own package, and the GEE using R package `geepack` and our own package (see Supplemental Material). With each procedure, we fit the model with an AR(1) working block correlation structure. Results for the GEE

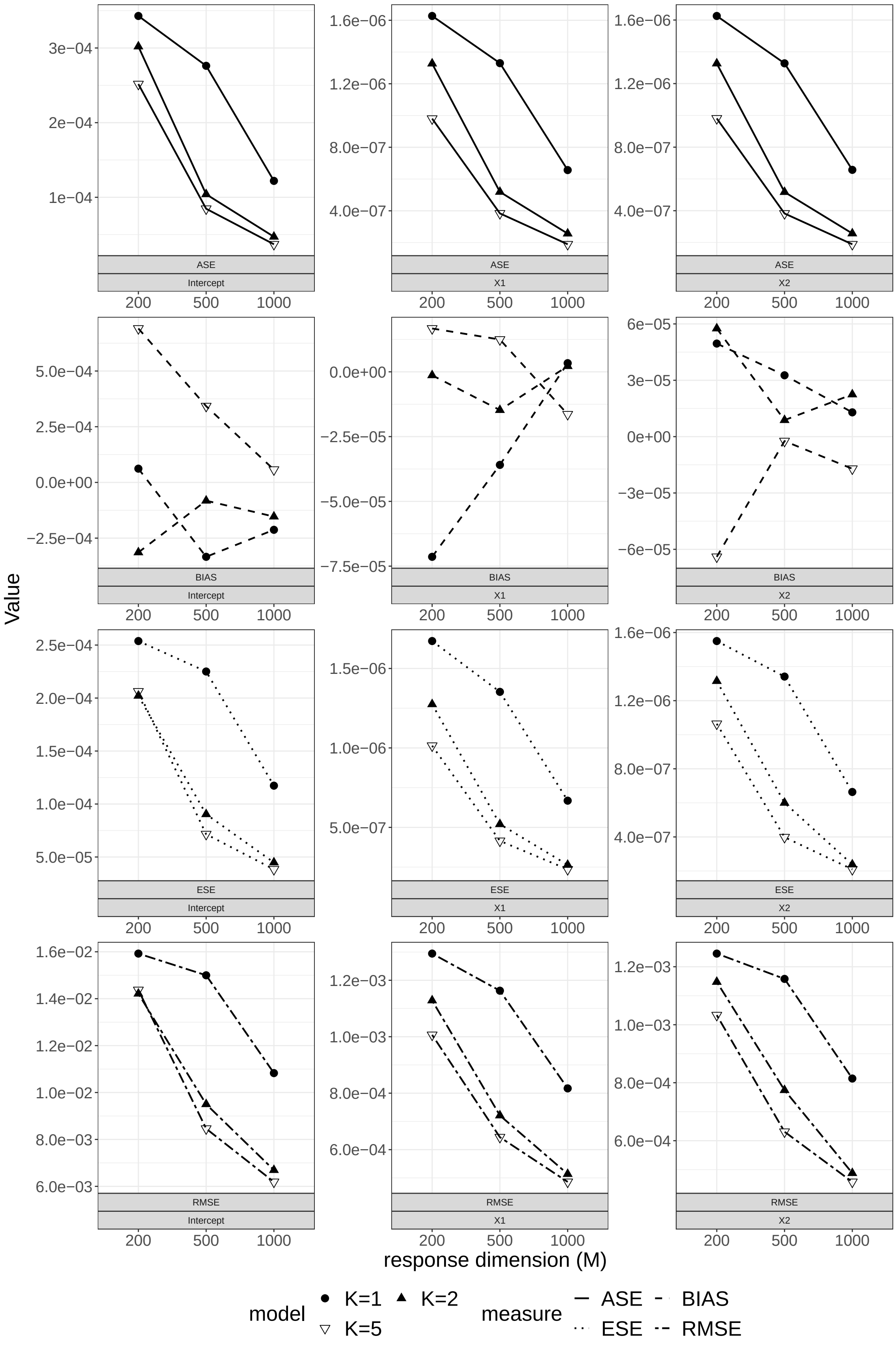


Figure 4.1: Plot of simulation metrics for GEE, averaged over 1,000 simulations.

are in Figure 4.1; results for the pairwise composite likelihood (CL) are in Appendix F. We see that the mean asymptotic standard error (ASE) of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ approximates the empirical standard error (ESE) for all models, with slight variations due to the type of covariates simulated. This means the covariance formula in Theorem IV.5 is correct. Additionally, $\widehat{\boldsymbol{\theta}}_{DDIMM}$ appears consistent since root mean squared error (RMSE), ASE and ESE are approximately equal. Moreover, we notice the ASE of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ decreases as the response dimension M increases. This makes intuitive sense, since an increase in M corresponds to an increase in overall number of observations, resulting in increased power. We also see a decrease in the ASE as the number of groups increases. This is due to the heterogeneity of block covariance parameters. Lastly, we observe from Table 4.2 that the mean CPU time is very fast for the GEE, and decreases substantially as the number of subject groups increases.

Response dimension	Number of subject groups		
	K=1	K=2	K=5
M=200	45	23	11
M=500	351	184	87
M=1,000	1956	961	417

Table 4.2: Mean CPU time in seconds for each setting with the GEE block analysis, averaged over 1,000 simulations. Mean CPU time is computed as the maximum CPU time taken over parallelized block analyses added to the CPU time taken by the rest of the procedure.

We describe the second set of simulations, where we consider diverging sample size N and response dimension M, and diverging number of subject groups K and response blocks J. We consider two settings: in Setting I, we let the sample size $N = 5,000$ with number of response groups $K = 1$, and let response dimension $M = 4,500$ with number of response blocks $J = 6$; in Setting II, we let the sample size $N = 10,000$ with number of response groups $K = 2$, and let response dimension $M = 9,000$ with number of response blocks $J = 12$. Responses are simulated from a Multivariate Normal distribution with AR(1) covariance structure, with standard deviation $\sigma = 6$ and correlation $\rho = 0.8$. This means there are no heterogeneous block parameters, so we expect a slightly less efficient estimator since there is less variability in the outcome. We learn mean and covariance parameters using GEE with an AR(1) working block correlation structure. Mean bias (BIAS), RMSE, ESE and ASE of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ are in Table 4.3. We observe that RMSE, ESE and ASE are very close, indicating appropriate estimation of $\widehat{\boldsymbol{\theta}}_{DDIMM}$ and its covariance in Theorem IV.7. We also confirm DDIMM's ability to handle large sample size N and response dimension M.

Setting	Measure	Intercept	X_1	X_2
I	RMSE/BIAS	3.89/−1.77	0.64/0.09	0.60/−0.40
	ESE/ASE	3.89/3.78	0.64/0.59	0.60/0.59
II	RMSE/BIAS	1.86/−0.99	0.28/−0.03	0.28/−0.17
	ESE/ASE	1.86/1.70	0.28/0.27	0.28/0.27

Table 4.3: RMSE$\times 10^{-3}$, BIAS$\times 10^{-4}$, ESE$\times 10^{-3}$, ASE$\times 10^{-3}$ for each setting and each covariate, averaged over 500 simulations.

4.7 Discussion

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

The GMM approach to deriving the combined estimator $(\widehat{\boldsymbol{\theta}}_c, \widehat{\boldsymbol{\zeta}}_c)$ proposed in (4.4) requires only weak regularity of the estimating equations $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$. These assumptions are satisfied by a broad range of learning procedures. The closed-form estimator proposed in equation (4.9), on the other hand, requires local $n_k^{1/2}$-consistent estimators in individual blocks of size n_k, which is easily satisfied if $\boldsymbol{\Psi}_{jk}$ and $\boldsymbol{G}_{jk}$ are regular (see Song (2007) Chapter 3.5 for a definition of regular inference functions). This restricts the class of possible learning procedures, but still includes many analyses of interest.

A detailed discussion of the limitations and trade-offs of the single split DIMM with CL block analyses is featured in Hector and Song (2020a). As mentioned in Section 4.5, the DDIMM introduces additional flexibility in trading off between computational speed and inference: the number of subject groups K and the smallest block size $n_{\min}$ can be chosen by the investigator to attain the desired speed and efficiency.

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

CHAPTER V

Joint Integrative Analysis of Multiple Data Sources with Correlated Vector Outcomes

5.1 Introduction

Data integration methods have drawn increasing attention with the availability of massive data from multiple sources, with proposed methods spanning the gamut from the frequentist confidence distribution approach (Xie et al., 2011; Xie and Singh, 2013) to Bayesian hierarchical models (Smith et al., 1995), as well as several generalisations of Glass (1976)'s meta-analysis (Ioannidis, 2006; DerSimonian and Laird, 2015; Kundu et al., 2019). This chapter is substantially motivated by the analysis of the effect of smoking on metabolites that are upstream determinants of cardiovascular health. We consider the analysis of multiple independent studies that collect multiple correlated outcome vectors (metabolic sub-pathways) on each subject. Of interest is a joint integrative regression analysis of all studies and outcome vectors, yielding improvements in estimation efficiency. We propose a distributed quadratic inference function framework for this joint integrative regression analysis that addresses five major aspects of data integration: correlation of outcomes, heterogeneity of data sources, statistical efficiency, privacy concerns and computational speed.

Recent work has primarily focused on synthesizing evidence from independent data sources, as in Claggett et al. (2014) and Yang et al. (2014b). In practice, however, studies may collect correlated outcomes from different structural modalities, such as high-dimensional longitudinal phenotype, pathway-networked omics biomarkers, or brain imaging measurements, which collectively form one high-dimensional correlated response vector for each subject. Of interest is conducting inference integrated not only over the independent data sources but also over the structurally correlated outcomes. High-order moments of complex high-dimensional correlated data may be difficult to model or handle computationally, which has led many to use working independence assumptions at the cost of statistical efficiency resulting in potentially misleading statistical inference; see for example the composite likelihood approach in Caragea and Smith (2007) and Varin (2008). Chapter IV proposes a method to account for correlation between data sources without specifying a full parametric model, but their method is burdened by the estimation of a high-dimensional parameter related to the second-order moments, whose dimension can rapidly increase and exceed the sample size as the number of data sources increases. To relieve this burden, we propose a fast and efficient approach that avoids estimation of parameters in second-order moments with no loss of statistical efficiency.

Traditional data integration methods, such as meta-analysis and the confidence distribution approach, frequently assume parameter or even likelihood homogeneity across data sources, which often does not hold in practice. Data source heterogeneity can stem from differences in populations, study design, or associations, and can result in first- and higher-order moment heterogeneity. On the other hand, seemingly unrelated regression (Zellner, 1962) can be inefficient when some parameters are homogeneous. One approach to dealing with first-order moment heterogeneity is to include data source-specific random effects, which can be inefficient and may induce misspecified correlation. Another approach is to allow study-specific fixed-effects, as in Lin and Zeng (2010); Liu et al. (2015); Hector and Song (2020b). In the current literature there is a lack of computationally fast and statistically efficient methods to handle high-dimensional second- and higher-order moment parameters, which are regarded as nuisance parameters in a correlated data integrative setting. With only one data source, the quadratic inference function (Qu et al., 2000) is widely used to estimate regression parameters in first-order moments while avoiding estimation of second- and higher-order moments. Thus, the quadratic inference function minimizes the excessive burden of handling nuisance parameters. Our proposed distributed quadratic inference functions estimate regression parameters in mean models for each data source, thereby avoiding estimation of nuisance parameters in higher-order moments, and linearly updates the regression parameters according to different heterogeneity patterns across data sources. Not only does our approach combine the strengths from both meta-analysis and seemingly unrelated regression, but it is more flexible than these two methods.

For privacy reasons we may not have access to individual level data when integrating correlated data sources, in which case it becomes imperative to develop methods that can be implemented in a computationally distributed fashion. Even with access to individual level data, distributed algorithms are often preferred for their ability to significantly reduce the computational burden of traditional inference methods (Jordan, 2013; Fan et al., 2014). There is a need for distributed methods able to handle parameter heterogeneity for computationally and statistically efficient inference with multiple correlated data sources.

Our proposed distributed quadratic inference function approach estimates mean parameters for study- and outcome-specific models in the integrative analysis of correlated outcome vectors while avoiding estimation of second-order moments. It yields statistically efficient estimation within a broad class of models. Study- and outcome-specific models are then selectively combined via a meta-estimator similar in spirit to Hansen (1982)'s generalised method of moments according to some characterization of data heterogeneity. This new method has two major advantages over existing methods: the integrated estimator does not require access to individual-level data, and it can be computed non-iteratively to minimise computational costs. We illustrate the application of our proposed method through simulations and the integrative analysis of metabolite sub-pathways in a multi-cohort study.

5.2 Distributed and integrated quadratic inference functions

5.2.1 Model formulation

Consider K independent studies with respective sample sizes n_k, $k = 1, \ldots, K$. In each study we observe J correlated $m_{i,j}$-element vector outcomes $\boldsymbol{y}_{i,jk} = (y_{i1,jk}, \ldots, y_{im_{i,j},jk})^T$, $j = 1, \ldots, J$, for each subject i, $i = 1, \ldots, n_k$, with $\boldsymbol{x}_{i,jk}$ the corresponding $m_{i,j} \times p$ covariate matrix. Here $\boldsymbol{x}_{i,jk}$ is assumed to be the study- and outcome-specific observations on the same variables across outcomes and studies (e.g. age, sex). Subjects are assumed independent, and let $\boldsymbol{\Sigma}_{i,k}$ be the covariance matrix of $\boldsymbol{y}_{i,k} = (\boldsymbol{y}_{i,1k}, \ldots, \boldsymbol{y}_{i,Jk})^T$. We consider the model $E(y_{ir,jk}) = h_{jk}(\boldsymbol{x}_{ir,jk}\boldsymbol{\theta}_{jk})$, $r = 1, \ldots, m_{i,j}$, where h_{jk} is a known link function and $\boldsymbol{\theta}_{jk}$ is a $p \times 1$ parameter vector of interest. Suppose there exists a known partition $\mathcal{P} = \{\mathcal{P}_g\}_{g=1}^G$, $\mathcal{P}$ a set of disjoint non-empty subsets $\mathcal{P}_g$, of $\{(j,k)\}_{j,k=1}^{J,K}$ such that $\boldsymbol{\theta}_{jk} \equiv \boldsymbol{\theta}_g$ and $h_{jk} \equiv h_g$ for $(j,k) \in \mathcal{P}_g$. There are G unique values of $\boldsymbol{\theta}_{jk}$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. Let $\mathcal{P}_g$ have cardinality d_g such that $\sum_{g=1}^G d_g = JK$. We want to estimate and make inference about the true value $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0,g})_{g=1}^G \in \mathbb{R}^{Gp}$ of $\boldsymbol{\theta} = (\boldsymbol{\theta}_g)_{g=1}^G \in \mathbb{R}^{Gp}$ based on all JK sources of information. We give an example from Section 5.4 to fix ideas. For $K = 4$ cohorts, we quantify 24 metabolites from $J = 5$ carbohydrate sub-pathways: the glycolysis, gluconeogenesis, and pyruvate metabolism sub-pathway, the pentose metabolism sub-pathway, the aminosugar metabolism sub-pathway, the fructose, mannose and galactose metabolism sub-pathway, and the glycogen metabolism sub-pathway. Given the biological function of these sub-pathways, we model the effect of smoking on the metabolites in the carbohydrate sub-pathways by integrating its effect over the four cohorts and the latter four sub-pathways, and integrating the effect of smoking on the first sub-pathway only over cohorts. This partition corresponds to $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$, where $\mathcal{P}_1 = \{(2,k),(3,k),(4,k),(5,k)\}_{k=1}^{K}$ and $\mathcal{P}_2 = \{(1,k)\}_{k=1}^{K}$. The proposed method creates a set of moment conditions on $\boldsymbol{\theta}$, with corresponding estimators, from each data source. We propose an efficient and computationally attractive estimator that linearly updates data source-specific estimators by weighting them as a function of their covariance.

We introduce some notation to facilitate the description of the proposed method in sections 5.2.2 and 5.2.3. For ease of exposition, we henceforth use the term "studies" to refer to the K disjoint and independent subject groups, "block" to refer to the J correlated vector outcomes. We refer to study k and block j as data source (j,k). Let $\|\cdot\|$ denote the L_2 norm on vectors and the Frobenius norm on matrices. Define the stacking operator $\mathbb{S}(\cdot)$ for vectors $\{\boldsymbol{a}_{jk}\}_{j=1,k=1}^{J,K}$, $\boldsymbol{a}_{jk} \in \mathbb{R}^D$ and matrices $\{\boldsymbol{A}_{jk}\}_{j=1,k=1}$, $\boldsymbol{A}_{jk} \in \mathbb{R}^{D\times D}$, as

$$\mathbb{S}^g(\boldsymbol{a}_{jk}) = \boldsymbol{a}_g = \begin{pmatrix} \boldsymbol{a}_{j_1k_1}^T & \dots & \boldsymbol{a}_{j_{d_g}k_{d_g}}^T \end{pmatrix}^T \in \mathbb{R}^{d_gD}, \quad \mathcal{P}_g = \left\{(j_1,k_1),\dots,(j_{d_g},k_{d_g})\right\},$$

$$\mathbb{S}^g(\boldsymbol{A}_{jk}) = \boldsymbol{A}_g = \begin{pmatrix} \boldsymbol{A}_{j_1k_1}^T & \dots & \boldsymbol{A}_{j_{d_g}k_{d_g}}^T \end{pmatrix}^T \in \mathbb{R}^{(d_gD)\times D}, \quad \mathcal{P}_g = \left\{(j_1,k_1),\dots,(j_{d_g},k_{d_g})\right\},$$

$$\mathbb{S}^G(\boldsymbol{a}_g) = \begin{pmatrix} \boldsymbol{a}_1^T & \dots & \boldsymbol{a}_G^T \end{pmatrix}^T \in \mathbb{R}^{JKD}, \quad \mathbb{S}^G(\boldsymbol{A}_g) = \begin{pmatrix} \boldsymbol{A}_1^T & \dots & \boldsymbol{A}_G^T \end{pmatrix}^T \in \mathbb{R}^{(JKD)\times D}.$$

Let $\boldsymbol{a}^{\otimes 2}$ denote the outer product of a vector $\boldsymbol{a}$ with itself, namely $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$. Denote $N = \sum_{k=1}^{K} n_k$ the combined sample size over K studies. For each subject i, denote the combined M_i-dimensional response $\boldsymbol{y}_i = \left(\boldsymbol{y}_{i,1}, \ldots, \boldsymbol{y}_{i,J}\right)^T$ over the J blocks such that $\sum_{j=1}^{J} m_{i,j} = M_i$ for each $i = 1, \ldots, N$. Combination across blocks is not restricted to the order of data entry: responses may be grouped according to pre-specified block memberships, according to, say, substantive scientific knowledge. In this chapter, with no loss of generality, we use the order of data entry in the data combination procedure.

We remark that our proposed method can also be applied as a divide-and-conquer procedure to a large dataset with N samples on M correlated outcomes. Dividing this large dataset into JK sources of data with sample size n_k and m_j-dimensional outcomes yields the above framework with the simplification $M_i = M$, $m_{i,j} = m_j$.

5.2.2 Quadratic Inference Functions

We propose to first obtain Qu et al. (2000)'s quadratic inference function estimator of $\boldsymbol{\theta}_{jk}$ in data source (j,k). This is a standard analysis that is performed on each data source individually as if there was no other source of data to improve estimation. Consider an arbitrary data source (j,k). Let $\boldsymbol{\mu}_{i,jk} = E(\boldsymbol{y}_{i,jk})$ the $m_{i,j}$-dimensional mean of the outcome $\boldsymbol{y}_{i,jk}$ for $i = 1, \ldots, n_k$. Let $\dot{\boldsymbol{\mu}}^{\boldsymbol{\theta}}_{i,jk} = \partial \boldsymbol{\mu}_{i,jk} / \partial \boldsymbol{\theta}_{jk}$ be an $m_{i,j} \times p$-dimensional partial derivative matrix and let $\ddot{\boldsymbol{\mu}}^{\boldsymbol{\theta}}_{i,jk} = (\partial^2 \boldsymbol{\mu}_{i,jk} / \partial^2 \boldsymbol{\theta}_{jk})$. Following Qu et al. (2000), we approximate the inverse working correlation matrix of $\boldsymbol{y}_{i,jk}$ by $\sum_{s=1}^{s_{jk}} b_{s,jk} \boldsymbol{B}_{s,jk}$ where $b_{1,jk}, \ldots, b_{s_{jk},jk}$ are unknown constants and $\boldsymbol{B}_{1,jk}, \ldots, \boldsymbol{B}_{s_{jk},jk}$

are known basis matrices with elements 0 and 1. Let

$$\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}_{jk}) = \frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{jk}) = \frac{1}{n_k}\sum_{i=1}^{n_k}\begin{pmatrix} \dot{\boldsymbol{\mu}}^{\boldsymbol{\theta}\,T}_{i,jk}\boldsymbol{D}^{-\frac{1}{2}}_{i,jk}\boldsymbol{B}_{1,jk}\boldsymbol{D}^{-\frac{1}{2}}_{i,jk}(\boldsymbol{y}_{i,jk}-\boldsymbol{\mu}_{i,jk}) \\ \vdots \\ \dot{\boldsymbol{\mu}}^{\boldsymbol{\theta}\,T}_{i,jk}\boldsymbol{D}^{-\frac{1}{2}}_{i,jk}\boldsymbol{B}_{s_{jk},jk}\boldsymbol{D}^{-\frac{1}{2}}_{i,jk}(\boldsymbol{y}_{i,jk}-\boldsymbol{\mu}_{i,jk}) \end{pmatrix}, \quad (5.1)$$

where $\boldsymbol{D}_{i,jk}$ is the diagonal marginal covariance matrix of $\boldsymbol{y}_{i,jk}$, and s_{jk} is typically chosen as $s_{jk} = 2$. Let $\boldsymbol{C}_{jk} = (1/n_k)\sum_{i=1}^{n_k}\boldsymbol{\psi}^{\otimes 2}_{i,jk}(\boldsymbol{\theta}_{jk})$, which depends only on $\boldsymbol{\theta}_{jk}$. The quadratic inference function takes the form $Q_{jk}(\boldsymbol{\theta}_{jk}) = n_k\boldsymbol{\Psi}^T_{jk}(\boldsymbol{\theta}_{jk})\boldsymbol{C}^{-1}_{jk}\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}_{jk})$, and the data source-specific quadratic inference function estimator is $\widehat{\boldsymbol{\theta}}_{jk} = \arg\min_{\boldsymbol{\theta}_{jk}} Q_{jk}(\boldsymbol{\theta}_{jk})$. No nuisance correlation parameter is involved in the estimation. Under mild regularity conditions, $\widehat{\boldsymbol{\theta}}_{jk}$ is consistent and asymptotically normal (Hansen, 1982). When the working correlation structure is correctly specified by the basis matrix expansion, this estimator is semi-parametrically efficient, i.e. as efficient as the quasilikelihood; even when the working correlation structure is misspecified, this estimator is still efficient within a general family of estimators (Qu et al., 2000). These advantageous properties allow us to derive an efficient integrated estimator in section 5.2.3.

5.2.3 Integrated Estimator

Define the subject group indicator $\delta_i(k) = \mathbb{1}(\text{subject } i \text{ is in study } k)$ for $i = 1,\dots,N$, $k = 1,\dots,K$. For subject i, let

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

Then we can define $\boldsymbol{\Psi}_N(\boldsymbol{\theta}) = (1/N)\sum_{i=1}^{N}\boldsymbol{\psi}_i(\boldsymbol{\theta})$. It is easy to show that

$$\boldsymbol{\Psi}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{S}^G\left[\mathbb{S}^g\left\{\delta_i(k)\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{jk})\right\}\right] = \frac{1}{N}\mathbb{S}^G\left[\mathbb{S}^g\left\{n_k\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}_{jk})\right\}\right].$$

We define a few sample sensitivity matrices. For data source (j,k), define the $(ps_{jk}) \times p$-dimensional sample sensitivity matrix $\widehat{\boldsymbol{S}}_{jk} = -\{\nabla_{\boldsymbol{\theta}_{jk}}\boldsymbol{\Psi}_{jk}(\boldsymbol{\theta}_{jk})\}|_{\boldsymbol{\theta}_{jk}=\widehat{\boldsymbol{\theta}}_{jk}}$.

For the gth set $\mathcal{P}^g$, define $\widehat{\boldsymbol{S}}_g = \mathbb{S}^g(n_k \widehat{\boldsymbol{S}}_{jk})$ the matrix of stacked sensitivity matrices with row-dimension $\sum_{(j,k)\in\mathcal{P}^g} p s_{jk}$ and column-dimension p. Finally, let $\widehat{\boldsymbol{S}} = \text{blockdiag}\{\widehat{\boldsymbol{S}}_g\}_{g=1}^G$ the sample sensitivity matrix of $\boldsymbol{\Psi}_N$ with row-dimension $\sum_{g=1}^G \sum_{(j,k)\in\mathcal{P}^g} p s_{jk} = \sum_{j,k=1}^{J,K} p s_{jk}$ and column-dimension Gp.

Let $\widehat{\boldsymbol{\theta}}_g = \mathbb{S}^g(\widehat{\boldsymbol{\theta}}_{jk})$ and $\widehat{\boldsymbol{\theta}}_{list} = \mathbb{S}^G(\widehat{\boldsymbol{\theta}}_g)$. Define $\boldsymbol{\psi}_i(\widehat{\boldsymbol{\theta}}_{list}) = \mathbb{S}^G[\mathbb{S}^g\{\delta_i(k)\boldsymbol{\psi}_{i,jk}(\widehat{\boldsymbol{\theta}}_{jk})\}]$. Let $\widehat{\boldsymbol{V}}_N = (1/N)\sum_{i=1}^N \{\boldsymbol{\psi}_i(\widehat{\boldsymbol{\theta}}_{list})\}^{\otimes 2}$ be the sample covariance of $\boldsymbol{\Psi}_N(\boldsymbol{\theta}_0)$ with row- and column-dimension $\sum_{j,k=1}^{J,K} p s_{jk}$. Then we define the integrated estimator of $\boldsymbol{\theta}$ as

$$\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{S}}^T \widehat{\boldsymbol{V}}_N^{-1} \widehat{\boldsymbol{S}}\right)^{-1} \widehat{\boldsymbol{S}}^T \widehat{\boldsymbol{V}}_N^{-1} \mathbb{S}^G \left\{\mathbb{S}^g(n_k \widehat{\boldsymbol{S}}_{jk} \widehat{\boldsymbol{\theta}}_{jk})\right\}. \tag{5.2}$$

Following similar steps to Hector and Song (2020a), we can show this integrated estimator is asymptotically equivalent to the minimizer of the optimal combination of the moment conditions. Estimators from different sets $\mathcal{P}^g$ may not be combined but still benefit from correlation between data sources, captured by $\widehat{\boldsymbol{V}}_N$, leading to improved statistical efficiency. This is similar to the gain in efficiency in seemingly unrelated regression (Zellner, 1962). The closed-form estimator in (5.2) depends only on estimators, estimating equations and sample sensitivity matrices from each data source. It can be implemented in a fully parallelized MapReduce framework, where data sources are analyzed in parallel on distributed nodes using quadratic inference functions and results from the separate analyses are sent to a main node to compute the integrated estimator. This procedure is privacy-preserving, since the combination step does not require access to individual level data, and communication-efficient, since it does not require multiple rounds of communication between the main and distributed nodes. In addition, it is computationally efficient at each node since nuisance correlation parameters are not involved in the estimation.

Two special cases of interest arise when $\mathcal{P}_g$ are all singletons ($G = JK$, $d_g = 1$) and when $\mathcal{P} = \{(j,k)\}_{j,k=1}^{J,K}$ ($G = 1$, $d_1 = JK$). The former case is similar to seemingly unrelated regression, in which JK regression equations are used to estimate JK parameter vectors. Unlike seemingly unrelated regression, however, we do not make distributional assumptions on the outcomes since we use estimating equations. The latter case corresponds to a fully integrated analysis of all JK data sources similar in spirit to meta-analysis. The estimator $\widehat{\boldsymbol{\theta}}$ in (5.2) takes a special form: let $\widehat{\boldsymbol{V}}_N = \text{blockdiag}\{(n_k/N)\widehat{\boldsymbol{V}}_k\}_{k=1}^K$ with block matrices

$$\widehat{\boldsymbol{V}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\left\{\boldsymbol{\psi}_{i,k}(\widehat{\boldsymbol{\theta}}_k)\right\}^{\otimes 2} \quad (k = 1, \dots, K).$$

Let $[\widehat{\boldsymbol{V}}_k]^{i;j}$ denote the rows and columns of $\widehat{\boldsymbol{V}}_k^{-1}$ corresponding to blocks i and j respectively and define the sample Godambe information $\widehat{\boldsymbol{J}}_{ijk} = \widehat{\boldsymbol{S}}_{ik}\left[\widehat{\boldsymbol{V}}_k\right]^{i;j}\widehat{\boldsymbol{S}}_{jk}$ (Godambe and Heyde, 1987; Song, 2007). The integrated estimator simplifies to

$$\widehat{\boldsymbol{\theta}} = \left(\sum_{k=1}^{K}\sum_{i,j=1}^{J} n_k\widehat{\boldsymbol{J}}_{ijk}\right)^{-1}\sum_{k=1}^{K}\sum_{i,j=1}^{J} n_k\widehat{\boldsymbol{J}}_{ijk}\widehat{\boldsymbol{\theta}}_{jk}.$$

Remark 5. Inversion of $\widehat{\boldsymbol{V}}_N$ may be numerically unstable or undefined in some settings. When J, K and/or p are large, the large dimension of $\widehat{\boldsymbol{V}}_N$ can lead to numerical difficulties in its inversion. Using an equicorrelated structure for the block analysis can also lead to a rank-deficient weight matrix $\widehat{\boldsymbol{V}}_N$ (Hu and Song, 2012). To handle these cases we propose to reduce the number of estimating equations similarly to Cho and Qu (2015): principal components of $\boldsymbol{\Psi}_N$ with non-zero eigenvalues are selected so as to maximize the variability explained and eliminate between-component correlations. These linear combinations of the original estimating equations have lower dimension than $\boldsymbol{\Psi}_N$ and yield an invertible sample variability matrix $\widehat{\boldsymbol{V}}_N$. The method described in Section 5.2 remains unchanged with the substitution of the principal components for $\boldsymbol{\Psi}_N$.

5.2.4 Large sample theory

Let $n_{\min} = \min_{k=1,\ldots,K} n_k$. Define the sensitivity matrices $\boldsymbol{s}_{jk}(\boldsymbol{\theta}_{jk}) = -\nabla_{\boldsymbol{\theta}_{jk}} E_{\boldsymbol{\theta}_{0,g}}\{\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{jk})\}$ for $(j,k) \in \mathcal{P}^g$, $\boldsymbol{s}_g(\boldsymbol{\theta}_g) = \mathbb{S}^g\{(n_k/N)\boldsymbol{s}_{jk}(\boldsymbol{\theta}_{jk}), (j,k) \in \mathcal{P}^g\}$, and $\boldsymbol{s}(\boldsymbol{\theta}) = \text{blockdiag}\{\boldsymbol{s}_g(\boldsymbol{\theta}_g)\}_{g=1}^G$. Define the variability matrix $\boldsymbol{v}(\boldsymbol{\theta}) = Var_{\boldsymbol{\theta}_0}\{\boldsymbol{\psi}_i(\boldsymbol{\theta})\}$. Regularity conditions required to establish the consistency and asymptotic normality of the integrated estimator $\widehat{\boldsymbol{\theta}}$ in (5.2) are listed in Appendix H. In particular, assumption H.1 guarantees the consistency and asymptotic normality of the data source-specific estimators $\widehat{\boldsymbol{\theta}}_{jk}$, and assumption H.2 guarantees the consistency and asymptotic normality of the integrated estimator $\widehat{\boldsymbol{\theta}}$ in (5.2). These results are summarized in Theorem V.1.

Theorem V.1 (Consistency and asymptotic normality)**.** *Suppose assumptions H.1-H.2 hold. Let* $\boldsymbol{j}(\boldsymbol{\theta}) = \lim_{n_{\min}\to\infty} \boldsymbol{s}^T(\boldsymbol{\theta})\boldsymbol{v}^{-1}(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{\theta})$ *denote the Godambe information matrix of* $\boldsymbol{\Psi}_N$*. As* $n_{\min} \to \infty$, $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{j}^{-1}(\boldsymbol{\theta}_0))$*, and* $\boldsymbol{j}(\boldsymbol{\theta})$ *has* (r,t)*th block element* $\lim_{n_{\min}\to\infty}\{\boldsymbol{s}_r^T(\boldsymbol{\theta}_0)[\boldsymbol{v}(\boldsymbol{\theta}_0)]^{r;t}\boldsymbol{s}_t(\boldsymbol{\theta}_0)\}$ *where* $[\boldsymbol{v}(\boldsymbol{\theta}_0)]^{r;t}$ *is the submatrix of* $\boldsymbol{v}^{-1}(\boldsymbol{\theta}_0)$ *consisting of rows and columns corresponding to partitions* r *and* t *respectively,* $r,t = 1,\ldots,G$.

The proof of Theorem V.1 can be done similarly to Theorem 9 in Hector and Song (2020b) and is omitted. It is clear from Theorem V.1 that the asymptotic covariance of $\widehat{\boldsymbol{\theta}}$ can be consistently estimated by the sandwich covariance $N(\widehat{\boldsymbol{S}}^T\widehat{\boldsymbol{V}}_N^{-1}\widehat{\boldsymbol{S}})^{-1}$. A goodness-of-fit test is available from Theorem V.2 below to check the validity of modelling assumptions and appropriateness of the data source partition $\mathcal{P}$.

Theorem V.2 (Homogeneity test)**.** *Suppose assumptions H.1-H.2 hold with* $\widehat{\boldsymbol{\theta}}$ *defined in* (5.2)*. Then as* $n_{\min} \to \infty$*, the statistic* $Q_N(\widehat{\boldsymbol{\theta}}) = N\boldsymbol{\Psi}_N^T(\widehat{\boldsymbol{\theta}})\widehat{\boldsymbol{V}}_N^{-1}\boldsymbol{\Psi}_N(\widehat{\boldsymbol{\theta}})$ *converges in distribution to a* χ^2 *random variable with degrees of freedom* $\sum_{j,k=1}^{J,K} ps_{jk} -$

Gp.

The proof of Theorem V.2 follows from Hansen (1982) and Hector and Song (2020b). In practice, the computation of the quadratic test statistic in Theorem V.2 can be implemented in a distributed fashion despite requiring access to individual data sources to recompute $\boldsymbol{\Psi}_N(\widehat{\boldsymbol{\theta}})$. Since $\boldsymbol{\psi}_i(\widehat{\boldsymbol{\theta}})$ becomes available, we recommend recomputing $\widehat{\boldsymbol{V}}_N$ based on $\boldsymbol{\psi}_i(\widehat{\boldsymbol{\theta}})$ to improve numerical performance. Theorem V.2 is particularly useful to compare the fit from different data source partitions, and can be used to detect inappropriate modelling and strong data heterogeneity requiring modification of the integration partition. Let $\mathcal{P}^i = \{\mathcal{P}^i_g\}_{g=1}^{G^i}$ and $\mathcal{P}^h = \{\mathcal{P}^h_g\}_{g=1}^{G^h}$ two data source partitions such that $\mathcal{P}^i$ is itself a nested partition of $\mathcal{P}^h$; let $Q^i_N(\widehat{\boldsymbol{\theta}}^i)$ and $Q^h_N(\widehat{\boldsymbol{\theta}}^h)$ be the statistics from Theorem V.2 based on partitions $\mathcal{P}^i$ and $\mathcal{P}^h$ respectively, where the same working correlation structures and mean models are used for both. Then a test statistic of the null hypothesis of parameter homogeneity in partition $\mathcal{P}^i$,

$$H_0 : \boldsymbol{\theta}_{jk} = \boldsymbol{\theta}^i_g, \quad \forall (j,k) \in \mathcal{P}^i_g, \quad g = 1, \dots, G^i,$$

can be formulated as

$$Q^i_N(\widehat{\boldsymbol{\theta}}^i) - Q^h_N(\widehat{\boldsymbol{\theta}}^h), \tag{5.3}$$

which under H_0 is asymptotically χ^2 distributed with degrees of freedom $G^h p - G^i p$. Failure to reject the null hypothesis implies the smaller partition $\mathcal{P}^i$ fits as well or better than the larger partition $\mathcal{P}^h$.

Lastly, we discuss estimation efficiency of our proposed integrated estimator $\widehat{\boldsymbol{\theta}}$ in (5.2), which is asymptotically equivalent to Hansen (1982)'s optimal generalised method of moments estimator $\widehat{\boldsymbol{\theta}}_{opt} = \arg\min_{\boldsymbol{\theta}} \boldsymbol{\Psi}^T_N(\boldsymbol{\theta}) \widehat{\boldsymbol{V}}^{-1}_N \boldsymbol{\Psi}_N(\boldsymbol{\theta})$. The optimality of this estimator Hansen (1982) is achieved within the class of estimators minimizing the quadratic form $\boldsymbol{\Psi}_N^T(\boldsymbol{\theta})\boldsymbol{W}\boldsymbol{\Psi}_N(\boldsymbol{\theta})$ with positive semi-definite matrices $\boldsymbol{W}$. Additionally, in Theorem V.3 we show the efficiency gain from combining estimators over data sources for an arbitrary data source $(j,k) \in \mathcal{P}_g$. The asymptotic covariance of $\widehat{\boldsymbol{\theta}}_{jk}$ is larger than or equal to (in the Löwner partial ordering) the subvector of $\widehat{\boldsymbol{\theta}}$ corresponding to $\mathcal{P}_g$, $\widehat{\boldsymbol{\theta}}^g$.

Theorem V.3 (Efficiency gain)**.** *Suppose assumptions H.1-H.2 hold with $\widehat{\boldsymbol{\theta}}$ defined in* (5.2)*. Consider an arbitrary data source $(j,k) \in \mathcal{P}^g$ for some $g \in \{1,\dots,G\}$. The asymptotic covariances, denoted by Avar, of $\widehat{\boldsymbol{\theta}}_{jk}$ and $\widehat{\boldsymbol{\theta}}^g$ satisfy $Avar(\sqrt{N}\widehat{\boldsymbol{\theta}}^g) \leq \{\lim_{n_k\to\infty}(N/n_k)\}Avar(\sqrt{n_k}\widehat{\boldsymbol{\theta}}_{jk})$, where $\leq$ denotes Löwner's partial ordering in the space of nonnegative definite matrices.*

The proof of Theorem V.3 is given in Appendix H. The gain in statistical efficiency given by Theorem V.3 is due to the use of between-block correlation, captured by $\widehat{V}_N$, and to the combination of estimators within each $\mathcal{P}^g$.

Remark 6. The proposed method is a generalization of Wang et al. (2012), which only allows for combining over independent data sources. Here we introduce a non-diagonal weight matrix $\widehat{\boldsymbol{V}}_N$ to incorporate correlation between data sources, leading to improved statistical efficiency. We also propose a closed form integrated estimator that is more computationally advantageous than their iterative minimization procedure, leading to improved computational scalability.

5.3 Simulations

We examine the performance of the integrated estimator $\widehat{\boldsymbol{\theta}}$ through three sets of simulations. In the first and third sets, for simplicity $\mathcal{P} = \{(j,k)\}_{j,k=1}^{J,K}$ ($G = 1$, $d_1 = JK$) and $M_i = M$ for $i = 1,\dots,N$. The second set explores the performance of the selective combination scheme with a partition of $\{(j,k)\}_{j,k=1}^{J,K}$ and confirms the distribution of statistic $Q_N(\widehat{\boldsymbol{\theta}})$ in Theorem V.2. Simulations are conducted using R software on a standard Linux cluster. In all simulations, covariates consist of an intercept and two independent M-dimensional continuous variables drawn from Multivariate Normal distributions with non-diagonal covariance matrices. True values of the regression parameters are drawn from uniform distributions on $(-5,5)$. The first set of simulations considers the logistic regression $\log\{\mu_{ir,jk}/(1-\mu_{ir,jk})\} = \boldsymbol{X}_{ir,jk}\boldsymbol{\theta}$ with $\mu_{ir,jk} = E(Y_{ir,jk}|\boldsymbol{X}_{ir,jk},\boldsymbol{\theta})$, $r = 1,\dots,m_j$, where $\boldsymbol{Y}_i$ is a M-variate correlated Bernoulli random variable. We illustrate the finite sample performance of $\widehat{\boldsymbol{\theta}}$ in two settings: in Setting I, $K = 2$ with $n_1 = n_2 = 5000$, $J = 4$ with block response dimensions 163, 181, 260, 396 such that $M = 1000$; in Setting II, $K = 4$ with $n_1 = n_2 = n_3 = n_4 = 5000$, $J = 8$ with block response dimensions 227, 252, 357, 381, 368, 276, 226, 413 such that $M = 2500$. The true value of $\boldsymbol{\theta}$ is set to $\boldsymbol{\theta}_0 = (-4.44, 1.11, -2.22)$. $\boldsymbol{Y}_i$ is simulated using the `SimCorMultRes` R package (Touloumis, 2016) with data source-specific AR(1) correlation structures. We estimate $\boldsymbol{\theta}$ with an AR(1) working block correlation structure. Root mean squared error (RMSE), empirical standard error (ESE), asymptotic standard error (ASE), mean bias (B), 95% confidence interval coverage (CI), 95% confidence interval length (L) and type-I error (ERR) of $\widehat{\boldsymbol{\theta}}$ averaged over 500 simulations are presented in Table 5.1. We see from Table 5.1 that the ASE of $\widehat{\boldsymbol{\theta}}$ approximates the ESE, confirming the covariance formula in Theorem V.1. Additionally, $\widehat{\boldsymbol{\theta}}$ appears consistent since RMSE, ASE and ESE are approximately equal, and the bias B is negligible. We observe appropriate 95% confidence interval coverage and proper Type-I error control.

The second set of simulations again considers the logistic regression $\log(\mu_{ir,jk}/(1-\mu_{ir,jk})) = \boldsymbol{X}_{ir,jk}\boldsymbol{\theta}$ with $\mu_{ir,jk} = E(Y_{ir,jk}|\boldsymbol{X}_{ir,jk},\boldsymbol{\theta})$, where $\boldsymbol{Y}_i$ is a

Table 5.1: Logistic regression simulation results with homogeneity partition $\mathcal{P} = \{(j,k)\}_{j,k=1}^{J,K}$ ($G = 1$, $d_1 = JK$).

(a) Setting I: $K = 2$ with $n_1 = n_2 = 5000$, $J = 4$ with block response dimensions 163, 181, 260, 396 such that $M = 1000$.

	RMSE$\times 10^{-3}$	ESE$\times 10^{-3}$	ASE$\times 10^{-3}$	B$\times 10^{-4}$	CI	L$\times 10^{-3}$	ERR
Intercept	4.89	4.86	4.83	−5.84	0.95	18.74	0.05
X_1	1.43	1.42	1.40	1.79	0.94	5.45	0.06
X_2	2.45	2.43	2.48	−3.49	0.95	9.65	0.05

(b) Setting II: $K = 4$ with $n_1 = n_2 = n_3 = n_4 = 5000$, $J = 8$ with block response dimensions 227, 252, 357, 381, 368, 276, 226, 413 such that $M = 2500$.

	RMSE$\times 10^{-3}$	ESE$\times 10^{-3}$	ASE$\times 10^{-3}$	B$\times 10^{-4}$	CI	L$\times 10^{-3}$	ERR
Intercept	2.29	2.20	2.21	−6.31	0.95	8.62	0.05
X_1	0.66	0.64	0.63	1.76	0.95	2.48	0.05
X_2	1.19	1.14	1.13	−3.36	0.95	4.42	0.05

M-variate correlated Bernoulli random variable of dimension $M = 500$ from $J = 5$ blocks with $(m_1, \ldots, m_5) = (130, 75, 92, 115, 88)$. We consider the integration of two studies of respective sample sizes $n_1 = n_2 = 5000$. The underlying partition is $\mathcal{P} = \{\mathcal{P}_g\}_{g=1}^{G}$ with $G = 3$, $\mathcal{P}_1 = \{(1,k),(2,k)\}_{k=1}^{K}$, $\mathcal{P}_2 = \{(3,k)\}_{k=1}^{K}$ and $\mathcal{P}_3 = \{(4,k),(5,k)\}_{k=1}^{K}$ with respective true values $\boldsymbol{\theta}_{0,1} = (-4.44, 1.11, -2.22)$, $\boldsymbol{\theta}_{0,2} = (0.222, -0.888, -0.444)$ and $\boldsymbol{\theta}_{0,3} = (-1.554, -3.108, 0.777)$. $\boldsymbol{Y}_i$ is simulated as in the first set of simulations. We estimate $\boldsymbol{\theta}$ and present summary results averaged over 500 simulations in Table 5.2 for the exchangeable working block correlation structure and in the Supplementary Material for the AR(1) working block correlation structure. From Table 5.2 we again observe correct estimation of the asymptotic covariance, minimal bias and proper Type-I error control for regression parameters. The integrative procedure seems to work well with partial heterogeneity of mean effects.

The third set of simulations considers the linear regression $\boldsymbol{\mu}_{i,jk} = \boldsymbol{X}_{i,jk}\boldsymbol{\theta}$ with $\boldsymbol{\mu}_{i,jk} = E(\boldsymbol{Y}_{i,jk} | \boldsymbol{X}_{i,jk}, \boldsymbol{\theta})$, where $\boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\theta}, \boldsymbol{\Sigma})$. We illustrate the finite sample performance of $\widehat{\boldsymbol{\theta}}$ with $K = 10$ studies where $n_k = 10000$ for all k for a total sample size of $N = 100000$, and $J = 250$ response blocks where $m_j = 400$ for

Table 5.2: Logistic regression simulation results with $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3\}$, $\mathcal{P}_1 = \{(1,k),(2,k)\}_{k=1}^K$, $\mathcal{P}_2 = \{(3,k)\}_{k=1}^K$ and $\mathcal{P}_3 = \{(4,k),(5,k)\}_{k=1}^K$, and exchangeable working block correlation structure.

(a) Summary results for $\mathcal{P}_1 = \{(1,k),(2,k)\}_{k=1}^K$.

	RMSE$\times10^{-3}$	ESE$\times10^{-3}$	ASE$\times10^{-3}$	B$\times10^{-4}$	CI	L$\times10^{-3}$	ERR
Intercept	10.89	10.62	10.30	−24.48	0.93	40.06	0.07
X1	3.16	3.11	3.03	5.94	0.94	11.84	0.06
X2	5.58	5.44	5.36	−12.26	0.93	20.86	0.07

(b) Estimates for $\mathcal{P}_2 = \{(3,k)\}_{k=1}^K$.

	RMSE$\times10^{-3}$	ESE$\times10^{-3}$	ASE$\times10^{-3}$	B$\times10^{-4}$	CI	L$\times10^{-3}$	ERR
Intercept	3.22	3.22	3.37	−2.16	0.95	12.93	0.05
X1	2.25	2.25	2.21	−0.48	0.95	8.57	0.05
X2	1.51	1.51	1.53	−0.06	0.95	5.98	0.05

(c) Estimates for $\mathcal{P}_3 = \{(4,k),(5,k)\}_{k=1}^K$.

	RMSE$\times10^{-3}$	ESE$\times10^{-3}$	ASE$\times10^{-3}$	B$\times10^{-4}$	CI	L$\times10^{-3}$	ERR
Intercept	4.72	4.69	4.80	−6.06	0.95	18.66	0.05
X1	6.36	6.25	6.31	−11.73	0.94	24.48	0.06
X2	2.03	2.01	2.04	2.82	0.95	7.98	0.05

all j for a total response dimension $M = 100000$ of Y. The true value of $\boldsymbol{\theta}$ is set to $\boldsymbol{\theta}_0 = (1.1, 2.2, 3.3)^T$. Responses are simulated from a Multivariate Normal distribution with block-AR(1) covariance structure with data source-specific variance and correlation parameters. We estimate $\boldsymbol{\theta}$ with an AR(1) working block correlation structure. RMSE, ESE, ASE, B, CI, L and ERR of $\widehat{\boldsymbol{\theta}}$ averaged over 500 simulations are presented in Table 5.3. We observe in Table 5.3 slight inflation of Type-I error due to under-estimation of the asymptotic covariance. This is potentially due to the high-dimensionality of $\boldsymbol{\Psi}_N$ and $\widehat{\boldsymbol{V}}_N$, which have dimension $JKpd = 15000$, leading to numerical instability. This under-estimation is similar to the generalised method of moments case and is discussed in Section 5.5. The performance of our method in this ultra-high dimension is nonetheless remarkable: with 10^{10} data points with high-variability in both outcomes and covariates, the procedure is able to estimate and infer the true mean effects with minimal bias and only slight under-coverage.

In the Supplementary Material, a quantile-quantile plot of the chi-squared statistic from Theorem V.2 in the second set of simulations illustrates its appropriate

Table 5.3: Linear regression simulation results with homogeneity partition $\mathcal{P} = \{(j,k)\}_{j,k=1}^{J,K}$ ($G = 1$, $d_1 = JK$).

	RMSE×10^{-4}	ESE×10^{-4}	ASE×10^{-4}	B×10^{-6}	CI	L×10^{-4}	ERR
Intercept	2.26	2.25	1.99	−21.03	0.93	7.81	0.07
X_1	0.35	0.35	0.31	−0.15	0.92	1.22	0.08
X_2	0.35	0.35	0.31	1.07	0.93	1.22	0.07

asymptotic distribution. Lastly, we observe mean CPU times of 1.4 and 2.1 minutes for logistic regression Settings I and II respectively in simulation set one, 11.2 seconds for the selective logistic regression in simulation set two, and 17.9 hours for the linear regression setting in simulation set three, which is computationally very fast.

5.4 Real Data Analysis

We illustrate the application of the proposed method to an integrative analysis of four studies of untargeted metabolites. The Metabolic Syndrome in Men study is a population-based study of 10197 Finnish men with the aim of investigating nongenetic and genetic factors associated with the risk of Type 2 diabetes, cardiovascular disease, and cardiovascular risk factors (Laakso et al., 2017). The Centers for Disease Control and Prevention list smoking as a major cause of cardiovascular disease and the 2014 Surgeon General's Report on smoking and health reported that smoking was responsible for one of every four deaths from cardiovascular disease. Investigating the association between smoking and metabolites can provide insight into the etiology of metabolic diseases such as cardiovascular disease.

Using the Metabolon platform, the Metabolic Syndrome in Men study profiled $N = 6223$ men in $K = 4$ separate samples with sample sizes $n_1 = 1229$, $n_2 = 2950$,

$n_3 = 1045$ and $n_4 = 999$. They measured 1018 metabolites belonging to 112 sub-pathways grouped in eight pathways with distinct biological functions. To facilitate interpretations, we focus on the effect of smoking on each of the eight pathways one at a time. For each pathway $s = 1, \dots, 8$, we investigate the association between smoking and metabolites in pathway s using our distributed and integrated quadratic inference functions approach to account for heterogeneity and correlation in metabolite sub-pathways. To highlight this pathway-specific implementation of our method, we add a superscript s to M and J to emphasize that these are pathway-specific variables. A schematic of the metabolite data structure is given in Table 5.4.

Consider pathway $s \in \{1, \dots, 8\}$. To illustrate the statistical efficiency gains from accounting for correlation between sub-pathways and combining models over independent studies, we first estimate sub-pathway and study specific effects and integrate them over studies (but not over sub-pathways); we then create a partition of sub-pathways in s by selectively combining sub-pathways based on prior knowledge. More specifically, we first estimate a heterogeneous model with partition $\mathcal{P}^{s,h} =$

Table 5.4: Metabolite data structure schematic.

		k	1	$\dots$	4
		i	$1 \quad \dots \quad n_1$	$\dots$	$1 \quad \dots \quad n_4$
s	j	r			
	1	1 $\vdots$ m_1	$\boldsymbol{y}^1_{11}$	$\dots$	$\boldsymbol{y}^1_{14}$
1	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	J^1	1 $\vdots$ m_{J^1}	$\boldsymbol{y}^1_{J^1 1}$	$\dots$	$\boldsymbol{y}^1_{J^1 4}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	1	1 $\vdots$ m_1	$\boldsymbol{y}^8_{11}$	$\dots$	$\boldsymbol{y}^8_{14}$
8	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	J^8	1 $\vdots$ m_{J^8}	$\boldsymbol{y}^8_{J^8 1}$	$\dots$	$\boldsymbol{y}^8_{J^8 4}$

$\{\mathcal{P}_j^{s,h}\}_{j=1}^{J^s}$, $\mathcal{P}_j^{s,h} = \{(j,1),\dots,(j,K)$: sub-pathway j is in pathway $s\}$ that yields unique values of the regression coefficients for each sub-pathway. We then create a partition $\mathcal{P}^{s,i}$ of $\mathcal{P}^{s,h}$ with respective cardinalities G^s by selectively combining sub-pathways based on prior knowledge and estimate an integrative model. Details on the combination scheme can be found in Appendix I along with plots and tables of parameters estimates. Note that the energy pathway is only constituted of two sub-pathways which cannot be combined.

We describe the marginal model for metabolites in pathway s. Denote by J^s the number of sub-pathways and M^s the number of metabolites in pathway s. Let $y^s_{ir,jk}$ denote the value of metabolite $r \in \{1,\dots,m_j\}$ in sub-pathway $j \in \{1,\dots,J^s\}$ for subject $i \in \{1,\dots,n_k\}$ in study $k \in \{1,\dots,K\}$, and let $\boldsymbol{y}^s_{jk} = (y^s_{ir,jk})_{i,r=1}^{n_k,m_j}$. Consider the marginal regression model

$$\begin{aligned} E(y^s_{ir,jk}) &= \theta^s_{jk,0} + \theta^s_{jk,1} smoking_{i,k} + \theta^s_{jk,2} age_{i,k} + \theta^s_{jk,3} BMI_{i,k} + \theta^s_{jk,4} drinking_{i,k} + \\ &+ \theta^s_{jk,5} bpmeds_{i,k} + \theta^s_{jk,6} lipidmeds_{i,k}, \qquad (5.4) \\ i &= 1,\dots,n_k,\ r = 1,\dots,m_j,\ j = 1,\dots,J^s,\ k = 1,\dots,4, \end{aligned}$$

where $smoking_i$ is subject i's smoking status (0 for non-smoker, 1 for smoker), $age_{i.k}$ is subject i's age (range: 45.3 to 74.4 years), $BMI_{i,k}$ is subject i's BMI (range: 16.9 to 55.4 kg/m^2), $drinking_{i,k}$ is an indicator for subject's i's alcohol consumption (0 for non-consumer, 1 for consumer), $bpmeds_{i,k}$ is an indicator for subject i's blood pressure medication use (0 for no use, 1 for use), and $lipidmeds_{i,k}$ is an indicator for subject i's lipid medication use (0 for no use, 1 for use), at the time of data collection. Let $\boldsymbol{\theta}^s_{jk} = (\theta^s_{0,jk},\dots,\theta^s_{6,jk})$.

Based on the integrative models, we find that the effect of smoking is significant in multiple sub-pathways. Of note, in the Xenobiotics pathway, the Tobacco metabolite sub-pathway is combined with multiple sub-pathways. The estimated effect of smoking in this integrated sub-pathway is 0.13 with a standard error of 0.011 and p-value 1.21×10^{-30} (95% confidence interval: $0.11, 0.15$). We observe that by combining sub-pathways in the integrative model we are able to borrow information across sub-pathways and obtain more precise inference.

5.5 Discussion

The proposed method can be viewed as a generalization of both seemingly unrelated regression and meta-analysis, striking a balance between the two that leverages correlation and partial homogeneity of regression equations. The distributed quadratic inference approach is privacy-preserving and computationally appealing because data source analyses can be run simultaneously in parallel and only one round of communication is necessary to compute the integrated estimator in (5.2). The test of model fit proposed in Theorem V.2 and the χ^2 test of homogeneity in (5.3) are derived from the unique properties of the generalised method of moments and provide a principled approach to model building that is lacking with other state of the art high-dimensional correlated data analysis techniques, such as generalised estimating equations.

The selective combination scheme over a data source partition has also been studied in Wang et al. (2016) and Tang and Song (2016). While we require specification of the data partition, their methods learn the partition in a data-driven way, which can be advantageous. Inference with the fused lasso, however, is burdened by debiasing methods that can be ill-defined or computationally burdensome. Additionally, the fused lasso approaches do not provide a formal procedure to test the validity of the parameter fusion scheme, relying instead on visualization such as dendograms. Our method is clearly advantageous when an approximately known partition exists.

Limitations of the proposed method include the need for a pre-defined partition of data sources defining regions of parameter homogeneity, which typically is given by related scientific knowledge but may occasionally be lacking in practice. Data preprocessing and learning and the test in (5.3) may help in determining an appropriate partition. Additionally, standard errors tend to be underestimated in small sample sizes or when the dimension of the moment conditions is large; this phenomenon has been well documented in the generalised method of moments (see Hansen et al. (1996) and others in the same issue).

CHAPTER VI

Summary and Future Work

Substantially motivated by analytic and computing needs in the analysis of high-dimensional correlated outcome data, this dissertation proposes novel distributed statistical methods with accompanying theory and R software implementations that address several modelling, computational and theoretical challenges.

The proposed methods specify local first and second-order models and aggregate them to form a complex, flexible model on the entire data. This formulation alleviates the modelling challenges stemming from high-dimensional likelihoods by avoiding the specification of high-order moments and incorporating heterogeneity of modelled moments. In Chapter III, local models specify first and second-order moments via the pairwise composite likelihood, which uses a working independence assumption between pairs of observations to avoid modelling moments of order higher than two. In Chapter IV, local models are generalized from the pairwise composite likelihood to a broad class of estimating functions for learning and carrying out inference on homogeneous and heterogeneous first and second-order moment parameters. In Chapter V, the estimation of second-order moments in marginal regression models is avoided altogether through the use of quadratic inference functions. These local model specifications allow heterogeneity in first and second-order moments across subvectors of the high-dimensional response, and leverage this heterogeneity for more precise estimation of the parameter of interest. Computational challenges arising from inversion of large matrices and iterative procedures over large amounts of data are overcome by estimating parameters on sub-datasets and aggregating results using a closed-form estimator. This distributed and parallelizable procedure circumvents the need for access to individual datasets and meets data privacy needs. It is communication efficient, requiring only one round of communication between distributed datasets and the main computing node. As shown in simulations and data analyses, CPU computing times are greatly reduced, and the distributed architecture reduces demands on random-access and read-only memory (RAM and ROM respectively). The R software implementation in Rcpp and RcppParallel makes for a user-friendly platform for direct use by biomedical investigators.

Theoretical challenges arise when integrating estimating functions or estimators from correlated data. The generalized method of moments provides a natural framework by non-parametrically estimating correlation between estimating functions. This approach also has intuitive justifications from standard estimating function theory and confidence estimating functions. The generalized method of moments benefits from a wealth of established properties but is burdened by the high computational cost of iteratively minimizing a quadratic form over the entire data. One of the key contributions of this dissertation is an asymptotically equivalent closed-form estimator that benefits from the same asymptotic properties conducive to inference and that is much more computationally tractable. Finally, theoretical challenges of estimation and inference when the dimension of the outcome diverges are addressed in Chapters III and IV.

A summary of the relative advantages and limitations of the three methods proposed in Chapters III, IV and V is given in Table 6.1. We first discuss the importance of the data split. The proposed methods split response data into blocks with the goal of creating low-dimensional responses with simple second order moment structures. This data split should be done according to pre-existing biological knowledge of the group structure of the outcome. The validity of the division can be checked by verifying the validity of the moment conditions on the parameters through a χ^2 goodness-of-fit test, as in Theorems III.3, IV.3 and V.2. When the data split is invalid, the sub-responses in each block will not have simple second order moment structures, resulting in a loss of efficiency in Chapters IV and V and incorrect standard error estimates in Chapter III (hence the limitation "quality of $\widehat{\boldsymbol{\gamma}}_j$" in Table 6.1). This can be mitigated by assuming an unstructured second order moment structure in each block, given sufficient data. This highlights a trade-off between data split and inference. Future work should explore the consequences of a misspecified data split through sensitivity analyses. Next, we discuss the stability of the inverse of the sample covariance matrix of estimating equations, $\widehat{\boldsymbol{V}}_N$, which affects the stability of the estimators proposed in this dissertation. In Big Data settings, the sample size is typically large enough that the inverse is stable: $\widehat{\boldsymbol{V}}_N$ is symmetric positive definite and invertible. If sample size is not large enough, the inverse can be replaced by the generalized inverse as in Wang et al. (2012) or a ridge estimator as in Han and Song (2011). Alternatively one may take the principal components of the estimating equations as in Cho and Qu (2015) to reduce their dimension and impose orthogonality, resulting in an invertible sample covariance matrix. These steps may slow computations and reduce the computational advantages of our methods. Finally, we highlight that Chapters

IV and V do not require marginal distributional assumptions on the outcome, and Chapter III only requires the specification of bivariate marginals. Future work should explore through sensitivity analyses the robustness of the proposed estimators to highly skewed outcome data.

Three important estimation methods were considered in the distributed step of the methods proposed in this dissertation: the pairwise composite likelihood (CL), generalized estimating equations (GEE), and quadratic inference functions (QIF). Several considerations may guide the choice of appropriate method in practice. The first consideration should be the target parameter of interest: both the CL and GEE can be used when the target are second order moments, whereas the QIF can not. Next, an investigator should evaluate the degree of certainty on the form of the distribution of the outcomes. The CL assumes an explicit form for the bivariate distribution between pairs of observations, whereas the GEE and QIF do not. The use of distributional assumptions versus estimating equations is also sometimes more of a philosophical consideration. An investigator should also consider their knowledge of second order moment structure. All three methods provide consistent estimates of model parameters regardless of how well specified this structure is. When this structure is correctly specified, the GEE and QIF are semi-parametrically efficient, as is the CL with Normally distributed outcomes. The CL will result in a small loss of efficiency with non-Normal outcomes. When the second order moment structure is misspecified, the CL estimator will have incorrect estimates of standard errors, the GEE estimator will be inefficient, and the QIF estimator will still be efficient within a class of estimating equations. Computation time varies greatly depending on implementation and specific application. In various simulations considered in this dissertation, the general take-away is that the QIF tends to outperform the CL and GEE in terms of computational speed, likely because the QIF does not need to estimate parameters of second order moments.

Table 6.1: Comparison of advantages and limitations of methods proposed in Chapters III, IV and V.

	Chapter III	Chapter IV	Chapter V
Estimation	CL	zero-mean, "weakly" regular, additive estimating equation	QIF
Advantages	only estimate mean and covariance	general & flexible	only estimate mean
		homogeneous & heterogeneous parameters	combination over sets with parameter homogeneity
	fast	faster	fastest
Limitations	quality of $\widehat{\boldsymbol{\gamma}}_j$	high-dimensional $\boldsymbol{\zeta}$	inversion of $\widehat{\boldsymbol{V}}_N$

Future research should extend the methods proposed in this dissertation to the integration of multimodal data, an important analytic task in biomedical data analysis for personalized medicine. The method proposed in Chapter V jointly estimates marginal regression parameters and integrates them over regions of parameter homogeneity, so that integrated estimators may come from different model specifications. This framework needs only minor work to ensure the valid construction of the multivariate distribution of different outcome types, such as binary, count and continuous outcomes.

Also of interest is the extension to high-dimensional covariate settings where the dimension of the parameter of interest is larger than the individual sample sizes in each data source. One approach is the inclusion of penalty terms in the analysis of each data source followed by fusion of similar estimates in the integration step (Tang and Song, 2016). This framework relies on debiasing of the estimators from each data source, which can be challenging. Additionally, this does not accommodate the case where the dimension of the parameter of interest is larger than the combined sample size. Alternative approaches through random data splitting (Wang et al., 2020) show promise and should be explored.

Finally, extensions of the proposed methods to stochastic processes, such as spatial or time series data, are of particular interest since these settings display very large dimension, correlation and heterogeneity. The mean regression parameter can be expressed as an expansion of known basis matrices with unknown coefficients that are the parameter interest and can be estimated through the methods proposed in this dissertation. Careful consideration should be given to the construction of the joint distribution of the outcome in the theoretical justification.

APPENDICES

APPENDIX A

Chapter III: Proofs

Conditions for proofs of Chapter III

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

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

$$\left\|\boldsymbol{\psi}_{j.sub}(\boldsymbol{\beta}_1;\boldsymbol{y}_i,\boldsymbol{\gamma}_{j1})-\boldsymbol{\psi}_{j.sub}(\boldsymbol{\beta}_2;\boldsymbol{y}_i,\boldsymbol{\gamma}_{j2})\right\| \leq C\left(\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|+\left\|\boldsymbol{\gamma}_{j1}-\boldsymbol{\gamma}_{j2}\right\|\right)$$

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

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

C.3 Let $\widehat{\boldsymbol{\beta}_c}$ be as defined in (3.5), and $\boldsymbol{\beta}_0$ an interior point of Θ. Following Newey and McFadden (1994), assume $Q_N(\widehat{\boldsymbol{\beta}_c}) \leq \inf_{\boldsymbol{\beta}\in\Theta} Q_N(\boldsymbol{\beta}) + o_p(1)$, and, for any $\delta_N \to 0$,

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leq\delta_N} \frac{\sqrt{N}}{1+\sqrt{N}\,\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|}\,\|\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y})-\boldsymbol{\Psi}_N(\boldsymbol{\beta}_0;\boldsymbol{y})-E_{\boldsymbol{\beta}}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\| \xrightarrow{p} 0.$$

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

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

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 Proposition 1: We follow the argument of Cox and Reid (2004). Let

$$\begin{aligned}\boldsymbol{U}_{j.sub}(\boldsymbol{\beta},\boldsymbol{\gamma}_j;\boldsymbol{y}_{i,j}) &= \sum_{r=1}^{m_j-1}\sum_{t=r+1}^{m_j} \boldsymbol{U}_{j.sub,rt}(\boldsymbol{\beta},\boldsymbol{\gamma}_j;y_{ir,j},y_{it,j}) \\ &= \sum_{r=1}^{m_j-1}\sum_{t=r+1}^{m_j} \nabla_{\boldsymbol{\beta},\boldsymbol{\gamma}_j} \log f_j(y_{ir,j},y_{it,j};\boldsymbol{\beta},\boldsymbol{\gamma}_j,\boldsymbol{X}_{i,j}).\end{aligned}$$

Let $(\boldsymbol{\beta}^*,\boldsymbol{\gamma}_j^*)$ be between $(\widehat{\boldsymbol{\beta}_j},\widehat{\boldsymbol{\gamma}_j})$ and $(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{j0})$. A Taylor expansion of $\boldsymbol{U}_{j.sub}(\widehat{\boldsymbol{\beta}_j},\widehat{\boldsymbol{\gamma}_j};\boldsymbol{y}_{i,j})$ around $(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{j0})$ yields

$$\begin{aligned}\boldsymbol{U}_{j.sub}(\widehat{\boldsymbol{\beta}_j},\widehat{\boldsymbol{\gamma}_j};\boldsymbol{y}_{i,j}) = \mathbf{0} = \frac{1}{m_j^2}\sum_{r=1}^{m_j-1}\sum_{t=r+1}^{m_j} \boldsymbol{U}_{j.sub,rt}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{j0};y_{ir,j},y_{it,j}) + \\ \begin{pmatrix} \widehat{\boldsymbol{\beta}_j} - \boldsymbol{\beta}^* \\ \widehat{\boldsymbol{\gamma}_j} - \boldsymbol{\gamma}_j^* \end{pmatrix}^T \frac{1}{m_j^2}\sum_{r=1}^{m_j-1}\sum_{t=r+1}^{m_j} \nabla_{\boldsymbol{\beta},\boldsymbol{\gamma}_j}\boldsymbol{U}_{j.sub,rt}(\boldsymbol{\beta},\boldsymbol{\gamma}_j;y_{ir,j},y_{it,j})\Big|_{\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{j0}}.\end{aligned}$$

The first term in the expansion has mean $\mathbf{0}$ and variance that has leading terms in m_j corresponding to

$$\frac{1}{4m_j^4}\sum_{\substack{r,t,v,w=1 \\ r\neq t\neq v\neq w}}^{m_j} E\left\{\boldsymbol{U}_{j.sub,rt}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{j0};y_{ir,j},y_{it,j})\boldsymbol{U}_{j.sub,vw}^T(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{j0};y_{iv,j},y_{iw,j})\right\} = O_p(1).$$

As $N \to \infty$, the CL score function $(1/N)\sum_{i=1}^N \boldsymbol{U}_{j.sub}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{j0};\boldsymbol{y}_{i,j})$ converges in probability to $\mathbf{0}$. Similarly, one can show $\nabla_{\boldsymbol{\beta},\boldsymbol{\gamma}_j}\boldsymbol{U}_{j.sub}(\boldsymbol{\beta},\boldsymbol{\gamma}_j;\boldsymbol{y}_{i,j})\big|_{\boldsymbol{\beta}_0,\boldsymbol{\gamma}_{j0}}$ has bounded variance. Consequently, consistency of the block-specific MCLE's $\widehat{\boldsymbol{\beta}_j}$ and $\widehat{\boldsymbol{\gamma}_j}$ can be established using Theorem 3.4 of Song (2007). □

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

and $\widehat{\gamma}_j - \boldsymbol{\gamma}_{j0} = o_p(1)$. Since J, p finite, $\left\|\widehat{\boldsymbol{\beta}}_{MCLE} - \boldsymbol{\beta}_0\right\| = o_p(1)$ and $\|\widehat{\boldsymbol{\gamma}}_{MCLE} - \boldsymbol{\gamma}_0\| = o_p(1)$. Then by C.1,

$$\left\|\boldsymbol{\psi}(\widehat{\boldsymbol{\beta}}_{MCLE}; \boldsymbol{y}_i) - \boldsymbol{\psi}(\boldsymbol{\beta}_0; \boldsymbol{y}_i)\right\| \leq C\left(\left\|\widehat{\boldsymbol{\beta}}_{MCLE} - \boldsymbol{\beta}_0\right\| + \|\widehat{\boldsymbol{\gamma}}_{MCLE} - \boldsymbol{\gamma}_0\|\right) = o_p(1).$$

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

Proof of Theorem III.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 $Q_0(\boldsymbol{\beta})$.

$$\begin{aligned}
&\left\|\frac{1}{N}Q_N(\boldsymbol{\beta}) - Q_0(\boldsymbol{\beta})\right\| \\
&= \Big\| \boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{y})\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y}) \\
&- 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}) + 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}) \\
&- 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) + 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) \\
&- E_{\boldsymbol{\beta}}\left(\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{Y})\right)\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)E_{\boldsymbol{\beta}}\left(\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right)\Big\| \\
&= \Big\| \boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{y})\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y}) - \boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{y})\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}E_{\boldsymbol{\beta}}\left(\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right) \\
&- E_{\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}) + E_{\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) \\
&+ 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}) - 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) \\
&+ E_{\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) - E_{\boldsymbol{\beta}}\left(\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{Y})\right)\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)E_{\boldsymbol{\beta}}\left(\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right)\Big\| \\
&\le \left\| \left[\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y}) - E_{\boldsymbol{\beta}}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right]^T \widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\left[\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y}) - E_{\boldsymbol{\beta}}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right]\right\| \\
&+ 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}}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right]\right\| \\
&+ \left\| E_{\boldsymbol{\beta}}\left(\boldsymbol{\Psi}_N^T(\boldsymbol{\beta};\boldsymbol{Y})\right)\left[\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1} - \boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)\right]E_{\boldsymbol{\beta}}\left(\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right)\right\| \\
&\le \left\|\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y}) - E_{\boldsymbol{\beta}}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right\|^2 \left\|\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\right\| \\
&+ 2\left\|E_{\boldsymbol{\beta}}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right\| \left\|\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{y}) - E_{\boldsymbol{\beta}}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right\| \left\|\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\right\| \\
&+ \left\|E_{\boldsymbol{\beta}}\boldsymbol{\Psi}_N(\boldsymbol{\beta};\boldsymbol{Y})\right\|^2 \left\|\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1} - \boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)\right\| \\
&= O_p(N^{-1/2}) + o_p(1).
\end{aligned}$$

It follows that $\sup_{\boldsymbol{\beta}\epsilon\Theta}\left\|\frac{1}{N}Q_N(\boldsymbol{\beta}) - Q_0(\boldsymbol{\beta})\right\| \xrightarrow{p} 0$ 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 $N \to \infty$. □

Proof of Theorem III.3: We take a Taylor expansion of $E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\boldsymbol{\beta};\boldsymbol{y}_i)$ about $\boldsymbol{\beta}_0$:

$$E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\widehat{\boldsymbol{\beta}_c};\boldsymbol{y}_i) - \underbrace{E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\boldsymbol{\beta}_0;\boldsymbol{y}_i)}_{=0} = \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\widehat{\boldsymbol{\beta}_c} - \boldsymbol{\beta}_0), \tag{A.1}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}_c}$. By condition C.3,

$$\begin{aligned}\boldsymbol{\Psi}_N(\widehat{\boldsymbol{\beta}_c};\boldsymbol{y}) - \boldsymbol{\Psi}_N(\boldsymbol{\beta}_0;\boldsymbol{y}) - E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\widehat{\boldsymbol{\beta}_c};\boldsymbol{y}_i) &= o_p(1)\frac{1+\sqrt{N}\left\|\widehat{\boldsymbol{\beta}_c}-\boldsymbol{\beta}_0\right\|}{\sqrt{N}} \\ &= o_p\left(N^{-1/2}+\left\|\widehat{\boldsymbol{\beta}_c}-\boldsymbol{\beta}_0\right\|\right). \qquad \text{(A.2)}\end{aligned}$$

Then adding (A.1) and (A.2) yields

$$\boldsymbol{\Psi}_N(\widehat{\boldsymbol{\beta}_c};\boldsymbol{y}) = \boldsymbol{\Psi}_N(\boldsymbol{\beta}_0;\boldsymbol{y}) + \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\widehat{\boldsymbol{\beta}_c}-\boldsymbol{\beta}_0) + o_p\left(N^{-1/2}+\left\|\widehat{\boldsymbol{\beta}_c}-\boldsymbol{\beta}_0\right\|\right). \qquad \text{(A.3)}$$

As the minimizer of $Q_N(\boldsymbol{\beta})$, $\widehat{\boldsymbol{\beta}_c}$ satisfies $\nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\boldsymbol{\Psi}_N(\widehat{\boldsymbol{\beta}_c};\boldsymbol{y}) = 0$. We premultiply (A.3) by $\nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}$ and obtain

$$\begin{aligned}0 &= \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\boldsymbol{\Psi}_N(\widehat{\boldsymbol{\beta}_c};\boldsymbol{y}) \\ &= \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\boldsymbol{\Psi}_N(\boldsymbol{\beta}_0;\boldsymbol{y}) \\ &\quad + \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\widehat{\boldsymbol{\beta}_c}-\boldsymbol{\beta}_0) \\ &\quad + o_p\left(N^{-1/2}+\left\|\widehat{\boldsymbol{\beta}_c}-\boldsymbol{\beta}_0\right\|\right)\nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1} \\ &= \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\boldsymbol{\Psi}_N(\boldsymbol{\beta}_0;\boldsymbol{y}) \\ &\quad + \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\widehat{\boldsymbol{\beta}_c}-\boldsymbol{\beta}_0) \\ &\quad + o_p\left(N^{-1/2}+\left\|\widehat{\boldsymbol{\beta}_c}-\boldsymbol{\beta}_0\right\|\right),\end{aligned}$$

since $\nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\boldsymbol{\beta};\boldsymbol{y}_i)$ is smooth in a neighbourhood of $\boldsymbol{\beta}_0$. Rearranging yields

$$\begin{aligned}\widehat{\boldsymbol{\beta}_c}-\boldsymbol{\beta}_0 = -\Big[\nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\Big]^{-1} \times \\ \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\boldsymbol{\Psi}_N(\boldsymbol{\beta}_0;\boldsymbol{y}) + o_p\left(N^{-1/2}+\left\|\widehat{\boldsymbol{\beta}_c}-\boldsymbol{\beta}_0\right\|\right).\end{aligned}$$

We plug this back into (A.3) to get

$$\begin{aligned}\sqrt{N}\boldsymbol{\Psi}_N(\widehat{\boldsymbol{\beta}_c};\boldsymbol{y}) = \Big[I - \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\left\{\nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right\}^{-1} \\ \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\Big] \times \sqrt{N}\boldsymbol{\Psi}_N(\boldsymbol{\beta}_0;\boldsymbol{y}) + o_p(1).\end{aligned}$$

By the Central Limit Theorem, $\sqrt{N}\boldsymbol{\Psi}_N(\boldsymbol{\beta}_0;\boldsymbol{y}) \overset{d}{\to} \mathcal{N}\left(0, \boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)\right)$. Moreover, since $\widehat{\boldsymbol{\beta}_c} \overset{p}{\to} \boldsymbol{\beta}_0$,

$$\begin{aligned}&\nabla_{\boldsymbol{\beta}} E_{\boldsymbol{\beta}} \boldsymbol{\psi}(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \left\{\nabla_{\boldsymbol{\beta}} E_{\boldsymbol{\beta}} \boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}} \widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1} \nabla_{\boldsymbol{\beta}} E_{\boldsymbol{\beta}} \boldsymbol{\psi}(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right\}^{-1} \nabla_{\boldsymbol{\beta}} E_{\boldsymbol{\beta}} \boldsymbol{\psi}^T(\boldsymbol{\beta};\boldsymbol{y}_i)|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}_c}} \widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1} \\ &\overset{p}{\to} \boldsymbol{s}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)\left\{\boldsymbol{s}_{\boldsymbol{\psi}}^T(\boldsymbol{\beta}_0)\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)\boldsymbol{s}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)\right\}^{-1} \boldsymbol{s}_{\boldsymbol{\psi}}^T(\boldsymbol{\beta}_0)\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0) = G(\boldsymbol{\beta}_0).\end{aligned}$$

Then by Slutsky's theorem, $\sqrt{N}\boldsymbol{\Psi}_N(\widehat{\boldsymbol{\beta}_c};\boldsymbol{y}) \overset{d}{\to} \mathcal{N}\left(0, \left(I - G(\boldsymbol{\beta}_0)\right)\boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)\left(I - G(\boldsymbol{\beta}_0)\right)^T\right)$. Finally, write $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}} = \widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{1/2}\left(\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{1/2}\right)^T$ such that $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{1/2} \overset{p}{\to} \boldsymbol{v}_{\boldsymbol{\psi}}^{1/2}(\boldsymbol{\beta}_0)$ and $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{1/2}$ nonsingular (some normalization may be required). Then

$$\sqrt{N}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1/2}\boldsymbol{\Psi}_N(\widehat{\boldsymbol{\beta}_c};\boldsymbol{y}) \overset{d}{\to} \mathcal{N}\left(0, \left\{\boldsymbol{v}_{\boldsymbol{\psi}}^{-1/2}(\boldsymbol{\beta}_0)(I - G(\boldsymbol{\beta}_0))\right\}\boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)\left\{\boldsymbol{v}_{\boldsymbol{\psi}}^{-1/2}(\boldsymbol{\beta}_0)(I - G(\boldsymbol{\beta}_0))\right\}^T\right).$$

We note that this covariance matrix is idempotent with rank $Jp - p$. Then

$$Q_N(\widehat{\boldsymbol{\beta}_c}) = N\boldsymbol{\Psi}_N^T(\widehat{\boldsymbol{\beta}_c};\boldsymbol{y})\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\boldsymbol{\Psi}_N(\widehat{\boldsymbol{\beta}_c};\boldsymbol{y}) \overset{d}{\to} \chi^2_{(J-1)p} \text{ as } N \to \infty. \qquad \square$$

Proof of Theorem III.4: We proceed with the proof in three steps: we first show that $\widehat{\boldsymbol{\beta}}_{DIMM}$ is consistent, then we show it has the same asymptotic distribution as $\widehat{\boldsymbol{\beta}_c}$. Finally, we show that $\widehat{\boldsymbol{\beta}_c}$ and $\widehat{\boldsymbol{\beta}}_{DIMM}$ are asymptotically equivalent. Some important results we use include:

- We showed in Lemma III.1 that $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}} = \boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0) + o_p(1)$. Under condition C.4, we can show that $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}} = \boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0) + O_p(N^{-1/2})$. Indeed, using notation from the proof of Lemma III.1,

$$\left\|\boldsymbol{\psi}(\widehat{\boldsymbol{\beta}}_{MCLE};\boldsymbol{y}_i) - \boldsymbol{\psi}(\boldsymbol{\beta}_0;\boldsymbol{y}_i)\right\| \le C\left(\left\|\widehat{\boldsymbol{\beta}}_{MCLE} - \boldsymbol{\beta}_0\right\| + \left\|\widehat{\boldsymbol{\gamma}}_{MCLE} - \boldsymbol{\gamma}_0\right\|\right) = O_p(N^{-1/2}).$$

Plugging in to the formula for $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ yields

$$\begin{aligned}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}} &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\psi}(\widehat{\boldsymbol{\beta}}_{MCLE};\boldsymbol{y}_i)\boldsymbol{\psi}^T(\widehat{\boldsymbol{\beta}}_{MCLE};\boldsymbol{y}_i)\\ &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\psi}(\boldsymbol{\beta}_0;\boldsymbol{y}_i)\boldsymbol{\psi}^T(\boldsymbol{\beta}_0;\boldsymbol{y}_i) + O_p(N^{-1/2})\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\psi}(\boldsymbol{\beta}_0;\boldsymbol{y}_i) + O_p(N^{-1})\\ &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\psi}(\boldsymbol{\beta}_0;\boldsymbol{y}_i)\boldsymbol{\psi}^T(\boldsymbol{\beta}_0;\boldsymbol{y}_i) + O_p(N^{-1})\\ &= v_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0) + O_p(N^{-1/2}).\end{aligned}$$

Therefore $\left[\widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}}\right]_{i,j} = \left[\boldsymbol{v}^{-1}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)\right]_{i,j} + O_p(N^{-1/2})$.

- We require that $\boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j) = \boldsymbol{s}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}_j}) + O_p(N^{-1/2})$. Then by C.4, $\boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j) = \boldsymbol{s}_{j,\boldsymbol{\psi}_j}(\boldsymbol{\beta}_0) + O_p(N^{-1/2})$. Similarly, $\boldsymbol{S}_{\boldsymbol{\psi}}(\widehat{\boldsymbol{\beta}_c};\boldsymbol{y}) = \boldsymbol{s}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0) + O_p(N^{-1/2})$.

Consistency of $\widehat{\boldsymbol{\beta}}_{DIMM}$: Define

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

By definition, $\lambda(\widehat{\boldsymbol{\beta}}_{DIMM}) = 0$. Moreover,

$$\begin{aligned}\lambda(\boldsymbol{\beta}_0) &= \sum_{i,j=1}^{J}\boldsymbol{S}_{i,\boldsymbol{\psi}_i}(\widehat{\boldsymbol{\beta}}_i;\boldsymbol{y}_i)\left[\widehat{\boldsymbol{V}}^{-1}_{N,\boldsymbol{\psi}}\right]_{i,j}\boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}}_j;\boldsymbol{y}_j)\left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}_j}\right)\\ &= \sum_{i,j=1}^{J}\left(\boldsymbol{s}_{i,\boldsymbol{\psi}_i}(\boldsymbol{\beta}_0) + O_p(N^{-1/2})\right)\left(\left[\boldsymbol{v}^{-1}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)\right]_{i,j} + O_p(N^{-1/2})\right)\times\\ &\quad\left(\boldsymbol{s}_{j,\boldsymbol{\psi}_j}(\boldsymbol{\beta}_0) + O_p(N^{-1/2})\right)o_p(1)\\ &= \sum_{i,j=1}^{J}o_p(1) = o_p(1).\end{aligned}$$

Since $\nabla_{\boldsymbol{\beta}}\lambda(\boldsymbol{\beta})$ exists and is nonsingular, for some $\boldsymbol{\beta}^*$ between $\widehat{\boldsymbol{\beta}}_{DIMM}$ and $\boldsymbol{\beta}_0$ we can write $\lambda(\widehat{\boldsymbol{\beta}}_{DIMM}) - \lambda(\boldsymbol{\beta}_0) = \nabla_{\boldsymbol{\beta}}\lambda(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\widehat{\boldsymbol{\beta}}_{DIMM} - \boldsymbol{\beta}_0) = o_p(1)$. Therefore $\widehat{\boldsymbol{\beta}}_{DIMM} = \boldsymbol{\beta}_0 + o_p(1)$.

<u>Distribution of $\widehat{\boldsymbol{\beta}}_{DIMM}$:</u> We can rewrite

$$0 = \lambda(\widehat{\boldsymbol{\beta}}_{DIMM}) = \sum_{i,j=1}^{J} \boldsymbol{S}_{i,\psi_i}(\widehat{\boldsymbol{\beta}}_i;\boldsymbol{y}_i)\left[\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\right]_{i,j}\boldsymbol{S}_{j,\psi_j}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j)\left(\widehat{\boldsymbol{\beta}}_{DIMM} - \boldsymbol{\beta}_0 + \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}_j}\right).$$

Rearranging yields

$$\widehat{\boldsymbol{\beta}}_{DIMM} - \boldsymbol{\beta}_0 = \left\{\sum_{i,j=1}^{J} \boldsymbol{S}_{i,\psi_i}(\widehat{\boldsymbol{\beta}}_i;\boldsymbol{y}_i)\left[\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\right]_{i,j}\boldsymbol{S}_{j,\psi_j}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j)\right\}^{-1}\times \left\{\sum_{i,j=1}^{J} \boldsymbol{S}_{i,\psi_i}(\widehat{\boldsymbol{\beta}}_i;\boldsymbol{y}_i)\left[\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\right]_{i,j}\boldsymbol{S}_{j,\psi_j}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j)\left(\widehat{\boldsymbol{\beta}_j} - \boldsymbol{\beta}_0\right)\right\}. \tag{A.4}$$

By Taylor expansion we have

$$E_{\boldsymbol{\beta}}\boldsymbol{\psi}_{j.sub}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_{i,j},\boldsymbol{\gamma}_{j0}) = E_{\boldsymbol{\beta}}\boldsymbol{\psi}_{j.sub}(\boldsymbol{\beta}_0;\boldsymbol{y}_{i,j},\boldsymbol{\gamma}_{j0}) + \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}_{j.sub}(\boldsymbol{\beta};\boldsymbol{y}_{i,j},\boldsymbol{\gamma}_{j0})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\widehat{\boldsymbol{\beta}_j} - \boldsymbol{\beta}_0) \tag{A.5}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}_j}$. By condition C.4,

$$\boldsymbol{\Psi}_{j.sub}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j,\widehat{\boldsymbol{\gamma}_j}) - \boldsymbol{\Psi}_{j.sub}(\boldsymbol{\beta}_0;\boldsymbol{y}_j,\boldsymbol{\gamma}_{j0}) - E_{\boldsymbol{\beta}}\boldsymbol{\psi}_{j.sub}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_{i,j},\boldsymbol{\gamma}_{j0}) = O_p(N^{-1}). \tag{A.6}$$

Adding (A.5) and (A.6), we have

$$\begin{aligned}\underbrace{\boldsymbol{\Psi}_{j.sub}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j,\widehat{\boldsymbol{\gamma}_j})}_{=0} - \boldsymbol{\Psi}_{j.sub}(\boldsymbol{\beta}_0;\boldsymbol{y}_j,\boldsymbol{\gamma}_{j0}) - \underbrace{E_{\boldsymbol{\beta}}\boldsymbol{\psi}_{j.sub}(\boldsymbol{\beta}_0;\boldsymbol{y}_{i,j},\boldsymbol{\gamma}_{j0})}_{=0}&\\ = \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}_{j.sub}(\boldsymbol{\beta};\boldsymbol{y}_{i,j},\boldsymbol{\gamma}_{j0})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\widehat{\boldsymbol{\beta}_j} - \boldsymbol{\beta}_0) + O_p(N^{-1}).&\end{aligned}$$

Rearranging yields

$$\boldsymbol{\Psi}_{j.sub}(\boldsymbol{\beta}_0;\boldsymbol{y}_j,\boldsymbol{\gamma}_{j0}) = \nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}_{j.sub}(\boldsymbol{\beta};\boldsymbol{y}_{i,j},\boldsymbol{\gamma}_{j0})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}_j}) + O_p(N^{-1}).$$

Finally, note that

$$\boldsymbol{S}_{j,\psi_j}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j) = \boldsymbol{s}_{j,\psi_j}(\boldsymbol{\beta}_0) + O_p(N^{-1/2}) = -\nabla_{\boldsymbol{\beta}}E_{\boldsymbol{\beta}}\boldsymbol{\psi}_{j.sub}(\boldsymbol{\beta};\boldsymbol{y}_{i,j},\boldsymbol{\gamma}_{j0})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} + O_p(N^{-1/2}),$$

so that

$$\begin{aligned}\boldsymbol{\Psi}_{j.sub}(\boldsymbol{\beta}_0;\boldsymbol{y}_j,\boldsymbol{\gamma}_{j0}) &= \left(\boldsymbol{S}_{j,\psi_j}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j) + O_p(N^{-1/2})\right)(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}_j}) + O_p(N^{-1})\\ &= \boldsymbol{S}_{j,\psi_j}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j)(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}_j}) + O_p(N^{-1}).\end{aligned} \tag{A.7}$$

Recall that, by the Central Limit Theorem, $\sqrt{N}\boldsymbol{\Psi}_N(\boldsymbol{\beta}_0;\boldsymbol{y}) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)\right)$. Then by equation (A.7), we can write

$$\sqrt{N}\begin{pmatrix} \boldsymbol{S}_{1,\boldsymbol{\psi}_1}(\widehat{\boldsymbol{\beta}_1};\boldsymbol{y}_1)(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}_1}) \\ \vdots \\ \boldsymbol{S}_{J,\boldsymbol{\psi}_J}(\widehat{\boldsymbol{\beta}_J};\boldsymbol{y}_J)(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}_J}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{v}_{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)\right).$$

We have that $\boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j) = \boldsymbol{s}_{j,\boldsymbol{\psi}_j}(\boldsymbol{\beta}_0) + O_p(N^{-1/2})$. By Slutsky we can write,

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{MCLE} - \boldsymbol{\beta}_0) = \sqrt{N}\begin{pmatrix} \widehat{\boldsymbol{\beta}_1} - \boldsymbol{\beta}_0 \\ \vdots \\ \widehat{\boldsymbol{\beta}_J} - \boldsymbol{\beta}_0 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \left\{\sum_{i,j=1}^{J} \boldsymbol{s}_{i,\boldsymbol{\psi}_i}(\boldsymbol{\beta}_0)\left[\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)\right]_{i,j} \boldsymbol{s}_{j,\boldsymbol{\psi}_j}(\boldsymbol{\beta}_0)\right\}^{-1}\right).$$

The sum of jointly Normal variables is also normal. Using this and Slutsky again, we have

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{DIMM} - \boldsymbol{\beta}_0) = \sqrt{N}\left\{\sum_{i,j=1}^{J} \boldsymbol{S}_{i,\boldsymbol{\psi}_i}(\widehat{\boldsymbol{\beta}}_i;\boldsymbol{y}_i)\left[\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\right]_{i,j} \boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j)\right\}^{-1}$$
$$\times \sum_{i,j=1}^{J} \boldsymbol{S}_{i,\boldsymbol{\psi}_i}(\widehat{\boldsymbol{\beta}}_i;\boldsymbol{y}_i)\left[\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}\right]_{i,j} \boldsymbol{S}_{j,\boldsymbol{\psi}_j}(\widehat{\boldsymbol{\beta}_j};\boldsymbol{y}_j)(\widehat{\boldsymbol{\beta}_j} - \boldsymbol{\beta}_0)$$

is asymptotically Normally distributed with mean 0 and variance

$$\boldsymbol{j}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0) = \left\{\sum_{i,j=1}^{J} \boldsymbol{s}_{i,\boldsymbol{\psi}_i}(\boldsymbol{\beta}_0)\left[\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)\right]_{i,j} \boldsymbol{s}_{j,\boldsymbol{\psi}_j}(\boldsymbol{\beta}_0)\right\}^{-1}.$$

Asymptotic equivalency of $\widehat{\boldsymbol{\beta}_c}$ and $\widehat{\boldsymbol{\beta}}_{DIMM}$: We can write

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}_c} - \boldsymbol{\beta}_0\right) = Z + o_p(1)$$
$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{DIMM} - \boldsymbol{\beta}_0\right) = Z + o_p(1),$$

where $Z \sim \mathcal{N}\left(0, \boldsymbol{j}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)\right)$. Rearranging yields

$$\widehat{\boldsymbol{\beta}_c} - \boldsymbol{\beta}_0 = \frac{1}{\sqrt{N}}Z + o_p\left(N^{-1/2}\right) \tag{A.8}$$

$$\widehat{\boldsymbol{\beta}}_{DIMM} - \boldsymbol{\beta}_0 = \frac{1}{\sqrt{N}}Z + o_p\left(N^{-1/2}\right). \tag{A.9}$$

Subtracting (A.9) from (A.8), we get $\widehat{\boldsymbol{\beta}_c} - \widehat{\boldsymbol{\beta}}_{DIMM} = o_p(N^{-1/2})$. Then $\left\|\widehat{\boldsymbol{\beta}_c} - \widehat{\boldsymbol{\beta}}_{DIMM}\right\| \overset{p}{\to} 0$ as $N \to \infty$. □

APPENDIX B

Chapter III: Regularization of Weight Matrix

Additional details on regularizing $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$

To tackle the potential difficulty of inverting $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ which can arise with large Jp ($Jp \approx 5000$), we propose to use a regularized modified Cholesky decomposition of $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ following Pourahmadi (1999). The modified Cholesky decomposition of $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ can be written as $\boldsymbol{T}\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}\boldsymbol{T}^T = \boldsymbol{D}$ where $\boldsymbol{T}$ is lower triangular with 1's as diagonal entries and $\boldsymbol{D}$ is diagonal. Entries of $\boldsymbol{T}$ are the negatives of the regression coefficients from regressing each row of $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}$ on the previous rows. To achieve sparsity in the estimate of $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}$ and to speed up computation, this regression can be regularized with an L_2 norm penalty depending on the choice of a regularization parameter λ. A regularized estimate of $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}$ is then $\widehat{\boldsymbol{W}}(\lambda) = \boldsymbol{T}^T\boldsymbol{D}^{-1}\boldsymbol{T}$. This computation requires the computation of the inverse of a diagonal matrix, which is fast to compute, and the selection of λ, which can be done by cross validation. We can compute an estimate of $\boldsymbol{\beta}$ using this regularized inverse as

$$\widehat{\boldsymbol{\beta}}_{reg} = \left(\sum_{i,j=1}^{J} \boldsymbol{S}_{i,\psi_i}(\widehat{\boldsymbol{\beta}}_i; \boldsymbol{y}_i) \left[\widehat{\boldsymbol{W}}(\lambda) \right]_{i,j} \boldsymbol{S}_{j,\psi_j}(\widehat{\boldsymbol{\beta}_j}; \boldsymbol{y}_j) \right)^{-1} \sum_{i,j=1}^{J} \boldsymbol{S}_{i,\psi_i}(\widehat{\boldsymbol{\beta}}_i; \boldsymbol{y}_i) \left[\widehat{\boldsymbol{W}}(\lambda) \right]_{i,j} \boldsymbol{S}_{j,\psi_j}(\widehat{\boldsymbol{\beta}_j}; \boldsymbol{y}_j) \widehat{\boldsymbol{\beta}_j}.$$

It follows from Newey and McFadden (1994) that $\widehat{\boldsymbol{\beta}}_{reg}$ is a consistent estimate of $\boldsymbol{\beta}$ and has an asymptotically normal distribution under mild conditions on $\boldsymbol{\Psi}_N$, Q_0

in C.2 and $\widehat{\boldsymbol{W}}(\lambda)$. Moreover, if $\widehat{\boldsymbol{W}}(\lambda)$ is a $\sqrt{N}$-consistent estimate of $\boldsymbol{v}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\beta}_0)$ and conditions C.1-C.4 hold with $\widehat{\boldsymbol{W}}(\lambda)$ instead of $\widehat{\boldsymbol{V}}_{N,\boldsymbol{\psi}}^{-1}$, then it clearly follows from the proofs in this chapter that $\widehat{\boldsymbol{\beta}}_{reg}$ is a consistent estimator of $\boldsymbol{\beta}$ and follows the same asymptotic distribution as $\widehat{\boldsymbol{\beta}_c}$ and $\widehat{\boldsymbol{\beta}}_{DIMM}$.

APPENDIX C

Chapter III: Additional Simulation Results

Additional simulation results

We present plots of type-1 error rates for each scenario considered in Section 5 in Figure C.1, chi-squared Q-Q plot of goodness-of-fit test statistics in Figures C.2-C.5.

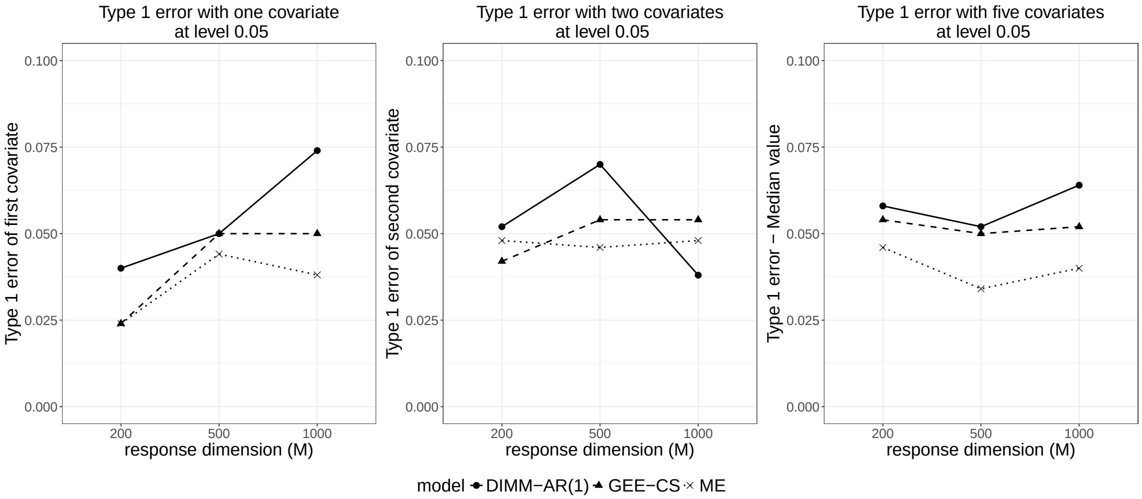


Figure C.1: Comparison of type-1 error rate for three methods for varying dimension M based on 500 simulations. Left column has $X_1 \sim \mathcal{N}(0,1)$; middle column has $X_1 \sim \mathcal{N}_M(0,S)$, where S is a positive-definite $M \times M$ matrix, and X_2 a vector of alternating 0's and 1's; 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.

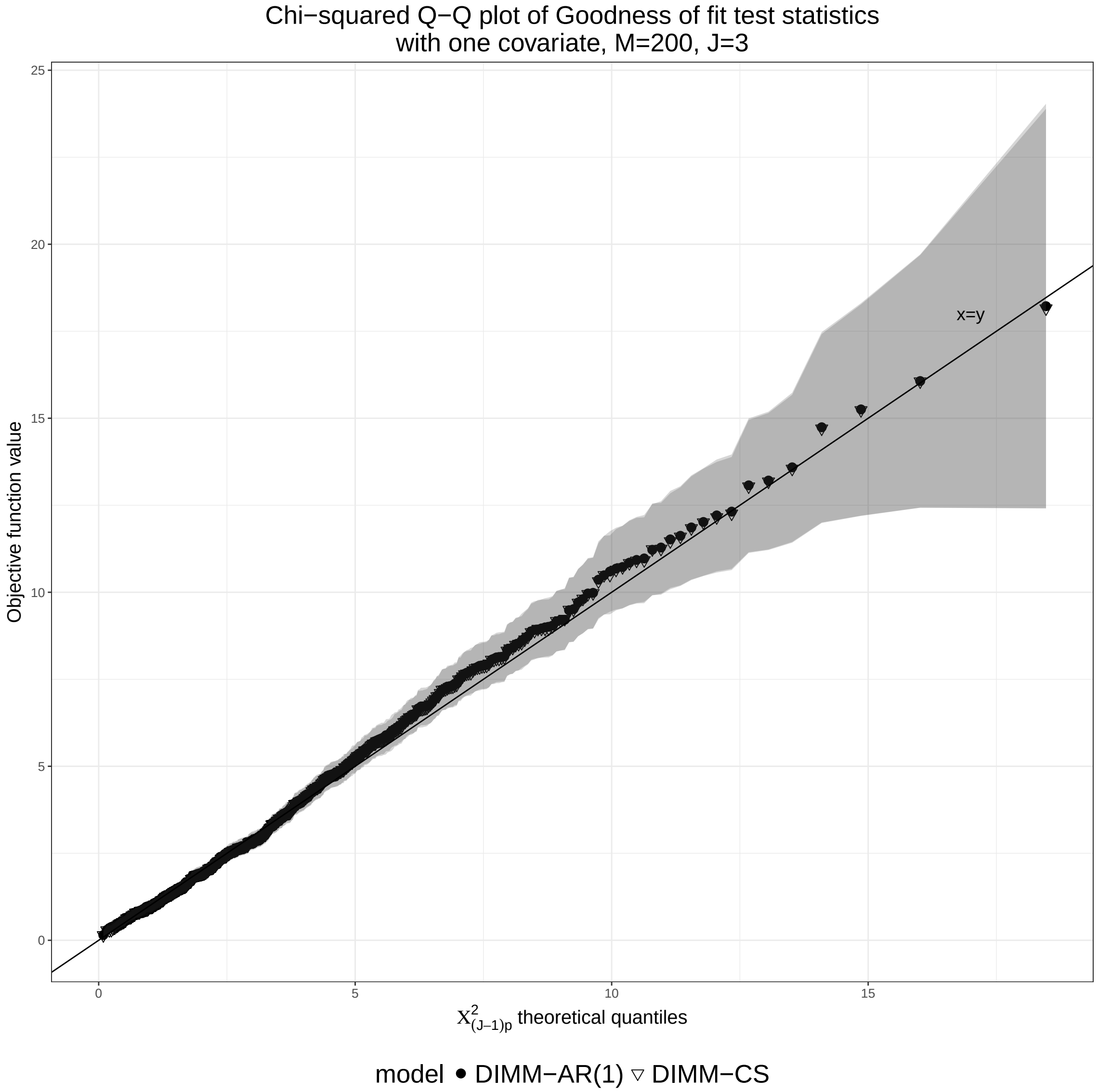


Figure C.2: Chi-squared Q-Q plot of goodness-of-fit test statistics with theoretical 95% confidence bands based on 500 simulations with one covariate $X_1 \sim \mathcal{N}(0,1)$, $M = 200$, $J = 3$, under correct and incorrect covariance structure specification.

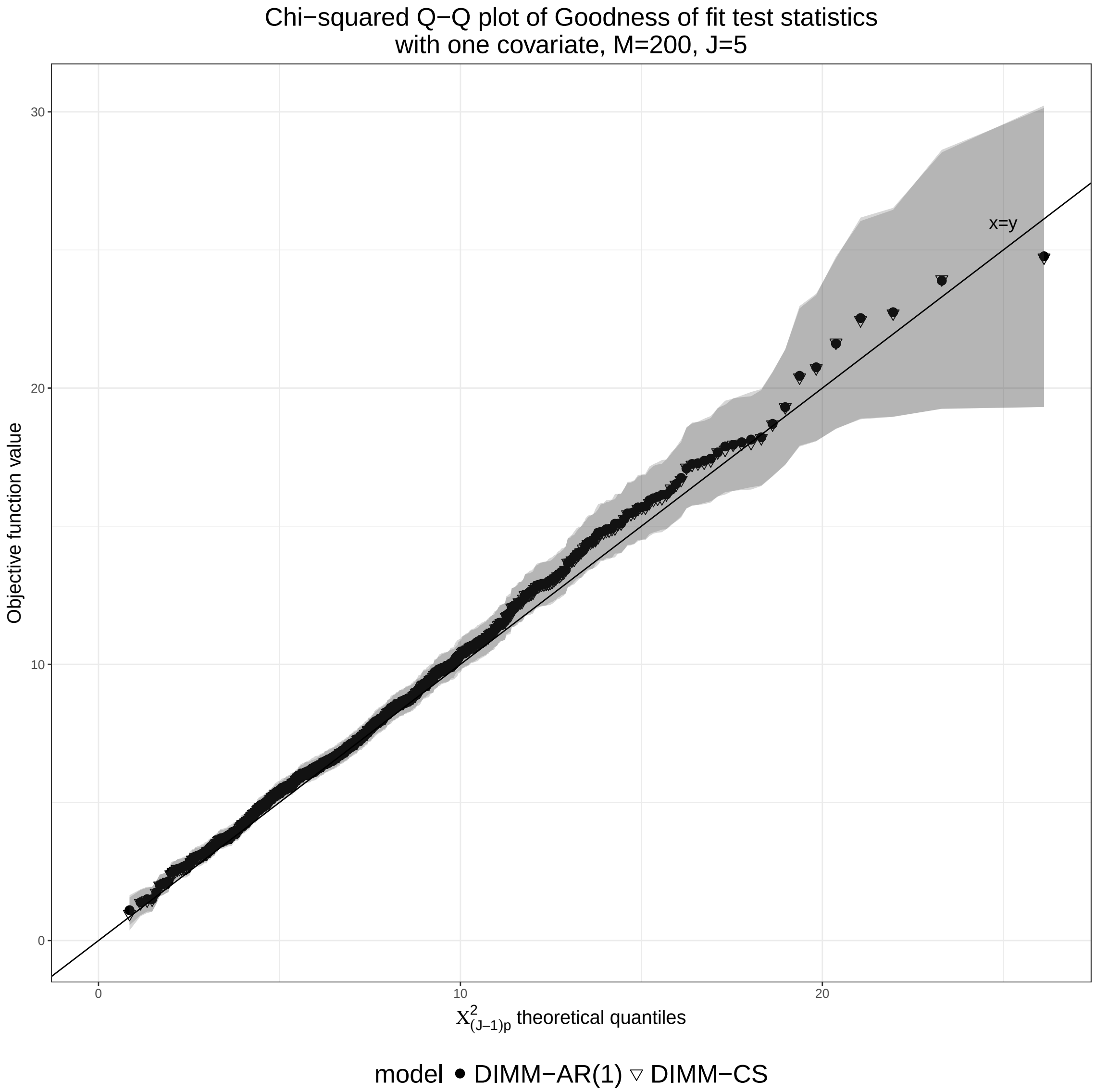


Figure C.3: Chi-squared Q-Q plot of goodness-of-fit test statistics with theoretical 95% confidence bands based on 500 simulations with one covariate $X_1 \sim \mathcal{N}(0,1)$, $M = 200$, $J = 5$, under correct and incorrect covariance structure specification.

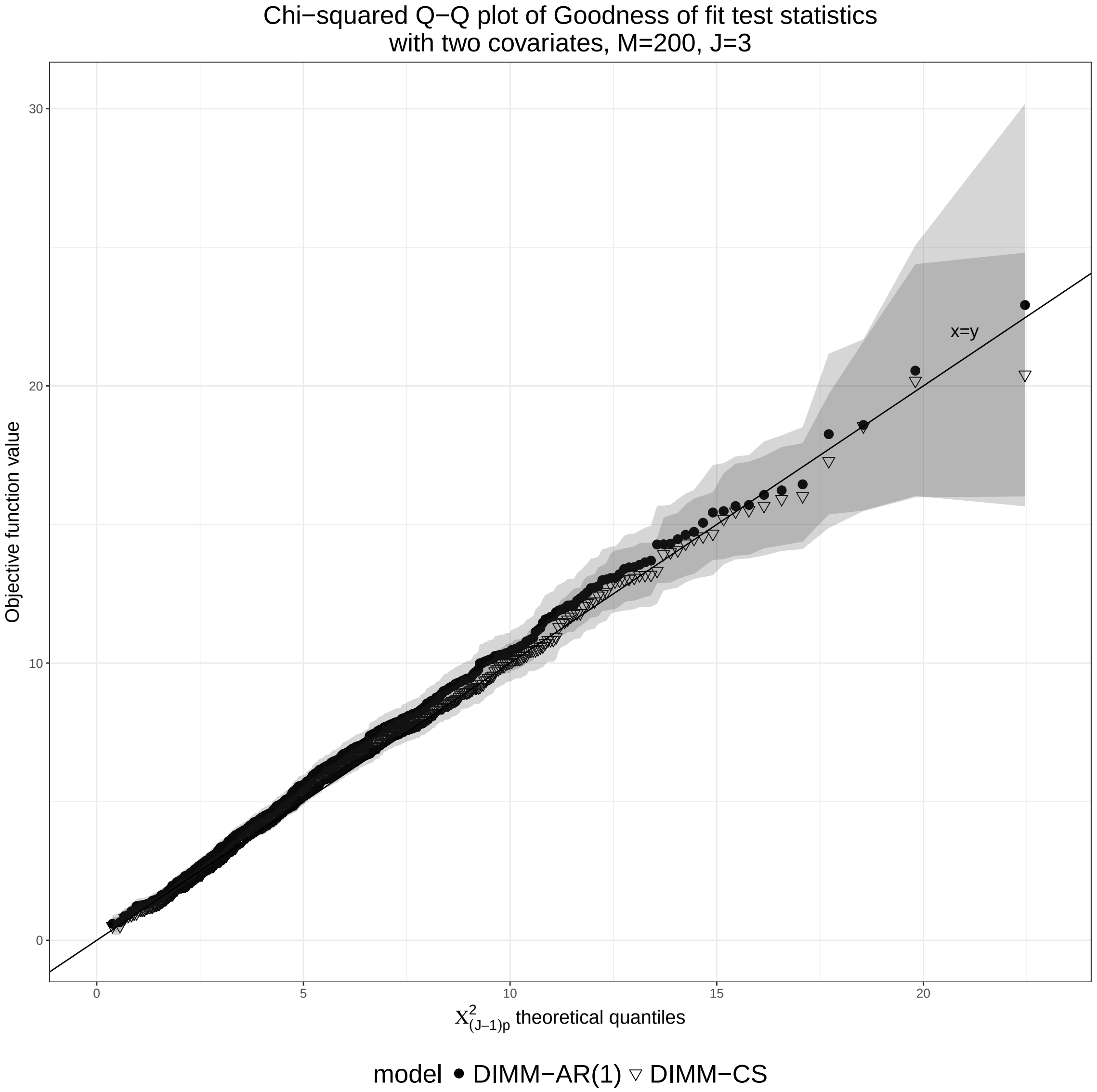


Figure C.4: Chi-squared Q-Q plot of goodness-of-fit test statistics with theoretical 95% confidence bands based on 500 simulations with two covariates $X_1 \sim Normal_M(0, S)$, where S is a positive-definite $M \times M$ matrix, and X_2 a vector of alternating 0's and 1's to imitate an exposure, $M = 200$, $J = 3$, under correct and incorrect covariance structure specification.

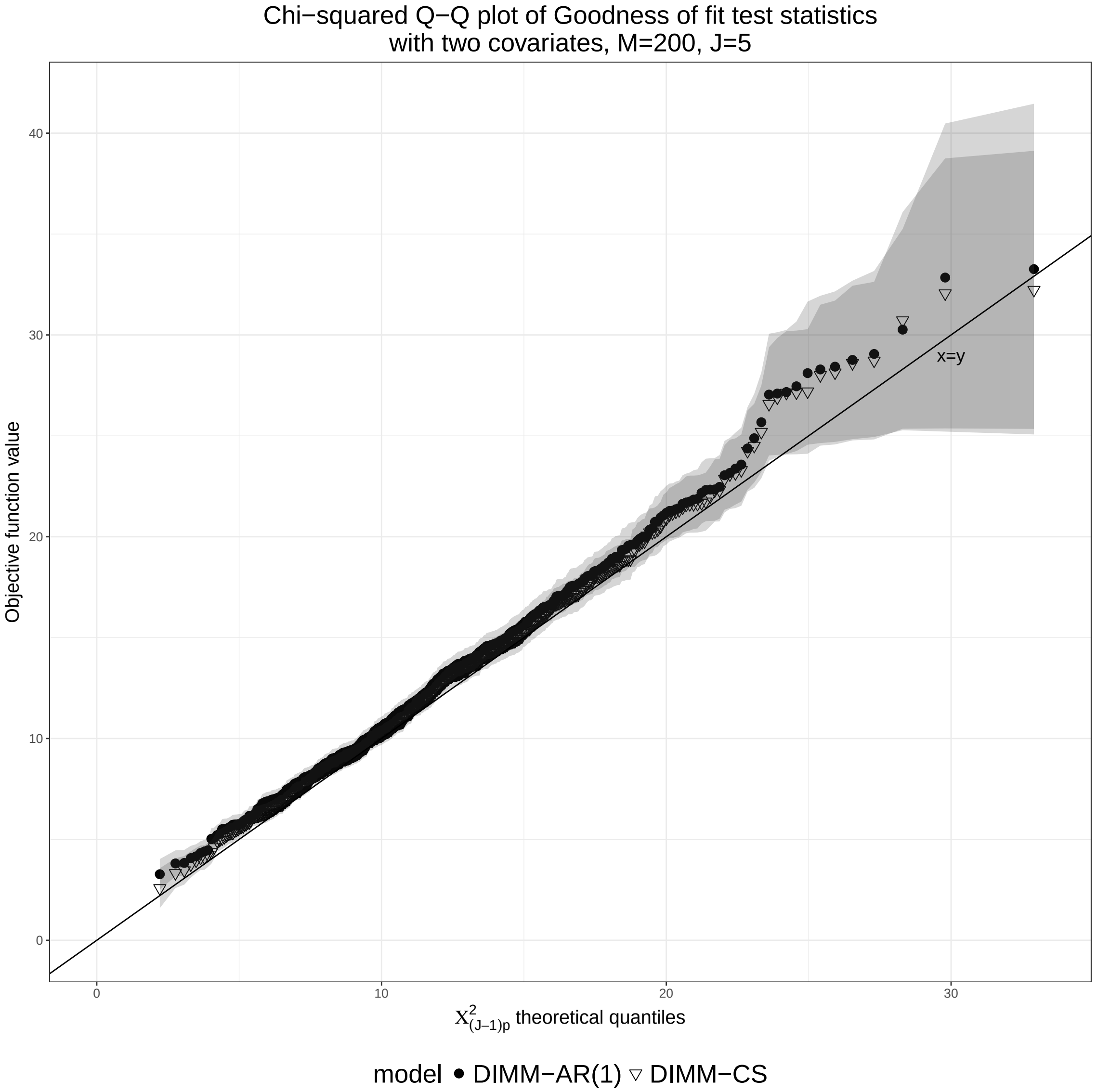


Figure C.5: Chi-squared Q-Q plot of goodness-of-fit test statistics with theoretical 95% confidence bands based on 500 simulations with two covariates $X_1 \sim Normal_M(0, S)$, where S is a positive-definite $M \times M$ matrix, and X_2 a vector of alternating 0's and 1's to imitate an exposure, $M = 200$, $J = 5$, under correct and incorrect covariance structure specification.

APPENDIX D

Chapter III: Additional Data Analysis Results

Additional data analysis simulations and results

We present additional average amplitude density maps in Figures D.1, D.2 and D.3, additional correlation heatmaps in Figures D.4, D.5, D.6 and D.7, data simulation results in Table D.1 to ensure sufficient power in the analysis presented in Section 6, block specific MCLE's from the data analysis in Table D.2, and full data analysis results in Table D.3.

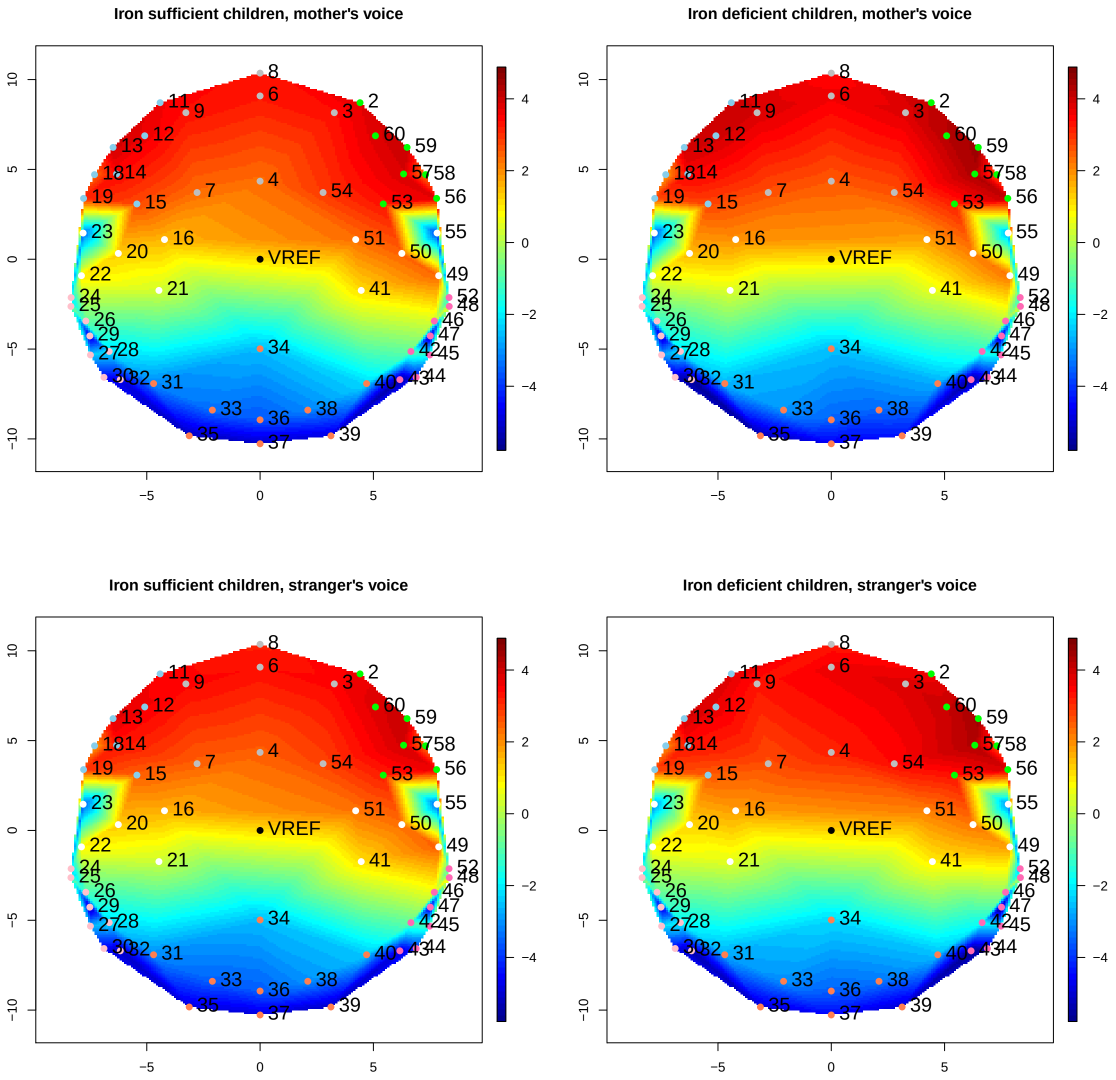
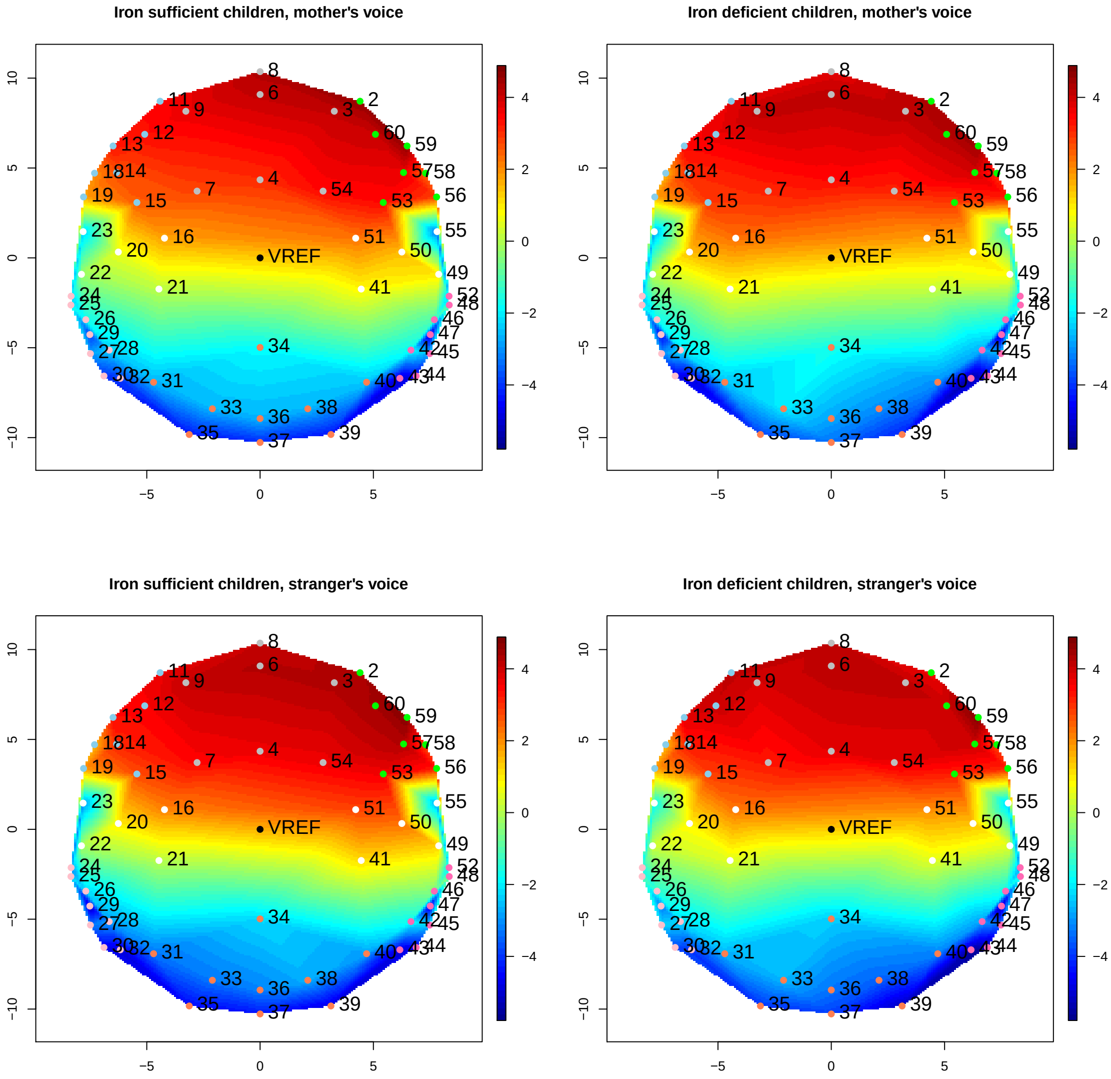


Figure D.1: Average P2 amplitude for iron sufficient and deficient children (left and right panels respectively) under stimulus of mother and stranger's voice (top and bottom panels respectively).



Iron sufficient children, mother's voice

Iron deficient children, mother's voice

Iron sufficient children, stranger's voice

Iron deficient children, stranger's voice

Figure D.2: Average P750 amplitude for iron sufficient and deficient children (left and right panels respectively) under stimulus of mother and stranger's voice (top and bottom panels respectively).

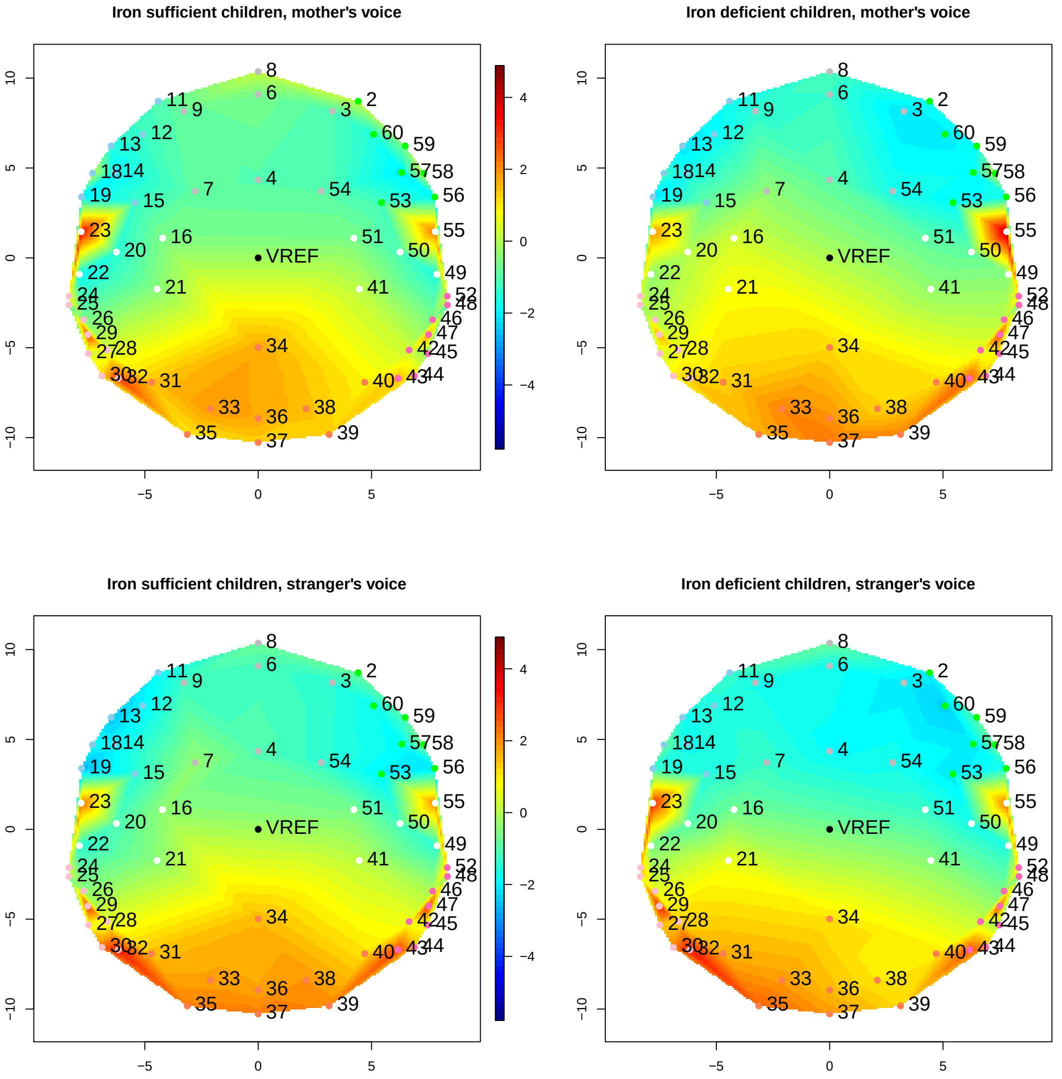


Figure D.3: Average LSW amplitude for iron sufficient and deficient children (left and right panels respectively) under stimulus of mother and stranger's voice (top and bottom panels respectively).

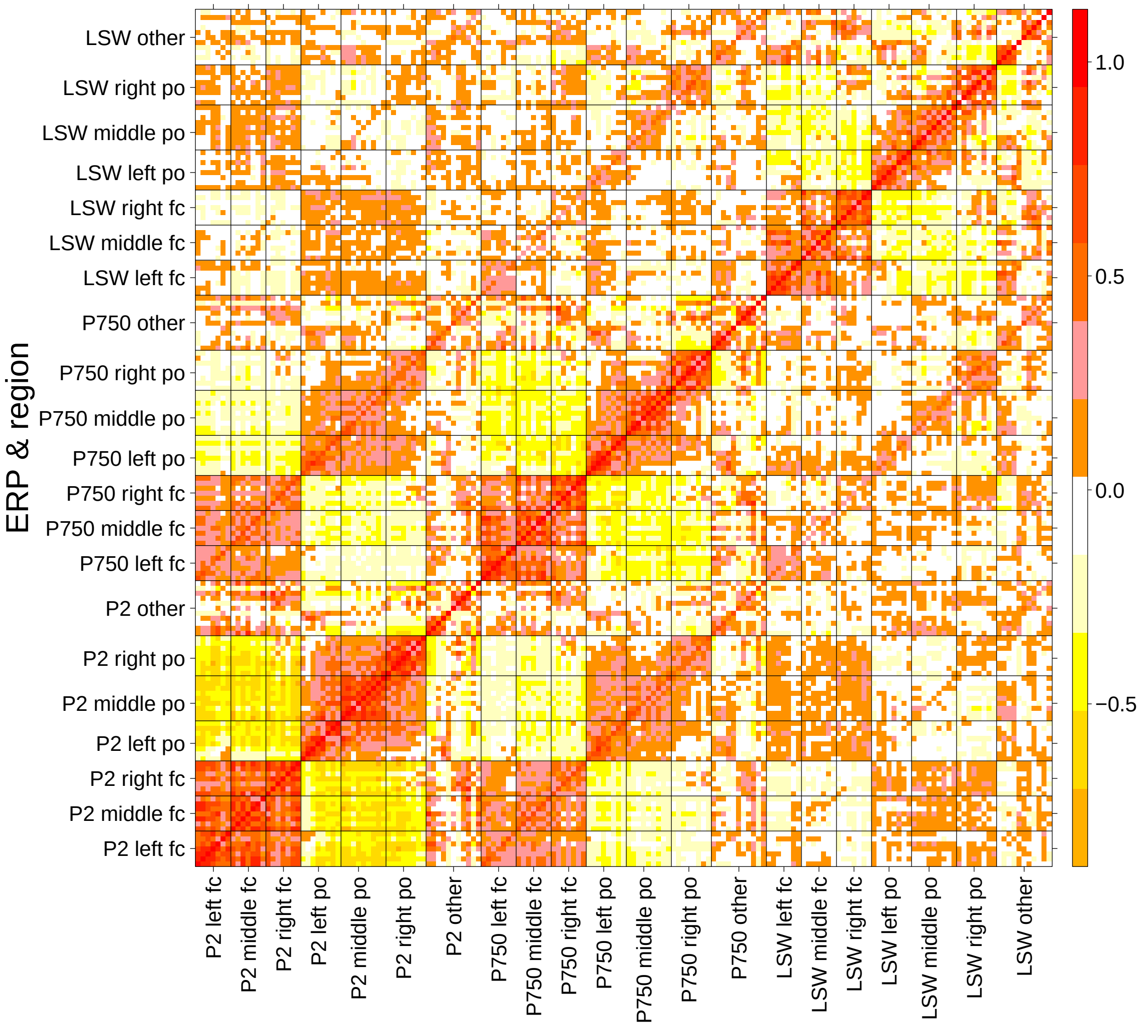


Figure D.4: Correlation of electrical amplitude at three ERP's for iron sufficient children under stimulus of mother's voice.

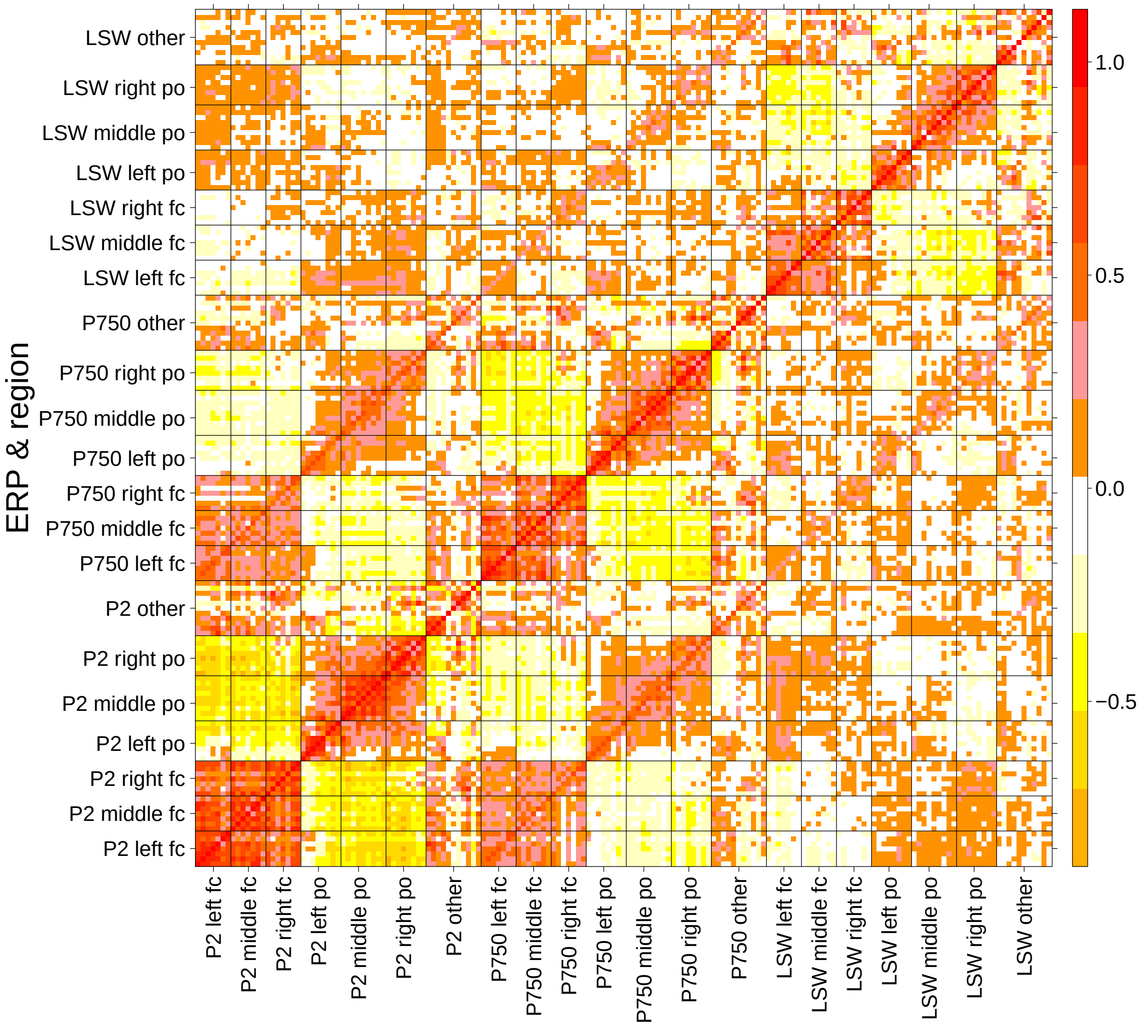


Figure D.5: Correlation of electrical amplitude at three ERP's for iron sufficient children under stimulus of stranger's voice.

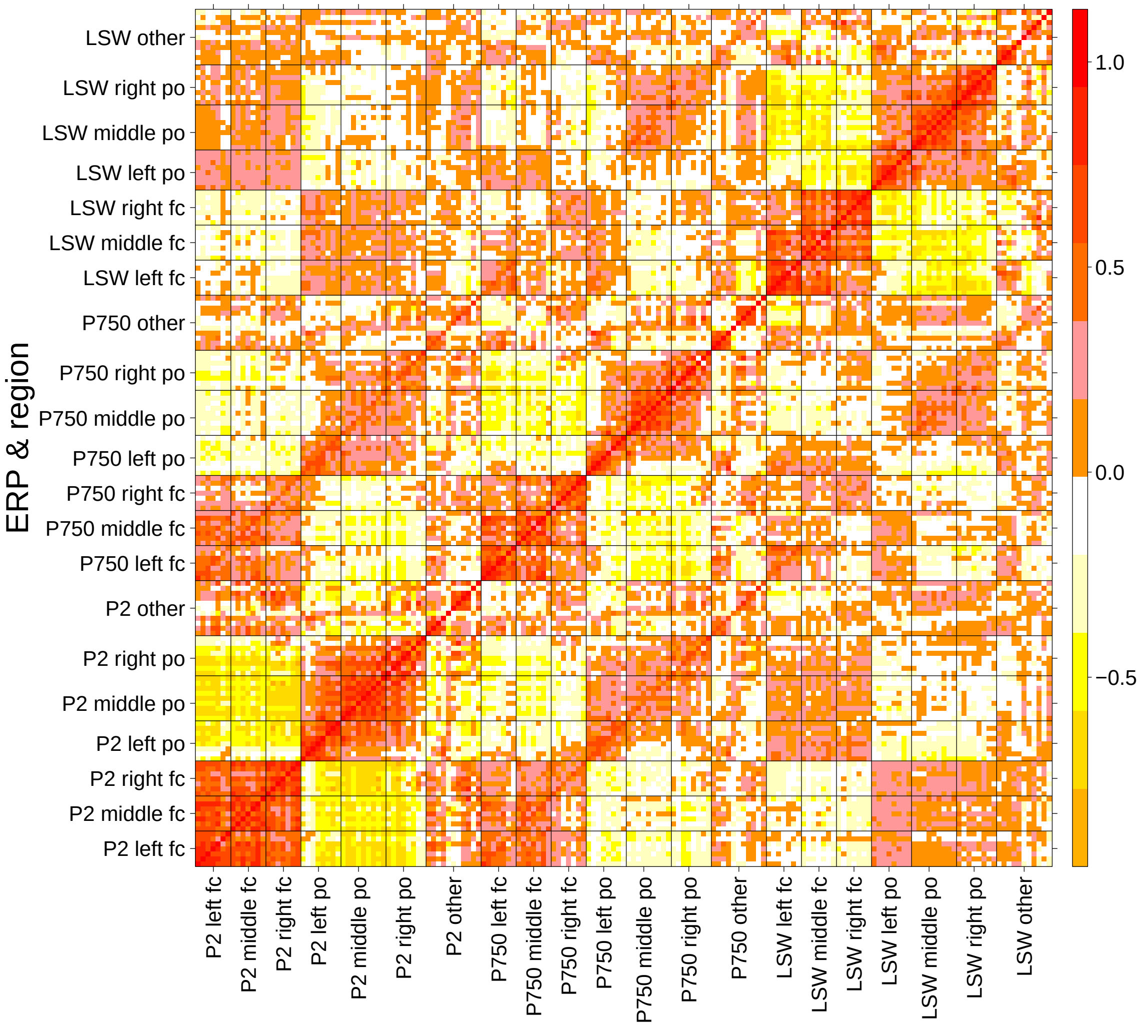


Figure D.6: Correlation of electrical amplitude at three ERP's for iron deficient children under stimulus of mother's voice.

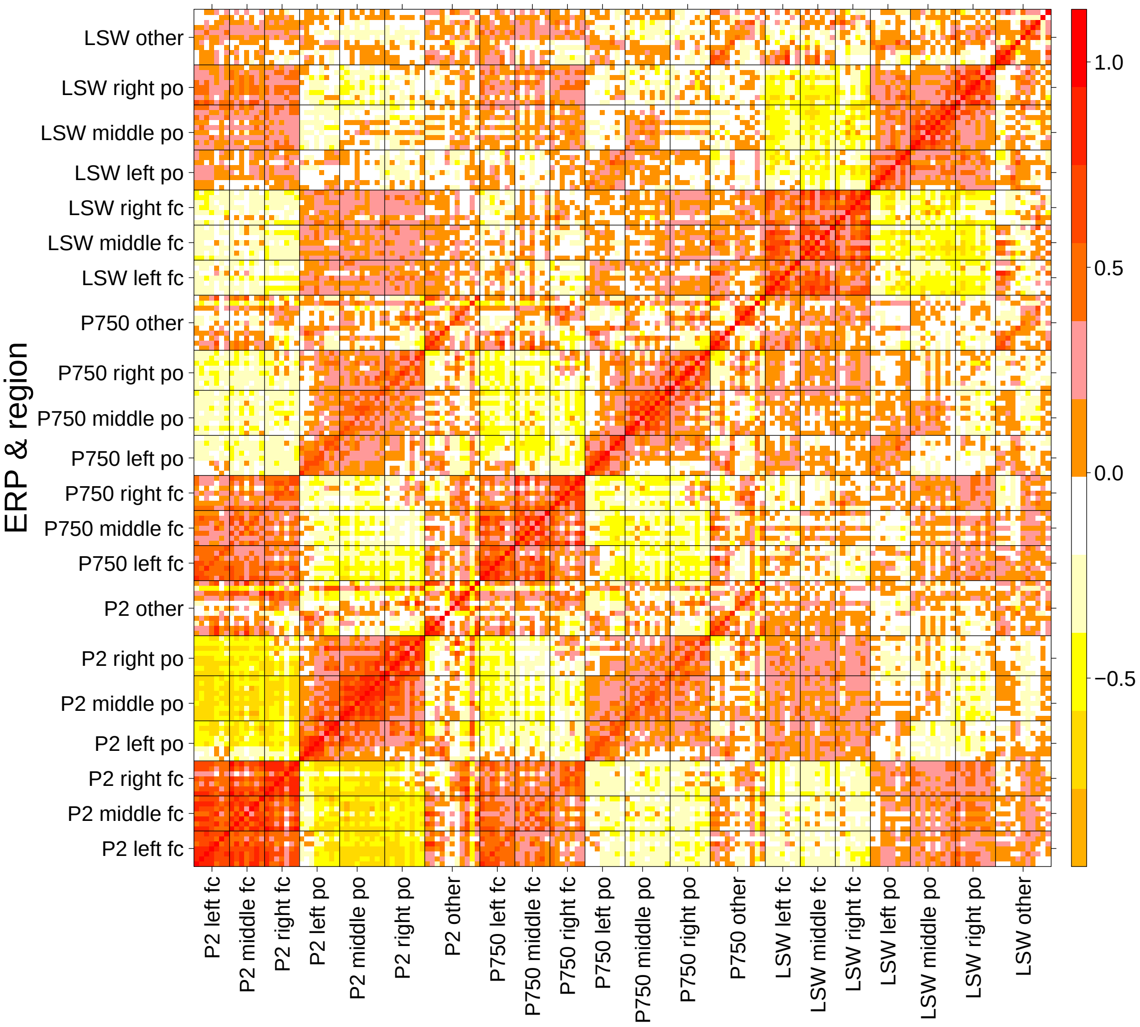


Figure D.7: Correlation of electrical amplitude at three ERP's for iron deficient children under stimulus of stranger's voice.

Table D.1: Iron sufficiency status effect mean squared error (MSE×10^{-2}) and mean variance (mean var×10^{-2}), 95% confidence interval (CI) coverage, type-1 error, and mean CPU time in seconds for each combination scheme based on 500 simulations.

combine region, ERP	method	MSE×10^{-2} (mean var×10^{-2})	95% CI coverage	type 1 error	mean CPU time
left, middle and right fc, P2	GEE-CS	1.4 (1.4)	0.942	0.058	0.35
	ME	1.4 (1.4)	0.958	0.042	3.59
	DIMM	1.4 (1.3)	0.934	0.066	0.22
left, middle and right fc, P750	GEE-CS	0.9 (0.9)	0.95	0.05	0.35
	ME	0.9 (0.9)	0.954	0.046	3.74
	DIMM	0.9 (0.8)	0.946	0.054	0.21
left, middle and right fc, LSW	GEE-CS	0.8 (0.8)	0.946	0.054	0.35
	ME	0.8 (0.8)	0.948	0.052	3.81
	DIMM	0.8 (0.7)	0.938	0.062	0.17
left po, P2 & P750	GEE-CS	0.7 (0.7)	0.952	0.048	0.19
	ME	0.7 (0.7)	0.956	0.044	1.65
	DIMM	0.7 (0.7)	0.942	0.058	0.24
middle and right po, P2	GEE-CS	1.1 (1.1)	0.946	0.054	0.22
	ME	1.1 (1.1)	0.952	0.048	3.37
	DIMM	1.1 (1.0)	0.94	0.06	0.25
middle and right po, P750	GEE-CS	0.8 (0.7)	0.94	0.06	0.22
	ME	0.7 (0.7)	0.948	0.052	3.17
	DIMM	0.8 (0.7)	0.948	0.052	0.25
left, middle and right po, LSW	GEE-CS	0.6 (0.6)	0.94	0.06	0.52
	ME	0.6 (0.6)	0.944	0.056	4.26
	DIMM	0.5 (0.5)	0.932	0.068	0.21

fc, frontal-central; po, parietal-occipital. Real covariate values were used to simulate response data. Response data was simulated with mean parameter values set to values estimated in Table 3.3, and covariance set to the sample covariance of the observed response.

Table D.2: Block specific MCLE's of $\boldsymbol{\beta}$.

Block: functional region and ERP	intercept	age	voice stimulus	sufficiency status
left frontal-central, P2	0.04	−0.04	0.04	1.14
middle frontal-central, P2	−0.1	−0.04	−0.01	0.13
right frontal-central, P2	−0.08	−0.01	0.06	0.03
left parietal-occipital, P2	0.16	0.02	−0.02	−0.17
middle parietal-occipital, P2	−0.09	0.01	−0.01	0.00
right parietal-occipital, P2	0.13	0.03	−0.00	−0.07
left frontal-central, P750	0.11	0.01	−0.02	−0.04
middle frontal-central, P750	0.14	0.02	−0.07	−0.02
right frontal-central, P750	−0.18	0.03	−0.03	0.02
left parietal-occipital, P750	−0.02	0.02	−0.03	−0.18
middle parietal-occipital, P750	−0.11	−0.05	0.11	0.09
right parietal-occipital, P750	−0.03	−0.02	0.05	0.11
left frontal-central, LSW	−0.04	0.02	0.12	−0.08
middle frontal-central, LSW	0.06	0.04	0.09	−0.14
right frontal-central, LSW	−0.14	0.02	0.13	0.02
left parietal-occipital, LSW	−0.04	0.00	−0.13	0.05
middle parietal-occipital, LSW	0.12	−0.04	−0.05	0.03
right parietal-occipital, LSW	0.03	−0.03	−0.13	0.05

Table D.3: Iron sufficiency status effect estimates and statistics for each combination scheme.

combine region, ERP	method	estimate (s.d.×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
	ME	0.103 (11.8)	0.38	1.97	1.49
	DIMM	0.087(11.9)	0.47	1.321	1
left, middle and right fc, P750	GEE-CS	−0.013 (9.6)	0.90	0.37	1.16
	ME	−0.013 (9.7)	0.90	1.99	6.24
	DIMM	−0.038 (9.0)	0.67	0.319	1
left, middle and right fc, LSW	GEE-CS	−0.064 (10.4)	0.54	0.37	1.17
	ME	−0.064 (9.0)	0.48	1.78	5.62
	DIMM	−0.073 (9.8)	0.46	0.317	1
left po, P2 & P750	GEE-CS	−0.174 (8.3)	0.04	0.22	0.43
	ME	−0.174 (8.3)	0.04	1.47	2.86
	DIMM	−0.226 (8.1)	0.005	0.514	1
middle and right po, P2	GEE-CS	−0.032 (10.6)	0.76	0.40	0.86
	ME	−0.034 (10.5)	0.75	2.81	6.02
	DIMM	0.009 (10.4)	0.93	0.467	1
middle and right po, P750	GEE-CS	0.096 (8.7)	0.27	0.24	0.51
	ME	0.096 (8.6)	0.26	1.46	3.12
	DIMM	0.106 (8.7)	0.22	0.468	1
left, middle and right po, LSW	GEE-CS	0.041 (8.7)	0.64	0.55	1.41
	ME	0.041 (7.4)	0.58	3.53	9.07
	DIMM	0.087(8.4)	0.30	0.389	1

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

APPENDIX E

Chapter IV: Technical Details

Summary of sensitivity matrix formulas

Sensitivity matrices are summarized in Table E.1.

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

Table E.1: Summary of sensitivity formulas. Formulas that are not used are marked "—". *"w.r.t." shorthand for "with respect to".

Subsetting operation on variability matrices

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

and columns $1 + D^{jk}$ to $d_{jk} + D^{jk}$, where d_{jk} is the dimension of $\boldsymbol{\zeta}_{jk}$ and D^{jk} is defined in Section 4.5.1.

Cumulative sum of dimensions of $\boldsymbol{\zeta}$

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

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

Let $k \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, J\}$. Recall the definitions of $\widehat{\boldsymbol{A}}^{\boldsymbol{\theta}}_{k,ij}$, $\widehat{\boldsymbol{A}}^{\boldsymbol{\zeta}}_{k,ij}$, $\widehat{\boldsymbol{B}}^{\boldsymbol{\theta}}_{k,ij}$ and $\widehat{\boldsymbol{B}}^{\boldsymbol{\zeta}}_{k,ij}$ in Section 4.5.1. Define

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

APPENDIX F

Chapter IV: Proofs

Proof [Proof of Lemma IV.1] Under conditions (A.2) (i) and (A.5),

$$\begin{aligned}\left\|\boldsymbol{\tau}_i(\widehat{\boldsymbol{\theta}}_{list}, \widehat{\boldsymbol{\zeta}}_{list}) - \boldsymbol{\tau}_i(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right\| &\leq \sum_{k=1}^{K}\sum_{j=1}^{J}(c_{jk} + b_{jk})\left\|\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix}\right\| \\ &= \sum_{k=1}^{K}\sum_{j=1}^{J}(c_{jk} + b_{jk})O_p(n_k^{-1/2}) = O_p(n_{\min}^{-1/2}).\end{aligned}$$

Plugging this into the formula for $\widehat{\boldsymbol{V}}_N$ yields

$$\begin{aligned}\widehat{\boldsymbol{V}}_N &= \frac{1}{N}\sum_{i=1}^{N}\{\boldsymbol{\tau}_i(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\}^{\otimes 2} + O_p(n_{\min}^{-1/2})\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\tau}_i(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p(n_{\min}^{-1}) \\ &= \boldsymbol{v}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p(N^{-1/2}).\end{aligned}$$

Recall that we require that $\boldsymbol{S}_{jk}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})$ is a $n_k^{1/2}$-consistent sample estimate of $\boldsymbol{s}_{jk}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})$ in Section 4.5. Then $\widehat{\boldsymbol{S}}_{jk} = \boldsymbol{s}_{jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk}) + O_p(n_k^{-1/2})$. Then by (A.2), (A.5), and a Taylor expansion:

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

It follows from the above that

$$\begin{aligned}\widehat{\boldsymbol{S}} &= \begin{pmatrix} \mathbb{S}^{JK}\left(\frac{n_k}{N}\widehat{\boldsymbol{S}}^{\boldsymbol{\theta}}_{\boldsymbol{\psi}_{jk}}\right) & \mathrm{diag}\left\{\frac{n_k}{N}\widehat{\boldsymbol{S}}^{\boldsymbol{\zeta}}_{\boldsymbol{\psi}_{jk}}\right\}_{j=1,k=1}^{J,K} \\ \mathbb{S}^{JK}\left(\frac{n_k}{N}\widehat{\boldsymbol{S}}^{\boldsymbol{\theta}}_{g_{jk}}\right) & \mathrm{diag}\left\{\frac{n_k}{N}\widehat{\boldsymbol{S}}^{\boldsymbol{\zeta}}_{g_{jk}}\right\}_{j=1,k=1}^{J,K} \end{pmatrix} \\ &= \boldsymbol{s}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p\left(n_{\max}^{1/2}N^{-1}\right) = \boldsymbol{s}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p(N^{-1/2}).\end{aligned}$$

Then $\widehat{\boldsymbol{S}}^T\widehat{\boldsymbol{V}}_N^{-1}\widehat{\boldsymbol{S}} = \boldsymbol{j}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p(N^{-1/2})$. Lastly, it can easily be shown that

$$\widehat{\boldsymbol{S}}^T\widehat{\boldsymbol{V}}_N^{-1}\widehat{\boldsymbol{S}} = \widehat{\boldsymbol{S}}^T\begin{pmatrix} \widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi}} & \widehat{\boldsymbol{V}}_N^{\boldsymbol{\psi g}} \\ \widehat{\boldsymbol{V}}^{\boldsymbol{\psi g}\ T} & \widehat{\boldsymbol{V}}_N^{\boldsymbol{g}} \end{pmatrix}\widehat{\boldsymbol{S}} = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{i=1}^{J} n_k^2\widehat{\boldsymbol{C}}_{k,i}. \quad \square$$

Proof [Proof of Theorem IV.4] By consistency of block estimates, we have $\widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \overset{p}{\to} 0$ and $\widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \overset{p}{\to} 0$ as $n_k \to \infty$. By (A.2),

$$\left\|\boldsymbol{\tau}_i(\widehat{\boldsymbol{\theta}}_{list}, \widehat{\boldsymbol{\zeta}}_{list}) - \boldsymbol{\tau}_i(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right\| \leq \sum_{k=1}^{K}\sum_{j=1}^{J}(c_{jk} + b_{jk})\left\|\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix}\right\|.$$

Then $\left\|\boldsymbol{\tau}_i(\widehat{\boldsymbol{\theta}}_{list}, \widehat{\boldsymbol{\zeta}}_{list})\right\| = \|\boldsymbol{\tau}_i(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\| + o_p(1)$. Plugging this into the formula for $\widehat{\boldsymbol{V}}_N$,

$$\begin{aligned}\widehat{\boldsymbol{V}}_N &= \frac{1}{N}\sum_{i=1}^{N}\left\{\boldsymbol{\tau}_i(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + o_p(1)\right\}^{\otimes 2} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{\boldsymbol{\tau}_i(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right\}^{\otimes 2} + o_p(1)\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\tau}_i(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + o_p(1) \\ &= \boldsymbol{v}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) + O_p(N^{-1/2}) + o_p(1). \quad \square\end{aligned}$$

Proof of Theorem IV.5:

The following lemmas complete the proof of Theorem IV.5 given in the chapter, under the assumed conditions.

Lemma F.0.0.1. Define $\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})$ as in (4.11) in the proof of Theorem IV.5. Then $\lambda(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0) \xrightarrow{p} 0$ as $n_{\min} \to \infty$.

Proof Using Lemma IV.1,

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

Lemma F.0.0.2. The following relationship holds:

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

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

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

By first-order Taylor expansion,

$$\begin{aligned}
E_{\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}} \left\{ \boldsymbol{\tau}_{i,jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk}) \right\} &= E_{\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}} \left\{ \boldsymbol{\tau}_{i,jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) \right\} + \\
&\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}} \left\{ \boldsymbol{\tau}_{i,jk}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk}) \right\} |_{\boldsymbol{\theta}^*, \boldsymbol{\zeta}_{jk}^*} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix}, \qquad \text{(F.1)}
\end{aligned}$$

where $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}_{jk}^*)$ lies between $(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0})$ and $(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk})$. By condition (A.5),

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

In other words, the norm of the difference between $\boldsymbol{T}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0})$ and $\boldsymbol{T}_{jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk}) - E_{\boldsymbol{\theta},\boldsymbol{\zeta}_{jk}}\{\boldsymbol{\tau}_{i,jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk})\}$ goes to 0 at a rate faster than $(Nn_k)^{-1/2}$. Adding (F.1) and (F.2), we have

$$\begin{aligned} -\boldsymbol{T}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) &= \boldsymbol{T}_{jk}(\widehat{\boldsymbol{\theta}}_{jk}, \widehat{\boldsymbol{\zeta}}_{jk}) - \boldsymbol{T}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) - E_{\boldsymbol{\theta},\boldsymbol{\zeta}_{jk}}\boldsymbol{\tau}_{i,jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) \\ &= \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta},\boldsymbol{\zeta}_{jk}}\boldsymbol{\tau}_{i,jk}(\boldsymbol{\theta}, \boldsymbol{\zeta}_{jk})|_{\boldsymbol{\theta}^*,\boldsymbol{\zeta}^*_{jk}} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix} + O_p(n_k^{-1/2}N^{-1/2}) \\ &= -\boldsymbol{s}_{jk}(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*_{jk}) \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix} + O_p(n_k^{-1/2}N^{-1/2}). \end{aligned}$$

Rearranging yields

$$\boldsymbol{T}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) = \boldsymbol{s}_{jk}(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*_{jk}) \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{jk} - \boldsymbol{\zeta}_{jk0} \end{pmatrix} + O_p(n_k^{-1/2}N^{-1/2}). \tag{F.3}$$

Finally, note that $\widehat{\boldsymbol{S}}_{jk} = \boldsymbol{s}_{jk}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_{jk0}) + O_p(n_k^{-1/2}) = \boldsymbol{s}_{jk}(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*_{jk}) + O_p(n_k^{-1/2})$. Then plugging this into (F.3), we have:

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

Proof [Proof of Corollary IV.2:] From Theorems IV.2 and IV.5, we can write

$$\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{opt} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{opt} - \boldsymbol{\zeta}_0 \end{pmatrix} = \frac{1}{N^{1/2}}\boldsymbol{Z} + \frac{1}{N^{1/2}}\boldsymbol{c}_{N1} \tag{F.4}$$

$$\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{DDIMM} - \boldsymbol{\theta}_0 \\ \widehat{\boldsymbol{\zeta}}_{DDIMM} - \boldsymbol{\zeta}_0 \end{pmatrix} = \frac{1}{N^{1/2}}\boldsymbol{Z} + \frac{1}{N^{1/2}}\boldsymbol{c}_{N2}, \tag{F.5}$$

where $\boldsymbol{c}_{N1}, \boldsymbol{c}_{N2} \overset{p}{\to} 0$ as $n_{\min} \to \infty$, and $\boldsymbol{Z} \sim \mathcal{N}\left(0, \boldsymbol{j}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\right)$. Subtracting (F.5) from (F.4), we get

$$\begin{pmatrix} \widehat{\boldsymbol{\theta}}_{opt} - \widehat{\boldsymbol{\theta}}_{DDIMM} \\ \widehat{\boldsymbol{\zeta}}_{opt} - \widehat{\boldsymbol{\zeta}}_{DDIMM} \end{pmatrix} = \frac{1}{N^{1/2}} (\boldsymbol{c}_{N2} - \boldsymbol{c}_{N1}).$$

The result follows. □

Proof of Theorem IV.6:

The following lemmas complete the proof of Theorem IV.6 given in the chapter, under the assumed conditions.

Lemma F.0.0.3. Define $\lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})$ as in (4.11) in the proof of Theorem IV.5. Then $\|\lambda(\boldsymbol{\theta}_0, \boldsymbol{\zeta}_0)\| = O_p(N^{-1/2-\delta} n_{\max}^{1/2})$ and $\left\| \{\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}} \lambda(\boldsymbol{\theta}, \boldsymbol{\zeta})\}^{-1} \right\| = O_p\left(N^{1/2+\delta} n_{\max}^{-1}\right)$.

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

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

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

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

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

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

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

Lemma F.0.0.4. For some matrices $\boldsymbol{E}_k$, $k=1,\ldots,K$, of $\boldsymbol{0}$'s and $\boldsymbol{1}$'s, the following asymptotic properties hold:

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

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

Proof Recall that $\widehat{\boldsymbol{C}}_{k,i}(\widehat{\boldsymbol{\theta}}^T_{ik}-\boldsymbol{\theta}^T_0,\widehat{\boldsymbol{\zeta}}^T_{list}-\boldsymbol{\zeta}^T_0)^T=\widehat{\boldsymbol{C}}^*_{k,i}(\widehat{\boldsymbol{\theta}}^T_{ik}-\boldsymbol{\theta}^T_0,\widehat{\boldsymbol{\zeta}}^T_{ik}-\boldsymbol{\zeta}^T_{ik0})^T$. Let $\left[\boldsymbol{v}_k^{-1}(\boldsymbol{\theta},\boldsymbol{\zeta}_k)\right]_{ij}$ subset the rows for the parameters corresponding to block (i,k) and the columns for the parameters corresponding to block (j,k) of matrix $\boldsymbol{v}_k^{-1}(\boldsymbol{\theta},\boldsymbol{\zeta}_k)$. Define $\boldsymbol{j}_{jik}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk},\boldsymbol{\zeta}_{ik})=\boldsymbol{s}_{jk}(\boldsymbol{\theta},\boldsymbol{\zeta}_{jk})\left[\boldsymbol{v}_k^{-1}(\boldsymbol{\theta},\boldsymbol{\zeta}_k)\right]_{ji}\boldsymbol{s}_{ik}(\boldsymbol{\theta},\boldsymbol{\zeta}_{ik})$, and $\left[\boldsymbol{j}_k^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0})\right]_i$ the submatrix of $\boldsymbol{j}_k^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0})$ corresponding to parameters in block (i,k), such that

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

Then using the results in the proof of Lemma F.0.0.3, let $\boldsymbol{E}_k$ and $\boldsymbol{E}_{k,i}$ matrices of $\mathbf{0}$'s and $\mathbf{1}$'s such that

$$
\begin{aligned}
&\frac{n_k^2}{N^2}\sum_{i=1}^{J}\widehat{\boldsymbol{C}}_{k,i} = \frac{n_k}{N}\boldsymbol{E}_k\left\{\boldsymbol{j}_k(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0}) + O_p\left(n_k^{-1/2}\right)\right\}\boldsymbol{E}_k^T\\
&= \frac{n_k}{N}\boldsymbol{E}_k\boldsymbol{j}_k(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{k0})\boldsymbol{E}_k^T + O_p\left(n_k^{1/2}N^{-1}\right), \text{ and}\\
&\frac{n_k^2}{N^2}\sum_{i=1}^{J}\widehat{\boldsymbol{C}}_{k,i}\begin{pmatrix}\widehat{\boldsymbol{\theta}}_{ik}-\boldsymbol{\theta}_0\\ \widehat{\boldsymbol{\zeta}}_{list}-\boldsymbol{\zeta}_0\end{pmatrix}\\
&= \frac{n_k}{N}\boldsymbol{E}_k\sum_{i=1}^{J}\boldsymbol{E}_{k,i}\left\{\sum_{j=1}^{J}\boldsymbol{j}_{jik}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{jk0},\boldsymbol{\zeta}_{ik0}) + O_p\left(n_k^{-1/2}\right)\right\}\begin{pmatrix}\widehat{\boldsymbol{\theta}}_{ik}-\boldsymbol{\theta}_0\\ \widehat{\boldsymbol{\zeta}}_{ik}-\boldsymbol{\zeta}_{ik0}\end{pmatrix}\\
&= \frac{n_k}{N}\boldsymbol{E}_k\sum_{i=1}^{J}\boldsymbol{E}_{k,i}\sum_{j=1}^{J}\boldsymbol{j}_{jik}(\boldsymbol{\theta}_0,\boldsymbol{\zeta}_{jk0},\boldsymbol{\zeta}_{ik0})\begin{pmatrix}\widehat{\boldsymbol{\theta}}_{ik}-\boldsymbol{\theta}_0\\ \widehat{\boldsymbol{\zeta}}_{ik}-\boldsymbol{\zeta}_{ik0}\end{pmatrix} + O_p\left(N^{-1}\right).
\end{aligned}
$$

To obtain the desired result, define

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

Lemma F.0.0.5. $N^{1/2}\boldsymbol{H}\left(\widehat{\boldsymbol{\theta}}_{DDIMM}^T-\boldsymbol{\theta}_0^T,\widehat{\boldsymbol{\zeta}}_{DDIMM}^T-\boldsymbol{\zeta}_0^T\right)$ can be rewritten as

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

Proof

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

APPENDIX G

Chapter IV: Additional Simulation Results

Additional simulation results

Simulation metrics for the pairwise composite likelihood (CL) can be found in Figure G.1.

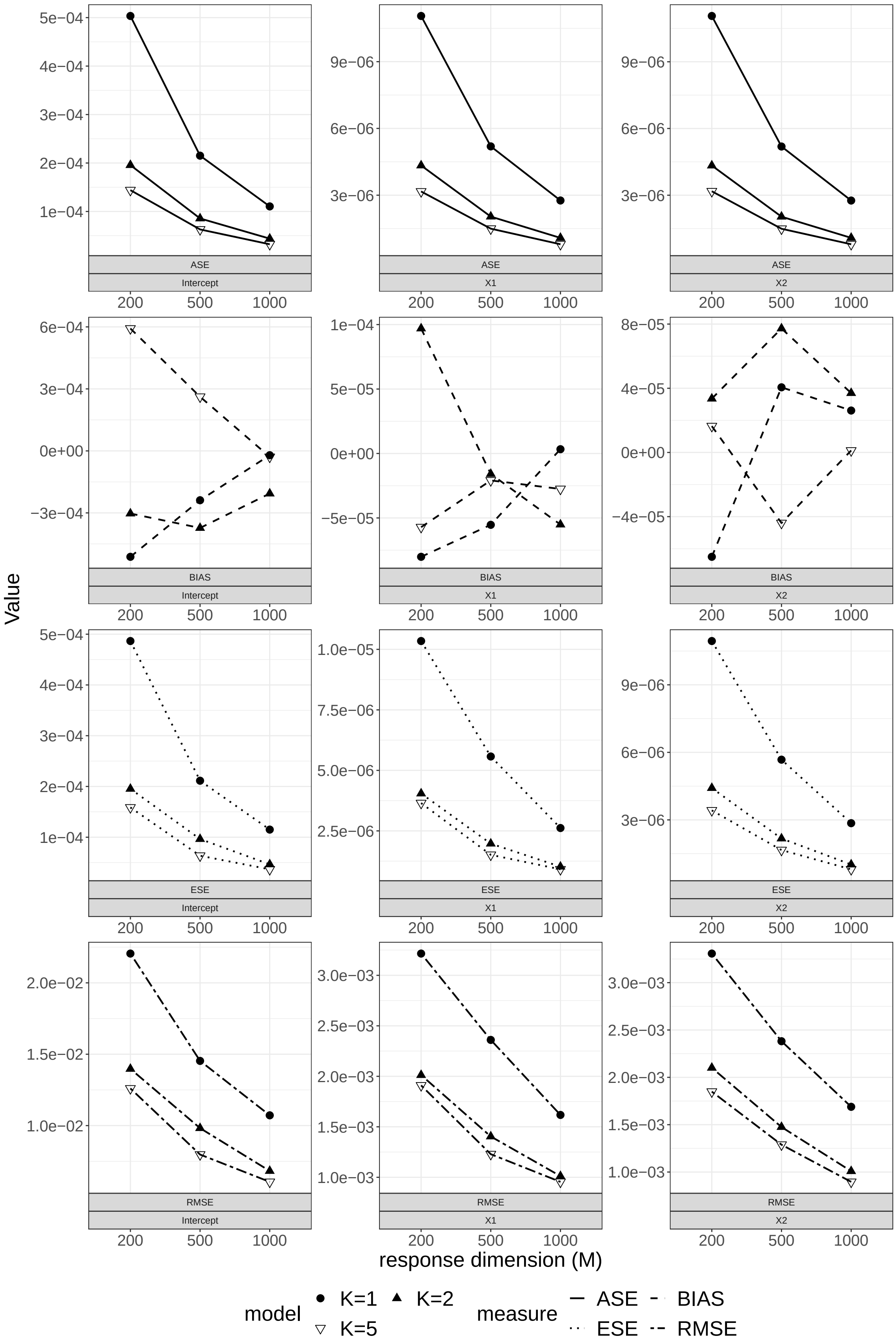


Figure G.1: Plot of simulation metrics for CL, averaged over 1,000 simulations.

APPENDIX H

Chapter V: Proofs

Conditions for proofs of Chapter V

Condition H.1. Conditions for consistency and asymptotic normality of $\widehat{\boldsymbol{\theta}}_{jk}$ for data source (j,k) in gth partition set $\mathcal{P}^g$: $\boldsymbol{C}_{jk}^{-1}$ is positive semi-definite and $\boldsymbol{C}_{jk}^{-1}E\{\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{jk})\} = 0$ if and only if $\boldsymbol{\theta}_{jk} = \boldsymbol{\theta}_{g0}$; $\boldsymbol{\theta}_0 = \mathbb{S}^G(\boldsymbol{\theta}_{g0})$ is an interior point of $\boldsymbol{\theta} = \times_{g=1}^{G}\boldsymbol{\theta}_g$, and $\boldsymbol{\theta}_g$ are compact; $\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{jk})$ is continuous at each $\boldsymbol{\theta}_{jk}$ with probability one; $\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{jk})$ is continuously differentiable in a neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}_{g0}$ with probability approaching one; $E\{\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{g0})\} = 0$ and $E\{\left\|\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{g0})\right\|^2\}$ is finite and positive-definite; $E\{\sup_{\boldsymbol{\theta}_{jk}\in\boldsymbol{\theta}_g}\left\|\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{jk})\right\|\} < \infty$ and $E\{\sup_{\boldsymbol{\theta}_{jk}\in\mathcal{N}}\left\|\nabla_{\boldsymbol{\theta}_{jk}}\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{jk})\right\|\} < \infty$; $[E\{\nabla_{\boldsymbol{\theta}_{g0}}\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{g0})\}]^T\boldsymbol{C}_{jk}^{-1}E\{\nabla_{\boldsymbol{\theta}_{g0}}\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{g0})\}$ is nonsingular.

Condition H.2. For any $\delta_N \to 0$,

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_N}\frac{N^{1/2}}{1+N^{1/2}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\right\|}\left\|\boldsymbol{\Psi}_N(\boldsymbol{\theta})-\boldsymbol{\Psi}_N(\boldsymbol{\theta}_0)-E_{\boldsymbol{\theta}_0}\boldsymbol{\Psi}_N(\boldsymbol{\theta})\right\| = O_p(N^{-1/2}).$$

Proof of Theorem V.3. Let $\boldsymbol{v}_{jk}(\boldsymbol{\theta}_{jk}) = Var_{\boldsymbol{\theta}_{0,g}}(\boldsymbol{\psi}_{i,jk}(\boldsymbol{\theta}_{jk}))$. From assumption H.1

and Theorem V.1, $Avar(\sqrt{n_k}\widehat{\boldsymbol{\theta}}_{jk}) = \{\boldsymbol{s}_{jk}(\boldsymbol{\theta}_{0,g})\boldsymbol{v}_{jk}^{-1}(\boldsymbol{\theta}_{0,g})\boldsymbol{s}_{jk}^T(\boldsymbol{\theta}_{0,g})\}^{-1}$, and

$$Avar(\sqrt{N}\widehat{\boldsymbol{\theta}}) = \left\{\boldsymbol{s}^T(\boldsymbol{\theta}_0)\boldsymbol{v}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{s}(\boldsymbol{\theta}_0)\right\}^{-1},$$

$$Avar(\sqrt{N}\widehat{\boldsymbol{\theta}}^g) = \left[\left\{\boldsymbol{s}^T(\boldsymbol{\theta}_0)\boldsymbol{v}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{s}(\boldsymbol{\theta}_0)\right\}^{-1}\right]_g,$$

where $[\boldsymbol{A}]_g$ denotes the submatrix of a matrix $\boldsymbol{A}$ consisting of rows and columns corresponding to blocks in $\mathcal{P}^g$. Let $[\boldsymbol{v}(\boldsymbol{\theta})]_{(j,k)}$ denote submatrix of $\boldsymbol{v}(\boldsymbol{\theta})$ with consisting of rows and columns corresponding to block (j,k), and let $[\boldsymbol{v}(\boldsymbol{\theta})]_{-(j,k);}$, $[\boldsymbol{v}(\boldsymbol{\theta})]_{;-(j,k)}$ and $[\boldsymbol{v}(\boldsymbol{\theta})]_{-(j,k)}$ denote the submatrices of $\boldsymbol{v}(\boldsymbol{\theta})$ eliminating respectively rows, columns, and rows and columns corresponding to block (j,k). Clearly $(n_k/N)\boldsymbol{v}_{jk}(\boldsymbol{\theta}_{jk})$ is a submatrix of $\boldsymbol{v}(\boldsymbol{\theta})$: $(n_k/N)\boldsymbol{v}_{jk}(\boldsymbol{\theta}_{jk}) = [\boldsymbol{v}(\boldsymbol{\theta})]_{(j,k)}$.
Consider $(j,k) = (1,1)$, which is in partition set $\mathcal{P}^{g_1}$ for some $g_1 \in \{1,\dots,G\}$ (this is without loss of generality since we can reorganize the rows of $\boldsymbol{\psi}_i$ for $(j,k) \neq (1,1)$ to make $(j,k) = (1,1)$). We write

$$\boldsymbol{v}(\boldsymbol{\theta}) = \begin{pmatrix} [\boldsymbol{v}(\boldsymbol{\theta})]_{(1,1)} & [\boldsymbol{v}(\boldsymbol{\theta})]_{;-(1,1)} \\ [\boldsymbol{v}(\boldsymbol{\theta})]_{-(1,1);} & [\boldsymbol{v}(\boldsymbol{\theta})]_{-(1,1)} \end{pmatrix}.$$

By Corollary 7.7.4. in Horn and Johnson (1990),

$$\boldsymbol{s}_{11}(\boldsymbol{\theta}_{g_0,1})\boldsymbol{v}_{11}^{-1}(\boldsymbol{\theta}_{g_0,1})\boldsymbol{s}_{11}^T(\boldsymbol{\theta}_{g_0,1}) \prec \boldsymbol{s}_{11}(\boldsymbol{\theta}_{g_0,1})\frac{n_1}{N}\left[\boldsymbol{v}^{-1}(\boldsymbol{\theta}_0)\right]_{(1,1)}\boldsymbol{s}_{11}^T(\boldsymbol{\theta}_{g_0,1})$$

where $\preceq$ denotes Löwner's partial ordering in the space of nonnegative definite matrices. By the definition of $\boldsymbol{s}_{g_1}(\boldsymbol{\theta}_{g_0,1})$ in Section 5.2.4, this implies that

$$\begin{aligned}\left\{\boldsymbol{s}_{g_1}^T(\boldsymbol{\theta}_{g_0,1})\left[\boldsymbol{v}^{-1}(\boldsymbol{\theta}_0)\right]_{g_1}\boldsymbol{s}_{g_1}(\boldsymbol{\theta}_{g_0,1})\right\}^{-1} &\prec \left\{\frac{n_1}{N}\boldsymbol{s}_{11}(\boldsymbol{\theta}_{g_0,1})\boldsymbol{v}_{11}^{-1}(\boldsymbol{\theta}_{g_0,1})\boldsymbol{s}_{11}^T(\boldsymbol{\theta}_{g_0,1})\right\}^{-1} \\ &= \lim_{n_1\to\infty}\frac{N}{n_1}Avar(\sqrt{n_1}\widehat{\boldsymbol{\theta}}_{11}).\end{aligned}$$

Again by Corollary 7.7.4. in Horn and Johnson (1990), we have that

$$Avar(\sqrt{N}\widehat{\boldsymbol{\theta}}^{g_1}) = \left[\left\{\boldsymbol{s}^T(\boldsymbol{\theta}_0)\boldsymbol{v}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{s}(\boldsymbol{\theta}_0)\right\}^{-1}\right]_{g_1} \prec \left\{\boldsymbol{s}_{g_1}^T(\boldsymbol{\theta}_{g_0,1})\left[\boldsymbol{v}^{-1}(\boldsymbol{\theta}_0)\right]_{g_1}\boldsymbol{s}_{g_1}(\boldsymbol{\theta}_{g_0,1})\right\}^{-1},$$

implying $Avar(\sqrt{N}\widehat{\boldsymbol{\theta}}^{g_1}) \preceq \{\lim_{n_k\to\infty}(N/n_k)\}Avar(\sqrt{n_1}\widehat{\boldsymbol{\theta}}_{11})$. □

APPENDIX I

Chapter V: Additional Simulation and Data Analysis Results

Additional simulation results

We present additional simulation results for setting two in Section 5.3: simulation metrics with an AR(1) working block correlation structure in Table I.1 and quantile-quantile plots of test statistics from Theorem V.2 in Figure I.1.

Table I.1: Logistic regression simulation setting two results with $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3\}$ and AR(1) working block correlation structure.

(a) Estimates for $\mathcal{P}_1 = \{(1,k),(2,k)\}_{k=1}^{K}$.

	RMSE×10^3	ESE×10^3	ASE×10^3	BIAS×10^4	CI	LEN×10^3	ERR
Intercept	10.86	10.62	10.29	−23.40	0.93	40.00	0.07
X1	3.13	3.08	3.02	5.49	0.94	11.80	0.06
X2	5.53	5.42	5.34	−11.34	0.93	20.81	0.07

(b) Estimates for $\mathcal{P}_2 = \{(3,k)\}_{k=1}^{K}$.

	RMSE×10^3	ESE×10^3	ASE×10^3	BIAS×10^4	CI	LEN×10^3	ERR
Intercept	3.15	3.14	3.33	−2.21	0.96	12.78	0.04
X1	2.16	2.16	2.16	−0.29	0.95	8.38	0.05
X2	1.43	1.44	1.47	0.01	0.96	5.77	0.04

(c) Estimates for $\mathcal{P}_3 = \{(4,k),(5,k)\}_{k=1}^{K}$.

	RMSE×10^3	ESE×10^3	ASE×10^3	BIAS×10^4	CI	LEN×10^3	ERR
Intercept	4.72	4.69	4.77	−5.60	0.95	18.57	0.05
X1	6.36	6.28	6.28	−10.46	0.95	24.36	0.05
X2	2.03	2.01	2.03	2.82	0.95	7.92	0.05

Additional data analysis results

A dictionary for the short-hand names of the sub-pathways is given in Table I.2.

Regression parameter estimates from the heterogeneous model with the partition $\mathcal{P}^h = \{\mathcal{P}_g^h\}_{g=1}^J$, $\mathcal{P}_g^h = \{(g,1),\ldots,(g,K)\}$ are displayed in Figures I.2-I.9. Regression parameter estimates from the integrative model with sub-partition $\mathcal{P}^i$ of $\mathcal{P}^h$ are displayed in Figures I.10-I.17. Scatterplots of smoking effects for heterogeneous and integrative models are shown in Figures I.18-I.25. Smoking effect estimates for the heterogeneous and integrative models are reported in Table I.3 and I.4.

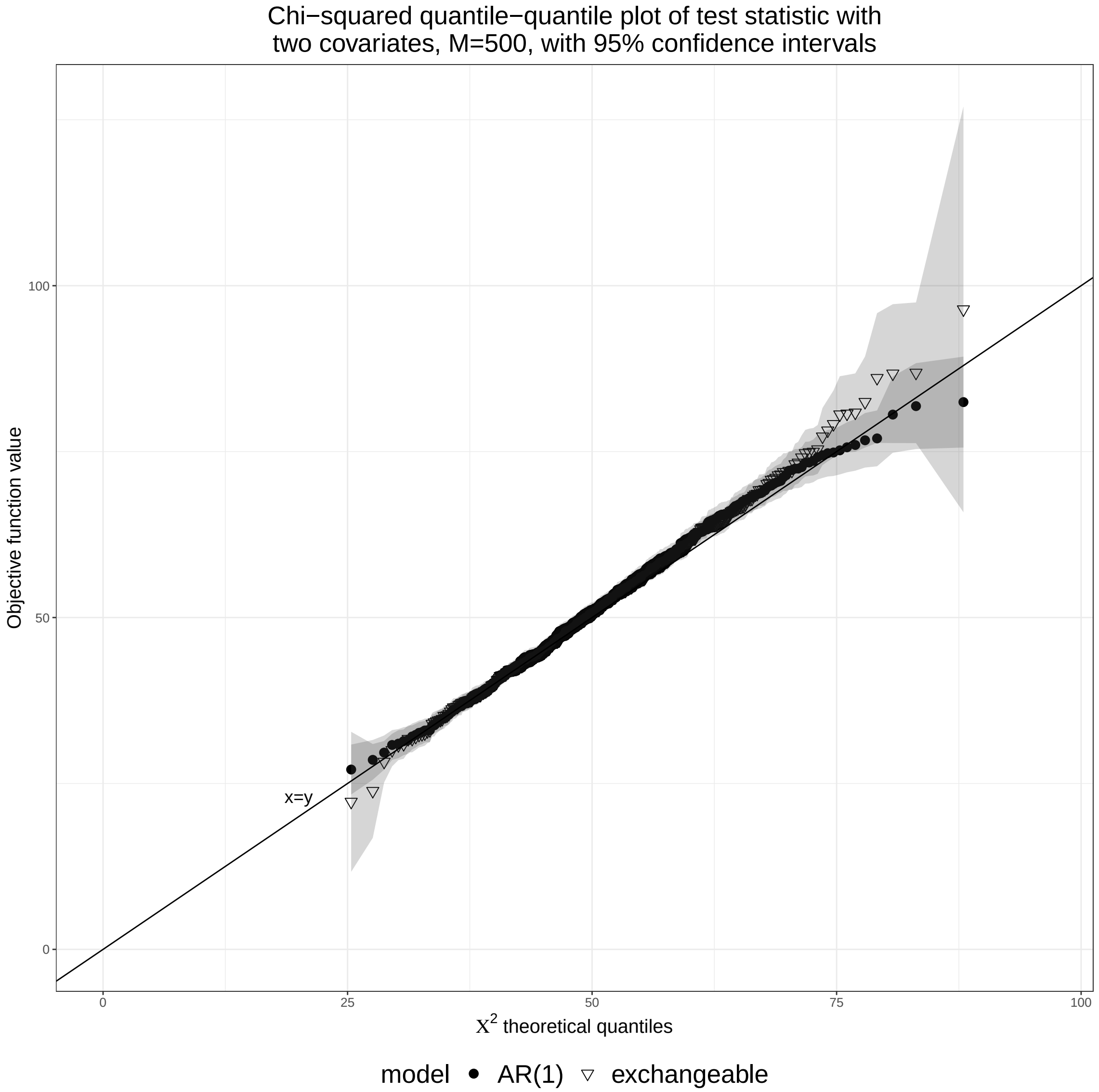


Figure I.1: Chi-squared quantile-quantile plot of test statistics in Theorem 2 with theoretical 95% confidence bands based on 500 simulations under correct and incorrect working block covariance structure. The simulation set-up is that of the second set of simulations ($J = 5$, $\mathcal{P} = \{\mathcal{P}_g\}_{g=1}^{3}$).

Table I.2: Dictionary for the short-hand names of the sub-pathways.

Pathway	Sub-pathway	Short name
Amino Acid	Polyamine Metabolism	SP1
Amino Acid	Leucine, Isoleucine and Valine Metabolism	SP2
Amino Acid	Phenylalanine Metabolism	SP3
Amino Acid	Tyrosine Metabolism	SP4
Amino Acid	Histidine Metabolism	SP5
Amino Acid	Lysine Metabolism	SP6
Amino Acid	Glutathione Metabolism	SP7
Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism	SP8
Amino Acid	Glycine, Serine and Threonine Metabolism	SP9
Amino Acid	Urea cycle; Arginine and Proline Metabolism	SP10
Amino Acid	Tryptophan Metabolism	SP11
Amino Acid	Guanidino and Acetamido Metabolism	SP12
Amino Acid	Glutamate Metabolism	SP13
Amino Acid	Alanine and Aspartate Metabolism	SP14
Amino Acid	Creatine Metabolism	SP15
Carbohydrate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	SP16
Carbohydrate	Pentose Metabolism	SP17
Carbohydrate	Aminosugar Metabolism	SP18
Carbohydrate	Fructose, Mannose and Galactose Metabolism	SP19
Carbohydrate	Glycogen Metabolism	SP20
Cofactors and Vitamins	Nicotinate and Nicotinamide Metabolism	SP21
Cofactors and Vitamins	Ascorbate and Aldarate Metabolism	SP22
Cofactors and Vitamins	Tocopherol Metabolism	SP23
Cofactors and Vitamins	Vitamin A Metabolism	SP24
Cofactors and Vitamins	Hemoglobin and Porphyrin Metabolism	SP25
Cofactors and Vitamins	Pantothenate and CoA Metabolism	SP26
Cofactors and Vitamins	Vitamin B6 Metabolism	SP27
Energy	TCA Cycle	SP28
Energy	Oxidative Phosphorylation	SP29
Lipid	Fatty Acid, Branched	SP30
Lipid	Medium Chain Fatty Acid	SP31
Lipid	Fatty Acid Metabolism (Acyl Carnitine, Hydroxy)	SP32
Lipid	Plasmalogen	SP33
Lipid	Lysoplasmalogen	SP34
Lipid	Lysophospholipid	SP35
Lipid	Monoacylglycerol	SP36
Lipid	Phosphatidylcholine (PC)	SP37
Lipid	Phosphatidylethanolamine (PE)	SP38
Lipid	Phosphatidylinositol (PI)	SP39
Lipid	Phosphatidylserine (PS)	SP40
Lipid	Long Chain Monounsaturated Fatty Acid	SP41
Lipid	Androgenic Steroids	SP42
Lipid	Fatty Acid, Monohydroxy	SP43
Lipid	Pregnenolone Steroids	SP44

Continued on next page.

Table I.2: *Continued from previous page.*

Pathway	Sub-pathway	Short name
Lipid	Fatty Acid, Amino	SP45
Lipid	Fatty Acid Metabolism (Acyl Glycine)	SP46
Lipid	Fatty Acid, Dicarboxylate	SP47
Lipid	Fatty Acid Metabolism (also BCAA Metabolism)	SP48
Lipid	Fatty Acid, Dihydroxy	SP49
Lipid	Fatty Acid Metabolism (Acyl Carnitine, Monounsaturated)	SP50
Lipid	Mevalonate Metabolism	SP51
Lipid	Ketone Bodies	SP52
Lipid	Secondary Bile Acid Metabolism	SP53
Lipid	Sterol	SP54
Lipid	Fatty Acid Metabolism (Acyl Glutamine)	SP55
Lipid	Progestin Steroids	SP56
Lipid	Fatty Acid Metabolism (Acyl Carnitine, Short Chain)	SP57
Lipid	Fatty Acid Metabolism (Acyl Carnitine, Dicarboxylate)	SP58
Lipid	Long Chain Polyunsaturated Fatty Acid (n3 and n6)	SP59
Lipid	Long Chain Saturated Fatty Acid	SP60
Lipid	Fatty Acid Metabolism (Acyl Carnitine, Polyunsaturated)	SP61
Lipid	Fatty Acid Metabolism (Acyl Choline)	SP62
Lipid	Dihydrosphingomyelins	SP63
Lipid	Sphingomyelins	SP64
Lipid	Short Chain Fatty Acid	SP65
Lipid	Carnitine Metabolism	SP66
Lipid	Ceramides	SP67
Lipid	Fatty Acid Metabolism (Acyl Carnitine, Long Chain Saturated)	SP68
Lipid	Primary Bile Acid Metabolism	SP69
Lipid	Phospholipid Metabolism	SP70
Lipid	Corticosteroids	SP71
Lipid	Fatty Acid Metabolism (Acyl Carnitine, Medium Chain)	SP72
Lipid	Diacylglycerol	SP73
Lipid	Estrogenic Steroids	SP74
Lipid	Glycerolipid Metabolism	SP75
Lipid	Hexosylceramides (HCER)	SP76
Lipid	Lactosylceramides (LCER)	SP77
Lipid	Endocannabinoid	SP78
Lipid	Fatty Acid Synthesis	SP79
Lipid	Inositol Metabolism	SP80
Lipid	Ceramide PEs	SP81
Lipid	Sphingolipid Synthesis	SP82
Lipid	Sphingosines	SP83
Nucleotide	Pyrimidine Metabolism, Uracil containing	SP84
Nucleotide	Pyrimidine Metabolism, Cytidine containing	SP85
Nucleotide	Pyrimidine Metabolism, Thymine containing	SP86
Nucleotide	Purine Metabolism, Guanine containing	SP87
Nucleotide	Purine Metabolism, Adenine containing	SP88

Continued on next page.

Table I.2: *Continued from previous page.*

Pathway	Sub-pathway	Short name
Nucleotide	Purine Metabolism, (Hypo)Xanthine/Inosine containing	SP89
Nucleotide	Pyrimidine Metabolism, Orotate containing	SP90
Peptide	Acetylated Peptides	SP91
Peptide	Polypeptide	SP92
Peptide	Fibrinogen Cleavage Peptide	SP93
Peptide	Gamma-glutamyl Amino Acid	SP94
Peptide	Dipeptide	SP95
Peptide	Modified Peptides	SP96
Xenobiotics	Food Component/Plant	SP97
Xenobiotics	Drug - Other	SP98
Xenobiotics	Xanthine Metabolism	SP99
Xenobiotics	Chemical	SP100
Xenobiotics	Drug - Analgesics, Anesthetics	SP101
Xenobiotics	Benzoate Metabolism	SP102
Xenobiotics	Tobacco Metabolite	SP103
Xenobiotics	Drug - Topical Agents	SP104
Xenobiotics	Drug - Antibiotic	SP105
Xenobiotics	Drug - Cardiovascular	SP106
Xenobiotics	Drug - Neurological	SP107
Xenobiotics	Drug - Respiratory	SP108
Xenobiotics	Drug - Psychoactive	SP109
Xenobiotics	Drug - Gastrointestinal	SP110
Xenobiotics	Bacterial/Fungal	SP111
Xenobiotics	Drug - Metabolic	SP112

Figure I.2: Estimated regression parameters for the amino acid pathway from the heterogeneous model.

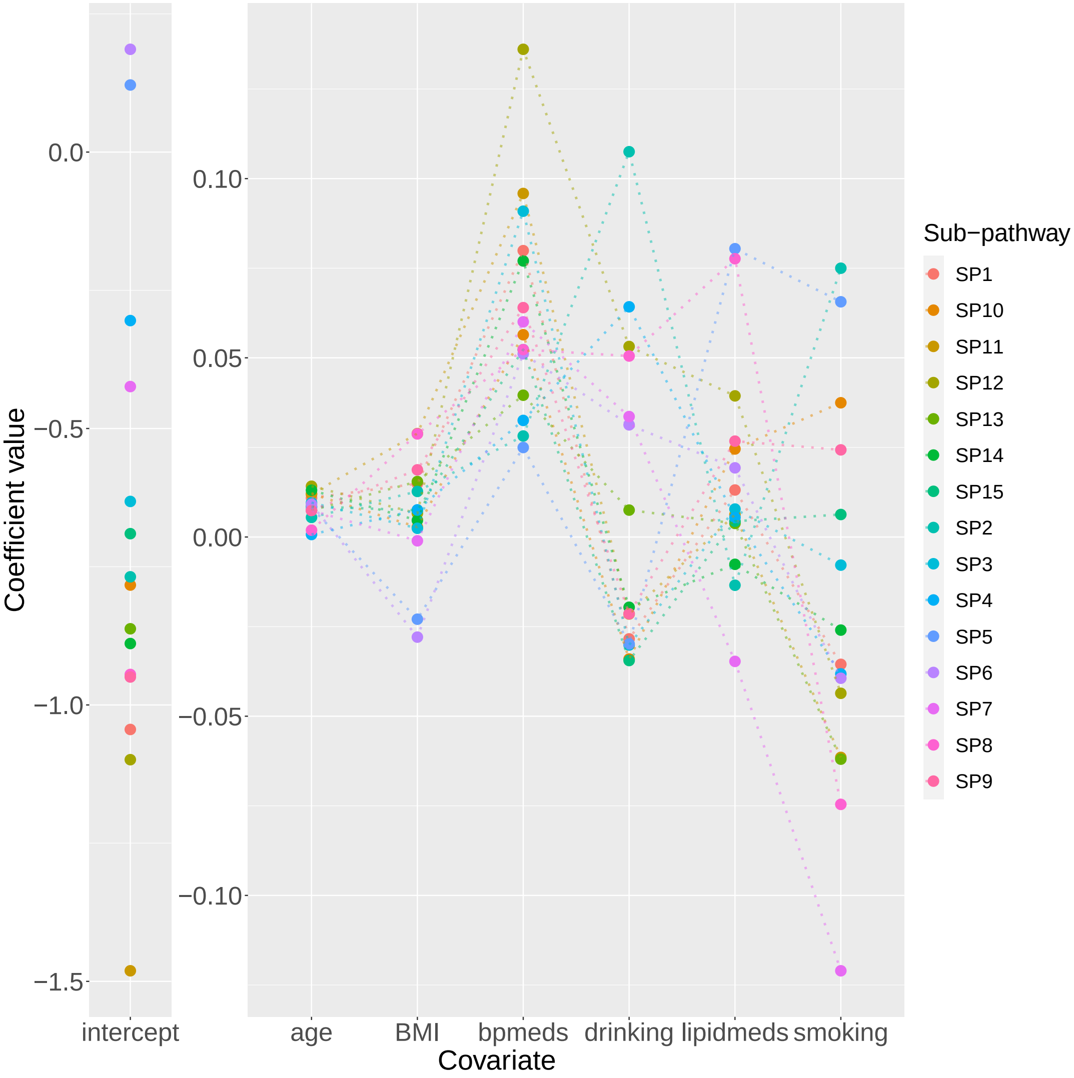


Figure I.3: Estimated regression parameters for the carbohydrate pathway from the heterogeneous model.

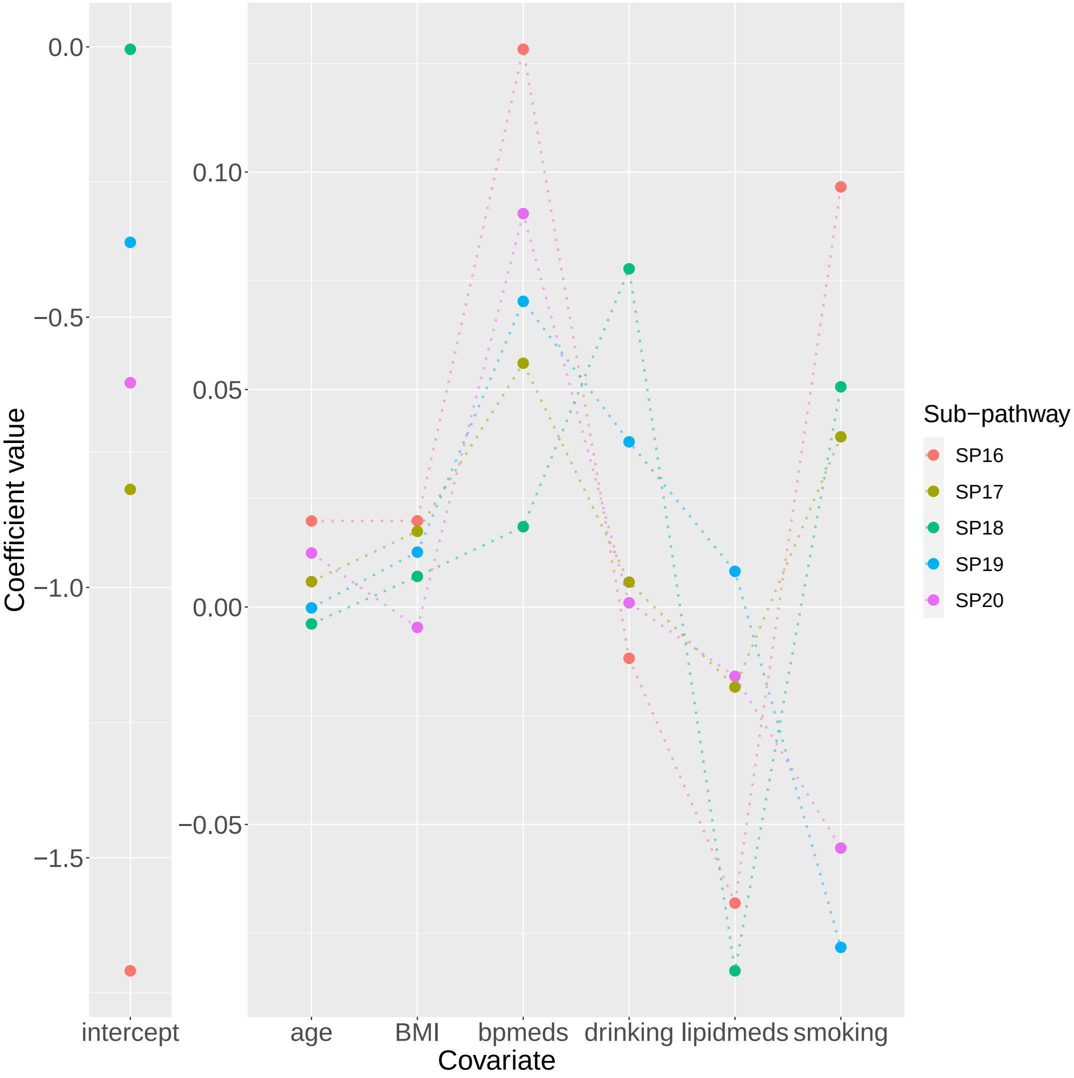


Figure I.4: Estimated regression parameters for the cofactors and vitamins pathway from the heterogeneous model.

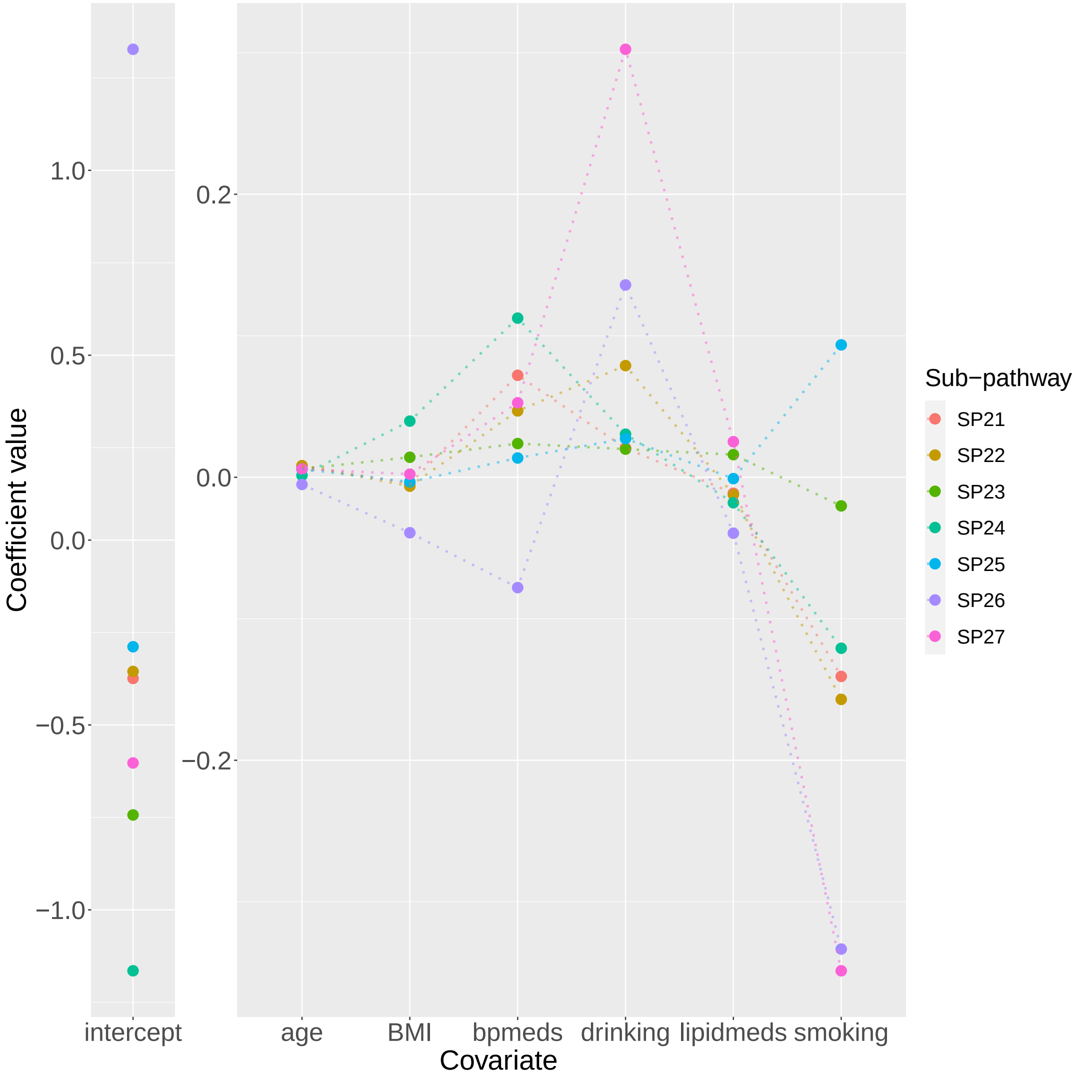


Figure I.5: Estimated regression parameters for the energy pathway from the heterogeneous model.

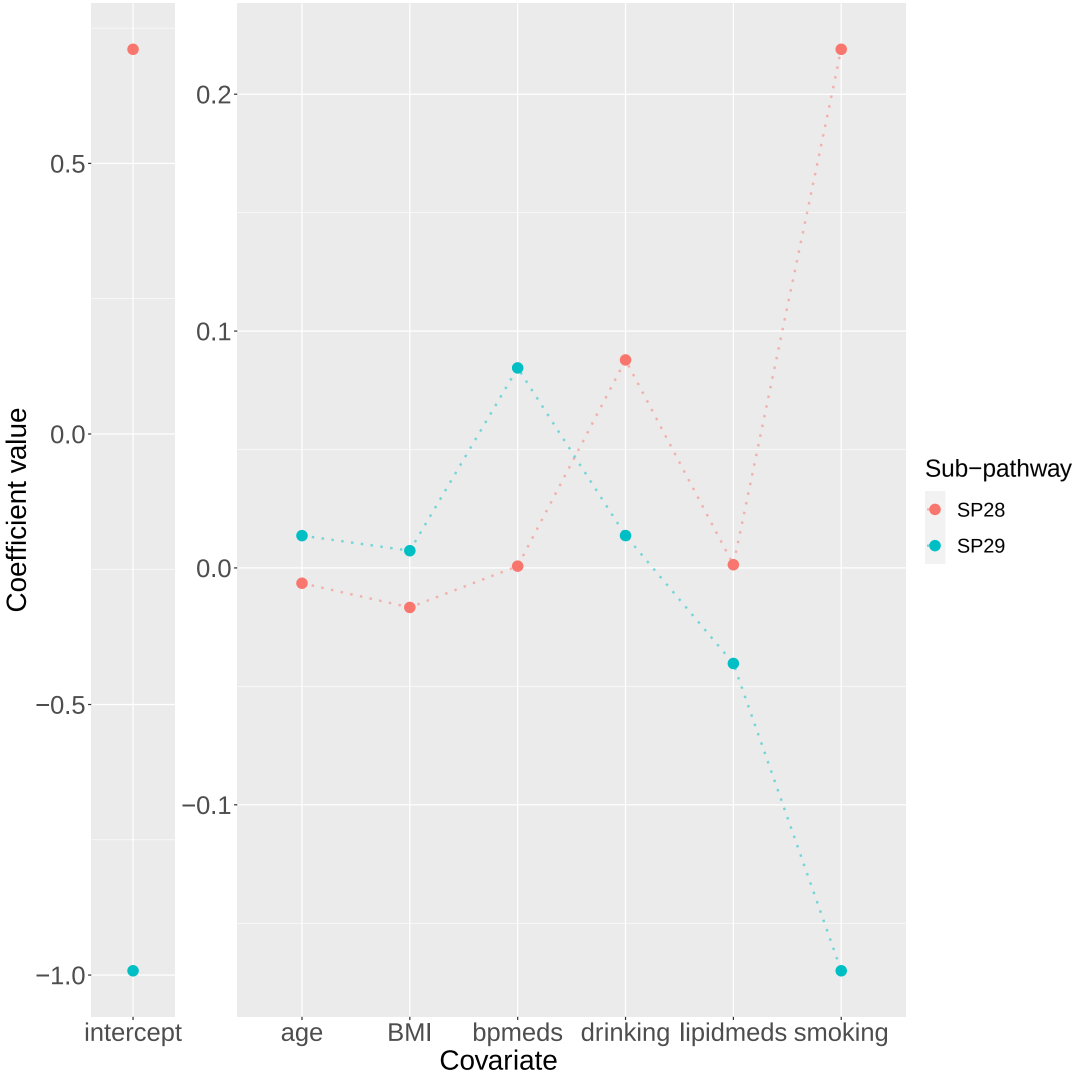


Figure I.6: Estimated regression parameters for the lipid pathway from the heterogeneous model.

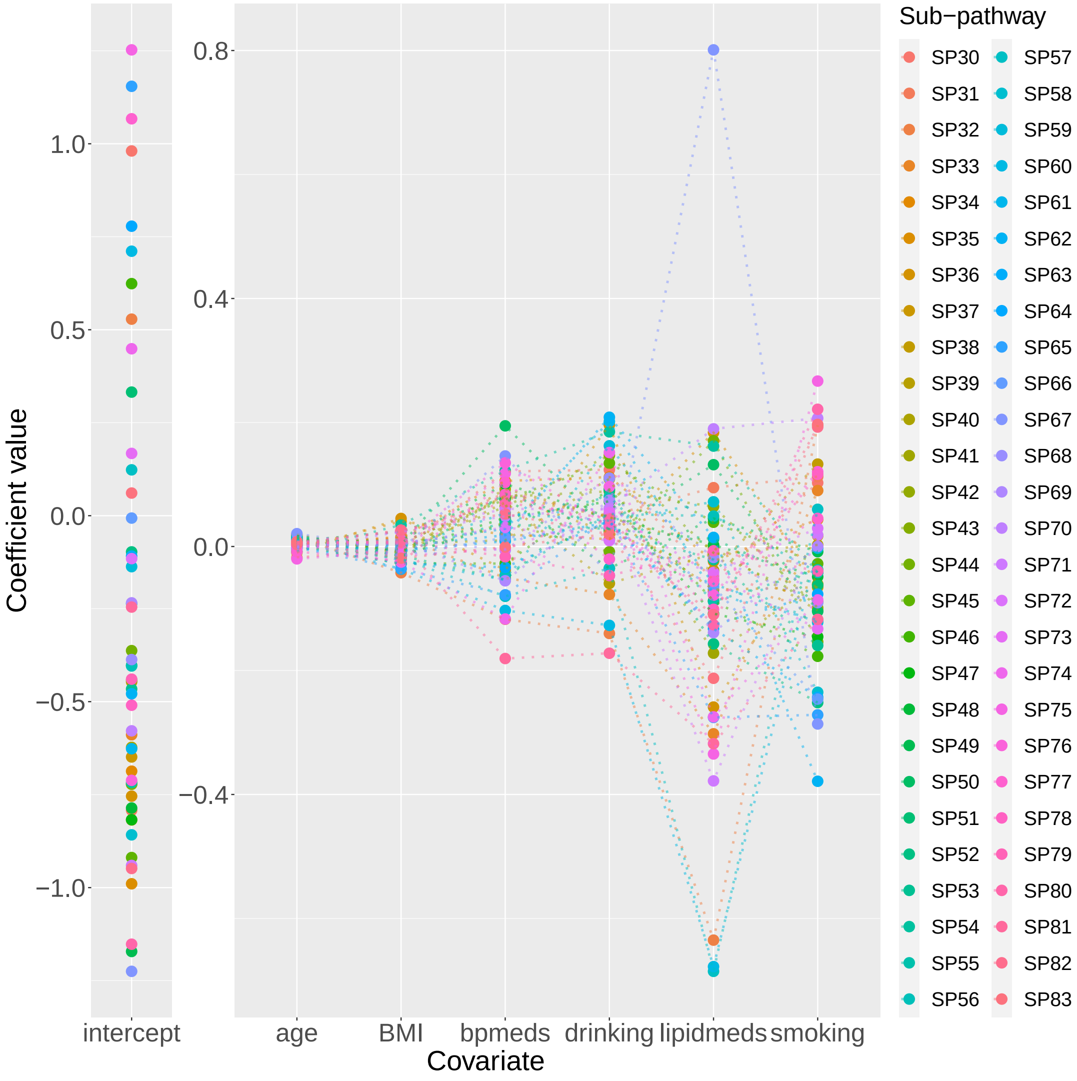


Figure I.7: Estimated regression parameters for the nucleotide pathway from the heterogeneous model.

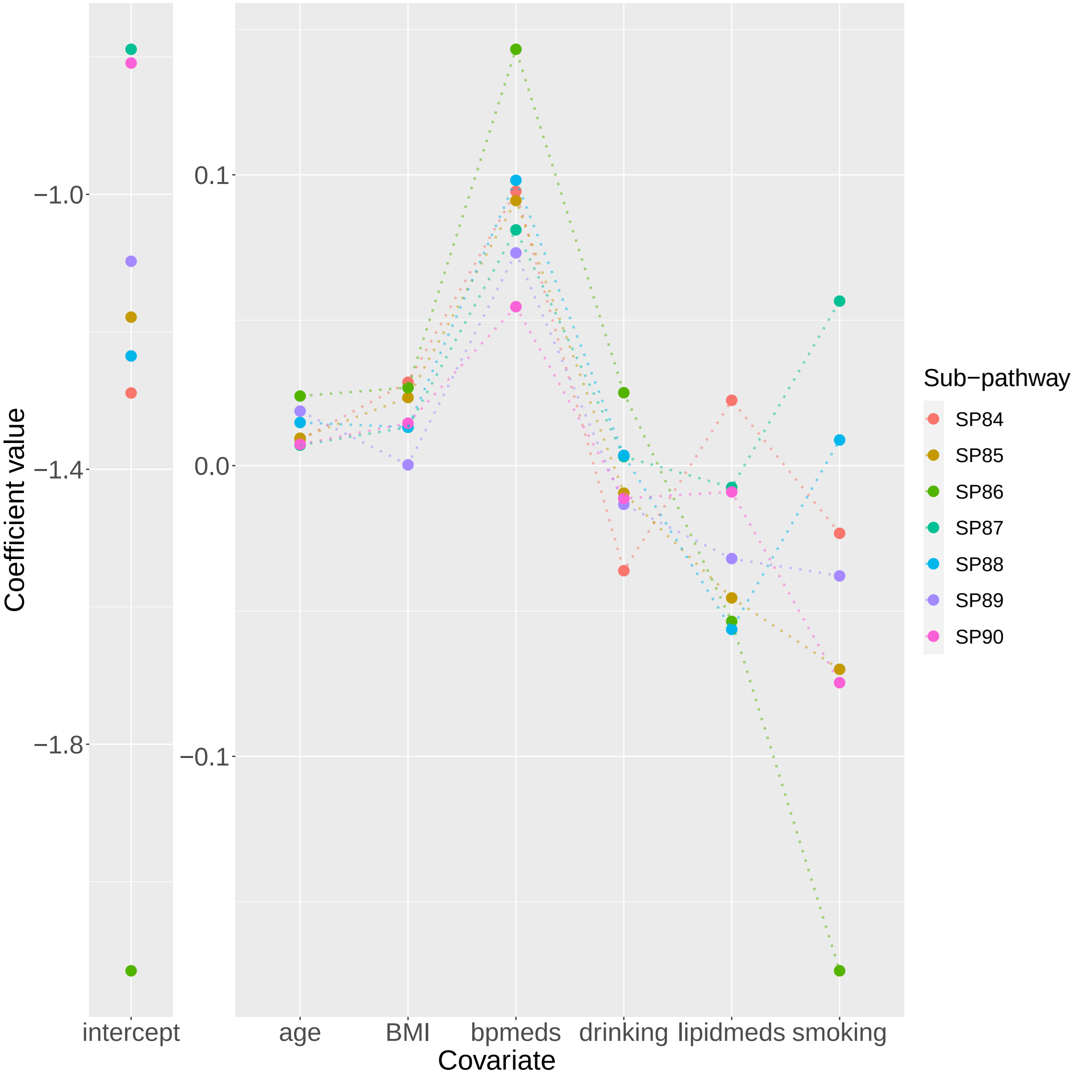


Figure I.8: Estimated regression parameters for the peptide pathway from the heterogeneous model.

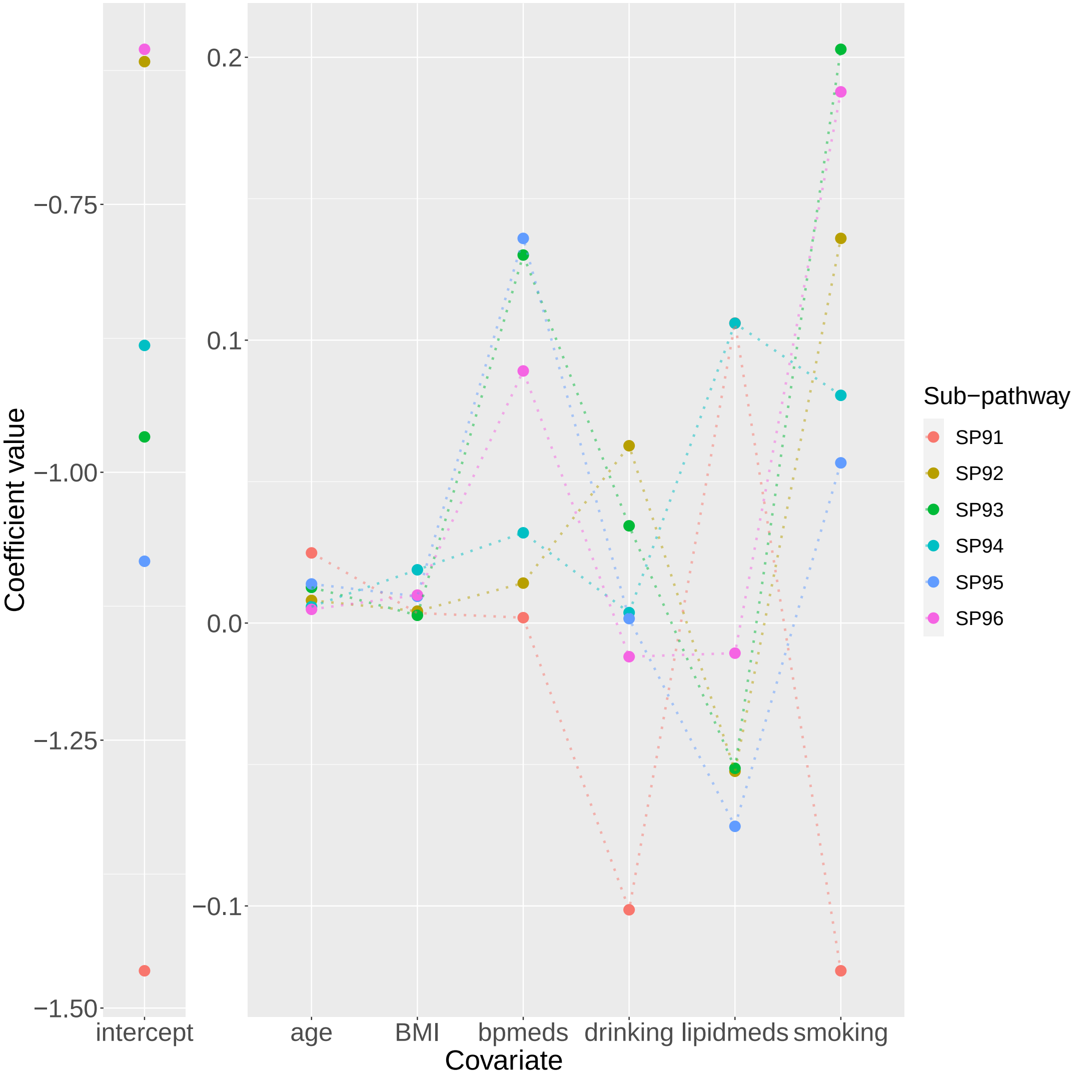


Figure I.9: Estimated regression parameters for the xenobiotics pathway from the heterogeneous model.

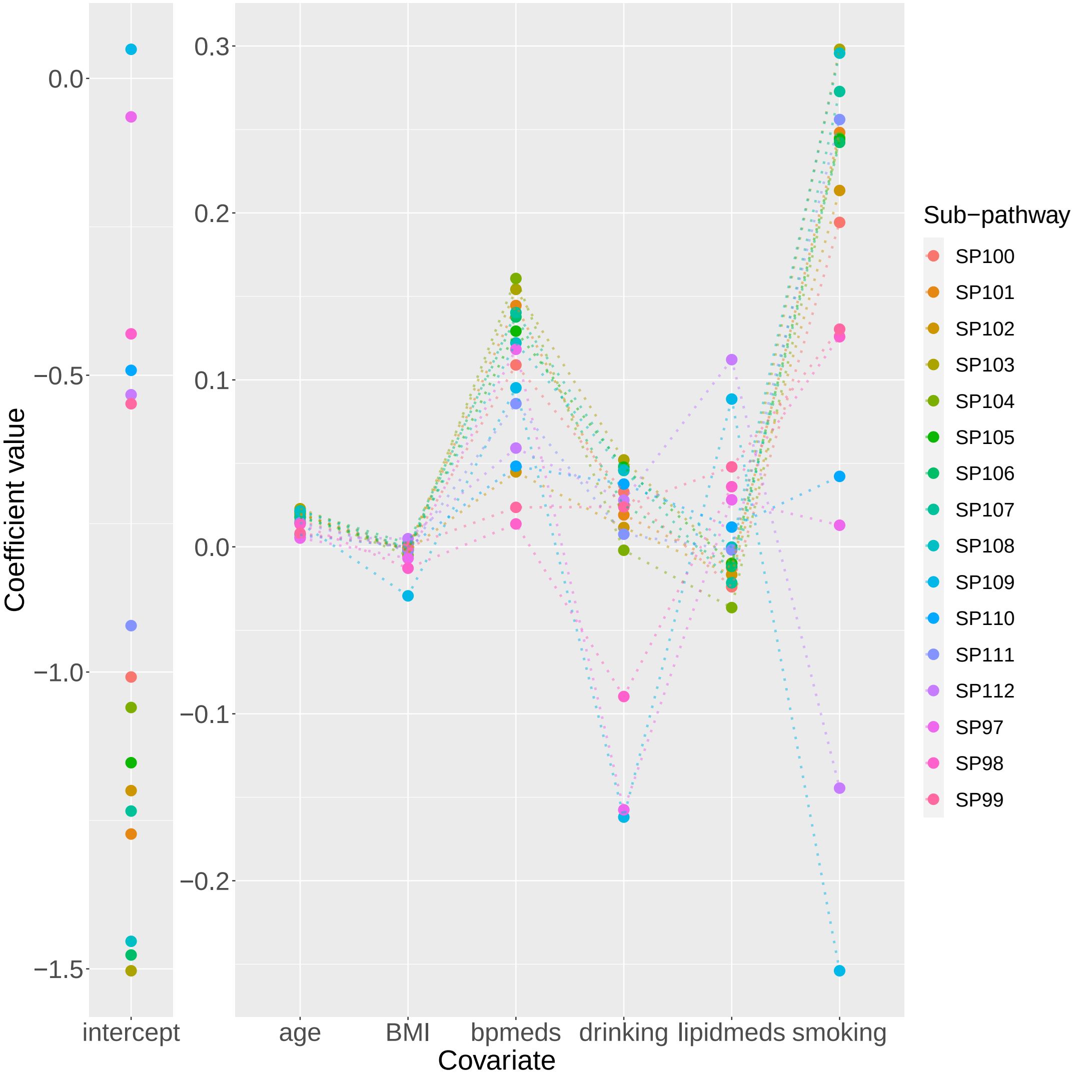


Figure I.10: Estimated regression parameters for the amino acid pathway from the integrative model.

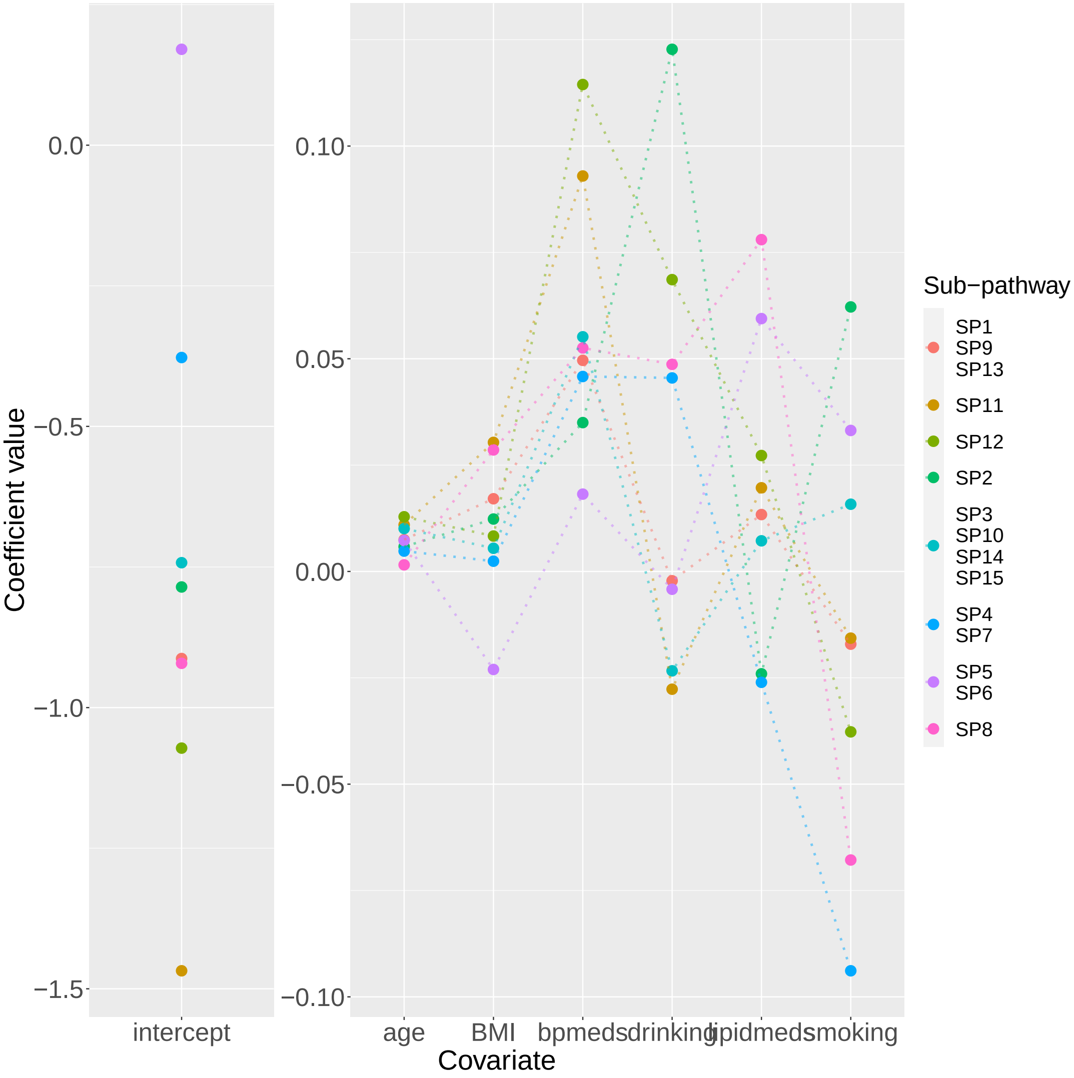


Figure I.11: Estimated regression parameters for the carbohydrate pathway from the integrative model.

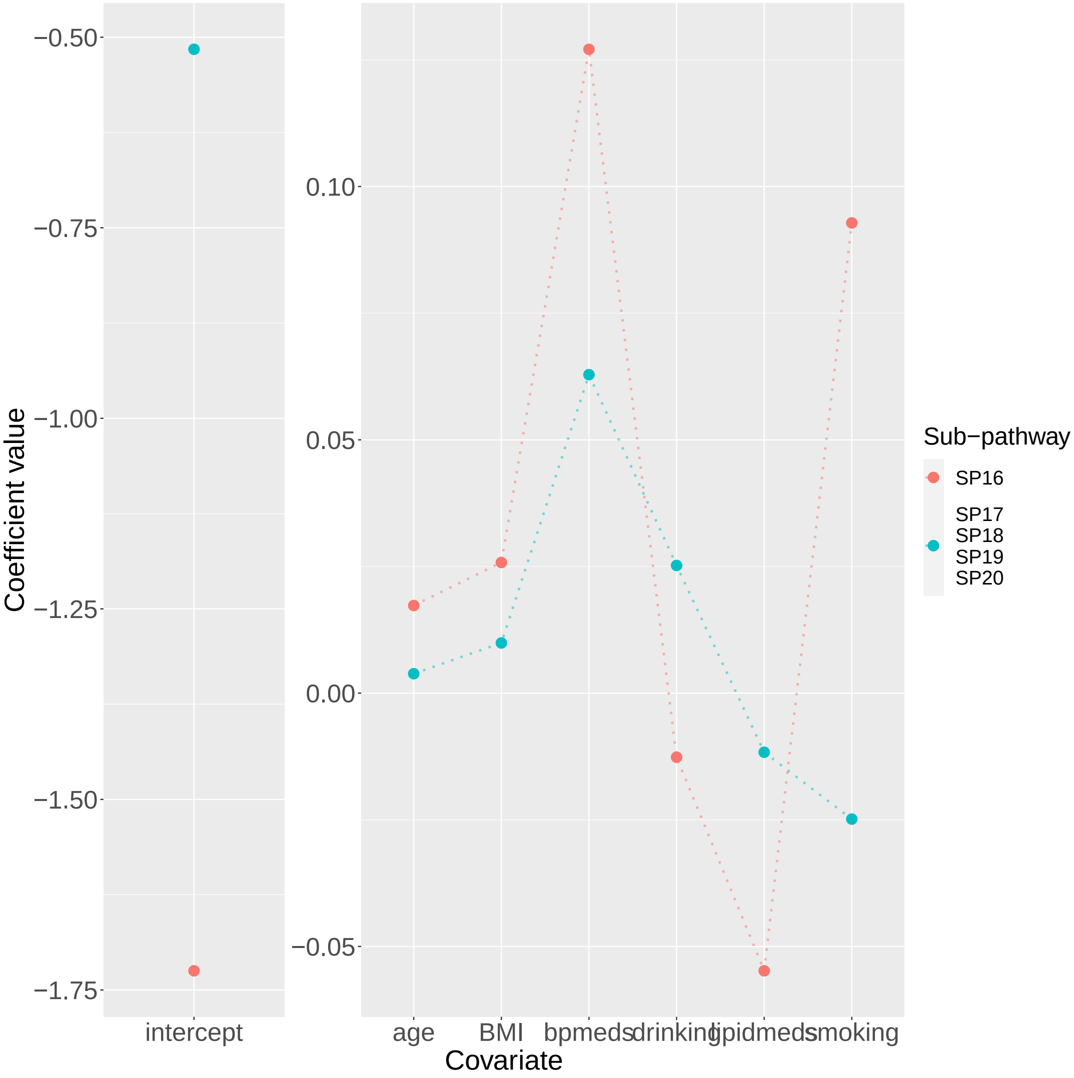


Figure I.12: Estimated regression parameters for the cofactors and vitamins pathway from the integrative model.

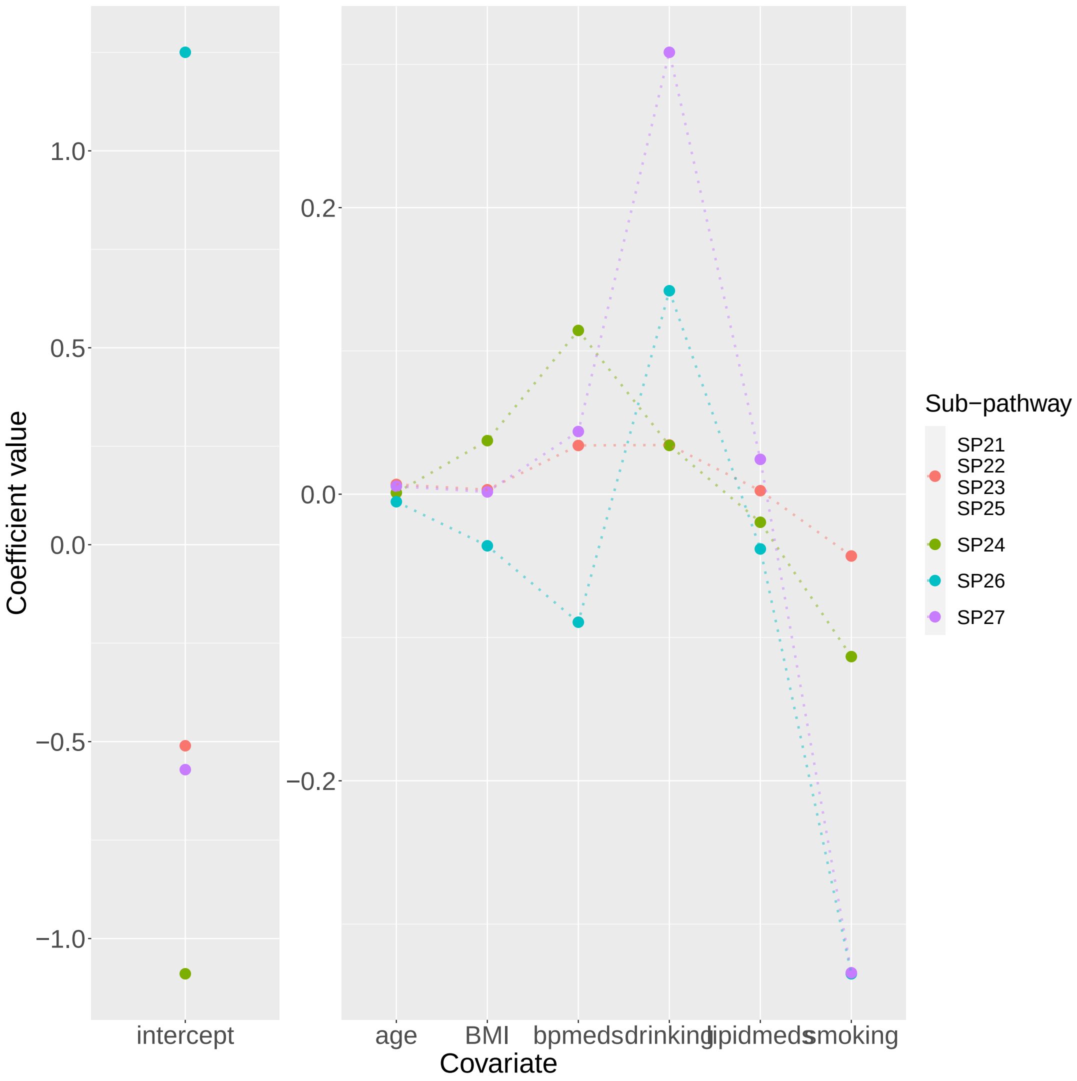


Figure I.13: Estimated regression parameters for the energy pathway from the integrative model.

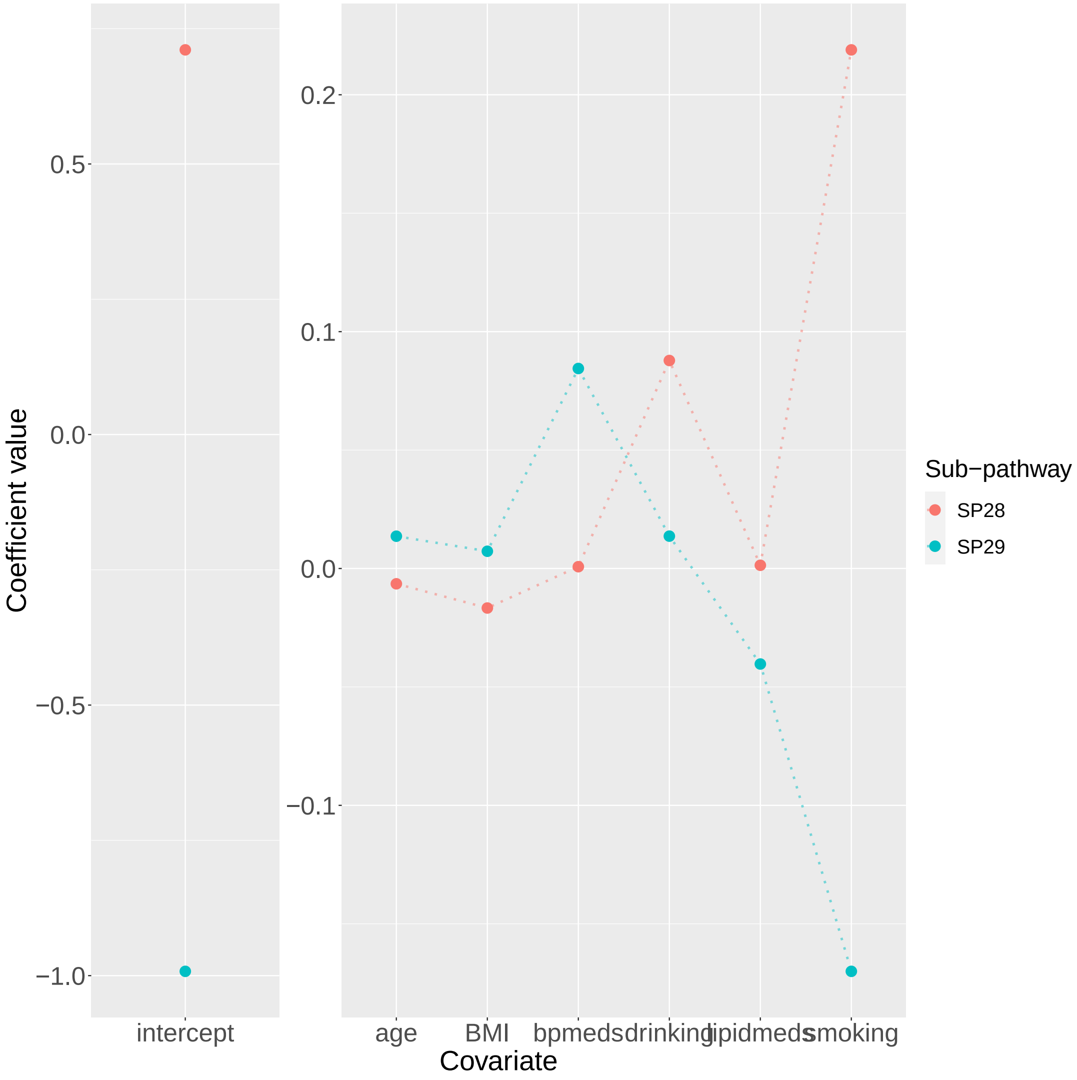


Figure I.14: Estimated regression parameters for the lipid pathway from the integrative model.

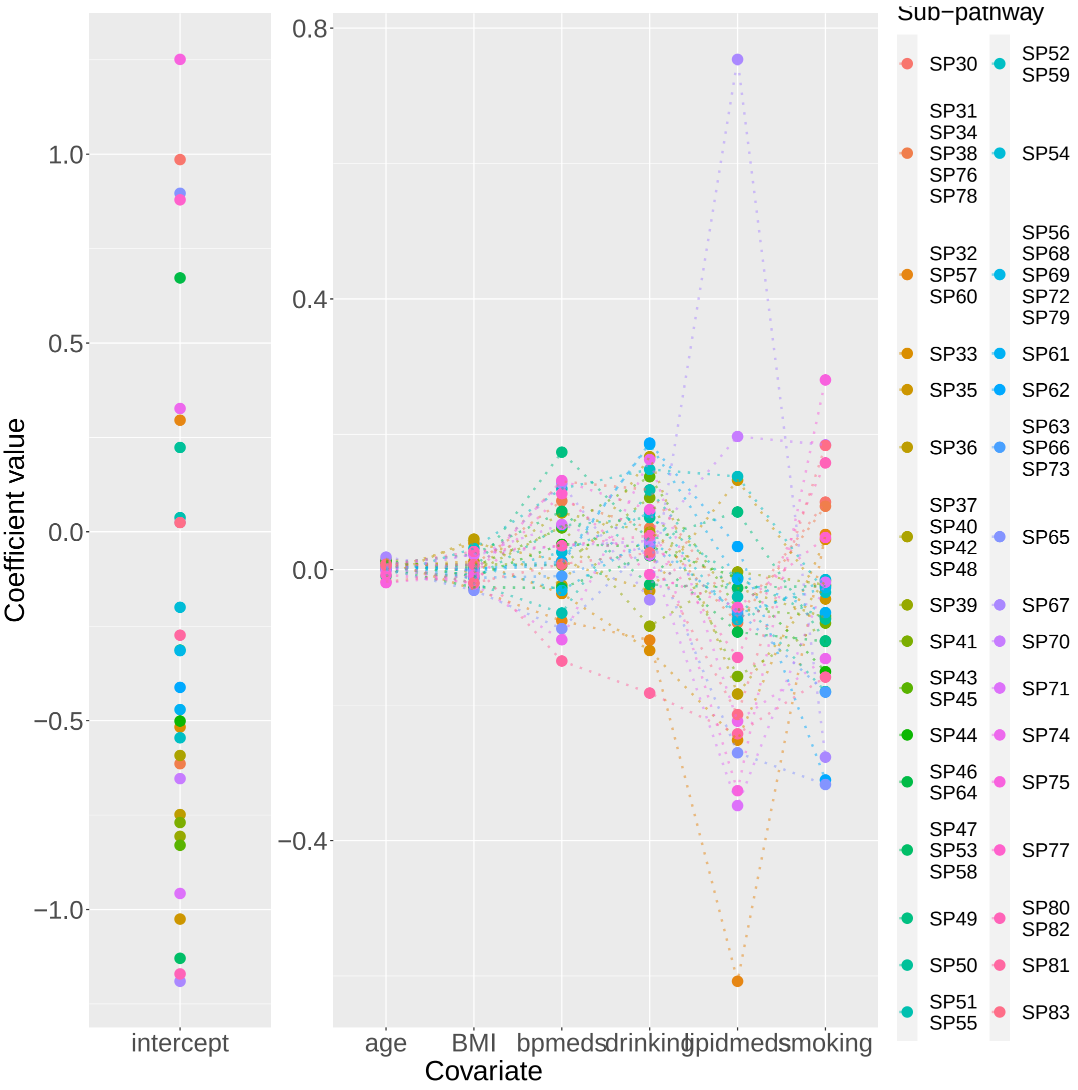


Figure I.15: Estimated regression parameters for the nucleotide pathway from the integrative model.

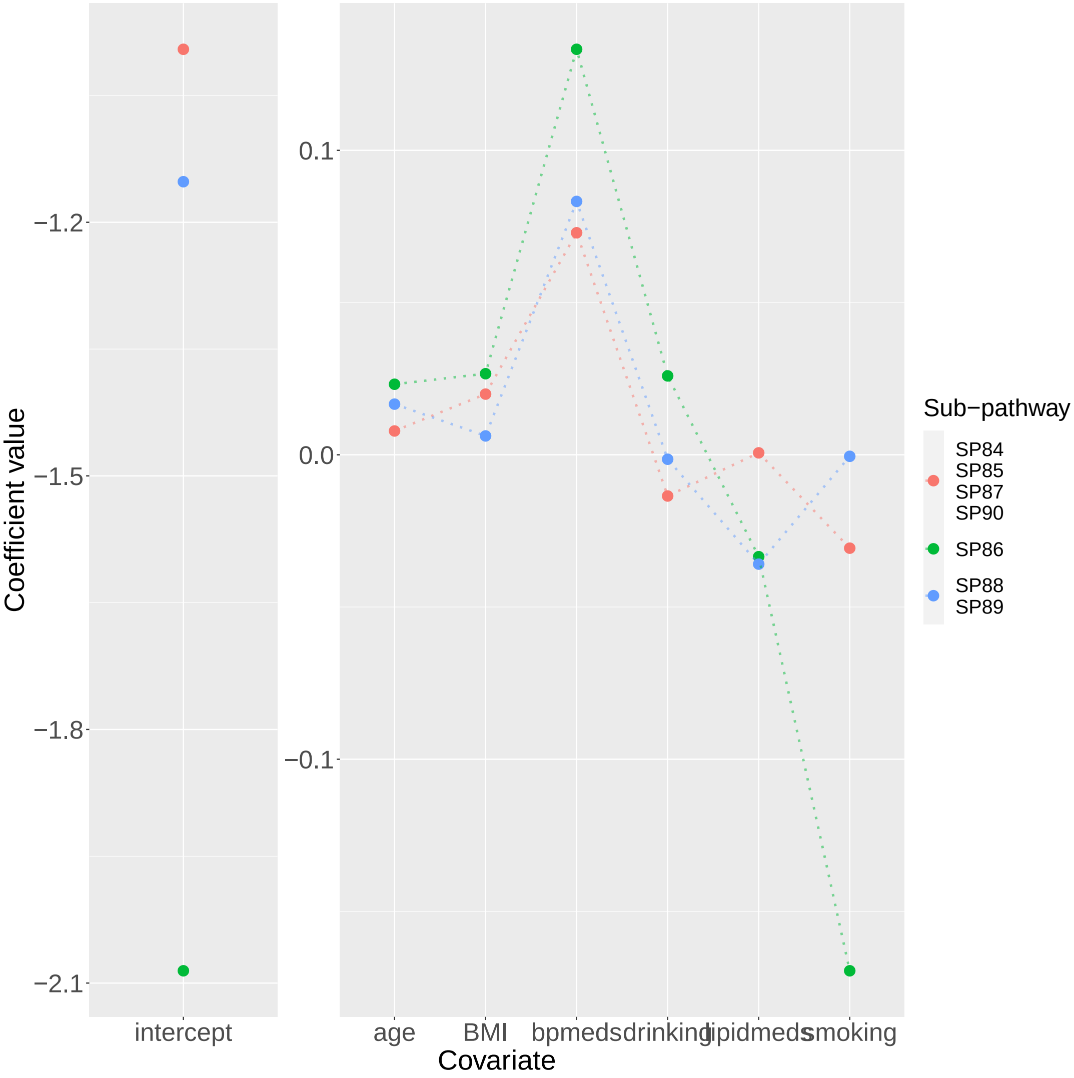


Figure I.16: Estimated regression parameters for the peptide pathway from the integrative model.

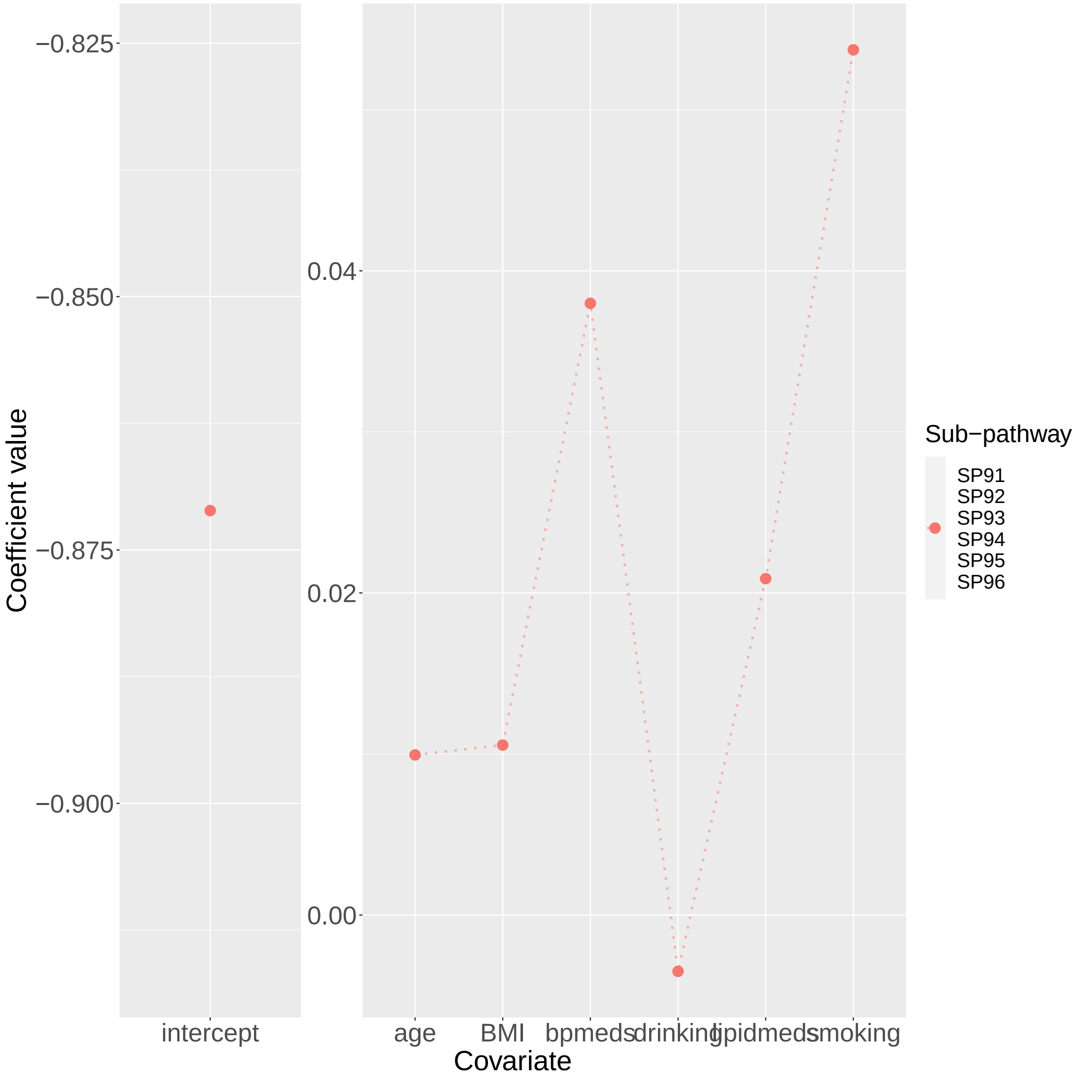


Figure I.17: Estimated regression parameters for the xenobiotics pathway from the integrative model.

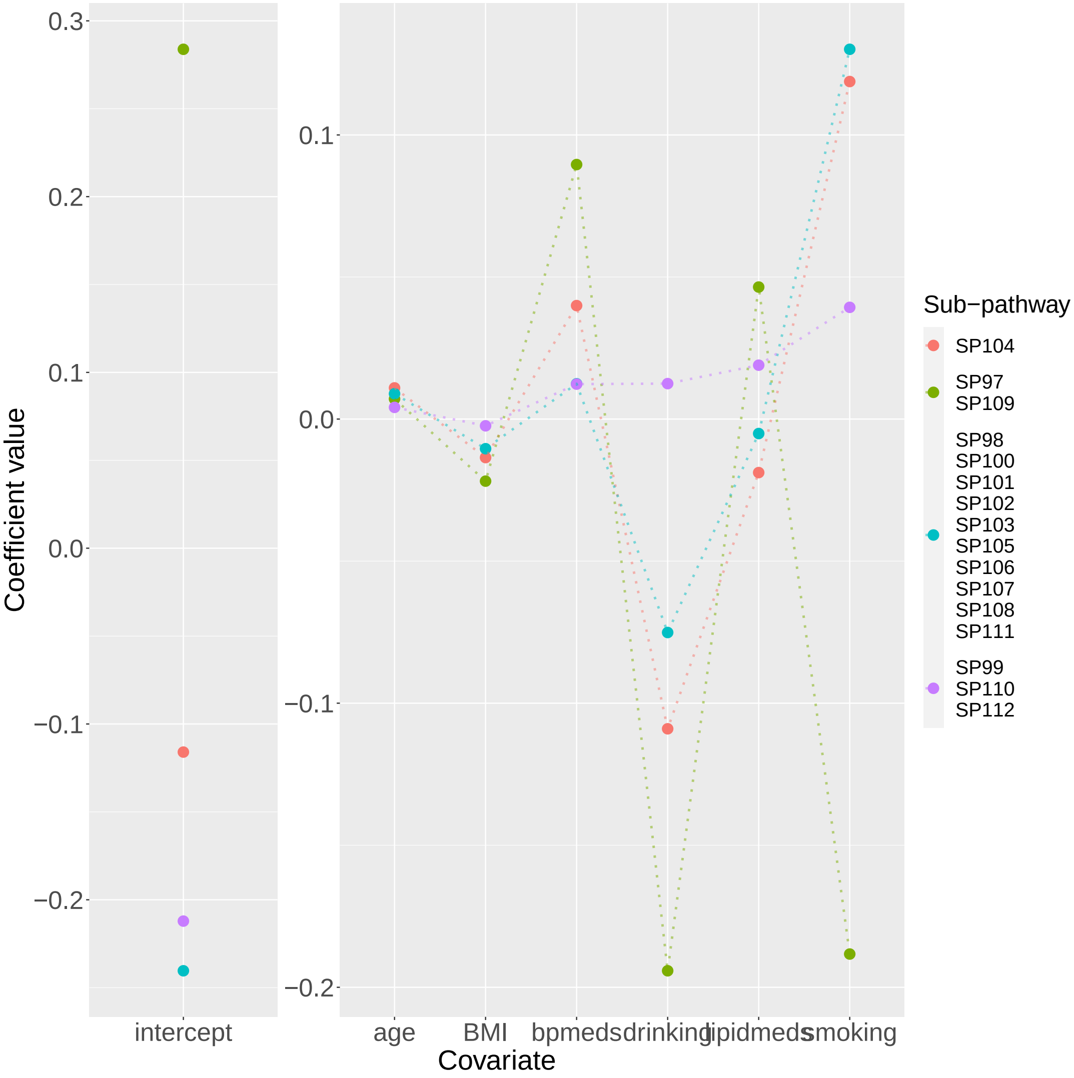


Figure I.18: Estimated smoking effect for the amino acid pathway from the heterogeneous and integrative models categorized by significance at the 0.05/8 level.

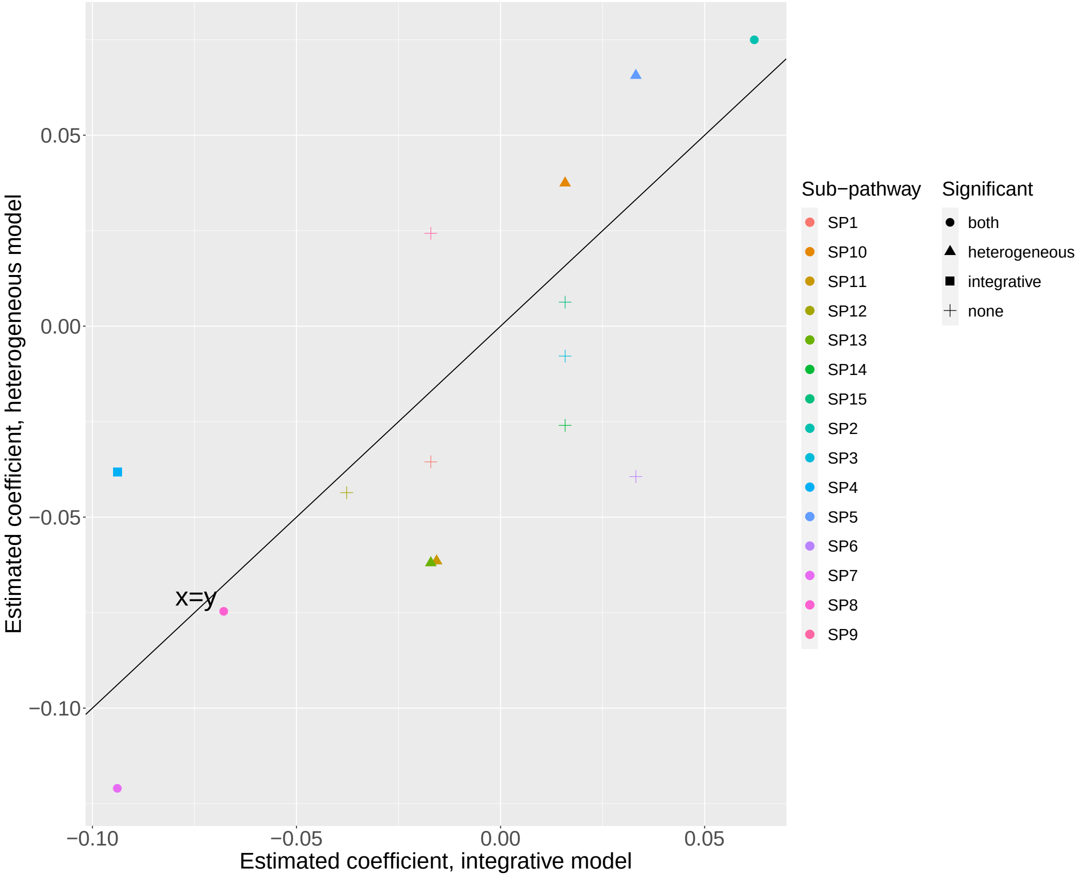


Figure I.19: Estimated smoking effect for the carbohydrate pathway from the heterogeneous and integrative models categorized by significance at the 0.05/8 level.

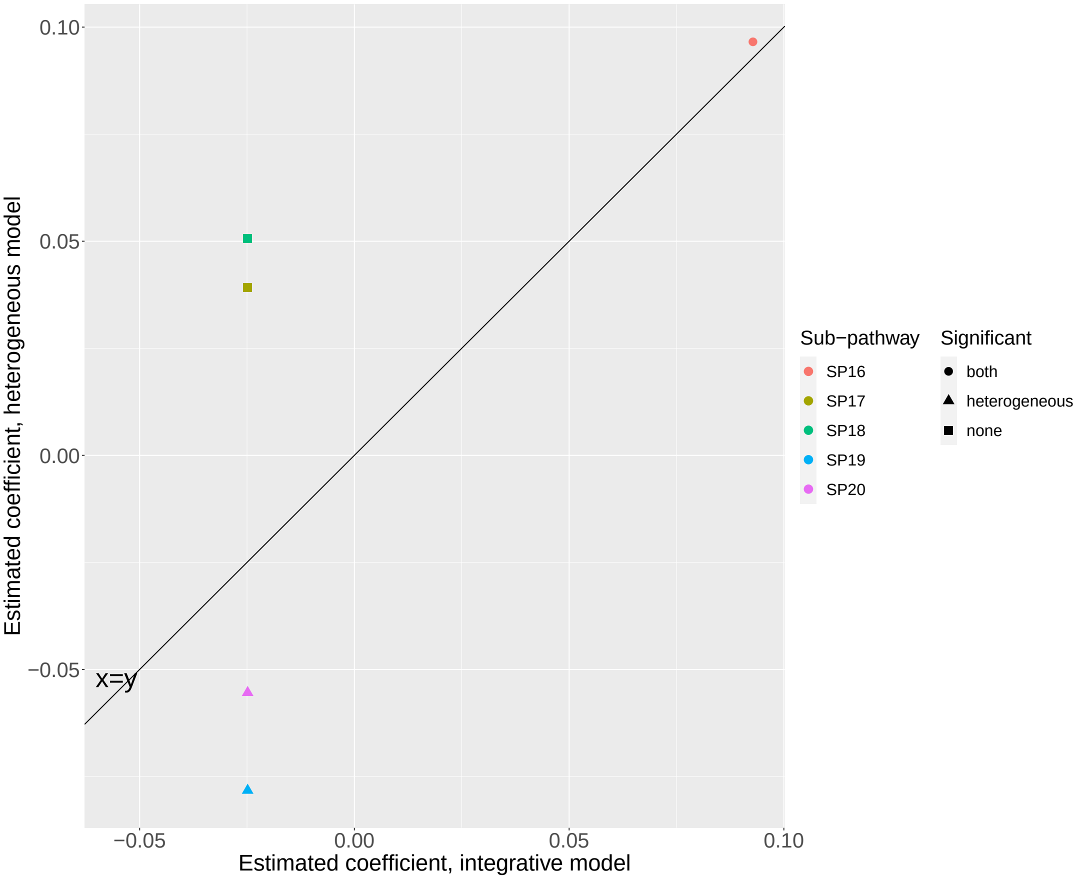


Figure I.20: Estimated smoking effect for the cofactors and vitamins pathway from the heterogeneous and integrative models categorized by significance at the 0.05/8 level.

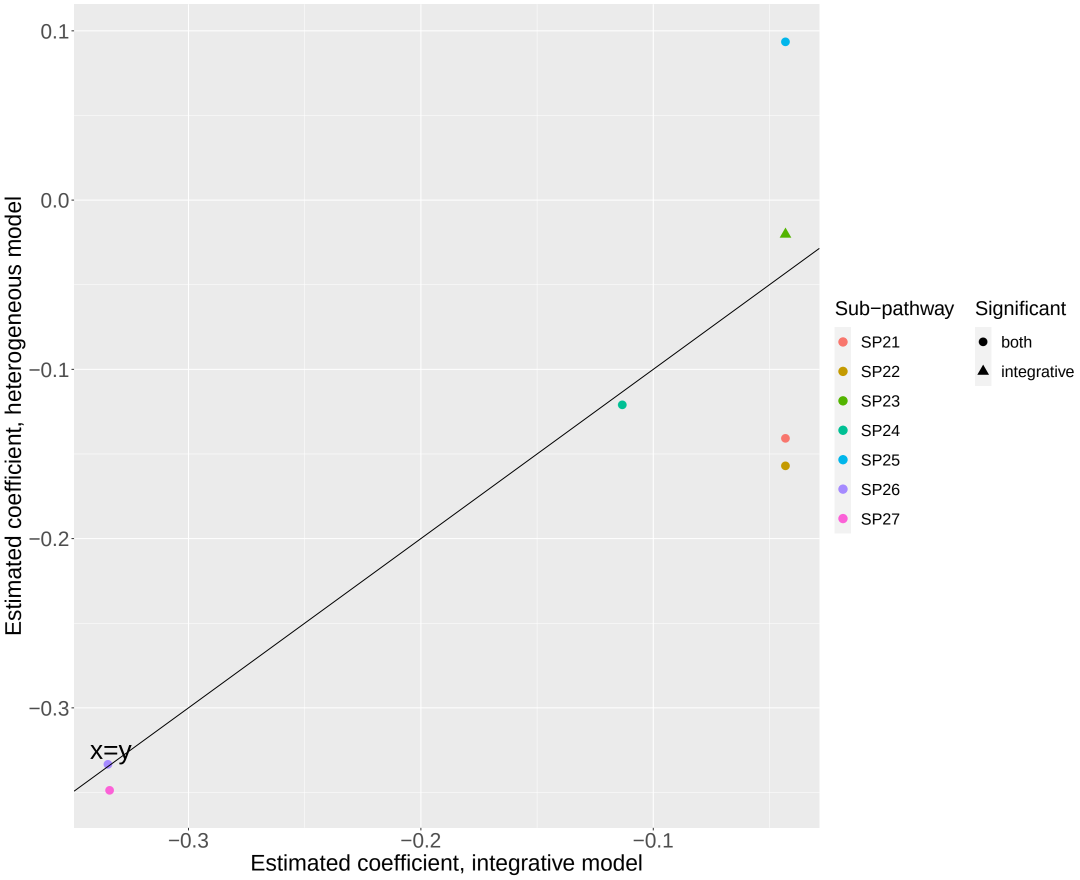


Figure I.21: Estimated smoking effect for the energy pathway from the heterogeneous and integrative models categorized by significance at the 0.05/8 level.

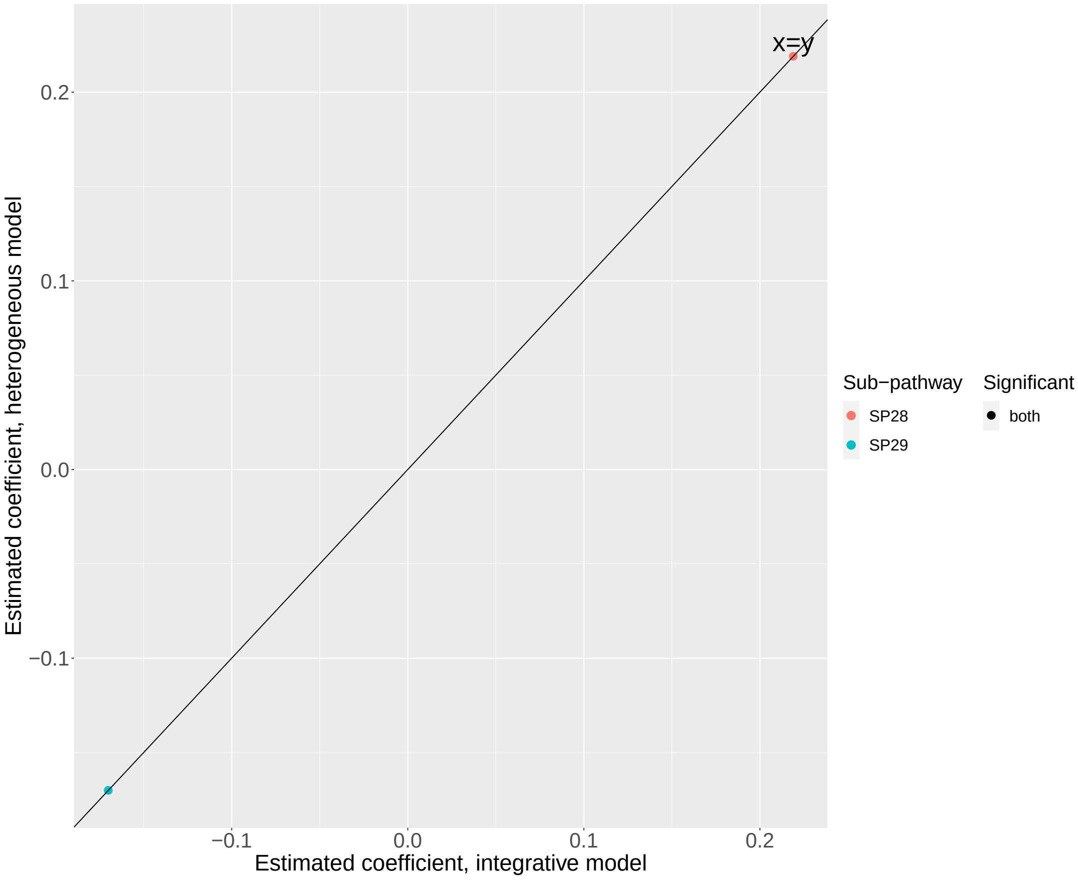


Figure I.22: Estimated smoking effect for the lipid pathway from the heterogeneous and integrative models categorized by significance at the 0.05/8 level.

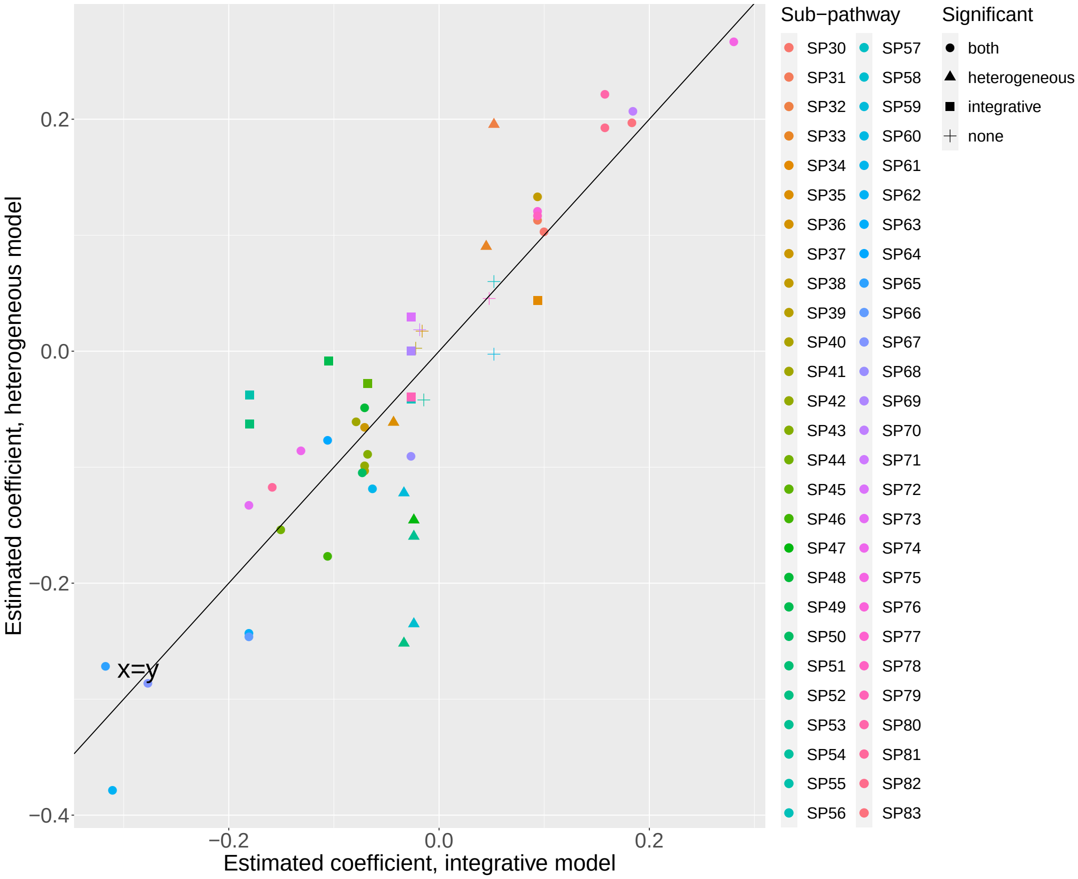


Figure I.23: Estimated smoking effect for the nucleotide pathway from the heterogeneous and integrative models categorized by significance at the 0.05/8 level.

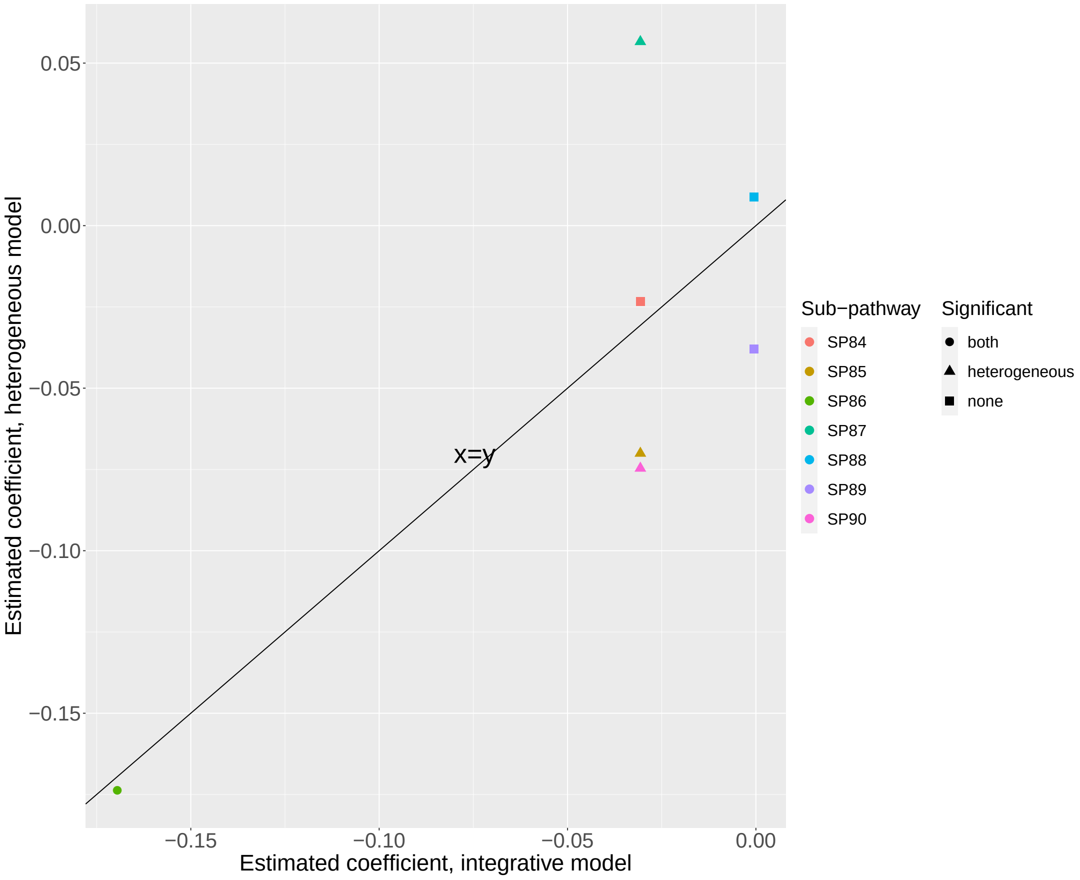


Figure I.24: Estimated smoking effect for the peptide pathway from the heterogeneous and integrative models categorized by significance at the 0.05/8 level.

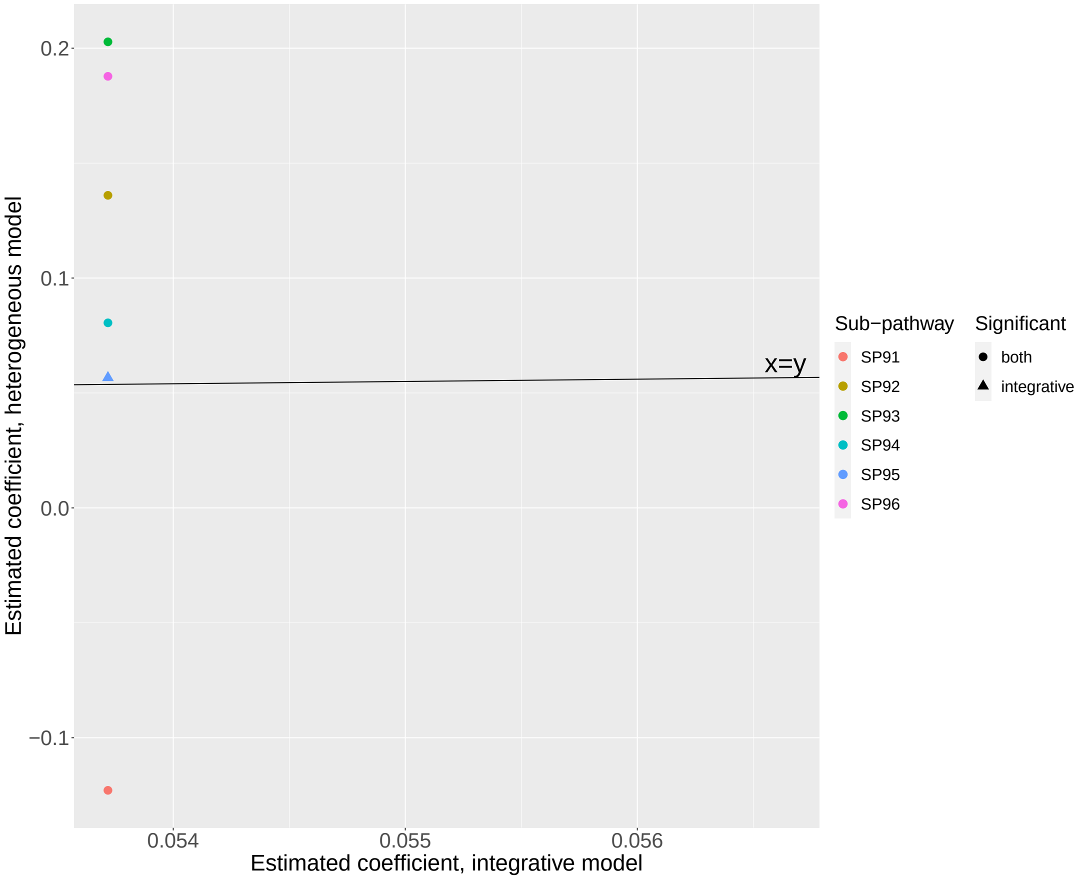


Figure I.25: Estimated smoking effect for the xenobiotics pathway from the heterogeneous and integrative models categorized by significance at the 0.05/8 level.

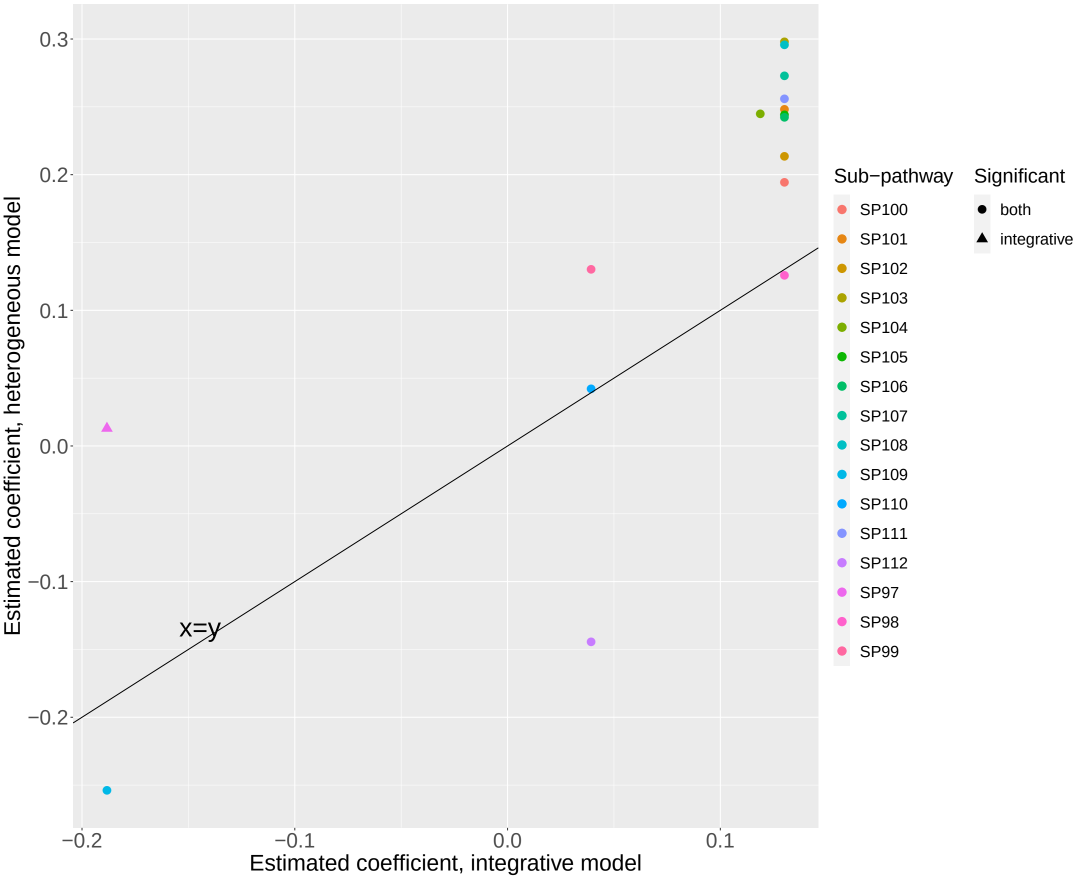


Table I.3: Estimated effects of smoking for the heterogeneous model. Starred sub-pathways have a significant effect of smoking at level 0.05/8. s.e. standard error.

Sub-pathway	Estimate (s.e.$\times10^2$)	p-value
SP1	−0.036 (1.6)	0.022
SP2*	0.075 (1.9)	6.3×10^{-5}
SP3	−0.0078 (1.3)	0.56
SP4	−0.038 (1.8)	0.031
SP5*	0.066 (1.6)	2.5×10^{-5}
SP6	−0.039 (2.5)	0.11
SP7*	−0.12 (1.4)	1.3×10^{-17}
SP8*	−0.075 (1.5)	8.9×10^{-7}
SP9	0.024 (1.3)	0.071
SP10*	0.037 (1.2)	0.0022
SP11*	−0.061 (1.8)	0.00058
SP12	−0.044 (2)	0.026
SP13*	−0.062 (1.3)	7.1×10^{-7}
SP14	−0.026 (1.3)	0.043
SP15	0.0063 (1.4)	0.65
SP16*	0.097 (2)	2.1×10^{-6}
SP17	0.039 (1.8)	0.031
SP18	0.051 (3)	0.087
SP19*	−0.078 (1.7)	4.2×10^{-6}
SP20*	−0.055 (1.9)	0.0037
SP21*	−0.14 (2.4)	7.7×10^{-9}
SP22*	−0.16 (1.9)	1.1×10^{-16}
SP23	−0.02 (1.6)	0.21
SP24*	−0.12 (3.5)	0.00047
SP25*	0.094 (1.7)	3.8×10^{-8}
SP26*	−0.33 (2.4)	6.1×10^{-45}
SP27*	−0.35 (3.5)	4.9×10^{-23}
SP28*	0.22 (3.5)	2.7×10^{-10}
SP29*	−0.17 (2)	4.9×10^{-18}
SP30*	0.1 (1.6)	1.7×10^{-10}
SP31*	0.11 (2.2)	4.2×10^{-7}
SP32*	0.2 (2.7)	3.6×10^{-13}
SP33*	0.09 (2.3)	8.4×10^{-5}
SP34	0.044 (1.9)	0.024
SP35*	−0.061 (2.1)	0.0031
SP36	0.017 (2.3)	0.45
SP37*	−0.066 (2)	0.0009
SP38*	0.13 (2.4)	1.7×10^{-8}
SP39	0.0026 (2.2)	0.91
SP40*	−0.1 (2.2)	3.5×10^{-6}
SP41*	−0.061 (2.1)	0.004
SP42*	−0.099 (2.5)	7.8×10^{-5}
SP43*	−0.089 (2)	9.9×10^{-6}
SP44*	−0.15 (2.4)	1.1×10^{-10}

Continued on next page.

Table I.3: *Continued from previous page.*

Sub-pathway	Estimate (s.e.×10^2)	p-value
SP45	−0.028 (2.8)	0.32
SP46*	−0.18 (2.6)	7.1×10^{-12}
SP47*	−0.15 (2.4)	6×10^{-10}
SP48*	−0.049 (1.6)	0.0019
SP49	−0.0084 (1.7)	0.62
SP50*	−0.1 (2.8)	0.00015
SP51	−0.063 (2.3)	0.0064
SP52*	−0.25 (2.4)	2.2×10^{-26}
SP53*	−0.16 (1.2)	6.3×10^{-41}
SP54	−0.042 (2.1)	0.049
SP55	−0.038 (1.4)	0.0091
SP56	−0.041 (2.1)	0.046
SP57	0.06 (2.3)	0.01
SP58*	−0.24 (2.6)	2×10^{-19}
SP59*	−0.12 (2.9)	1.9×10^{-5}
SP60	−0.0026 (2.8)	0.93
SP61*	−0.12 (2.4)	8×10^{-7}
SP62*	−0.38 (2.1)	1.1×10^{-73}
SP63*	−0.24 (2.3)	1.8×10^{-25}
SP64*	−0.077 (1.6)	1.3×10^{-6}
SP65*	−0.27 (2.1)	2.8×10^{-39}
SP66*	−0.25 (1.9)	1.6×10^{-39}
SP67*	−0.29 (2.7)	3.4×10^{-26}
SP68*	−0.091 (1.9)	2.1×10^{-6}
SP69	0.00011 (1.4)	0.99
SP70*	0.21 (2.3)	8.5×10^{-20}
SP71	0.019 (2.2)	0.4
SP72	0.029 (2.6)	0.26
SP73*	−0.13 (1.6)	3.9×10^{-17}
SP74*	−0.086 (1.8)	1.1×10^{-6}
SP75*	0.27 (2.4)	1.3×10^{-29}
SP76*	0.12 (1.5)	1.7×10^{-15}
SP77	0.045 (2.2)	0.042
SP78*	0.12 (1.2)	5.2×10^{-22}
SP79	−0.04 (2.1)	0.06
SP80*	0.22 (2.6)	1.8×10^{-17}
SP81*	−0.12 (2.1)	1.1×10^{-8}
SP82*	0.19 (2.5)	2.9×10^{-14}
SP83*	0.2 (1.6)	4.9×10^{-35}
SP84	−0.023 (1.7)	0.18
SP85*	−0.07 (2.1)	0.00096
SP86*	−0.17 (2.9)	2.3×10^{-9}
SP87*	0.057 (2)	0.0038
SP88	0.0088 (2.1)	0.68

Continued on next page.

Table I.3: *Continued from previous page.*

Sub-pathway	Estimate (s.e.$\times 10^2$)	p-value
SP89	−0.038 (2.1)	0.076
SP90*	−0.075 (1.6)	2.5×10^{-6}
SP91*	−0.12 (2.7)	6×10^{-6}
SP92*	0.14 (2.1)	4.9×10^{-11}
SP93*	0.2 (2.8)	2.9×10^{-13}
SP94*	0.081 (2.4)	0.00066
SP95	0.057 (2.6)	0.032
SP96*	0.19 (2.2)	5.2×10^{-18}
SP97	0.013 (3.2)	0.69
SP98*	0.13 (1.5)	6.9×10^{-18}
SP99*	0.13 (1.1)	3.1×10^{-33}
SP100*	0.19 (2)	3×10^{-22}
SP101*	0.25 (2.5)	6.2×10^{-24}
SP102*	0.21 (2.2)	2.3×10^{-21}
SP103*	0.3 (2.8)	6.2×10^{-27}
SP104*	0.24 (2.5)	6.4×10^{-22}
SP105*	0.24 (2.3)	2.9×10^{-26}
SP106*	0.24 (3.1)	4.9×10^{-15}
SP107*	0.27 (2.5)	2.8×10^{-28}
SP108*	0.3 (2.9)	8.3×10^{-25}
SP109*	−0.25 (2.2)	1.7×10^{-30}
SP110*	0.042 (1)	2.9×10^{-5}
SP111*	0.26 (2.8)	1.5×10^{-20}
SP112*	−0.14 (2.6)	1.7×10^{-8}

Table I.4: Estimated effects of smoking for the integrative model. Sub-pathways separated by a semi-colon have been combined in the integrative analysis. Starred sub-pathway combinations have a significant effect of smoking at level 0.05/8. s.e. standard error.

Sub-pathway	Estimate (s.e.$\times 10^2$)	p-value
SP1; SP9; SP13	−0.017 (1.1)	0.11
SP2*	0.062 (1.8)	0.00048
SP3; SP10; SP14; SP15	0.016 (1)	0.12
SP4; SP7*	−0.094 (1.2)	5.2×10^{-15}
SP5; SP6	0.033 (1.4)	0.015
SP8*	−0.068 (1.5)	6.3×10^{-6}
SP11	−0.016 (1.7)	0.35
SP12	−0.038 (1.9)	0.047
SP16*	0.093 (2)	2.1×10^{-6}
SP17; SP18; SP19; SP20	−0.025 (1.3)	0.055
SP21; SP22; SP23; SP25*	−0.043 (1.1)	7.4×10^{-5}
SP24*	−0.11 (3.4)	0.00094
SP26*	−0.33 (2.3)	3×10^{-49}
SP27*	−0.33 (3.5)	6.9×10^{-22}
SP28*	0.22 (3.5)	2.7×10^{-10}
SP29*	−0.17 (2)	4.9×10^{-18}
SP30*	0.1 (1.5)	8.6×10^{-11}
SP31; SP34; SP38; SP76; SP78*	0.094 (0.98)	1.1×10^{-21}
SP32; SP57; SP60	0.052 (1.9)	0.007
SP33	0.045 (2)	0.029
SP35	−0.043 (1.7)	0.011
SP36	−0.016 (2.1)	0.45
SP37; SP40; SP42; SP48*	−0.071 (1.1)	1.5×10^{-10}
SP39	−0.022 (1.9)	0.24
SP41*	−0.079 (1.6)	1.2×10^{-6}
SP43; SP45*	−0.068 (1.2)	1.8×10^{-8}
SP44*	−0.15 (0.96)	8.6×10^{-56}
SP46; SP64*	−0.11 (1.2)	8.2×10^{-19}
SP47; SP53; SP58	−0.024 (1.5)	0.099
SP49*	−0.11 (2.6)	5.5×10^{-5}
SP50*	−0.073 (2.1)	0.00068
SP51; SP55*	−0.18 (1.3)	2.8×10^{-41}
SP52; SP59	−0.033 (2)	0.094
SP54	−0.015 (1.1)	0.19
SP56; SP68; SP69; SP72; SP79*	−0.027 (0.88)	0.0025
SP61*	−0.063 (1.5)	1.9×10^{-5}
SP62*	−0.31 (1.5)	6.2×10^{-92}
SP63; SP66; SP73*	−0.18 (1.1)	1.3×10^{-64}
SP65*	−0.32 (1.8)	3.2×10^{-68}
SP67*	−0.28 (2.5)	7.7×10^{-28}
SP70*	0.18 (2)	1.8×10^{-20}
SP71	−0.018 (2)	0.35

Continued on next page.

Table I.4: *Continued from previous page.*

Sub-pathway	Estimate (s.e.$\times10^2$)	p-value
SP74*	-0.13 (1.6)	2.1×10^{-16}
SP75*	0.28 (2.3)	1.5×10^{-35}
SP77	0.048 (2.2)	0.027
SP80; SP82*	0.16 (1.8)	5.8×10^{-18}
SP81*	-0.16 (1.8)	4.7×10^{-19}
SP83*	0.18 (1.5)	2.3×10^{-35}
SP84; SP85; SP87; SP90	-0.031 (1.4)	0.025
SP86*	-0.17 (2.7)	4.7×10^{-10}
SP88; SP89	-5×10^{-4} (1.8)	0.98
SP91; SP92; SP93; SP94; SP95; SP96*	0.054 (1.4)	0.00021
SP97; SP109*	-0.19 (1.8)	1.5×10^{-24}
SP98; SP100; SP101; SP102; SP103; SP105; SP106; SP107; SP108; SP111*	0.13 (1.1)	1.2×10^{-30}
SP99; SP110; SP112*	0.039 (0.75)	1.8×10^{-7}
SP104*	0.12 (1.6)	2.5×10^{-13}

BIBLIOGRAPHY

BIBLIOGRAPHY

Arbia, G. (2014), Pairwise likelihood inference for spatial regressions estimated on very large datasets, *Spatial Statistics*, *7*, 21–39.

Bahadur, R. R. (1961), A representation of the joint distribution of responses on n dichotomous items, *Studies in Item Analysis and Prediction, Stanford Mathematical Studies in the Social Sciences (H. Solomon, ed.)*, *VI*, 158–168.

Bai, Y., P. X. K. Song, and T. E. Raghunathan (2012), Joint composite estimating functions in spatiotemporal models: joint composite estimating functions, *Journal of the Royal Statistical Society, Series B*, *74*(5), 799–824.

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

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008), Gaussian predictive process models for large spatial data sets, *Journal of the Royal Statistical Society, Series B*, *70*(4), 825–848.

Battey, H., J. Fan, H. Liu, J. Lu, and Z. Zhu (2015), Distributed estimation and inference with statistical guarantees, *arXiv preprint arXiv:1509.05457*.

Bevilacqua, M., C. Gaetan, J. Mateu, and E. Porcu (2012), Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach, *Journal of the American Statistical Association*, *107*(497), 268–280.

Bishop, Y. M., S. E. Fienberg, and P. W. Holland (1974), *Discrete multivariate analysis: theory and practice*, MIT Press, Cambridge.

Bodnar, O., T. Bodnar, and A. K. Gupta (2010), Estimation and inference for dependence in multivariate data, *Journal of multivariate analysis*, *101*(4), 869–881.

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

Caragea, P., and R. L. Smith (2007), Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models, *Journal of Multivariate Analysis*, *98*(7), 1417–1440.

Carey, V., S. L. Zeger, and P. Diggle (1993), Modelling multivariate binary data with alternating logistic regressions, *Biometrika*, *80*(3), 517–526.

Chaganty, N. R., and H. Joe (2004), Efficiency of generalized estimating equations for binary responses, *Journal of the Royal Statistical Society, Series B*, *66*(4), 851–860.

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

Chang, W., M. Haran, R. Olson, and K. Keller (2015), A composite likelihood approach to computer model calibration with high-dimensional spatial data, *Statistica Sinica*, *25*(1), 243–259.

Chen, X., and M. Xie (2014), A split-and-conquer approach for analysis of extraordinarily large data, *Statistica Sinica*, *24*(4), 1655–1684.

Cho, H., and A. Qu (2015), Efficient estimation for longitudinal data by combining large-dimensional moment conditions, *Electronic Journal of Statistics*, *9*, 1315–1334.

Claggett, B., M. Xie, and L. Tian (2014), Meta-analysis with fixed, unknown, study-specific parameters, *Journal of the American Statistical Association*, *109*(508), 1660–1671.

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

Cressie, N., and G. Johannesson (2008), Fixed rank kriging for very large spatial data sets, *Journal of the Royal Statistical Society, Series B*, *70*(1), 209–226.

Crowder, M. (1987), On linear and quadratic estimating functions, *Biometrika*, *74*(3), 591–597.

Crowder, M. (1995), On the use of a working correlation matrix in using generalised linear models for repeated measures, *Biometrika*, *82*(2), 407–410.

DerSimonian, R., and N. Laird (2015), Meta-analysis in clinical trials revisited, *Contemporary Clinical Trials*, *45*, 139–145.

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

Durbin, J. (1960), Estimation of parameters in time-series regression models, *Journal of the Royal Statistical Society, Series B*, *22*(1), 139–153.

Fan, J., F. Han, and H. Liu (2014), Challenges of big data analysis, *National Science Review*, *1*(2), 293–314.

Fisher, R. A. (1930), Inverse probability, in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 26, pp. 528–535, Cambridge University Press.

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

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

Glass, G. V. (1976), Primary, secondary, and meta-analysis of research, *Educational Researcher*, *5*(10), 3–8.

Godambe, V. P. (1960), An optimum property of regular maximum likelihood estimation, *The Annals of Mathematical Statistics*, *31*(4), 1208–1211.

Godambe, V. P. (1991), *Estimating functions*, Oxford Science Publications, Oxford.

Godambe, V. P., and C. C. Heyde (1987), Quasi-likelihood and optimal estimation, *International Statistical Review*, *55*(3), 231–244.

Hall, A. R. (2004), *Generalized method of moments*, Oxford University Press, Oxford.

Han, P., and P. X.-K. Song (2011), A note on improving quadratic inference functions using a linear shrinkage approach, *Statistics and probability letters*, *81*(3), 438–445.

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

Hansen, L. P., J. Heaton, and A. Yaron (1996), Finite-sample properties of some alternative GMM estimators, *Journal of Business and Economic Statistics*, *14*(3), 262–280.

Heagerty, P. J., and S. R. Lele (1998), A composite likelihood approach to binary spatial data, *Journal of the American Statistical Association*, *93*(443), 1099–1111.

Hector, E. C., and P. X.-K. Song (2020a), 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, 1–14.

Hector, E. C., and P. X.-K. Song (2020b), Doubly distributed supervised learning and inference with high-dimensional correlated outcomes, *Invited for minor revisions, Journal of Machine Learning Research.*

Heyde, C. C. (1997), *Quasi-likelihood and its application: a general approach to optimal parameter estimation*, Springer Series in Statistics.

Højsgaard, S., U. Halekoh, and J. Yan (2006), The R package geepack for generalized estimating equations, *Journal of Statistical Software*, *15*(2), 1–11.

Horn, R. A., and C. R. Johnson (1990), *Matrix analysis*, New York: Cambridge University Press.

Hu, Y., and P. X.-K. Song (2012), Sample size determination for quadratic inference functions in longitudinal design with dichotomous outcomes, *Statistics in Medicine*, *31*(8), 787–800.

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

Huber, P. J. (2009), *Robust statistics*, 2nd ed., Wiley Series in Probability and Statistics.

Ioannidis, J. P. (2006), Meta-analysis in public health: potentials and problems, *Italian Journal of Public Health*, *3*(2), 9–14.

Jin, Z. (2011), Aspects of composite likelihood inference, University of Toronto.

Joe, H. (1997), *Multivariate models and dependence concepts*, 1 ed., Chapman & Hall.

Joe, H. (2014), *Dependence modeling with copulas*, 1 ed., Chapman & Hall.

Johnstone, I. M., and D. M. Titterington (2009), Statistical challenges of high-dimensional data, *Philosophical transactions of the royal society A: mathematical, physical and engineering sciences*, *367*(1906), 4237–4253.

Jordan, M. I. (2013), On statistics, computation and scalability, *Bernoulli*, *19*(4), 1378–1390.

Jørgensen, B. (1987), Exponential dispersion models, *Journal of the Royal Statistical Society, Series B*, *49*(2), 127–162.

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

Khezr, S. N., and N. J. Navimipour (2017), Mapreduce and its applications, challenges and architecture: a comprehensive review and directions for future research, *Journal of grid computing*, *15*(3), 295–321.

Koenker, R., and G. Bassett, Jr. (1978), Regression quantiles, *Econometrica*, *46*(1), 33–50.

Kong, X., M.-C. Wang, and R. Gray (2015), Analysis of longitudinal multivariate outcome data from couples cohort studies: application to HPV transmission dynamics, *Journal of the American Statistical Association*, *110*(510), 472–485.

Kuk, A. Y., and D. J. Nott (2000), A pairwise likelihood approach to analyzing correlated binary data, *Statistics and Probability Letters*, *47*(4), 329–335.

Kundu, P., R. Tang, and N. Chatterjee (2019), Generalized meta-analysis for multiple regression models across studies with disparate covariate information, *Biometrika*, *106*(3), 567–585.

Laakso, M., J. Kuusisto, A. Stančáková, T. Kuulasmaa, P. Pajukanta, A. J. Lusis, F. S. Collins, K. L. Mohlke, and M. Boehnke (2017), The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases, *Journal of Lipid Research*, *58*(3), 481–493.

Laird, N. M., N. Lange, and D. Stram (1987), Maximum likelihood computations with repeated measures: applications of the EM algorithm, *Journal of the American Statistical Association*, *82*(397), 97–105.

Larribe, F., and P. Fearnhead (2011), On composite likelihoods in statistical genetics, *Statistica Sinica*, *21*(1), 43–69.

Li, C. (2017), Fusion learning of dependent studies by confidence distribution (CD): theory and applications, Rutgers University.

Liang, K.-Y., and S. L. Zeger (1986), Longitudinal data analysis using generalized linear models, *Biometrika*, *73*(1), 13–22.

Liang, K.-Y., S. L. Zeger, and B. Qaqish (1992), Multivariate regression analyses for categorical data, *Journal of the Royal Statistical Society, Series B*, *54*(1), 3–40.

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

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

Lindsay, B. G. (1988), Composite likelihood methods, *Contemporary Mathematics*, *80*, 220–239.

Lindstrom, M. J., and D. M. Bates (1988), Newton-raphson and EM algorithms for linear mixed-effects models for repeated-measures data, *Journal of the American Statistical Association*, *83*(404), 1014–1022.

Lipsitz, S. R., G. M. Fitzmaurice, L. Sleeper, and L. P. Zhao (1995), Estimation methods for the joint distribution of repeated binary observations, *Biometrics*, *51*(2), 562–570.

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

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

Mackey, L., A. Talwalkar, and M. I. Jordan (2015), Distributed matrix completion and robust factorization, *Journal of Machine Learning Research*, *16*(1), 913–960.

Masarotto, G., and C. Varin (2012), Gaussian copula marginal regression, *Electronic journal of statistics*, *6*, 1517–1549.

McLeish, D. L., and C. G. Small (1988), *The theory and applications of statistical inference functions*, Lecture Notes in Statistics, Springer, New York.

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

Newey, W. K., and D. McFadden (1994), Large sample estimation and hypothesis testing, *Handbook of Econometrics*, *4*, 2111–2245.

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

Peligrad, M. (1986), Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables (a survey), in *Dependence in probability and statistics. Progress in probability and statistics*, vol. 11, edited by E. Eberlein and M. S. Taqqu, Birkhäuser, Boston, MA.

Perry, P. O. (2017), Fast moment-based estimation for hierarchical models, *Journal of the Royal Statistical Society, Series B*, *79*(1), 267–291.

Pourahmadi, M. (1999), Joint mean-covariance models with applications to longitudinal data: unconstrained parametrisation, *Biometrika*, *86*(3), 677–690.

Qu, A., B. G. Lindsay, and B. Li (2000), Improving generalised estimating equations using quadratic inference functions, *Biometrika*, *87*(4), 823–836.

Qu, A., J. J. Lee, and B. G. Lindsay (2008), Model diagnostic tests for selecting informative correlation structure in correlated data, *Biometrika*, *95*(4), 891–905.

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

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

Smith, T. C., D. J. Spiegelhalter, and A. Thomas (1995), Bayesian approaches to randomeffects metaanalysis: A comparative study, *Statistics in Medicine*, *14*(24), 2685–2699.

Song, P. X.-K. (2000), Multivariate dispersion models generated from gaussian copula, *Scandinavian Journal of Statistics*, *27*(2), 305–320.

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

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

Tang, L., and P. X.-K. Song (2016), Fused lasso approach in regression coefficients clustering – learning parameter heterogeneity in data integration, *Journal of Machine Learning Research*, *17*(113), 1–23.

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

Varin, C. (2008), On composite marginal likelihoods, *AStA Advances in Statistical Analysis*, *92*(1), 1–28.

Varin, C., N. Reid, and D. Firth (2011), An overview of composite likelihood methods, *Statistica Sinica*, *21*(1), 5–42.

Wang, F., L. Wang, and P. X.-K. Song (2012), Quadratic inference function approach to merging longitudinal studies: validation and joint estimation, *Biometrika*, *99*(3), 755–762.

Wang, F., L. Wang, and P. X.-K. Song (2016), Fused lasso with the adaptation of parameter ordering in combining multiple studies with repeated measurements, *Biometrics*, *72*(4), 1184–1193.

Wang, J., X. He, and G. Xu (2020), Debiased inference on treatment effect in a high-dimensional model, *Journal of the American Statistical Association*, *115*(529), 442–454.

Wang, Y.-G., and V. Carey (2003), Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance, *Biometrika*, *90*(1), 29–41.

Wang, Y.-G., X. Lin, and M. Zhu (2005), Robust estimating functions and bias correction for longitudinal data analysis, *Biometrics*, *61*(3), 684–691.

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

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

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

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

Yang, G., D. Liu, R. Y. Liu, M. Xie, and D. C. Hoaglin (2014a), Efficient network meta-analysis: a confidence distribution approach, *Statistical Methodology*, *20*, 105–125.

Yang, G., D. Liu, R. Y. Liu, M. Xie, and D. C. Hoaglin (2014b), Efficient network meta-analysis: a confidence distribution approach, *Statistical Methodology*, *20*, 105–125.

Yang, Y., and X. He (2015), Quantile regression for spatially correlated data: an empirical likelihood approach, *Statistica Sinica*, *25*(509), 261–274.

Zaharia, M., M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica (2010), Spark: cluster computing with working sets, in *Proceedings of the 2Nd USENIX conference on hot topics in cloud computing*, HotCloud'10 USENIX Association, Berkeley, USA.

Zellner, A. (1962), An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *Journal of the American Statistical Association*, *57*(298), 348–368.

Zhang, W., C. Leng, and C. Y. Tang (2015a), A joint modelling approach for longitudinal studies, *Journal of the Royal Statistical Society, Series B*, *77*(1), 219–238.

Zhang, Y., J. Duchi, and M. Wainwright (2015b), Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates, *Journal of Machine Learning Research*, *16*, 3299–3340.

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

Zhou, Y., and P. X.-K. Song (2016), Regression analysis of networked data, *Biometrika*, *103*(2), 287–301.