Copula Regression Models for the Analysis of Correlated Data with Missing Values

by

Wei Ding

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2015

Doctoral Committee:

Professor Peter X-K Song, Chair
Assistant Professor Veronica Berrocal
Assistant Professor Hui Jiang
Associate Professor Qiaozhu Mei

To my parents

ACKNOWLEDGEMENTS

It was the best of time. It was the worst of time. For Charles Dickinson, it was French Revolution, but for me, it was my life in the past five years as a phd student in Ann Arbor. When I first came here, I was with curiosity, with fear, with doubt, and with uncertainty. I wish I can leave here with confidence, with independence, and with hope, to a bright future.

First and foremost, I am deeply grateful to my wonderful advisor, Dr. Peter Song, who encouraged me, inspired me, and helped me with great tolerance and patience. Without him, this dissertation is never made possible. I was moved by his rigorous thinking, broad vision, and his passion to statistics. Moreover, Dr. Peter Song is not just an advisor, but also a mentor, and a role model. I learned a lot from his thoughtfulness and kindness.

My thanks go out to Dr. Veronica Berrocal, Dr. Hui Jiang, and Dr. Qiaozhu Mei for serving on my dissertation committee and providing valuable suggestions and support on my research.

I am thankful for all my best friends in Ann Arbor and elsewhere around the world. With them, I had a great time, and never felt lonely.

The most important of all, I must thank my parents, who love me unconditionally, support me no matter what, and always have absolute faith in me. This thesis is dedicated to my beloved parents.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Copula Regression Models for the Analysis of Correlated Data with Missing Values

by
Wei Ding

Advisor: Peter X-K Song

The class of Gaussian copula regression models provides a unified modeling framework to accommodate various marginal distributions and flexible dependence structures. In the presence of missing data, the Expectation-Maximization (EM) algorithm plays a central role in parameter estimation. This classical method is greatly challenged by multilevel correlation, a large dimension of model parameters, and a misaligned missing data mechanism encountered in the analysis of data from collaborative projects. This dissertation will develop a series of new methodologies to enhance the effectiveness of the EM algorithm in dealing with complex correlated data analysis via a combination of new concepts, estimation approaches, and computing procedures. The dissertation consists of three major projects given as follows.

The focus of Project 1 is on the development of an effective EM algorithm in Gaussian copula regression models with missing values, in which univariate location-scale family distributions are utilized for marginal regression models and Gaussian copula for dependence. The proposed class of regression models includes the classical multivariate normal model as a special case and allows both Pearson correlation and rank-based correlations (e.g. Kendall's tau and Spearman's rho). To improve the

implementation of the EM algorithm, following Meng and Rubin (1993), we establish an effective peeling procedure in the M-step to sequentially maximize the observed log-likelihood with respect to regression parameters and dependence parameters. In addition, the Louis formula is provided for the calculation of the Fisher information. The EM algorithm is tailored for misaligned missing data mechanism under structured correlation structures (e.g. exchangeable and first-order autoregression). We run simulation studies to evaluate the proposed model and algorithm, and to compare with both model-based multiple imputation and hot-deck imputation methods.

Project 2 is devoted to a critical extension of Project 1, where the assumption of structured correlation structure is relaxed, so the resulting model and algorithm can be applied to deal with complex correlated data with missing values. The key new contribution in the extension concerns the development of EM algorithm for composite likelihood estimation in the presence of misaligned missing data. We propose the complete-case composite likelihood, which is more general than the classical pairwise composite likelihood, to handle both point-identifiable and partially identifiable parameters in the Gaussian copula regression model. Estimation of a partially identifiable correlation parameter is given by an estimated interval. Both estimation properties and algorithmic convergences are discussed. The proposed method is evaluated and illustrated by simulation studies and a quality-of-life data set.

Motivated by an electroencephalography (EEG) data collected from 128 electrodes on the scalps of 9 months old infants, Project 3 concerns the regression analysis of multilevel correlated data. Indeed multilevel correlated data are pervasive in practice, which is routinely modeled by the hierarchical modeling system using random effects. We develop an alternative class of parametric regression models using Gaussian copulas and implement the maximum likelihood estimation. The proposed model is very

flexible; in the aspect of regression model, it can accommodate continuous outcomes, discrete outcomes or outcomes of mixed types; and in the aspect of dependence, it can allow temporal (e.g. AR), spatial (e.g. Matern), clustered (e.g. exchangeable), or combined dependence structures. Parameters in the proposed model have marginal interpretation, which is absent in the hierarchical model when outcomes of interest are non-normal (e.g. binary or ordinal categorical). Moreover, it allows the presence of missing data. The proposed EM algorithm with peeling procedure provides a fast and stable parameter estimation algorithm. The proposed model and algorithm are assessed by simulation studies, and further illustrated by the analysis of EEG data for the adverse effect of iron deficiency on infants' visual recognition memory.

CHAPTER I

Introduction

1.1 Summary

The class of Gaussian copula regression models provides a unified modeling framework to accommodate various marginal distributions and flexible dependence structures. In the presence of missing data, the Expectation-Maximization (EM) algorithm plays a central role in parameter estimation. This seminal method is greatly challenged by complex data structures, such as multilevel correlation, large dimension of model parameters, and a misaligned missing data pattern that we have encountered in our collaborative projects at University of Michigan. This dissertation aims to develop a set of new statistical methodologies and algorithms to enhance the applications of the EM algorithm to deal with complex correlated data analysis. Based on new concepts, estimation approaches, and computing procedures as well as their combinations, we hope to yield more flexible and effective analytic tools to analyze complex correlated data. The dissertation consists of three major projects described as follows.

Project 1 focus on the development of an effective EM algorithm in Gaussian copula regression models with missing values, in which univariate location-scale family distributions are utilized for marginal regression models and Gaussian copula for dependence. The proposed class of regression models includes the classical multivariate normal model as a special case and allows both Pearson correlation and rank-based correlations (e.g. Kendall's tau and Spearman's rho). To improve the implementation of the EM algorithm, following Meng and Rubin (1993), we establish an effective peeling procedure in the M-step to sequentially maximize the observed log-likelihood with respect to regression parameters and dependence parameters. In addition, Louis' formula is provided for the calculation of the Fisher information. The EM algorithm is particularly tailored for the so-called misaligned missing data mechanism under structured correlation structures (e.g. exchangeable and first-order autoregression). We run extensive simulation studies to evaluate the proposed model and algorithm, and to compare our method with both model-based multiple imputation and hot-deck imputation methods.

Project 2 is devoted to a critical extension of Project 1, where the assumption of structured correlation structure is relaxed, so the resulting model and algorithm can be applied to deal with complex correlated data with missing values. The key new contribution in the extension concerns the development of the peeling algorithm for composite likelihood estimation in the presence of misaligned missing data pattern. We propose the complete-case composite likelihood for estimation, which is more general than the classical pairwise composite likelihood. The proposed method is intended to handle both point-identifiable and partially identifiable parameters in the Gaussian copula regression model. Estimation of a partially identifiable correlation parameter is given by an estimated interval. Both estimation properties and algorithmic convergences are discussed. The proposed method is evaluated and illustrated by simulation studies and a quality-of-life data set.

Motivated by an electroencephalography (EEG) data collected from 128 electrodes on the scalps of 9 months old infants, Project 3 concerns the regression analysis of multilevel correlated data. Arguably, multilevel correlated data are pervasive in practice, which is routinely modeled by the hierarchical modeling system using random effects. We develop an alternative class of parametric regression models using Gaussian copulas and implement the maximum likelihood estimation. The proposed model is very flexible; in the aspect of regression model, it can accommodate continuous outcomes, discrete outcomes or outcomes of mixed types; and in the aspect of dependence, it can allow temporal (e.g. AR), spatial (e.g. Matérn), clustered (e.g. exchangeable), or a mixture of dependence structures. Parameters in the proposed model have marginal interpretation, which is absent in the hierarchical model when outcomes of interest are non-normal (e.g. binary or ordinal categorical). Moreover, it allows the presence of missing data. The proposed EM algorithm with peeling procedure provides a fast and stable iterative procedure for parameter estimation algorithm. The proposed model and algorithm are assessed by simulation studies, and further illustrated by the analysis of EEG data for the adverse effect of iron deficiency on infant's visual recognition memory.

1.2 Objectives

The Objective of Chapter II is to develop the Gaussian copula regression model (Song (2000); Song et al. (2009a)) to analyze correlated data with missing values. The proposed class of multidimensional regression models for correlated data have various meritorious features that have led to its popularity in practical studies. First, the copula regression model allows to define, evaluate and interpret correlations between variables in a full probability manner, in a very similar way to that of the classical multivariate normal distribution which has been extensively studied in the statistical literature and widely applied in the analysis of multivariate data. Second, from the copula regression model various types of correlations are furnished to address different questions related to a joint regression analysis. For example, depending on if the marginal distributions are normal or skewed, it provides Pearson linear correlation or rank-based nonlinear correlations (e.g., Kendall's tau or Spearman's rho). Moreover, these correlations types may be represented either in a form of unconditional pairwise correlation, or in a form of conditional pairwise correlation. Third, the copula regression model has the flexibility to incorporate marginal location-scale families to adjust for confounding factors, which is of practical importance. Last, the availability of the full joint probability model gives rise to the great ease of implementing powerful EM algorithm to handle missing data in a broad range of multi-dimensional models where the regression parameters in the mean model and the correlation parameters can be estimated simultaneously under one objective function. In such a framework, both estimation and inference are safeguarded by the well-established classical maximum likelihood theory.

Largely motivated by a collaborative project concerns a quality of life study on children with nephrotic syndrome, the objective of Chapter III centers on a further extension of the Gaussian copula regression analysis methodology proposed in Chapter II by addressing two challenging problems. One is the difficulty of estimating correlation parameters when misaligned missing pattern occurs between variables. By misaligned missingness we mean a missing data pattern in which two variables are measured in disjoint subsets of subjects and have no overlapped observations. The other is the issue of parameter identifiability, which is a serious consequence from misaligned missing data pattern encountered in the estimation of unstructured correlation matrix. Note that estimating correlation matrix is indeed required in a joint regression analysis of multiple correlated outcomes. We propose a complete-case composite likelihood method to perform estimation and inference for the model parameters, in which the above two major methodological challenges are handled via a composition of are marginal distributions of observed variables. Also, the correlation parameters that are not point-identified are estimated by both lower and upper bounds that form interval estimation for the partially identifiable parameters. For implementation, the effective peeling optimization procedure is modified for the composite likelihood to estimate point-identifiable parameters. We investigate the performance of complete case composite likelihood method, and compare it with the maximum likelihood estimation given in Chapter II through simulation studies.

The objective of Chapter IV focuses on the development of Gaussian regression models for multilevel correlated data. This work is motivated by a collaborative project that aims to assess the adverse effect of prenatal iron deficiency on infant's visual recognition memory. In this study, memory is measured by electroencephalography (EEG) sensor net of 128 electrodes, from which event-related potential (ERP) such as low slow wave is extracted to quantify the capacity of memory. A major technical challenge arises from a multilevel dependence structure, including temporal, spatial and clustered correlations. When an ERP outcome is skewed, multilevel rank-based correlations are appealing, which are naturally supplied by the Gaussian copula model. Thus, in this project we extend the framework of Gaussian copula regression models by accommodating multiple types of correlations. This flexibility of dependence modeling allows us to analyze complex data structures in the regression analysis and to provide more comprehensive results than those obtained by a subset of data with one-level correlation. This extension of copula model to multilevel correlation is established by the utility of Kronecker product of correlation matrices. We also extend the peeling procedure to carry out estimation of the model parameters, which is particularly useful to deal with potentially a large number of correlation parameters. Both simulation studies and data analysis examples will be provided to illustrate the proposed methodology. In the presence of missing data, the EM algorithm with the peeling procedure is used in implementation.

1.3 Literature Review

The amount of the literature related to my dissertation research topics is so vast that it is not possible to review all major articles in this chapter. Instead, below I attempt to provide my review based on the set of references that I have actually read in a reasonable detail.

1.3.1 Correlated Data

Multi-dimensional regression models for correlated data involve typically the specification of both correlation structures and marginal mean models that can be formulated by the classical univariate generalized linear model (GLM) (Nelder and Baker (1972)). Although the great popularity of quasi-likelihood approaches to analyzing correlated data, such as generalized estimating equation (GEE) (Liang and Zeger (1986)) and quadratic inference function (QIF) (Qu et al. (2000)), a fully specified probability model with interpretable correlation structures is actually a desirable formulation to address the need of evaluating correlations between variables. It is known that in the quasi-likelihood method correlations are treated as nuisance parameters, so that their estimation and interpretation are not of primary interest in data analysis. This treatment may not always be desirable and can be improved by some will-behaved dependence models such as copula models (Joe (1997)).

1.3.2 Missing Data

Missing data is an important issue in statistics, and can deliver a significant influence on the conclusions. There are many reasons for the occurrence of missing data: no information is collected for some subjects, certain subjects are unwilling to provide sensitive or private information, subject's dropout due to moving, or researcher cannot collect the whole data due to time or budgetary limitations.Three mechanisms of missing data (Rubin, 1976, Rubin (1976)) are commonly considered in the data analysis, including (i) missing completely at random (MCAR) when the missing mechanism is independent from both observed and missing data; (ii) missing at random (MAR) when the missing mechanism is not related to missing data; and (iii) missing not at random (MNAR) when the missing mechanism depends on missing.

In terms of handling missing data, the complete case analysis, which is often used in practice for convenience, simply discards any cases with missing values on those of the variables selected and proceeds with the analysis using standard methods. Obviously, the data attrition reduces the sample size, resulting potentially in a great loss of estimation efficiency. EM algorithm (Dempster et al. (1977)) is a widely used iterative algorithm to carry out the maximum likelihood estimation in a statistical analysis with incomplete data. Multiple Imputation (Rubin (2004)) provides an alternative approach useful to deal with statistical analysis with missing values. Instead of filling in a single value for each missing value, (Rubin (2004)) multiple imputation procedure actually replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. When data come from skewed distributions, hot-deck Imputation (Andridge and Little (2010)) is also widely used, where a missing value is imputed with a randomly drawn similar record in terms of the nearest neighbors criterion. One caveat of hot-deck imputation is that it is a single imputation method, which may fail to provide desirable uncertainty associated with missing values. In addition, the number of imputed data sets is critical to obtain proper data analysis results, and a small number may lead to inappropriate inference. Some researchers have recommended 20 to 100 imputation data sets or even more (Graham et al. (2007)), which appears computationally costly in practice. The imputation methods may become nontrivial and no longer straightforward when data distributions are skewed and adjusting for confounding factors is needed.

1.3.3 Copula

Copula is a joint multivariate probability distribution of random variables, and the marginal probability distribution of each variable is uniform, and is used to model the dependence between random variables. Sklar's Theorem (Sklar (1959)) states that for a multivariate joint distribution, there exists a suitable copula that not only links the univariate marginal distribution functions, but also captures the dependence. The representation of a copula model separates the marginal models and the dependence model.

Most of recently published works on the copula regression models have been focused on analyzing fully observed data; for example, Song (2007); Czado (2010); Joe et al. (2010); Genest et al. (2011); Masarotto et al. (2012); Acar et al. (2012). There is little knowledge available concerning how the analysis may be done in the presence of missing data.

Gaussian copula is a generated model from multivariate normal distribution by inverse normal transformation, where the correlation matrix under Gaussian copula is the Pearson correlation matrix of the normal distributed quantiles. Gaussian copula regression model (Song (2000); Song et al. (2009a)) is an useful probability model for the correlated data because of the following meritorious features. First, the Gaussian copula regression model allows us to define, evaluate and interpret correlations between variables in a full probability manner, and the classical multivariate normal distribution is a special the Gaussian copula regression model. Second, from the copula model various types of correlations are provided to answer for different questions. For example, it provides Pearson linear correlation or rank-based nonlinear correlations (Kendall's tau or Spearman's rho), depending on if the marginal distributions are normal or skewed. Moreover these correlations may be obtained either in a form of unconditional marginal pairwise correlation, or in a form of conditional pairwise correlation. Third, the copula model has the flexibility to incorporate marginal GLMs to adjust for confounding factors, which is of practical importance. Last, the availability of the full probability model gives rise to the great ease of implementing powerful EM algorithm to handle missing data in a broad range of multi-dimensional models where the regression parameters in the mean model and the correlation parameters can be estimated simultaneously under one objective function. In such a framework, both estimation and inference are safeguarded by the well-established classical maximum likelihood theory.

1.3.4 EM Algorithm

The EM algorithm proposed by Dempster et al. (1977) is widely used to find the maximum likelihood estimators of a statistical model in cases where the equations cannot be solved directly, or with the presence of missing data. It contains two iterative steps. Expectation step (E-step) calculates the expectation of the observed log likelihood function, based on the conditional distribution of missing data given observed data under estimate of the parameters of current iteration, and maximization step (M-step) finds the parameter that maximize the observed log likelihood function.

In the expectation conditional maximization (ECM) algorithm proposed by Meng and Rubin (1993), each M-step is replaced with a sequence of conditional maximization steps (CM-steps) where one or a group of parameters are maximized sequentially, conditionally on the other parameters being fixed.

1.3.5 Composite Likelihood Method

Composite likelihood (Lindsay (1988)) has received increasing attention in the recent statistical literature. It is also known as a pseudo likelihood (Molenberghs and Verbeke (2005)) in longitudinal data setting, or an approximate likelihood (Stein et al. (2004)) in spatial data setting, or a quaisi-likelihood (Hjort et al. (1994); Glasbey (2001); Hjort and Varin (2008)) in spatial and time series data settings.

As composite likelihood may be treated as a special class of inference functions, statistical inference can be established by an application of the standard theory of inference functions (Chapter 3, Song (2007)). For example, Godambe information matrix (Godambe (1960)) is typically used to obtain the asymptotic variance of a composite likelihood estimator, and in the presence of missing data, Godambe information matrix is calculated according to an empirical procedure suggested by Gao and Song (2011).

1.3.6 Partial Identification

For the case of completely misaligned missingness considered in this thesis, for the unstructured correlation matrix, some of correlation parameters may not be fully identifiable. Manski (2003) proposed several approaches to address such a partial identification problem in parameter estimation. A parameter is said to be partially identifiable if the true parameter is not point-identifiable but a range of parameter values containing the true value is identifiable. Fan and Zhu (2009) provided a method to determine the lower and upper bounds of the parameter range in the setting of bivariate copula models, where the pairwise correlation parameter is partially identified by an estimated parameter range. However, Fan and Zhu (2009)'s method does not work for a general d-dimensional copula model, and it is not clear at this moment how easily an extension of their method may be accomplished analytically. This thesis aims to develop a new composite likelihood method to overcome this estimation difficulty.

1.3.7 Multilevel Model

Multilevel data, also known as hierarchical data, clustered data, and nested data, are a common type of data structure in spatio-temporal analysis, or when subjects are grouped by some specific clusters. For example, Aitkin and Longford (1986) designed a two-level model for educational data, in which students are clustered in schools. Random effects model, also known as variance components model, is one of the most popular methods to estimate parameters in multilevel models.

Random effects model was introduced by Laird and Ware (1982), where both "fixed" and "random" effects are respectively referred to as the population-average and subject-specific effects. Related theories and applications of random effects models in data analysis may be found in Verbeke et al. (2010); Liang and Zeger (1986); Zeger et al. (1988), and Zeger and Liang (1986), among others.

1.3.8 Motivating data I: Quality of Life Study

Nephrotic Syndrome (NS) is a common disease in pediatric patients with kidney disease. The typical symptom of this disease is characterized by the presence of edema that significantly affects the health-related quality of life in children and adolescents. The PROMIS (Fries et al. (2005); Gipson et al. (2013)) is a well-validated instrument to assess pediatric patient's quality of life. The instrument consists of 7 domains, including pain interference, fatigue, depression, anxiety, mobility, social peer relationship, and upper extremity functioning. In the data, two QoL scores, pain and fatigue, are measured on two exclusive sets of subjects due to some logistic difficulty at the clinic. That is, out of 224 subjects, 107 subjects have QoL measurements of pain, but no QoL measurements of fatigue, while the other 117 subjects have QoL measurements of fatigue but no QoL measurements of pain. Interestingly, QoL measurements of anxiety have been fully recorded on all 224 individuals with no missing data.

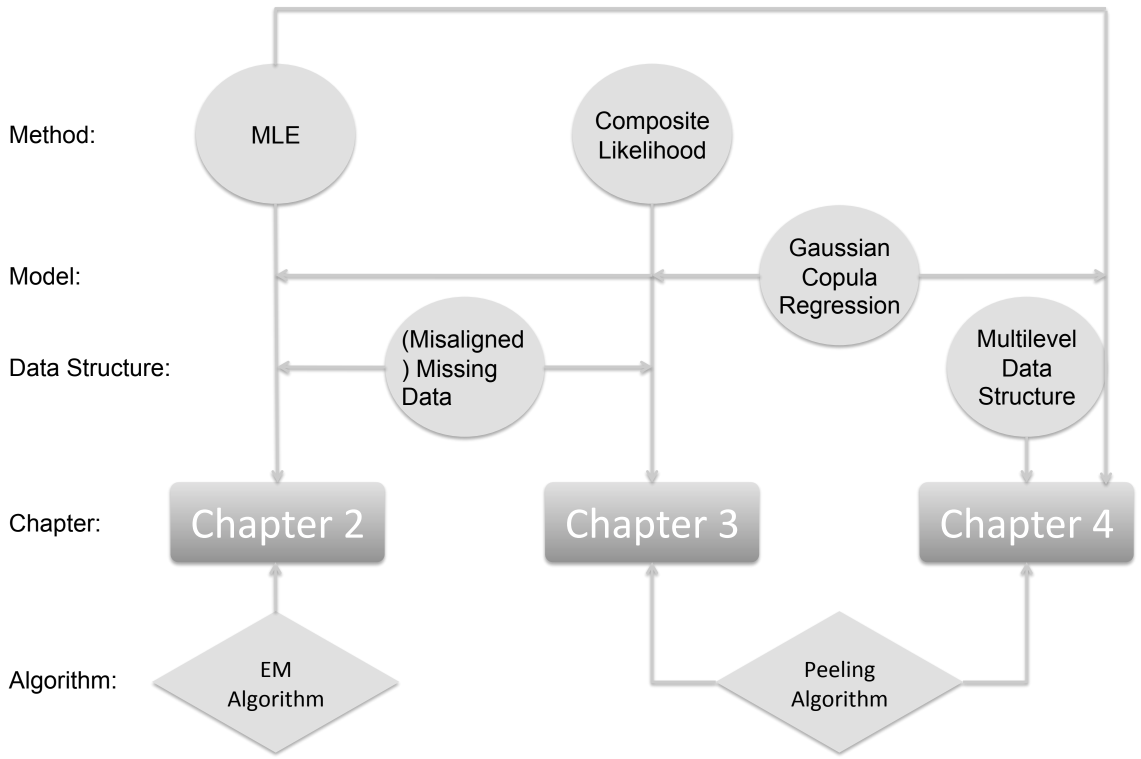
1.3.9 Motivating data II: Infants' Visual Recognition Memory Study

Infants' visual recognition memory study aims to evaluate whether or not, and if so, how, iron deficiency affects visual recognition memory for infants. We refer to some of important related work that has been summarized in de Haan et al. (2003). Infants' memory capability is measured by the activity of the brain during a period of 1700 milliseconds using electroencephalograph (EEG) net with 128-channel sensors on the scalp (Reynolds et al. (2011)). The data collection occurs at two time points: when an infant sees his/her mother's picture and when he or she sees a stranger's picture. At each time point, an event-related potential (ERP) of interest, late slow wave (LSW), is extracted from after the standard data processing, which is widely used as primary outcomes of visual recognition memory. In total, there are 91 children in this study, with fully observed data. 20 out of 128 electrodes are of interest with 5 in each of the four subregions.

1.4 Outline of Dissertation

This dissertation is organized as follows: In Chapter I, we give an overview about my dissertation and an introduction to related works. In Chapter II, we develop a peeling algorithm in Gaussian copula regression model and provide a solution to the misaligned missing data pattern. In Chapter III, we present a complete-case composite likelihood method as an alternative solution to the analysis of missing data with the misaligned missing pattern, which Gaussian copula regression model is used as an example to illustrate this approach. In Chapter IV, a multilevel Gaussian copula regression model is developed with peeling algorithm. In concluding some discussions and future work are presented in Chapter V. The connection and structure between Chapter II, III, are IV are displayed in Figure 1.1.

Figure 1.1: Connection and Structure of the Dissertation



CHAPTER II

EM Algorithm in Gaussian Copula with Missing Data

2.1 Summary

Rank-based correlation is widely used to measure dependence between variables when their marginal distributions are skewed. Estimation of such correlation is challenged by both the presence of missing data and the need for adjusting for confounding factors. In this paper, we consider a unified framework of Gaussian copula regression that enables us to estimate either Pearson correlation or rank-based correlation (e.g. Kendall's tau or Spearman's rho), depending on the types of marginal distributions. To adjust for confounding covariates, we utilize marginal regression models with univariate location-scale family distributions. We establish the EM algorithm for estimation of both correlation and regression parameters with missing values. For implementation, we propose an effective peeling procedure to carry out iterations required by the EM algorithm. We compare the performance of the EM algorithm method to the traditional multiple imputation approach through simulation studies. For structured types of correlations, such as exchangeable or first-order auto-regressive (AR-1) correlation, the EM algorithm outperforms the multiple imputation approach in terms of both estimation bias and efficiency.

2.2 Introduction

Estimation of rank-based correlation is frequently required in practice to evaluate relationships between variables when they follow marginally skewed distributions. However, estimation of such correlation becomes a great challenge in the presence of missing data and with the need of adjusting for confounders. Most of recently published works on the copula models have been focused on analyzing fully observed data, e.g., Czado (2010); Joe et al. (2010); Genest et al. (2011); Masarotto et al. (2012); Acar et al. (2012), and there is little knowledge available concerning how the analysis may be done in the presence of missing data.

In terms of handling missing data, the complete case analysis, which is often used in practice for convenience, simply discards any cases with missing values on those of the variables selected and proceeds with the analysis using standard methods. Obviously, the data attrition reduces the sample size, resulting potentially in a great loss of estimation efficiency. EM algorithm (Dempster et al. (1977)) is a widely used iterative algorithm to carry out the maximum likelihood estimation in a statistical analysis with incomplete data. Multiple Imputation (Rubin (2004)) provides an alternative approach useful to deal with statistical analysis with missing values. Instead of filling in a single value for each missing value, (Rubin (2004)) multiple imputation procedure actually replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. When data come from skewed distributions, Hot-Deck Imputation (Andridge and Little (2010)) is also widely used, where a missing value is imputed with a randomly drawn similar record in terms of the nearest neighbor criterion. One caveat of Hot-Deck imputation is that it is a single imputation method, which may fail to provide desirable uncertainty associated with missing values. In addition, the number of imputed data sets is critical to obtain proper data analysis results, and a small number may lead to inappropriate inference. Some researchers have recommended 20 to 100 imputation data sets or even more (Graham et al. (2007)), which appears computationally costly in practice. The imputation methods may become nontrivial and no longer straightforward when data distributions are skewed and adjusting for confounding factors is needed.

Multi-dimensional regression models for correlated data involve typically the specification of both correlation structures and marginal mean models that can be formulated by the classical univariate generalized linear model (GLM) (Nelder and Baker (1972)). Although the great popularity of quasi-likelihood approaches to analyzing correlated data, such as generalized estimating equation (GEE) (Liang and Zeger (1986)) and quadratic inference function (QIF) (Qu et al. (2000)), a fully specified probability model with interpretable correlation structures is actually a desirable device to achieve the objective of evaluating correlations between variables. It is known that in the quasi-likelihood method correlations are treated as nuisance parameters, so that their estimation and interpretation are not of primary interest in data analysis.

In this paper we consider the Gaussian copula regression model (Song (2000); Song et al. (2009a)) as the probability model for the correlated data because of the following meritorious features. First, the copula model allows us to define, evaluate and interpret correlations between variables in a full probability manner, very similar to the classical multivariate normal distribution. Second, from the copula model various types of correlations are provided to answer for different questions. For example, it provides Pearson linear correlation or rank-based nonlinear correlations (Kendall's tau or Spearman's rho), depending on if the marginal distributions are normal or skewed. Moreover these correlations may be obtained either in a form of unconditional marginal pairwise correlation, or in a form of conditional pairwise correlation. Third, the copula model has the flexibility to incorporate marginal GLMs to adjust for confounding factors, which is of practical importance. Last, the availability of the full probability model gives rise to the great ease of implementing powerful EM algorithm to handle missing data in a broad range of multi-dimensional models where the regression parameters in the mean model and the correlation parameters can be estimated simultaneously under one objective function. In such a framework, both estimation and inference are safeguarded by the well-established classical maximum likelihood theory.

It is of interest in the context of copula models to investigate and compare the two principled methods of handling missing data, EM algorithm and multiple imputation, as well as their computational complexity. Since the development of the EM algorithm is not trivial in the framework of Gaussian copula models, we propose an efficient peeling procedure to update model parameters in the M-step due to the involvement of a multi-dimensional integral. To adjust for confounding factors in the marginals, we focus on the location-scale family distribution in marginal regression models to embrace the flexibility of marginal distributions.

We compare the performance of the EM algorithm to the multiple imputation approach through simulation studies. For structured types of correlations, such as exchangeable or first-order auto-regressive (AR-1) correlation matrix, the EM algorithm method outperforms the multiple imputation approach in both aspects of estimation bias and efficiency. These two approaches perform similarly when the correlation matrix is unstructured.

This paper is organized as follows. Section 2.3 describes the Gaussian copula model. Together with some examples of practically useful models, Section 2.4 presents the details of the EM algorithm and Louis' formula (Louis (1982)) for standard error calculation. Section 2.5 presents simulation study, and a data analysis is included in Section 2.6. Section 2.7 provides some concluding remarks.

2.3 Model

The focus of this paper is on using EM algorithm in Gaussian copula to estimate of correlation with missing data. We assume that there are n partially observed subjects. For a subject, let $Y = (y_1, y_2, \cdots, y_d)'$ be a d-dimensional random vector of continuous outcomes, part of which is observed and the other part is missing. Denote by R_j as a missing data indicator, where $R_j = 0$ or 1 if the j^{th} element y_j is missing or observed. Note that this indicator is known but varies for different subjects. Let y_{mis} be the set of variables with missing data, and y_{obs} be the set of variables with observed data of a subject.

2.3.1 Location-Scale Family Distribution Marginal Model

Suppose $\theta = (\theta_1, \theta_2, \cdots, \theta_d)'$, where each θ_j denotes a set of marginal parameters associated with the $j^{th}(j = 1, \cdots, d)$ marginal density function, $f_j(y_j|\theta_j)$. Denote by $u_j = F_j(y_j|\theta_j)$ the marginal cumulative distribution function(CDF) corresponding to the j^{th} margin, where F_j is a location-scale family distribution parametrized by a location parameter μ_j and a positive scale parameter σ_j, $\theta_j = (\mu_j, \sigma_j)$. More specifically, the marginal location-scale density function is given by

$$f_j(y_j|\theta_j) = \frac{1}{\sigma_j}\tilde{f}\left(\frac{y_j - \mu_j}{\sigma_j}\right), j = 1, \cdots, d, \tag{2.1}$$

where $\tilde{f}(\cdot)$ is the standard kernel density with $\int_R y\tilde{f}(y)dy = 0$, and $\int_R y^2\tilde{f}(y)dy = 1$. In this paper, $\tilde{f}$ may be taken as a parametric or a nonparametric kernel density, and parameter μ_j or σ_j may be modelled as a function of confounding covariates.

2.3.2 Gaussian Copula

A copula is a multivariate probability distribution in which the marginal probability distribution of each variable is uniform on $(0,1)$. Sklar's theorem (Sklar (1959)) states that every multivariate cumulative distribution function of a continuous random vector $Y = (y_1, y_2, \cdots, y_d)'$ with marginals $F_j(y_j|\theta_j)$ can be written as $F(y_1, \ldots, y_d) = C\left(F_1(y_1), \ldots, F_d(y_d)\right)$, where C is a certain copula. In this paper, Y is assumed to follow a d-dimensional distribution generated by a Gaussian copula (Song (2000)), whose density function is given by

$$f(Y|\theta, \Gamma) = c(u|\Gamma) \prod_{j=1}^{d} f_j(y_j|\theta_j), u = (u_1, u_2, \cdots, u_d)' \in [0,1]^d, \tag{2.2}$$

where $c(u|\Gamma) = c(u_1, \cdots, u_d|\Gamma), u \in [0,1]^d$, is the Gaussian copula density, with $u_j = F_j(y_j|\theta_j)$, $i = 1, \cdots, d$, and Γ is an $d \times d$ matrix of correlation.

Let $q_j = q_j(u_j) = \Phi^{-1}(u_j)$ be the j^{th} marginal normal quantile, where Φ is CDF of the standard normal distribution. According to Song (2007), the joint density of a Gaussian copula function $c(\cdot|\Gamma)$ takes the form:

$$c(u|\Gamma) = |\Gamma|^{-\frac{1}{2}} \exp\left\{\frac{1}{2} Q(u)^T (\mathrm{I} - \Gamma^{-1}) Q(u)\right\}, u \in [0,1]^d \tag{2.3}$$

where $\Gamma = [\gamma_{j_1 j_2}]_{d \times d}$ is the Pearson correlation matrix of $Q(u) = (q_1(u_1), \cdots, q_d(u_d))'$, and I is the $d \times d$ identity matrix. Here $|\cdot|$ denotes the determinant of a matrix. Marginally, $u_j \sim$ Uniform$(0,1)$, and $q_j \sim$ Normal$(0,1)$. When y_j is marginally normal distributed, matrix Γ gives the Pearson correlation matrix of Y; otherwise, Γ represents as a matrix of pairwise rank-based correlations. In fact, given a matrix Γ in equation (4.3), two types of pairwise rank-based correlations, Kendall's tau ($[\tau_{j_1 j_2}]_{d \times d}$) and Spearman's rho ($[\rho_{j_1 j_2}]_{d \times d}$) can be obtained as follows: $\tau_{j_1 j_2} = \dfrac{2}{\pi} \arcsin(\gamma_{j_1 j_2})$, and $\rho_{j_1 j_2} = \frac{6}{\pi}\arcsin(\frac{\gamma_{j_1 j_2}}{2})$ for $j_1, j_2 = 1, \cdots, d, j_1 \neq j_2$, respectively (McNeil et al. (2010)).

2.3.3 Examples of Marginal Models

Among many possible marginal models, here we present two examples of marginal models to illustrate our proposed method, with or without the inclusion of covariates. These two following models are practically useful.

Example-1: Marginal Parametric Distribution

To adjust for confounding factors in the mean marginal model, let $X_i = (1, x_i^T)^T$, $i = 1, \cdots, n$. For the j^{th} margin, the linear model is imposed on the location parameter in equation (3.11), $\mu_{ij} = \mathrm{E}(y_{ij}|X_i) = h(X_i^T \beta_j), j = 1, \cdots, d$, where $\beta_j = (\beta_{j0}, \beta_{j1}, \cdots, \beta_{jp})'$ is a $(p+1)$-element unknown regression vector, and h is a link function. For convenience, denote the resulting model by $Y_{ij} \sim F_j(y_j|\mu_{ij}(\beta_j), \sigma_j)$.

As an important special case, we consider $p = 0$ (no covariates), and thus $\mu_{ij} = h(\beta_{j0})$ is a common parameter for all subjects $i = 1, \cdots, n$. More generally, the marginal distribution model with the CDF $u_{ij} = F_j(y_j|\theta_j)$ may be a generalized location-scale family distribution, such as gamma distribution, of which the location parameter is 0, and the estimation procedure remains the same under a given marginal parametric distribution. This will be discussed as an example in simulation study in Section 2.5.1.

Example-2: Semi-parametric Marginal Distribution

If the type of the density function $f_j(y_j), j = 1, \cdots, d$ is unknown, there are several possible forms available to specify equation (3.11). In this paper, we consider an example of fully unspecified marginal distribution function $F_j(y_j)$, which will be estimated using the empirical distribution function. In this case, all the marginal parameter θ_j is absorbed into the CDF.

2.4 EM Algorithm

Our goal is to estimate the model parameter (θ, Γ) in the presence of missing data. This may be achieved by utilizing the EM algorithm. We propose an effective peeling procedure in the EM algorithm, which serves as a core engine to speed up the calculation of M-step in the copula model. Both E-step and M-step are discussed in detail in Section 2.4.1, and the examples will be revisited in Section 2.4.2, respectively.

2.4.1 Expectation and Maximization

Computing the likelihood of (θ, Γ) and iteratively updating the model parameter (θ, Γ) by maximizing the observed likelihood constitute the two essential procedures of the EM algorithm, corresponding respectively to the expectation step (E-step) and the maximization step (M-step). The details of these two steps are discussed below under the setting where the forms of parametric marginal location-scale distributions are given. When these marginal distribution of forms are unspecified, we replace them by the corresponding empirical CDFs (see Example-2 above), and the resulting approximate likelihood will be used in the EM algorithm.

E-step

Denote by u_{obs} the subvector of observed margins of u and u_{mis} the subvector of margins with missing values; similarly, q_{obs} and q_{mis} denote the corresponding subvectors of transformed quantiles. Let D_{obs} and D_{mis} be the sets of indices for components with observed data and missing data, respectively. Then $D = D_{\text{obs}} \cup D_{\text{mis}}$ is the set of all indices, and $D_{\text{obs}} \cap D_{\text{mis}}$ is an empty set. Note that both D_{obs} and D_{mis}

are subject-dependent, and its partition varies across subjects. Let $d_m = \dim(y_{\mathrm{mis}}) = |D_{\mathrm{mis}}|$.

At the E-step, the primary task is to calculate $\lambda(\theta, \Gamma|\theta^{(t)}, \Gamma^{(t)}, y_{\mathrm{obs}})$ for each subject, where the pair $(\theta^{(t)}, \Gamma^{(t)})$ are the updated values of (θ, Γ) obtained from the t-th iteration. For the ease of exposition, suppress index i in the following formulas. Given a subject, the λ-function $\lambda(\theta, \Gamma|\theta^{(t)}, \Gamma^{(t)}, y_{\mathrm{obs}})$ is the expected value of the log likelihood function of (θ, Γ) with respect to the conditional distribution of y_{mis} given y_{obs} and $(\theta^{(t)}, \Gamma^{(t)})$:

$$\begin{aligned}\lambda(\theta, \Gamma|\theta^{(t)}, \Gamma^{(t)}, y_{\mathrm{obs}}) &= \int_{R^{d_m}} \ln\{f(y|\theta, \Gamma)\} f\left(y_{\mathrm{mis}}|y_{\mathrm{obs}}, \theta^{(t)}, \Gamma^{(t)}\right) \mathrm{d}y_{\mathrm{mis}} \\ = \sum_{j \in D_{\mathrm{obs}}} \ln\{f_j(y_j|\theta_j)\} &+ \int_{(0,1)^{d_m}} \ln\{c(u|\theta, \Gamma)\} c\left(u_{\mathrm{mis}}|u_{\mathrm{obs}}, \theta^{(t)}, \Gamma^{(t)}\right) \mathrm{d}u_{\mathrm{mis}} \\ &+ \sum_{j \in D_{\mathrm{mis}}} \int_0^1 \ln\left[f_j\left\{F_j^{-1}(u_j|\theta_j)|\theta_j\right\}\right] c\left(u_j|u_{\mathrm{obs}}, \theta_j^{(t)}, \Gamma^{(t)}\right) \mathrm{d}u_j,\end{aligned} \tag{2.4}$$

where the right-hand side of equation (2.4) consists of three terms. The first term $\sum_{j \in D_{\mathrm{obs}}} \ln\{f_j(y_j|\theta_j)\}$ is a sum of marginal likelihoods over those observed margins $j \in D_{\mathrm{obs}}$, which can be evaluated directly. The second term is the observed likelihood, although it is of d_m dimension, its closed form expression can be analytically obtained. To do so, let $A = [A_{j_1 j_2}]_{d \times d} = \Gamma^{-1}$ be the precision matrix. The log copula density may be rewritten as follows:

$$\ln c(u|\theta, \Gamma) = \frac{1}{2}\ln|A| + \frac{1}{2}\sum_{j=1}^{d}(1 - A_{jj})q_j^2 - \frac{1}{2}\sum_{j_2 \neq j_1}^{d} A_{j_1 j_2} q_{j_1} q_{j_2}. \tag{2.5}$$

It follows from equation (2.5) that

$$
\begin{aligned}
&\int_{(0,1)^{d_m}} \ln\{c(u|\theta,\Gamma)\} c\left(u_{\text{mis}}|u_{\text{obs}},\theta^{(t)},\Gamma^{(t)}\right) \mathrm{d}u_{\text{mis}} \\
&= \frac{1}{2}\ln|A| + \frac{1}{2}\sum_{j\in D_{\text{obs}}}(1-A_{jj})q_j^2 + \frac{1}{2}\sum_{j\in D_{\text{mis}}}(1-A_{jj})\int_R q_j^2\phi(q_j|q_{\text{obs}},\theta^{(t)},\Gamma^{(t)})\mathrm{d}q_j \\
&\quad -\frac{1}{2}\sum_{j_1\neq j_2\in D_{\text{obs}}} A_{j_1j_2}q_{j_1}q_{j_2} - \sum_{j_1\in D_{\text{obs}}} q_{j_1}\sum_{j_2\in D_{\text{mis}}} A_{j_1j_2}\int_R q_{j_2}\phi(q_{j_2}|q_{\text{obs}},\theta^{(t)},\Gamma^{(t)})\mathrm{d}q_j \\
&\quad -\frac{1}{2}\sum_{j_1\neq j_2\in D_{\text{mis}}} A_{j_1j_2}\int_{R^2} q_{j_1}q_{j_2}\phi_2(q_{j_1},q_{j_2}|q_{\text{obs}},\theta^{(t)},\Gamma^{(t)})\mathrm{d}q_{j_1}\mathrm{d}q_{j_2} \\
&= \frac{1}{2}\ln|A| + \frac{1}{2}\sum_{j\in D_{\text{obs}}}(1-A_{jj})q_j^2 \\
&\quad +\frac{1}{2}\sum_{j\in D_{\text{mis}}}(1-A_{jj})\left[1-(\Gamma^{(t)}_{\text{obs},j})^T(\Gamma^{(t)}_{\text{obs,obs}})^{-1}\Gamma^{(t)}_{\text{obs},j} + \left\{(\Gamma^{(t)}_{\text{obs},j})^T(\Gamma^{(t)}_{\text{obs,obs}})^{-1}q^{(t)}_{\text{obs}}\right\}^2\right] \\
&\quad -\frac{1}{2}\sum_{j_1\neq j_2\in D_{\text{obs}}} A_{j_1j_2}q_{j_1}q_{j_2} + \sum_{j_1\in D_{\text{obs}}}\sum_{j_2\in D_{\text{mis}}} A_{j_1j_2}q_{j_1}\left\{(\Gamma^{(t)}_{\text{obs},j_2})^T(\Gamma^{(t)}_{\text{obs,obs}})^{-1}q^{(t)}_{\text{obs}}\right\} \\
&\quad -\frac{1}{2}\sum_{j_1\neq j_2\in D_{\text{mis}}} A_{j_1j_2}\left\{\Gamma^{(t)}_{j_1,j_2} - (\Gamma^{(t)}_{\text{obs},j_1})^T(\Gamma^{(t)}_{\text{obs,obs}})^{-1}\Gamma^{(t)}_{\text{obs},j_2}\right\} \\
&\quad -\frac{1}{2}\sum_{j_1\neq j_2\in D_{\text{mis}}} A_{j_1j_2}\left\{(\Gamma^{(t)}_{\text{obs},j_1})^T(\Gamma^{(t)}_{\text{obs,obs}})^{-1}q^{(t)}_{\text{obs}}\right\}\left\{(\Gamma^{(t)}_{\text{obs},j_2})^T(\Gamma^{(t)}_{\text{obs,obs}})^{-1}q^{(t)}_{\text{obs}}\right\},
\end{aligned}
$$

where $\Gamma_{\text{obs},j}$ is the j^{th} column of Γ with observed margins, and $\Gamma_{\text{obs,obs}}$ is a submatrix of Γ, whose columns and rows are observed margins. Also, $\phi(\cdot)$ is the univariate normal density, and $\phi_2(\cdot)$ is the bivariate normal density. The third term in equation (2.4) may be rewritten as follows:

$$
\begin{aligned}
&\sum_{j\in D_{\text{mis}}}\int_0^1 \ln\left[f_j\left\{F_j^{-1}(u_j|\theta_j)|\theta_j\right\}\right] c\left(u_j|u_{\text{obs}},\theta_j^{(t)},\Gamma^{(t)}\right)\mathrm{d}u_j \\
&= \sum_{j\in D_{\text{mis}}} E\left[\ln\left\{f_j\left(F_j^{-1}(u_j|\theta_j)|\theta_j\right)\right\}|y_{\text{obs}},\theta_j^{(t)},\Gamma^{(t)}\right],
\end{aligned}
\tag{2.6}
$$

where u_j is the CDF of normally distributed quantile q_j with mean $(\Gamma^{(t)}_{\text{obs},j})^T(\Gamma^{(t)}_{\text{obs,obs}})^{-1}q^{(t)}_{\text{obs}}$, and variance $\left\{1-(\Gamma^{(t)}_{\text{obs},j})^T(\Gamma^{(t)}_{\text{obs,obs}})^{-1}\Gamma^{(t)}_{\text{obs},j}\right\}$, and the expectation $E(\cdot)$ may be evaluated numerically using the method of Gaussian quadratures (Abramowitz and Stegun

(1972)). The observed likelihood for the full data of n subjects is expressed as:

$$\lambda(\theta, \Gamma | \theta^{(t)}, \Gamma^{(t)}, Y_{\text{obs}}) = \sum_{i=1}^{n} \lambda_i(\theta, \Gamma | \theta^{(t)}, \Gamma^{(t)}, y_{i,\text{obs}}), \tag{2.7}$$

where function $\lambda_i(\cdot)$ is given by equation (2.4). It is worth noting that equation (2.6) is of critical importance as it turns a d_m-dimensional integral a closed form expression, which ensures the E-step to be numerically feasible and stable. As a result, the evaluation of the E-step is computationally fast.

M-step

In the M-step we update parameters values by maximizing (2.7) with respect to θ and Γ. Following the ECM algorithm (Meng and Rubin (1993)), we will execute the M-step with several computationally simpler CM-steps. We propose a peeling procedure to facilitate the computation in the M-step, which consists of four routines given as follows.

Step M-1: Updating Marginal Parameters

For a specific marginal parameter θ_j, we obtain its update by sequentially maximizing the observed likelihood (2.7) as follows, for $j = 1, \cdots, d$,

$$\theta_j^{(t+1)} = \arg\max_{\theta_j} \sum_{i=1}^{n} \lambda_i(\theta_1^{(t+1)}, \cdots, \theta_{j-1}^{(t+1)}, \theta_j, \theta_{j+1}^{(t)}, \cdots, \theta_d^{(t)} | \Gamma^{(t)}, y_{i,\text{obs}}).$$

This optimization is carried out numerically by a quasi-Newton optimization routine available in R function *nlm*, and this step is computationally fast as the optimization involves only a set of low-dimensional parameters θ_j at one time.

Step M-2: Updating Correlation Parameters

If Γ is an unstructured correlation matrix, each off-diagonal element $\gamma_{j_1 j_2}$ is updated by maximizing the observed log-likelihood (2.7), which has a closed form expression. That is, for $j_1, j_2 = 1, \cdots, d, j_1 \neq j_2$,

$$\gamma_{j_1 j_2}^{(t+1)} = \frac{\sum_{i=1}^{n} q_{ij_1}^{(t)} q_{ij_2}^{(t)} 1(R_{ij_1} = 1) 1(R_{ij_2} = 1)}{\sum_{i=1}^{n} 1(R_{ij_1} = 1) 1(R_{ij_2} = 1)}, \tag{2.8}$$

where $1(\cdot)$ is an indicator function. Note that the diagonal elements $\gamma_{jj} = 1, j = 1, \cdots, d$.

If Γ is a structured correlation matrix such as exchangeable or first-order auto-regressive correlation, say $\Gamma = \Gamma(\gamma)$, we update the correlation parameter γ by maximizing equation (2.7). This can be done numerically by applying R function *optim* (Nelder & Mead, 1965). In both cases of exchangeable and first-order auto-regressive correlations, there is only one correlation parameter involved in optimization, and the related computing is fast.

Step M-3: Updating Quantiles

For each subject $i = 1, \cdots, n$, the quantiles are updated by the posterior mean for each margin $j = 1, \cdots, d$, as follows:

$$q_{ij}^{(t+1)} = \begin{cases} \Gamma_{j,-j}^{(t+1)} \left(\Gamma_{-j,-j}^{(t+1)} \right)^{-1} \left(q_{i,-j}^{(t+1)} \right)^{T}, j \in D_{i,\text{mis}} \\ \Phi^{-1} \left\{ F_j \left(y_{ij} | \theta_j^{(t+1)} \right) \right\}, j \in D_{i,\text{obs}}, \end{cases} \tag{2.9}$$

where $\Gamma_{j,-j}^{(t+1)}$ denotes the j^{th} row vector of matrix $\Gamma^{(t+1)}$ without the j^{th} element, $\Gamma_{-j,-j}^{(t+1)}$ is a submatrix of matrix $\Gamma^{(t+1)}$ without the j^{th} row and the j^{th} column, and $q_{i,-j}^{(t+1)}$ is the subvector of quantiles for subject i, $q_i^{(t+1)}$, with the j^{th} element deleted. Note that the quantile updating is carried out by borrowing information from the other correlated variables via matrix $\Gamma^{(t+1)}$.

Step M-4: Updating Outcome Values

Based on the updated parameter $\theta^{(t+1)}$ and quantiles $q_{ij}^{(t+1)}$, the outcome values are updated as follows:

$$y_{ij}^{(t+1)} = \begin{cases} F_j^{-1}\left\{\Phi\left(q_{ij}^{(t+1)}|\theta_{ij}^{(t+1)}\right)\right\}, j \in D_{i,\text{mis}} \\ y_{ij}, j \in D_{i,\text{obs}}. \end{cases} \tag{2.10}$$

2.4.2 Examples Revisited

Now we revisit the examples outlined in Section 2.3.3 in connection to the EM algorithm.

Example-1: Marginal Parametric Distribution

Example 1 is straightforward, and the marginal parameters and correlation parameters can be estimated by directly applying the above EM algorithm.

Example-2: Semi-parametric Marginal Distribution

Since the marginal CDFs are no longer parametric, the step of updating marginal parameters $\theta_1, \cdots, \theta_d$ in the EM algorithm is void. At each iteration, we need to update the missing values via Step M-4 and update matrix Γ via Step M-2. In addition, quantiles $q_{ij}, j \in D_{i,\text{mis}}$ are updated by Step M-3, and consequently the uniform variates $u_{ij}, j \in D_{i,\text{mis}}$ are updated as follows,

(2.11)
$$u_{ij}^{(t+1)} = \frac{1}{n}\left\{\sum_{k=1}^{n} 1(q_{kj}^{(t)} < q_{ij}^{(t)}) + \frac{1}{2}\right\}, \text{and } q_{ij}^{(t+1)} = \Phi^{-1}\left(u_{ij}^{(t+1)}\right), j = 1, 2, \cdots, d,$$

where the term $\frac{1}{2}$ in equation (2.11) is used to avoid $u_{ij}^{(t+1)} = 0$ leading to $q_{ij}^{(t+1)} = -\infty$, which causes numerical problem in the EM algorithm.

2.4.3 Standard Error Calculation

Louis' formula (Louis (1982)) is a well-known procedure useful to obtain standard errors of the estimates from the EM algorithm. As shown in equation (2.12) below, the observed Fisher Information matrix can be obtained via two information matrices. The first term in equation (2.12) is the expected full-data information matrix, while the second is the expected missing data information matrix. For the ease of exposition, suppress index i in the following formulas.

$$\begin{aligned} I(\hat{\theta},\hat{\Gamma}) &= -\nabla^2 \ln\left\{f(y_{\text{obs}}|\theta,\Gamma)\right\}|_{\theta=\hat{\theta},\Gamma=\hat{\Gamma}} \\ &= -I_{\text{full}} + I_{\text{mis}} \\ &= -\int \nabla^2 \ln\left\{f(y_{\text{mis}},y_{\text{obs}}|\theta,\Gamma)\right\}|_{\theta=\hat{\theta},\Gamma=\hat{\Gamma}} f(y_{\text{mis}}|y_{\text{obs}},\hat{\theta},\hat{\Gamma})dy_{\text{mis}} \\ &\quad + \int \nabla^2 \ln\left\{f(y_{\text{mis}}|y_{\text{obs}},\theta,\Gamma)\right\}|_{\theta=\hat{\theta},\Gamma=\hat{\Gamma}} f(y_{\text{mis}}|y_{\text{obs}},\hat{\theta},\hat{\Gamma})dy_{\text{mis}} \end{aligned} \tag{2.12}$$

where ∇^2 denotes the second order derivative with respect to the model parameters, and ($\hat{\theta}$, $\hat{\Gamma}$) are the estimates obtained as the final outputs of the EM algorithm. Therefore, the Fisher Information matrix is

$$I(\hat{\theta},\hat{\Gamma}) = \sum_{i=1}^{n} I_i(\hat{\theta},\hat{\Gamma}), \tag{2.13}$$

where $I_i(\hat{\theta},\hat{\Gamma}) = -\nabla^2 l_i(\hat{\theta},\hat{\Gamma})$, and $l_i(\hat{\theta},\hat{\Gamma})$ is the observed log likelihood evaluated at the estimates for subject i, which can be calculated numerically via the following expression:

$$\begin{aligned} l_i(\hat{\theta},\hat{\Gamma}) &= \frac{1}{2}\ln(|\hat{A}_i|) + \frac{1}{2}\sum_{j\in D_{i,\text{obs}}}\left(1-\hat{A}_{i,jj}\right)\hat{q}_{ij}^2 - \frac{1}{2}\sum_{j_1\neq j_2\in D_{i,\text{obs}}}\hat{A}_{i,j_1j_2}\hat{q}_{ij_1}\hat{q}_{ij_2} \\ &\quad + \sum_{j\in D_{i,\text{obs}}}\ln\left\{f_j(y_{ij}|\hat{\theta}_j)\right\}. \end{aligned} \tag{2.14}$$

Here $A_i = (\Gamma_i)^{-1} = [A_{i,j_1j_2}]_{d_{m,i}\times d_{m,i}}$, where Γ_i is the submatrix of Γ whose columns correspond to the observed variables in y_i for subject i, and $d_{m,i}$ counts the dimensions. By R function *hessian*, the Hessian function of equation (2.14) can both be numerically carried out. This provides the observed Fisher information matrix I, and moreover the asymptotic variance for $(\hat{\theta}, \hat{\Gamma})$ is $\mathrm{I}(\hat{\theta}, \hat{\Gamma})^{-1}$.

2.4.4 Initialization

It is known that the quality of initial values is critical to the accuracy and efficiency of the EM algorithm. The initial parameters values $(\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_d^{(0)}, \Gamma^{(0)})$ may be given by the estimates obtained from the complete case analysis. Although theoretically the initial values may be set arbitrary, all numerical experiences have suggested that the closer initial values are to the true values, the faster the algorithm converges.

2.5 Simulation Study

We conduct simulation experiments to evaluate and compare the performance of the EM algorithm with the multiple imputation method. In our experiments, the dimension of outcomes is set as $d = 3$, and $d_m = 1$ or 2 for different subjects. Three types of the correlation matrices Γ are considered: unstructured, exchangeable, and first-order autoregressive. Both Multiple Imputation (Little and Rubin (2002)) and Hot-deck Imputation (Andridge and Little (2010)) are included in the comparison.

Note that the R package of Multiple Imputation (R Package "*MI*") applied here is developed under multivariate normal distributions, so the skewness of the marginal distributions for outcomes may result in estimation bias. In Hot-Deck Imputation (R Package "*HotDeckImputation*"), as discussed above, each missing value is imputed by a randomly drawn similar record in terms of the nearest neighbor criterion. To adjust for confounders, Hot-Deck Imputation is adopted through the following steps. First, we run regression on the complete cases; second, impute residuals of the missing data, and then finally obtain imputed missing outcomes that will be used to run regression analysis on the “full” outcomes to yield the estimates of model parameters.

A naive approach is to use marginal data to obtain estimate of CDF $F_j(y_j|\hat{\theta}_j), j = 1, \cdots, d$, if F_j is a parametric model, or $\hat{F}_j(y_j), j = 1, \cdots, d$ by empirical CDF if F_j is nonparametric model, and make inverse-normal transformation $\hat{q}_j = \Phi^{-1}(F_j(y_j|\hat{\theta}_j))$ or $\hat{q}_j = \Phi^{-1}(\hat{F}_j(y_j))$, which are used to calculate $\mathrm{cor}(\hat{q}_{j_1}, \hat{q}_{j_2})$. Since the naive approach only uses marginal information and available data. Because it is inferior to imputation methods that replace the missing data with plausible values. So in this section, we did not include the naive approach in the comparison.

2.5.1 Skewed Marginal Model

We first examine the EM algorithm in the setting of the semi-parametric model discussed in Example-2, Section 2.3.3. In this case, only correlation parameters (Kendall’s tau) are updated. To generate data, the marginal distributions are set as gamma distribution with the shape parameter $\alpha = 0.2$ and rate parameter $\beta = 0.1$, leading to the skewness 4.47. The correlation matrix Γ with $\gamma_{12} = 0.3, \gamma_{13} = 0.5, \gamma_{23} = 0.4$ is used, with the corresponding Kendall’s tau being $(0.1940, 0.3333, 0.2620)$. We compare the results obtained from the full data without missingness (regarded as the gold standard) to the results obtained by the EM algorithm, Multiple Imputation, and Hot-Deck Imputation with incomplete data. The missingness percent varies from 20% to 50%. The sample size is fixed at 200, while 1000 replicates are run to draw summary statistics.

As shown in Table 2.1, with no surprise, in such a case of highly skewed distributions, the estimates of three Kendall’s tau parameters obtained from Multiple Imputation are more biased. The estimation results from the EM algorithm and Hot-Deck Imputation are comparable, but the EM algorithm method provides smaller empirical standard errors. In both simple cases above, the EM algorithm works well.

Table 2.1: Simulation results of correlation (Kendall's tau) parameters estimation in copula model for marginal skewed distributed data obtained by full data likelihood, EM algorithm and Imputation methods with different missing percentage. (Standard error ratio is calculated by a ratio of two standard errors between a method and the gold standard.)

	Full Data		Copula&EM		Multiple Imputation		Hot Deck Imputation	
%mis	bias($\times 10^{-2}$)	std.err	bias($\times 10^{-2}$)	std.err table	bias($\times 10^{-2}$)	std.err table	bias($\times 10^{-2}$)	std.err table
20%	0.10	0.0440	-0.14	1.0250	-1.78	1.0727	0.04	1.1273
	0.00	0.0402	-0.46	0.9851	-3.12	1.0995	-0.05	1.1144
	0.05	0.0434	-0.20	1.0069	-2.36	1.1060	-0.01	1.0945
30%	-0.03	0.0454	-0.30	1.0639	-2.77	1.0771	-0.13	1.1718
	-0.10	0.0416	-0.63	1.0673	-4.88	1.1370	-0.31	1.1563
	0.04	0.0442	-0.32	1.0452	-3.65	1.0905	0.01	1.1516
50%	-0.15	0.0437	-0.09	1.1716	-4.32	1.2449	-0.15	1.2792
	-0.14	0.0413	-0.55	1.1840	-7.10	1.2736	-0.46	1.3099
	-0.11	0.0428	-0.43	1.1869	-5.99	1.2453	-0.40	1.2944

2.5.2 Misaligned Missing Data

Motivated from one of our collaborative projects on a quality of life study (see the detail in Section 2.6), we consider a rather challenging missing data pattern in this simulation study. That concerns the so-called misaligned missingness, which refers to a situation where two correlated variables have missing values on exclusive subsets of subjects. In a completely misaligned missing case, where there is no overlap between two margins, Hot-Deck imputation fails to work, and the method of multiple imputation cannot effectively capture between-variable correlations, resulting in poor estimation of correlation parameters. However, when the correlation matrix is specified by a structured form in the Gaussian copula model, the EM algorithm is able to utilize the correlation structure for information sharing, and consequently the resulting estimation of model parameters is highly satisfactory.

The simulation setup is given as follows. Following Example-1 in Section 2.3.3, we include two covariates $X_1 \sim$ Bin(1,0.5) and $X_2 \sim \Gamma$(2,1), and generate residuals ϵ in a linear model with $\mu_j = X^T\beta_j, j = 1, 2, 3$ from a tri-variate normal with the marginal N(0,1) and first-order autoregressive correlation matrix with parameter $\gamma = 0.5$.

The missing mechanism concerns missing at random (MAR) with a partially misaligned pattern with specified as follows. A tri-variate outcome $(Y_1, Y_2, Y_3)'$ is subject to be missing at random, where Y_1 is fully observed, while each of Y_2 and Y_3 has 45% missing data that are partially misaligned, with only 10% of subjects have an overlap on the observed parts of Y_2 and Y_3. The reason that a partial misalignment is considered here is to allow the Hot-Deck Imputation method possibly in the part of the comparison. The EM algorithm procedure and notations follow as discussed in Exmaple-1, Section 2.3.3. The missing probability in the marginal of Y_2 is,

$$P(R_2 = 0|X_1, Y_2) = \begin{cases} 0.45, \text{ if } X_1 = 1; \\ 0.81, \text{ if } X_1 = 0, \text{ and } Y_2 > \mu_2; \\ 0.09, \text{ if } X_1 = 0, \text{ and } Y_2 < \mu_2. \end{cases}$$

The missing probability in the third marginal Y_3 is given by,

$$P(R_3 = 0|R_2) = \begin{cases} 0, \text{ if } R_2 = 0; \\ \dfrac{0.45}{1 - 0.45}, \text{ if } R_2 = 1. \end{cases} \tag{2.15}$$

We compare the results obtained from the EM algorithm with those from the gold standard using the full data, the multiple imputation and the Hot-Deck imputation. In addition, this comparison includes two types of standard errors: the first type is the empirical standard error in four methods, and the other type is the average of 1000 model-based standard errors obtained from Louis' formula discussed in Section 2.4.3, which is only provided in the EM algorithm.

2.6 Data Example

Nephrotic Syndrome (NS) is a common disease in pediatric patients with kidney disease. The typical symptom of this disease is characterized by the presence of edema that significantly affects the health-related quality of life in children and adolescents. The PROMIS (Fries et al. (2005); Gipson et al. (2013)) is a well-validated instrument to assess pediatric patient's quality of life. The instrument consists 7 domains, but here we only choose 3 domains with missing misalignment pattern for illustration. In the data, two QoL measures, pain and fatigue, are measured on two exclusive sets of subjects due to some logistic difficulty at the clinic; out of 226 subjects, 107 subjects have measurements of pain, but no measurements of fatigue, while the other 117 subjects have measurements of fatigue but no measurements of

Table 2.2: Simulation results concerning estimation of Pearson correlation and marginal regression parameters in the copula model for partially misaligned missing at random data obtained EM algorithm, compared with the gold standard with full data, Multiple Imputation and Hot-Deck Imputation. (Standard error ratio is calculated by a ratio of two standard errors between a method and the gold standard.)

		Full Data		Copula&EM		Multiple Imputation		Hot-Deck Imputation	
parameter	true value	estimate	std.err	estimate	std.err	estimate	std.err	estimate	std.err
β_{10}	0	-0.0012	0.143	0.0596	0.1466 / 0.1488	-0.0012	0.143	-0.0012	0.143
β_{11}	1	1.0030	0.1424	0.9425	0.1451 / 0.1485	1.0030	0.1424	1.0030	0.1424
β_{12}	3	3.0003	0.0496	3.0000	0.0498 / 0.0528	3.0003	0.0496	3.0003	0.0496
σ_1	1	0.9975	0.051	1.0446	0.0566 / 0.0522	0.9975	0.051	0.9975	0.051
β_{20}	0	-0.0050	0.1435	-0.0087	0.189 / 0.2146	-0.0918	0.2134	-0.1405	0.2397
β_{21}	2	2.0049	0.1413	2.0098	0.1864 / 0.2157	2.1654	0.2053	2.1989	0.2104
β_{22}	2	2.0006	0.0514	2.0017	0.0648 / 0.0774	2.0015	0.0719	2.0020	0.0711
σ_2	1	0.9977	0.0493	1.0918	0.0883 / 0.0754	0.9836	0.0796	0.9398	0.1046
β_{30}	0	-0.0023	0.1474	0.1109	0.1915 / 0.1797	0.0503	0.2178	0.0424	0.2371
β_{31}	3	3.0046	0.1481	2.8902	0.1936 / 0.1776	2.9111	0.2155	2.9160	0.2191
β_{32}	1	1.0010	0.051	1.0005	0.0665 / 0.0634	0.9997	0.0745	1.0006	0.0737
σ_3	1	0.9965	0.0508	0.9379	0.0615 / 0.0622	0.9946	0.0752	0.9662	0.0956
γ	0.5	0.5019	0.0408	0.4951	0.0606 / 0.0557	0.4409	0.1212	0.4178	0.1149

pain. In addition, two subjects have neither measurements of pain nor measurements of fatigue. Interestingly, measurements of anxiety have been fully recorded on all 226 individuals with no missing data. In this case, Hot-Deck imputation does not work. We first apply the complete case univariate analysis of each QoL domain score (Y_1 = anexiety, Y_2 = pain, Y_3 = fatigue) on covariates of age, gender, edema, race (white, black, and other as reference), and estimate the linear correlation coefficient of the residuals as 0.6830 between anxiety and pain and 0.5106 between anxiety and fatigue, which turns out to be approximately the square of the correlation coefficient between anxiety and pain. This suggests us use first order autoregressive correlation for matrix Γ in the copula model.

The EM algorithm has two advantages to handle this misaligned missing data pattern. One is that we can estimate both marginal and correlations parameter adjusting for the confounders, where the information across the three QoL scores can be shared to improve efficiency. The other is the prediction of the missing QoL scores by using the correlated QoL scores together with the marginal regression models, which requires the availability of inverse correlation matrix, Γ^{-1}.

The observed data and predicted data from the EM algorithm are all shown in Figure 2.1. The triangles indicate patients with missing fatigue data, and the circles correspond to patients with missing pain data. Between pain and fatigue QoL scores, outcomes have no overlap. The circles and triangles are well distributed and appear to lie in elliptical in the first two scatter plots. In the third plot, the reason that the predicted triangles appear a straight line is the use of AR-1 correlation matrix, and the shape of these points may change to another pattern when a different correlation structure is used.

In Table 2.3, the standard errors for the estimates obtained by the multiple im-

Table 2.3: Estimation of correlation and marginal regression parameters in the copula model for quality of life study obtained by univariate analysis, EM algorithm and Multiple Imputation.

		Univariate Analysis		Copula&EM		Multiple Imputation	
Outcome	Covariates	estimate	std.err	estimate	std.err	estimate	std.err
Anxiety	Intercept	40.5406	4.1387	39.3566	4.4013	40.5406	4.1387
	Age	0.0068	0.2451	0.0879	0.2607	0.0068	0.2451
	Gender	1.8753	1.4392	1.8749	1.5306	1.8753	1.4392
	Edema	5.2453	2.1133	5.0332	2.2474	5.2453	2.1133
	White	0.4908	2.0272	0.8502	2.1558	0.4908	2.0272
	Black	4.7095	2.3491	4.5299	2.4981	4.7095	2.3491
	σ	10.33	*	10.9813	0.5165	10.2107	*
Pain	Intercept	41.3243	6.1947	36.9533	7.8865	45.1838	5.7978
	Age	0.0649	0.3575	0.3587	0.4551	-0.0029	0.3418
	Gender	0.3423	2.0304	0.1196	2.5849	0.3112	1.9516
	Edema	6.0597	3.3437	4.8594	4.2569	7.7273	3.2138
	White	1.2991	3.3331	2.8328	4.2434	-1.3024	3.1549
	Black	6.8171	3.7939	6.5658	4.8300	3.5309	3.5994
	σ	10.58	*	13.4695	0.8805	11.2709	*
Fatigue	Intercept	30.9033	5.5260	30.9030	4.6716	31.5289	4.6886
	Age	0.7043	0.3275	0.7043	0.2768	0.6351	0.2844
	Gender	0.6004	1.9962	0.6005	1.6876	0.7472	1.6722
	Edema	7.7774	2.6296	7.7774	2.2230	8.0216	2.2691
	White	3.6899	2.4766	3.6900	2.0937	3.2258	2.1573
	Black	7.1261	2.8784	7.1261	2.4334	7.0781	2.5138
	σ	9.568	*	8.0890	0.5504	9.2272	*
Correlation	γ	-	-	0.6851	0.0395	0.4001	-

* unavailable in R function *lm* for linear regression.

putation are calculated by the conventional method given by Little & Rubin (2002). Moreover, some findings in the results shown in Table 2.3 are noteworthy. First,

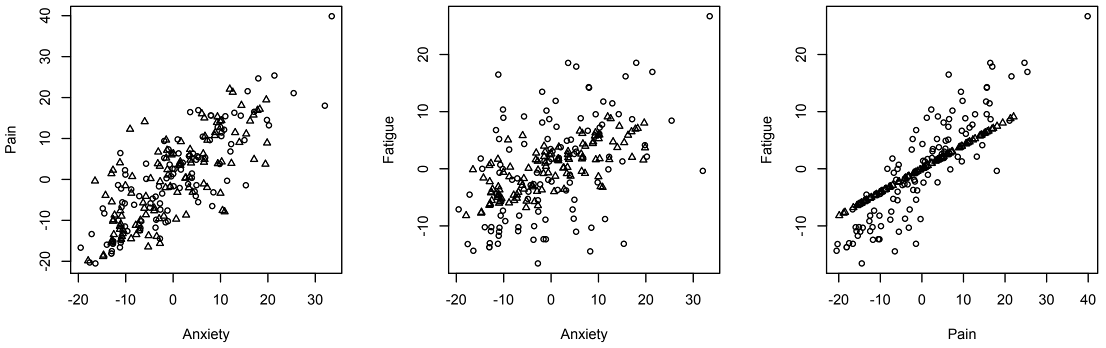


Figure 2.1: Plots of observed and predicted residuals from the EM algorithm.

the estimated rank-based correlations Kendall's τ and Spearman's ρ between anxiety and pain are, respectively, $\tau_{12} = 0.4805$ and $\rho_{12} = 0.6677$, between anxiety and fatigue are $\tau_{13} = 0.3110$ and $\rho_{13} = 0.4524$, and between pain and fatigue are $\tau_{23} = 0.4805$ and $\rho_{23} = 0.6677$. In addition, the estimated correlation parameter by the multiple imputation approach is clearly smaller than that obtained by the EM algorithm. This is because the key difference is that the EM algorithm makes use of the correlation structure to access the entire data, whereas the imputation method does not. Imputation methods are based on available observed information, but not on the correlation structure. Moreover, the EM algorithm provides a straightforward calculation of asymptotic standard error of the correlation parameter for inference; for example, p-value for $H_0 : \gamma = 0$ is of practical importance.

In addition, with regard to the effect of edema in pain, according to clinical information available on Mayo Clinic Website, pain is not regarded as one of key symptoms associated with edema. Both results obtained by the EM algorithm and the univariate analysis are in the agreement with this clinical information, indicating no significant effect of edema on pain score, while the multiple imputation method reports an opposite result.

2.7 Discussion

This paper presents a Gaussian copula framework that provides both marginal Pearson correlations, and marginal rank-based correlation in the presence of missing data. The EM-algorithm is developed and implemented to estimate both marginal parameters and correlation parameters. The proposed methodology allows to adjust for confounding factors via marginal regression models to obtain adjusted marginal correlation estimates, which are useful in practice. We propose a peeling procedure in the M-step to facilitate the computation of updating parameter values. In addition, missing values may also be updated as part of the EM-algorithm. The EM algorithm outperforms imputation-based methods when the marginal outcomes are skewed and/or missing data patterns are fully or severely misaligned.

For the completely misaligned missing data pattern, Hot-Deck Imputation does not work, and the multiple imputation cannot effectively utilize the correlation structure in data imputation and parameter estimation. When the correlation matrix is structured, EM algorithm can fully access the correlation structure in the Gaussian copula, and share information across different outcome variables, and therefore the resulting estimates from the EM algorithm are satisfactory. Note that structured correlation (e.g., exchangeable) is seen in other families of copulas, such as Archimedean copulas, in which expansion of the EM algorithm with misaligned missing data is feasible and worth a further study.

The EM algorithm developed in a parametric Gaussian copula framework may be sensitive to model misspecification. Model diagnostics are required before to draw final conclusions. Several authors have proposed diagnostic methods, such as Masarotto et al. (2012); Joe (1997); Genest et al. (1995); Ané and Kharoubi (2003), among others. However, how these diagnostic approaches may perform in the case of incomplete data remains unknown and is an interesting future work. Furthermore, Segers et al. (2013) proposed a one-step estimation for correlation parameters in the Gaussian copula, which is shown to be efficient after a novel one-step adjustment, and this approach may be applied to improve the M-step of the EM algorithm.

For the case of completely misaligned missingness, when the correlation matrix is unstructured, the correlation parameters are not fully identifiable. Manski (2003) introduced several approaches for partial identification problem, and Fan and Zhu (2009) developed a method to determine the bounds, within which the estimates of correlation parameters of a copula model are partially identified by a parameter set. Following the notation given in Fan and Zhu (2009), we consider $\mu(x, y) = xy$ for the problem of covariance estimation. This function is supermodular because its cross-derivative is 1, and this function is symmetric and marginal variances are finite. Thus, according to Fan and Zhu (2009) theory, we can establish a partial identification range for the correlation parameter in the presence of misaligned missing data with the lower and upper bounds, denoted by $\gamma^L_{j_1,j_2}$ and $\gamma^U_{j_1,j_2}$. They are the lower and upper bounds of correlation parameter γ_{j_1,j_2} given by $\gamma^L_{j_1,j_2} = \left[\int_0^1 \left\{F^{-1}_{j_1}(u|\theta_{j_1})F^{-1}_{j_2}(1-u|\theta_{j_2})\right\} du - \mu_{j_1}\mu_{j_2}\right] / \sigma_{j_1}\sigma_{j_2}$, and $\gamma^U_{j_1,j_2} = \left[\int_0^1 \left\{F^{-1}_{j_1}(u|\theta_{j_1})F^{-1}_{j_2}(u|\theta_{j_2})\right\} du - \mu_{j_1}\mu_{j_2}\right] / \sigma_{j_1}\sigma_{j_2}$, where quantiles functions $F^{-1}_{j_1}$ and $F^{-1}_{j_2}$ may be estimated by available data of y_{j_1} and y_{j_2}. This direction of research is worth a thorough exploration.

CHAPTER III

Composite Likelihood Approach in Gaussian Copula Regression Models with Missing Data

3.1 Summary

Misaligned missing data occur in many large-scale studies due to some impediments in data collection such as policy restriction, equipment limitation and budgetary constraint. By misaligned missingness we mean a missing data pattern in which two sets of variables are measured from disjoint subgroups of subjects with no overlapped observations. An analytic challenge arising from the analysis of such data is that some of correlation parameters related to those misaligned variables are not point identifiable but possibly partially identifiable. This parameter identification issue hinders us from utilizing classical multivariate models in the data analysis. To overcome this difficulty, we propose a composite likelihood approach based on marginal distributions of variables with full observations, so that the resulting pseudo likelihood is free of any unidentifiable parameters. After obtaining estimates of the point identifiable parameters, we further estimate the parameter range for partially identifiable parameters. For implementation, we develop an effective peeling optimization procedure to obtain estimates of point identifiable parameters. We investigate the performance of the proposed composite likelihood method through simulation studies, with comparisons to the classical maximum likelihood estimation

obtained from both EM algorithm and multiple imputation strategy. The proposed method is illustrated by one data example from our collaborative project.

3.2 Introduction

We consider a d-dimensional parametric model, $f(\mathbf{x};\boldsymbol{\xi})$, $\mathbf{x} \in \mathcal{X} \subset R^d$ and $\boldsymbol{\xi} \in \Xi \subset R^m$, where $\boldsymbol{\xi}$ is the parameter of interest. Suppose that an incomplete data is collected from n partially observed subjects. For a subject, let $\mathbf{y} = (y_1, y_2, \cdots, y_d)^T$ be a d-dimensional random vector of outcomes, part of which is observed and the other part is missing. Denote by R_j as a missing data indicator, where $R_j = 0$ or 1 if the j-th margin y_j is missing or observed. Note that this indicator is known and varies across different subjects. In this paper, we focus on a special type of missing data pattern, termed as misaligned missingness. In an example of misaligned missingness between two variables, say the first two margins y_1 and y_2, the sum of their missing data indictors on each subject is always 1, namely $R_1 + R_2 = 1$. This implies that the pair of variables is not observed simultaneously among subjects. An obvious difficulty in the analysis of such data is that the correlation parameter between these two variables is not point identifiable. This consequently gives rise to some analytic challenges that cannot be easily handled in the framework of the classical maximum likelihood estimation.

This misaligned missing data pattern was encountered in one of our collaborative projects concerning a quality of life (QoL) study on pediatric patients with nephrotic syndrome at the University of Michigan Children's Hospital. A typical symptom of this common kidney disease is characterized by the presence of edema that affects the quality of life in children and their families. PROMIS (Fries et al. (2005)) is a well-validated instrument widely used to assess QoL of patients with renal disease, which consists of seven QoL domains, including pain interference, fatigue, depression, anxiety, mobility, social peer relationship, and upper extremity functioning. These seven QoL scores are intrinsically correlated. A scientific objective of this study was to perform a joint regression analysis of the QoL scores on patient's characteristics such as edema condition. The difficulty in the required analysis arises from the carelessly designed data collection in that two QoL measures, pain interference and fatigue, were measured on two disjoint subgroups of children in order to save the study cost. Since the correlation parameter between pain interference and fatigue is not point identifiable due to the misaligned missingness, the required joint analysis of these seven QoL scores cannot be carried out straightforwardly by any existing methods in the literature.

In effect, such missing data pattern may occur in many other practical settings due to various reasons. For example, in the data generation by a high-throughput technology, variables (or features) are typically measured by allocating bio-samples into multiple batches, each containing a disjoint subset of samples. When certain batches partially fail due to technical limitations, some of features may be measured on exclusive subsets of bio-samples, leading to misaligned missingness. In practice, those features with misaligned missingness are routinely discarded in the data analysis. In the emerging field of data harmonization where data sets from multiple surveys are combined to form a mega data set, some of variables may be measured by different instruments that do not contain identical sets of questions for a trait measurement (e.g. cognitive function). In this case, although being highly correlated, the trait measures from different data sources are misaligned missing.

To handle missing data, the complete-case analysis is often employed in practice mostly for technical convenience, which simply discards any cases with missing values and proceeds with the analysis using standard methods. Obviously, the data attrition is a concern, because the reduced sample size may result potentially in a substantial loss of estimation efficiency. For remedy, EM algorithm (Dempster et al. (1977)) is a widely used iterative algorithm to carry out the maximum likelihood estimation with incomplete data. In the case of misaligned missingness, EM algorithm is no longer applicable under the full parametric model $f(\mathbf{x};\boldsymbol{\xi})$, because some parameters in the full model are not point identifiable. Multiple Imputation (Rubin (2004)) provides an alternative approach, which is extensively used in statistical analysis with missing values. Instead of filling in a single value for each missing value, multiple imputation procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. Being a non-model based imputation method, hot-deck imputation (e.g. Andridge and Little (2010)) is also widely used, where a missing value is imputed with a randomly drawn similar record in terms of the nearest neighbor criterion. However, in the case of the misaligned missing data pattern, the hot-deck imputation cannot work in full capacity, because misaligned missingness prohibits us from borrowing information between margins with no overlapped observations. Multiple imputation is based on a multivariate distribution assumption, often the multivariate normal distribution, which cannot work under the full distribution, either, when some of correlation parameters are not point identifiable due to the misaligned missingness. As a result, the efficacy of imputation may get harmed because only marginal distributions, rather than the full distribution, are used in the imputation. Some related numerical evidence is later provided in our simulation studies and data analysis examples in this paper.

There is little work available in the literature concerning statistical methods to handle partially identifiable parameters. A parameter is said to be partially identifiable if the true value of a parameter is not point identifiable but a range of parameter values containing its true value is identifiable. As a trivial example, a correlation parameter is always partially identifiable because interval $[-1, 1]$ is a valid range for a correlation parameter. Manski (2003) proposed approaches to addressing such a parameter identification problem in estimation. Fan and Zhu (2009) developed a method in the setting of bivariate copulas to determine both lower and upper bounds for the range of the pairwise dependence parameter. However, their method does not work for a general d-dimensional multivariate model.

In this paper, we develop a new approach to handling parameter estimation in the presence of partially identifiable parameters. Our new approach, termed as the complete-case composite likelihood, takes the advantage of composite likelihood that allows the composition of a pseudo likelihood function through the utility of only marginal distributions. Because only those marginal distributions with observed data are used, the resulting composite likelihood will not contain any unidentifiable parameters. Consequently, estimation of point identifiable parameters becomes feasible without invocation of the EM algorithm or the multiple imputation scheme. Composite likelihood, proposed first by Besag (1974, 1977) and later formally formulated by Lindsay (1988), has received increasing attention in the recent statistical literature because of its computational ease. This method has been successfully applied in many areas, including generalized linear mixed models (Renard et al. (2004)), statistical genetics (Fearnhead and Donnelly (2002)), spatial statistics (Hjort et al. (1994); Heagerty and Lele (1998); Stein et al. (2004); Varin and Vidoni (2005); Bai et al. (2012, 2014)), longitudinal data analysis (Molenberghs and Verbeke (2005)) and multivariate survival analysis (Parner (2001)), among others. It has demonstrated to possess good theoretical properties, such as estimation consistency, and can be utilized to establish hypothesis testing procedures, as well as model selection. For more detail, refer to Varin et al. (2011) and additional references therein.

Among extensive publications in composite likelihood methodology, there is very limited work concerning the statistical method with incomplete data in the current literature. Gao and Song (2011) developed the EM algorithm for composite likelihood estimation. Molenberghs et al. (2011) studied the double robustness of composite likelihood estimation for incomplete data under missing at random (MAR) mechanism. Much remains unknown. For example, whether or not the inverse probability weighting (IPW, Horvitz and Thompson (1952)) technique is needed in the composite likelihood when data are missing under MAR. The method of IPW has been shown of critical importance in generalized estimation equation (GEE, Liang and Zeger (1986)) to ensure estimation consistency. In this paper, we show that like the maximum likelihood estimation, IPW is not required in the proposed complete-case composite likelihood to establish estimation consistency under the MAR mechanism, including the misaligned missing pattern. Moreover, for partially identifiable parameters, we provide consistent estimation for both upper and lower bounds of the parameter range via certain constraints on the model validity. It turns out that in the setting of copula models, our method provides a narrower estimated range for the dependence parameter than that given by Fan and Zhu (2009).

This paper is organized as follows. Section 3.3 describes the complete-case composite likelihood estimation method, including statistical inference and properties. Section 3.4 presents the implementation via peeling algorithm. In Section 3.5, we consider an important application based on Gaussian copula regression model with location-scale family marginal distribution to illustrate the proposed methodology. Section 3.6 presents simulation results, and Section 3.7 presents a data analysis example. Section 3.8 provides some concluding remarks. All proofs of theories are included in the Appendix.

3.3 Method

3.3.1 Complete-Case Composite Likelihood

We propose to use the composite likelihood function to estimate point identifiable parameters in the model $f(\mathbf{x};\boldsymbol{\xi})$. First, we partition the set of parameters, $\boldsymbol{\xi}$, into two disjoint subsets, one containing all point identifiable parameters, denoted by $\boldsymbol{\eta}$, and the other containing all partially identifiable parameters, denoted by $\boldsymbol{\zeta}$. Obviously, $\boldsymbol{\xi} = \boldsymbol{\eta} \cup \boldsymbol{\zeta}$, $\boldsymbol{\eta} \in R^{m_1}$, and $\boldsymbol{\zeta} \in R^{m_2}$, with $m_1 + m_2 = m$.

Let S denote the collection of all possible subsets of $\{1, \cdots, d\}$, and S is the same for all subjects. Let $w_s \in \{0,1\}, s \in S$, with $\sum\limits_{s\in S} w_s = 1$, be an indicator of subset s. Let $\boldsymbol{\xi}_s = \{\boldsymbol{\xi}_{\{k\}}, k \in s\}$ be the subset of $\boldsymbol{\xi}$ corresponding to the margins indexed by set s; $\boldsymbol{\xi}_{\{k\}}, k = 1, \cdots, d$ is the single-element subset of $\boldsymbol{\xi}$ corresponding to the j-th univariate marginal distribution. Let f_s be the marginal density function with respect to the margins in set s, namely $\mathbf{y}_s = \{y_k, k \in s\}$. Denote $D_{\text{obs}} = \{k, R_k = 1\}$ is the subset of indices that correspond to the observed margins.

For one subject, we construct the complete-case composite likelihood as of the following form:

$$L_c(\boldsymbol{\xi}|\mathbf{y}, \mathbf{R}) = \prod_{s\in S} f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)^{w_s} = f_{\text{obs}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\eta}_{\text{obs}}), \tag{3.1}$$

where $w_s = w_s(\mathbf{R}) = \mathrm{I}\{s = D_{\text{obs}}\}$, and $\boldsymbol{\eta}_{\text{obs}} = \boldsymbol{\xi}_s|_{s=D_{\text{obs}}}$. For convenience, we use f_{obs} denote the marginal density of variables in D_{obs}, whose dimension is $|D_{\text{obs}}| = (d - d_m)$. Obviously, these three terms, f_{obs}, $\mathbf{y}_{\text{obs}}$, $\boldsymbol{\eta}_{\text{obs}}$ vary across subjects. In this paper, the subscript "obs" indicates those margins confined within set D_{obs}. It is easy to see that the set of all point identifiable parameters is the union of such $\boldsymbol{\eta}_{\text{obs}}$ parameters from all subjects, i.e., $\boldsymbol{\eta} = \cup_i \boldsymbol{\eta}_{i,\text{obs}}$, where $\boldsymbol{\eta}_{i,\text{obs}}$ is the set of point identifiable parameters for subject i. As a matter of fact, for each subject, the complete-case composite likelihood of $\boldsymbol{\xi}$ is a function of the reduced parameter, $\boldsymbol{\eta}_{\text{obs}}$, given $\mathbf{y}_{\text{obs}}$ and $\mathbf{R}$. Note that in general, we have $|D_{\text{obs}}| > 1$, and this dimension varies from subject to subject. It is worth pointing out that, for each subject, there is only one term w_s equal to 1, i.e., when $s = D_{\text{obs}}$. Moreover, the log complete-case composite likelihood function for one subject is given by

$$l_c(\boldsymbol{\eta}_{\text{obs}}|\mathbf{y}_{\text{obs}}, \mathbf{R}) = \ln f_{\text{obs}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\eta}_{\text{obs}}). \tag{3.2}$$

Therefore, the log complete-case composite likelihood function for the random sample of n subjects is given by,

$$l_c(\boldsymbol{\eta}|\mathbf{Y}_{\text{obs}}, \mathbf{R}) = \sum_{i=1}^{n} l_{c,i}(\boldsymbol{\eta}_{\text{obs}}|\mathbf{y}_{i,\text{obs}}, \mathbf{R}_i), \tag{3.3}$$

where subscript i indexes subject i, $\mathbf{Y}_{\text{obs}} = \cup_i^n \mathbf{y}_{i,\text{obs}}$ and $\mathbf{R} = \cup_{i=1}^n \mathbf{R}_i$.

The composite likelihood estimator of the point identifiable parameter, $\hat{\boldsymbol{\eta}}$, is obtained by maximizing the objective function in (3.3), namely $\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} l_c(\boldsymbol{\eta}|\mathbf{Y}_{\text{obs}}, \mathbf{R})$.

3.3.2 Likelihood Orthogonality

We first establish the likelihood orthogonality in the presence of misaligned missing data in Theorem III.1. We show that under the MAR mechanism, the complete-case composite likelihood estimation of $\boldsymbol{\eta}$ is not affected by the missing data mechanism. This is a well-known property in the classical maximum likelihood estimation (Rubin (1976)). This implies that unlike GEE, in the proposed complete-case composite likelihood estimation, IPW is not required for estimation consistency in the MAR mechanism, including the case of misaligned missing data pattern.

For one subject, denote $\mathbf{y}_{\text{mis}} = \{y_j, R_j = 0\}$, and for the entire sample, $\mathbf{Y}_{\text{mis}} = \cup_{i=1}^n \mathbf{y}_{i,\text{mis}}$.

Theorem III.1. (Validity) *Suppose that the MAR mechanism is governed by $f(\mathbf{R}|\mathbf{Y}_{\mathrm{obs}}, \phi)$, where ϕ is a parameter in the missing data mechanism and different from the parameters $\boldsymbol{\eta}$ in the data measurement mechanism. Then, the observed likelihood function for the random sample of n subjects is given by,*

$$L^{\mathrm{obs}}(\boldsymbol{\xi}, \phi|\mathbf{Y}_{\mathrm{obs}}, \mathbf{R}) = f(\mathbf{Y}_{\mathrm{obs}}|\boldsymbol{\eta})f(\mathbf{R}|\mathbf{Y}_{\mathrm{obs}}, \phi),$$

where $f(\mathbf{Y}_{\mathrm{obs}}|\boldsymbol{\eta}) = \prod_{i=1}^{n} f_{\mathrm{obs}}(\mathbf{y}_{i,\mathrm{obs}}|\boldsymbol{\eta}_{i,\mathrm{obs}})$, and $f(\mathbf{R}|\mathbf{Y}_{\mathrm{obs}}, \phi) = \prod_{i=1}^{n} f(\mathbf{R}_i|\mathbf{y}_{i,\mathrm{obs}}, \phi)$.

It is worth pointing out that this likelihood orthogonality is different form the classical result, due to the fact that the first factor $f(\mathbf{Y}_{\mathrm{obs}}, \boldsymbol{\eta})$ does not contain any unidentifiable parameters. Indeed, we here present a generalization of the classical observed likelihood of compete cases to overcome the hurdle of the misaligned missing data pattern. Moreover, this form of composite likelihood allows us to estimate parameter $\boldsymbol{\eta}$ and establish related large sample properties in the framework of composite likelihood estimation.

The result of Theorem III.1 holds because of the following arguments. By definition, the observed likelihood function takes the form,

$$\begin{aligned} L^{\mathrm{obs}}(\boldsymbol{\xi}, \phi|\mathbf{Y}_{\mathrm{obs}}, \mathbf{R}) &= \int f(\mathbf{Y}, \mathbf{R}|\boldsymbol{\xi}, \phi)\mathrm{d}\mathbf{Y}_{\mathrm{mis}} \\ &= \int \prod_{i=1}^{n} \left\{ f(\mathbf{y}_i|\boldsymbol{\xi})f(\mathbf{R}_i|\mathbf{y}_i, \phi)\mathrm{d}\mathbf{y}_{i,\mathrm{mis}} \right\}. \end{aligned}$$

The MAR mechanism implies that, for a subject, $f(\mathbf{R}|\mathbf{y}, \phi) = f(\mathbf{R}|\mathbf{y}_{\mathrm{obs}}, \phi)$.

Therefore,

$$\begin{aligned}
L^{\text{obs}}(\boldsymbol{\xi}, \phi|\mathbf{Y}, \mathbf{R}) &= \int \prod_{i=1}^{n} \{f(\mathbf{y}_{i,\text{obs}}, \mathbf{y}_{i,\text{mis}}|\boldsymbol{\xi}) f(\mathbf{R}_i|\mathbf{y}_{i,\text{obs}}, \phi) \mathrm{d}\mathbf{y}_{i,\text{mis}}\} \\
&= \int \prod_{i=1}^{n} \{f(\mathbf{y}_{i,\text{obs}}, \mathbf{y}_{i,\text{mis}}|\boldsymbol{\xi}) \mathrm{d}\mathbf{y}_{i,\text{mis}}\} \prod_{i=1}^{n} f(\mathbf{R}_i|\mathbf{y}_{i,\text{obs}}, \phi) \\
&= \prod_{i=1}^{n} \int f(\mathbf{y}_{i,\text{obs}}, \mathbf{y}_{i,\text{mis}}|\boldsymbol{\xi}) \mathrm{d}\mathbf{y}_{i,\text{mis}} f(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \phi) \\
&= \prod_{i=1}^{n} f_{\text{obs}}(\mathbf{y}_{i,\text{obs}}|\boldsymbol{\eta}_{i,\text{obs}}) f(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \phi) \\
&= f(\mathbf{Y}_{\text{obs}}|\boldsymbol{\eta}) f(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \phi).
\end{aligned}$$

Because the above likelihood orthogonality between $\boldsymbol{\eta}$ and ϕ, we can estimate the point identifiable parameters $\boldsymbol{\eta}$ simply using the first factor $f(\mathbf{Y}_{\text{obs}}|\boldsymbol{\eta}) = \prod_{i=1}^{n} f_{\text{obs}}(\mathbf{y}_{i,\text{obs}}|\boldsymbol{\eta}_{i,\text{obs}})$. This is the rationale for the formulation of complete-case composite likelihood as presented in equation (3.3).

Consequently, for the set of point identifiable parameters $\boldsymbol{\eta}$, the complete-case composite score function is given by,

$$\Psi(\mathbf{Y}, \mathbf{R}; \boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} l_c(\boldsymbol{\eta}|\mathbf{Y}_{\text{obs}}, \mathbf{R}) = \frac{\partial}{\partial \boldsymbol{\eta}} \ln f(\mathbf{Y}_{\text{obs}}|\boldsymbol{\eta}). \tag{3.4}$$

It follows from Theorem III.1 that it is easy to show that the function Ψ in equation (3.4) is an unbiased inference function, as stated in Theorem III.2.

Theorem III.2. (Unbiasedness)*Assume the full model $f(\mathbf{x}; \boldsymbol{\xi})$ is correctly specified at the true value $\boldsymbol{\xi}_0 = (\boldsymbol{\eta}_0, \boldsymbol{\zeta}_0)$. Let $\boldsymbol{\eta} \subset \boldsymbol{\xi}$ be the subset of point identifiable parameters. The complete-case composite score function Ψ is unbiased at the true values $\boldsymbol{\eta}_0$, i.e.,* $\mathrm{E}_{\boldsymbol{\eta}_0} \Psi(\mathbf{Y}, \mathbf{R}|\boldsymbol{\eta}_0) = \mathbf{0}$.

The proof of Theorem III.2 is given in the appendix. Because this property of unbiasedness, IPW is not required to establish estimation consistency for the parameters $\boldsymbol{\eta}$.

3.3.3 Large-Sample Properties

Large-sample properties may be easily established by following the theory of composite likelihood (Varin et al. (2011)), and the theory of inference function (Song (2007)). To make this paper self-contained, here we present three important properties, including consistency, asymptotic normality, and efficiency, which are essential for statistical inference.

Theorem III.3. (Consistency) *Under some mild regularity conditions (Def 3.5, Song (2007)), the complete-case composite likelihood method provides a consistent estimator, i.e.,* $\boldsymbol{\eta}_n \xrightarrow{p} \boldsymbol{\eta}_0$*, under* $p_{\boldsymbol{\eta}_0}$. □

Applying Theorem 3.8 in Song (2007), we establish the following theorem.

Theorem III.4. (Asymptotic Normality) *Under some mild regularity conditions (Def 3.5, Song (2007)), the asymptotic normality holds for the complete-case composite likelihood estimator* $\hat{\boldsymbol{\eta}}$*, and the asymptotic variance of* $\hat{\boldsymbol{\eta}}$ *is the inverse of Godambe information* $\mathbf{G}$*, i.e.,*

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{G}^{-1}(\boldsymbol{\eta}_0)), \text{under } p_{\boldsymbol{\eta}_0}, \tag{3.5}$$

where $\mathbf{G}(\boldsymbol{\eta}) = \mathbf{H}(\boldsymbol{\eta})^T \mathbf{J}(\boldsymbol{\eta})^{-1} \mathbf{H}(\boldsymbol{\eta})$*, with sensitivity matrix* $\mathbf{H}(\boldsymbol{\eta}) = \mathrm{E}\left\{-\nabla_{\boldsymbol{\eta}} \Psi(\mathbf{Y}, \mathbf{R}; \boldsymbol{\eta})\right\}$*, and variability matrix* $\mathbf{J}(\boldsymbol{\eta}) = \mathrm{var}\left\{\Psi(\mathbf{Y}, \mathbf{R}; \boldsymbol{\eta})\right\}$ *(Godambe (1960)).*

The asymptotic variance matrix of $\hat{\boldsymbol{\eta}}$ is consistently estimated by

$$\hat{\mathrm{var}}(\hat{\boldsymbol{\eta}}) = \frac{1}{n}\hat{\mathbf{G}}^{-1}(\hat{\boldsymbol{\eta}}) = \frac{1}{n}\hat{\mathbf{H}}^{-1}(\hat{\boldsymbol{\eta}})\hat{\mathbf{J}}(\hat{\boldsymbol{\eta}})\hat{\mathbf{H}}^{-1}(\hat{\boldsymbol{\eta}}), \tag{3.6}$$

with $\mathbf{H}$ and $\mathbf{J}$ being consistently estimated by,

$$\hat{\mathbf{H}}(\hat{\boldsymbol{\eta}}) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{s \in S} w_s \frac{\partial^2 \log f_s(\mathbf{y}_{i,s}|\boldsymbol{\xi}_s)}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T}\right\}, \tag{3.7}$$

and

$$\hat{\mathbf{J}}(\hat{\boldsymbol{\eta}}) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{s\in S} w_s \frac{\partial \log f_s(\mathbf{y}_{i,s}|\boldsymbol{\xi}_s)}{\partial \boldsymbol{\eta}}\right)\left(\sum_{s\in S} w_s \frac{\partial \log f_s(\mathbf{y}_{i,s}|\boldsymbol{\xi}_s)}{\partial \boldsymbol{\eta}}\right)^T. \tag{3.8}$$

Numerically, using an *R* function *hessian* in *R* package *numDeriv*, we can calculate this Hessian matrix easily, while using an *R* function *grad* in *R* package *numDeriv*, we can also compute the gradients conveniently.

Furthermore, we can establish the full asymptotic efficiency of the complete-case composite likelihood in Theorem III.5.

Theorem III.5. (Barlett Identity) *For the point identifiable parameters estimated from the complete-case composite likelihood function, under some mild regularity conditions (Def 3.5, Song (2007)), the sensitivity matrix and the variability matrix are equal, i.e.,* $\mathbf{H}(\boldsymbol{\eta}) = \mathbf{J}(\boldsymbol{\eta})$.

It follows from Theorem III.5, that the Godambe information $\boldsymbol{\Gamma}(\boldsymbol{\eta}) = \mathbf{H}(\boldsymbol{\eta})$, so the proposed $\hat{\boldsymbol{\eta}}$ achieves the full efficiency. According to Song et al. (2005), the sandwich form given by the Godambe information matrix demonstrates desirable numerical stability than either the sensitivity matrix or the variability matrix, and has been recommended for practical use, especially when both matrices are involved numerical derivatives as done by two R packages. Therefore, in this paper, we follow the recommendation and use the sandwich form to obtain model-based standard errors.

3.3.4 Estimation of Partially Identifiable Parameters

For a correctly specified model, the set of partially identifiable parameters $\boldsymbol{\zeta}$ may be restricted within certain bounds over which the assumed model is valid. For example, Fan and Zhu (2009) provided certain sharp bounds for the pairwise correlation parameter in a bivariate copula model. In this paper, we consider a more general setting of a d-dimension multivariate model and estimate bounds of partially identifiable correlation parameter under the misaligned missing data pattern.

For the ease of exposition, we first consider a simple case of a correlation parameter, say, for the first and second margins. Denote the correlation matrix of $\mathbf{y}$ by $\mathbf{\Gamma} = \text{var}(\mathbf{y}) = (\gamma_{ij})_{d\times d}$; then γ_{12} is the target correlation parameter. Because γ_{12} is not point identifiable, it is not involved in the complete-case composite likelihood above. To estimate the upper and lower bounds for γ_{12}, we utilize the constraint of non-negative definiteness for a correlation matrix; that is, determinant $|\mathbf{\Gamma}| \geq 0$. This implies that the upper and lower bounds for γ_{12} are given by,

$$(3.9)\quad \gamma_{12}^{U} = \mathbf{\Gamma}_{3:d,1}^{T}\mathbf{\Gamma}_{3:d,3:d}^{-1}\mathbf{\Gamma}_{3:d,2} + \left(1 - \mathbf{\Gamma}_{3:d,1}^{T}\mathbf{\Gamma}_{3:d,3:d}^{-1}\mathbf{\Gamma}_{3:d,1}\right)^{\frac{1}{2}}\left(1 - \mathbf{\Gamma}_{3:d,2}^{T}\mathbf{\Gamma}_{3:d,3:d}^{-1}\mathbf{\Gamma}_{3:d,2}\right)^{\frac{1}{2}},$$

and

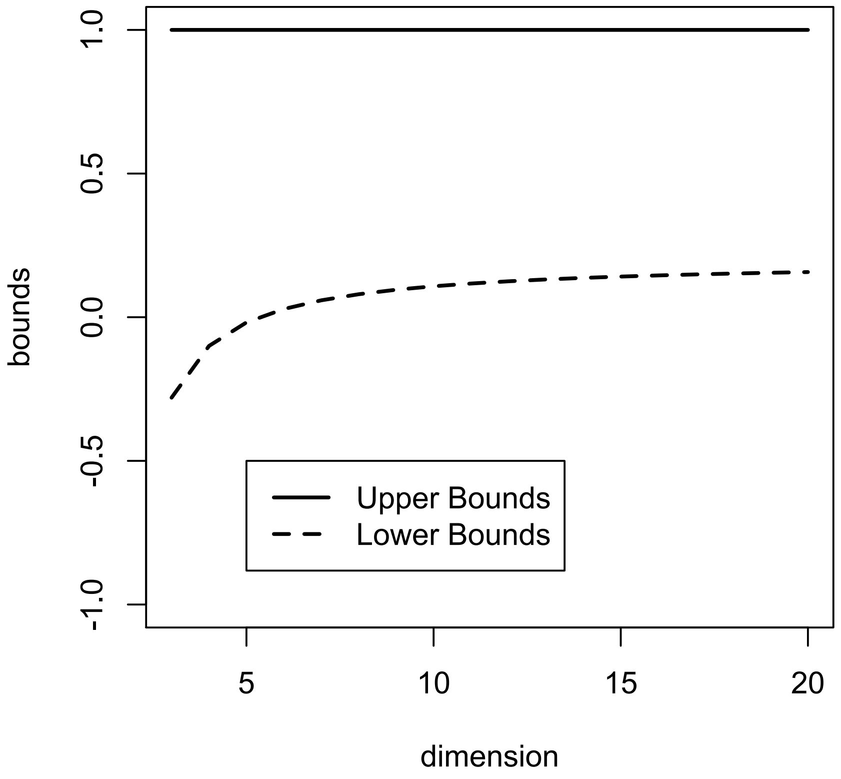
$$(3.10)\quad \gamma_{12}^{L} = \mathbf{\Gamma}_{3:d,1}^{T}\mathbf{\Gamma}_{3:d,3:d}^{-1}\mathbf{\Gamma}_{3:d,2} - \left(1 - \mathbf{\Gamma}_{3:d,1}^{T}\mathbf{\Gamma}_{3:d,3:d}^{-1}\mathbf{\Gamma}_{3:d,1}\right)^{\frac{1}{2}}\left(1 - \mathbf{\Gamma}_{3:d,2}^{T}\mathbf{\Gamma}_{3:d,3:d}^{-1}\mathbf{\Gamma}_{3:d,2}\right)^{\frac{1}{2}},$$

where $a : b$ denotes column a to column b.

It is important to note that both bounds, respectively, are functions of the other entries of the correlation matrix, all of which are supposedly point identifiable. To help visualize the bounds, an example is provided in Figure 3.1. It illustrates the upper and lower bounds of a partially identifiable parameter γ_{12} for a $d \times d$ exchangeable correlation matrix, $\mathbf{\Gamma}$ where all correlation parameters, except γ_{12} and γ_{21} are 0.6. The dimension d varies from 3 to 20. It is easy to see that as the dimension increases, the parameter range becomes narrower.

Generalizing the above idea, we propose a general method to estimate bounds of two or more partially identifiable correlation parameters. For example, if in addition to γ_{12}, $\gamma_{d-1,d}$ is also partially identifiable. Denote $\mathbf{\Gamma}_{-k,-k}$, $k = 1, \cdots, d$, as the submatrix of $\mathbf{\Gamma}$ with both the k-th column and the k-th row deleted. Applying the

Figure 3.1: Upper and Lower Bounds for a Partially Identifiable Correlation Parameter



procedure (3.9) and (3.10) to two submatrices $\mathbf{\Gamma}_{-d,-d}$ and $\mathbf{\Gamma}_{-(d-1),-(d-1)}$, both of which have no involvement of $\gamma_{d-1,d}$, we obtain two pairs of bounds for γ_{12}, denoted by $(\gamma^L_{12(d)}, \gamma^U_{12(d)})$ and $(\gamma^L_{12(d-1)}, \gamma^U_{12(d-1)})$, respectively. Then, the resulting bounds are $\gamma^L_{12} = \max(\gamma^L_{12(d)}, \gamma^L_{12(d-1)})$, and $\gamma^U_{12} = \min(\gamma^U_{12(d)}, \gamma^U_{12(d-1)})$. In the same way, we can obtain bounds $\gamma^L_{d-1,d}$ and $\gamma^U_{d-1,d}$ for $\gamma_{d-1,d}$. Furthermore, for any $\gamma_{12} \in (\gamma^L_{12}, \gamma^U_{12})$, applying the procedure discussed above, we can obtain even narrower bounds for $\gamma_{d-1,d}$, vice versa for γ_{12}.

To discuss the issue of inference for the bounds, we again start with a simple case of one partially identifiable parameter γ_{12}. Noting that the estimated bounds $\hat{\gamma}^U_{12}$ and $\hat{\gamma}^L_{12}$ are functions of all points identifiable parameters given by matrix $\mathbf{\Gamma}$. Denote $\mathbf{\Gamma}_{-12}$ as the vector of elements from the lower matrix of $\mathbf{\Gamma}$ without γ_{12}, and the asymptotic variances of $\mathbf{\Gamma}_{-12}$ are corresponding submatrix of $\dfrac{\hat{\mathbf{G}}^{-1}(\hat{\boldsymbol{\eta}})}{n}$. Applying the multivariate delta's method, we can derive the asymptotic variances as $\hat{\text{var}}(\hat{\gamma}_{12}^U) = \nabla\gamma_{12}^U(\hat{\mathbf{\Gamma}})^T\text{var}(\hat{\mathbf{\Gamma}}_{-12})\nabla\gamma_{12}^U(\hat{\mathbf{\Gamma}})$, and $\hat{\text{var}}(\hat{\gamma}_{12}^L) = \nabla\gamma_{12}^L(\hat{\mathbf{\Gamma}})^T\text{var}(\hat{\mathbf{\Gamma}}_{-12})\nabla\gamma_{12}^L(\hat{\mathbf{\Gamma}})$, where $\nabla\gamma_{12}^U(\mathbf{\Gamma})$ and $\nabla\gamma_{12}^L(\mathbf{\Gamma})$ are the gradient vectors of γ_{12}^U and γ_{12}^L with respect to the point identifiable elements in $\mathbf{\Gamma}$.

For models with two or more partially identifiable parameters, the derivation of the delta's method may become rather tedious in the general setting. At this moment, we carry out only need-based derivations in a given problem under investigation.

3.4 Implementation

Peeling algorithm is developed to obtain estimates of point identifiable parameters $\boldsymbol{\eta}$ that maximize the complete-case composite likelihood. This algorithm proceeds over a sequence of iterative steps on multiple subsets of parameters. Suppose we partition $\boldsymbol{\eta}$ into g disjoint groups, say, $(\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_g)$. Maximizing the complete-case composite likelihood with respect to $\boldsymbol{\eta}$ is carried out by sequentially updating each subset of parameters, from $\boldsymbol{\eta}_1$ to $\boldsymbol{\eta}_g$.

Suppose iteration t has been completed, and $\boldsymbol{\eta}^{(t)}$ is available. At the $(t+1)$-th iteration, given that the first $(j-1)$ subgroups of $\boldsymbol{\eta}$ have been updated, updating the jth subset $\boldsymbol{\eta}_j$ is obtained by maximizing the following the complete-case composite likelihood,

$$\boldsymbol{\eta}_j^{(t+1)} = \arg\max_{\boldsymbol{\eta}_j} \sum_{i=1}^{n} l_{c,i}\left(\boldsymbol{\eta}_1^{(t+1)}, \cdots, \boldsymbol{\eta}_{j-1}^{(t+1)}, \boldsymbol{\eta}_j, \boldsymbol{\eta}_{j+1}^{(t)}, \cdots, \boldsymbol{\eta}_g^{(t)} | \mathbf{Y}_{\text{obs}}, \mathbf{R}\right),$$

where $l_{c,i}$ is given in equation (3.3).

The above optimization may be done numerically using a quasi-Newton optimization routine available in R function *nlm*. This step of optimization is computationally fast as it involves only a set of low-dimensional parameters vector $\boldsymbol{\eta}_j$. At the completion of iteration $(t+1)$, g subsets of parameters are updated, denoted as $\boldsymbol{\eta}^{(t+1)}$.

After the peeling algorithm converges, the upper and lower bounds for each partially identifiable parameter in $\boldsymbol{\zeta}$ will be estimated according to suitable formulas determined in a given problem. For example, the bounds of γ_{12} are calculated through equations (3.9) and (3.10). Fill γ_{12} by its upper or lower bound results in the two estimated correlation matrices, denoted by $\boldsymbol{\Gamma}^U$ and $\boldsymbol{\Gamma}^L$, respectively.

3.5 Gaussian Copula Regression Model

We now apply the proposed method to an important example of Gaussian copula regression model with location-scale family marginal model. This development provides a needed preparation for the analysis of the quality-of-life data in the motivating example introduced in Section 3.2. A Gaussian copula regression model consists of two components: marginal regression model and Gaussian copula dependence model, both of which are presented in detail below.

3.5.1 Location-Scale Family Distribution Marginal Model

Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_d)^T$, where each $\boldsymbol{\theta}_k$ denotes a set of marginal parameters associated with the $k^{th}(k = 1, \cdots, d)$ marginal density function, $f_k(y_k|\boldsymbol{\theta}_k)$. Denote by $u_k = F_k(y_k|\boldsymbol{\theta}_k)$ the marginal cumulative distribution function(CDF) corresponding to the k^{th} margin, where F_k is a location-scale family distribution parametrized by a location parameter μ_k and a positive scale parameter σ_k, $\boldsymbol{\theta}_k = (\mu_k, \sigma_k)$. More specifically, the marginal location-scale density function is given by

$$f_k(y_k|\boldsymbol{\theta}_k) = \frac{1}{\sigma_k}\tilde{f}\left(\frac{y_k - \mu_k}{\sigma_k}\right), k = 1, \cdots, d, \tag{3.11}$$

where $\tilde{f}(\cdot)$ is the standard kernel density with $\int_R y\tilde{f}(y)dy = 0$.

To specify a marginal regression model, let $\mathbf{X}_i = (1, \mathbf{x}_i^T)^T$, $i = 1, \cdots, n$. For the k^{th} margin, the linear model is imposed on the location parameter in equation (3.11), $\mu_{ik} = \mathrm{E}(y_{ik}|\mathbf{X}_i) = h(\mathbf{X}_i^T \beta_k), k = 1, \cdots, d$, where $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \cdots, \beta_{kp})^T$ is a $(p+1)$-element unknown regression vector, and h is a known link function. For convenience, denote the resulting model by $Y_{ik} \sim F_k(y_k|\mu_{ik}(\boldsymbol{\beta}_k), \sigma_k)$. As an important special case, we may consider $p = 0$ (no covariates), namely $\mu_{ik} = h(\beta_{k0})$ is a common location parameter for all subjects $i = 1, \cdots, n$.

3.5.2 Gaussian Copula

A copula is a multivariate probability distribution in which the marginal probability distribution of each variable is uniform on $(0,1)$. According to Sklar's theorem (Sklar (1959)), a multivariate cumulative distribution function of a continuous random vector $\mathbf{Y} = (y_1, y_2, \cdots, y_d)^T$ with marginals $F_k(y_k)$ can be written as $F(y_1, \ldots, y_d) = C\left(F_1(y_1), \ldots, F_d(y_d)\right)$, where C is a suitable copula. In this paper, we assume that $\mathbf{Y}$ follows the d-dimensional distribution generated by a Gaussian copula (Song (2000)). The d-variate density function of $\mathbf{Y}$ is given by

$$f(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Gamma}) = c(\mathbf{u}|\boldsymbol{\Gamma}) \prod_{k=1}^{d} f_k(y_k|\boldsymbol{\theta}_k), \ \mathbf{u} = (u_1, u_2, \cdots, u_d)^T \in [0,1]^d, \tag{3.12}$$

with Gaussian copula function $c(\cdot|\boldsymbol{\Gamma})$ of the following form:

$$c(\mathbf{u}|\boldsymbol{\Gamma}) = |\boldsymbol{\Gamma}|^{-\frac{1}{2}} \exp\left\{\frac{1}{2}\mathbf{Q}(\mathbf{u})^T(\mathbf{I} - \boldsymbol{\Gamma}^{-1})\mathbf{Q}(\mathbf{u})\right\}, \ \mathbf{u} \in [0,1]^d, \tag{3.13}$$

where $\boldsymbol{\Gamma} = (\gamma_{k_1 k_2})_{d\times d}$ is the Pearson correlation matrix of normal quantiles $\mathbf{Q}(\mathbf{u}) = (q_1(u_1), \cdots, q_d(u_d))^T$, and $q_k = q_k(u_k) = \Phi^{-1}(u_k)$ is the k^{th} marginal normal quantile, $k = 1, \cdots, d$. Here Φ denotes the univariate CDF of the standard normal distribution, and I is the $d \times d$ identity matrix, and $|\cdot|$ denotes the determinant of a matrix. Marginally, $u_k \sim \mathrm{Unif}(0,1)$, and $q_k \sim \mathrm{N}(0,1)$. When all margins y_k are normal distributed, matrix $\mathbf{\Gamma}$ gives the Pearson correlation matrix of $\mathbf{Y}$; otherwise, $\mathbf{\Gamma}$ represents a matrix of pairwise rank-based correlations (e.g., Kendall's τ or Spearman's ρ).

3.5.3 Complete-Case Composite Likelihood

In the presence of misaligned missing data pattern, the complete-case composite likelihood in equation (3.1), for one subject, may be written as follows,

$$L_c^{\text{obs}}(\boldsymbol{\xi}|\mathbf{y},\mathbf{R}) = \prod_{s\in S} f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)^{w_s} = \prod_{s\in S}\left\{c(\mathbf{u}_s|\mathbf{\Gamma}_s)\prod_{k\in s} f_k(y_k|\boldsymbol{\theta}_k)\right\}^{w_s}, \tag{3.14}$$

where $c(\mathbf{u}_s|\mathbf{\Gamma}_s) = |\mathbf{\Gamma}_s|^{-\frac{1}{2}}\exp\left\{\frac{1}{2}\mathbf{q}_s^T(\mathbf{I}-\mathbf{\Gamma}_s^{-1})\mathbf{q}_s\right\}$ is the density function under Gaussian copula for the corresponding set s due to the property of marginal closure.

To estimate $(\boldsymbol{\theta},\mathbf{\Gamma})$ by using the complete-case composite likelihood of a random sample of n subjects, for each subject $i = 1,\cdots,n$ denote $\mathbf{u}_{\text{obs}} = \{u_k, k\in D_{\text{obs}}\}$, and $\mathbf{\Gamma}_{\text{obs}} = \{\boldsymbol{\gamma}_{k_1k_2}, k_1, k_2 \in D_{\text{obs}}\}$. The marginal density of Gaussian copula is given by, for $\mathbf{u}_{\text{obs}} \in [0,1]^{|D_{\text{obs}}|}$ is given by,

$$c(\mathbf{u}_{\text{obs}}|\mathbf{\Gamma}_{\text{obs}}) = |\mathbf{\Gamma}_{\text{obs}}|^{-\frac{1}{2}}\text{exp}\left\{\frac{1}{2}\mathbf{q}_{\text{obs}}(\mathbf{u}_{\text{obs}})^T\left(\mathrm{I}-\mathbf{\Gamma}_{\text{obs}}^{-1}\right)\mathbf{q}_{\text{obs}}(\mathbf{u}_{\text{obs}})\right\}, \tag{3.15}$$

and the logarithm of the complete-case composite likelihood function for one subject is,

$$l(\boldsymbol{\eta}|\mathbf{y}_{\text{obs}}) = \sum_{j\in D_{\text{obs}}}\text{ln}f_k(y_k|\boldsymbol{\theta}_k) - \frac{1}{2}\text{ln}|\mathbf{\Gamma}_{\text{obs}}| + \frac{1}{2}\mathbf{q}_{\text{obs}}(\mathbf{u}_{\text{obs}})^T\left(\mathrm{I}-\mathbf{\Gamma}_{\text{obs}}^{-1}\right)\mathbf{q}_{\text{obs}}(\mathbf{u}_{\text{obs}}). \tag{3.16}$$

Denote $\mathbf{A} = \mathbf{\Gamma}^{-1}$ as the precision matrix. The complete-case composite score function is given by,

$$\Psi(\mathbf{Y},\mathbf{R};\boldsymbol{\eta}) = \frac{\partial}{\partial(\boldsymbol{\theta},A)} l(\boldsymbol{\eta}|\mathbf{Y}_{\text{obs}},\mathbf{R}). \tag{3.17}$$

Proposition III.6. *(***Uniqueness***) Assume the Gaussian copula regression model is correctly specified. Then, the complete-case composite score function in equation (3.17) has a unique zero at the true values,* $(\boldsymbol{\theta}_0,\mathbf{\Gamma}_0)$, *i.e.,* $\mathrm{E}_{(\boldsymbol{\theta}_0,\mathbf{\Gamma}_0)}\Psi(\mathbf{Y},\mathbf{R}|\boldsymbol{\theta}_0,\mathbf{\Gamma}_0) = \mathbf{0}$.

This uniqueness property ensures that the complete-case composite likelihood estimator $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Gamma}})$ converges to the true values in probability.

We like to comment that the Fan and Zhu (2009)'s sharp bounds are actually wider than ours. According to Fan and Zhu (2009) theory, the lower and upper bounds of correlation parameter γ_{k_1,k_2} are given by $\gamma^L_{k_1,k_2} = \left[\int_0^1 \left\{F_{k_1}^{-1}(u|\theta_{k_1})F_{k_2}^{-1}(1-u|\theta_{k_2})\right\} du - \mu_{k_1}\mu_{k_2}\right] / \sigma_{k_1}\sigma$ and $\gamma^U_{k_1,k_2} = \left[\int_0^1 \left\{F_{k_1}^{-1}(u|\theta_{k_1})F_{k_2}^{-1}(u|\theta_{k_2})\right\} du - \mu_{k_1}\mu_{k_2}\right] / \sigma_{k_1}\sigma_{k_2}$. By some routine calculations, we find that in the Gaussian copula framework, the upper and lower bound are equal to -1 and 1, respectively.

3.5.4 Implementation

Below are some details of the peeling algorithm to update parameters in both location-scale family marginal model and Gaussian copula.

Step 1: To Update Marginal Parameters

Suppose iteration t has been completed, for a specific marginal parameter $\boldsymbol{\theta}_k, k = 1, \cdots, d$, we first denote $\mathbf{q}_{\text{obs}\backslash k} = \{q_{k'}, k' \in D_{\text{obs}}, k' \neq k\}$, which is the subvector of $\mathbf{q}_{\text{obs}}$ without the component q_k. Let $\mathbf{q}_{\text{obs}}^{(k-1|t+1)}$ be the vector where the elements $k' < k$ have been updated in the $(t+1)^{th}$ iteration where the elements $k' > k$ have been updated in the t^{th} iteration, and $\mathbf{q}_{\text{obs}\backslash k}^{(k-1|t+1)}$ be the subvector of $\mathbf{q}_{\text{obs}}^{(k-1|t+1)}$ excluding q_k. Thus, $\mathbf{q}_{\text{obs}}^{(t+1)} = \mathbf{q}_{\text{obs}}^{(d|t+1)} = \mathbf{q}_{\text{obs}}^{(0|t+2)}$. Therefore, the corresponding partial log complete-case composite likelihood may be written as, for $k = 1, \cdots, d$,

$$\boldsymbol{\theta}_k^{(t+1)} = \arg\max_{\boldsymbol{\theta}_k} \left\{ \sum_{i=1}^{n} \mathrm{I}\{R_{i,k} = 1\} \left[\ln f_k(y_{i,k}|\boldsymbol{\theta}_k) + \frac{1}{2} q_{i,k}^2 - (\mathbf{q}_{i,\text{obs}\backslash k}^{(k-1|t+1)})^T (\boldsymbol{\Gamma}_{i,\text{obs}\backslash k}^{(t)})^{-1} \mathbf{q}_{i,\text{obs}\backslash k}^{(k-1|t+1)} \right] \right\}.$$

Note that when $R_{ik} = 0$ and y_{ik} is missing, $l_{c,ik} = 0$.

After $\boldsymbol{\theta}^{(t+1)_k}$ have been updated, we obtain $u_{ik}^{(t+1)} = F_k(y_{ik}|\boldsymbol{\theta}_k^{(t+1)})$ and $q_{ik}^{(t+1)} = \Phi^{-1}(u_{ik}^{(t+1)})$, for $k = 1, \cdots, d$, and $i = 1, \cdots, n$. So the iteration $(t+1)$ for marginal parameters has been completed.

Step 2: To Update Point-Identifiable Correlation Parameters

To update the point identifiable correlation parameters,

$$\hat{\Gamma} = \arg\max_{\Gamma} \left\{ \sum_{i=1}^{n} \left[-\frac{1}{2}\ln|\mathbf{\Gamma}_{\text{obs}}| - \frac{1}{2}(\mathbf{q}_{i,\text{obs}}^{(t+1)})^T \mathbf{\Gamma}_{i,\text{obs}}^{-1} \mathbf{q}_{i,\text{obs}}^{(t+1)} \right] \right\},$$

when $\Gamma_{k_1 k_2}$ is partially identifiable, temporarily put a zero in the entry.

Step 3: To Estimate Partially Identifiable Correlation Parameters

Follow the instruction in Section 3.3.4.

3.6 Simulation Experiments

We conduct extensive simulation experiments to evaluate the performance of the complete-case composite likelihood method, and to compare it with the classical EM algorithm and imputation methods. In the first two experiments, the dimension of outcomes is set as $d = 3$, and one or two outcomes may be subject to missing (i.e., $|D_{\text{obs}}| = 1$ or 2) across subjects. In the third experiment, the dimension of outcomes is set as $d = 5$, in which two or three outcomes may be missing (i.e., $|D_{\text{obs}}| = 2$ or 3) across subjects, so that the misaligned missingness occurs in two pairs of outcomes.

3.6.1 Three-Dimensional Linear Regression Model

The setup of the first simulation is similar to the motivating example of 3-dimensional QoL scores, in which a pair of the outcomes is misaligned missing. In the scenario of 3-dimensional linear regression model, there is a partially identifiable correlation parameter for the first and second margins, while the third margin is set only under MAR. In the marginal regression models, we include two covariates $X_1 \sim$ Bin(1,0.5) and $X_2 \sim$N(0,1) in the linear model with $\mu_k = \mathbf{X}^T\boldsymbol{\beta}_k, k = 1, 2, 3$ and generate residuals from a tri-variate normal $N_3(\mathbf{0}, \mathbf{\Gamma})$ with the standard normal marginal N(0,1) and an unstructured correlation matrix $\mathbf{\Gamma}$ with parameters $(\gamma_{12}, \gamma_{13}, \gamma_{23}) = (0.5, 0.4, 0.3)$.

The following missing mechanism is set for the first margin:

$$P(R_1 = 0|X_1, Y_1) = \begin{cases} 0.3, \text{ if } X_1 = 1; \\ 0.42, \text{ if } X_1 = 0, \text{ and } Y_1 > \mu_1; \\ 0.18, \text{ if } X_1 = 0, \text{ and } Y_1 < \mu_1. \end{cases}$$

For the second margin, set $R_2 = 1 - R_1$ to generate the misaligned missingness between the first two outcomes. The above missing data model leads to on average 30% missing in the first variable and 70% missing in the second variable. MAR for the third variable is given by $P(R_3 = 0|X_2) = 0.2\Phi(X_2)$ leading to on average 10% missing.

The sample size n is fixed at 200, we run 1000 replicates to draw summary statistics. We compare the results obtained from the complete-case composite likelihood with the peeling algorithm to those obtained from the gold standard of the maximum likelihood estimation using the full data, the maximum likelihood estimation via EM algorithm, and the method of multiple imputations. In comparison, two types of standard errors are reported: the first type is the empirical standard error in the three methods under the misaligned missing data, and the other type is the average of 1000 model-based standard errors obtained from the inverse information matrix in equation (3.6) in the proposed method, from the Louis's formula in the EM algorithm, and from Raghunathan (2004)'s method in the multiple imputation.

As shown in Table 3.1, in the presence of misaligned missing data, the proposed complete-case composite likelihood method outperforms the EM algorithm. Taking a close look at the estimation results for the parameters of the second regression model where on average 70% of outcome observations are missing. The bias for the estimation of the scale parameter σ_2 is not ignorable and the difference between the empirical and the average model-based standard errors is large. The reason that the MLE with the EM algorithm performs poorly is that there is no information shared between the first two variables in the updating of γ_{12}, and such estimation instability propagates great damage on the estimation of the other two correlation parameters and standard error estimation. The multiple imputation method performs well for the estimation of the marginal regression parameters, but badly for the estimation of γ_{12} with large bias. The gap between the empirical standard error and the average model-based standard error is so substantial that inference for the parameters in the second marginal model would be problematic. Such large gap in the standard errors may be attributed to the fact that the method proposed in Raghunathan (2004) is based on marginal univariate regression model, rather than on a multivariate analysis. The same issue exacerbates in estimation of the second order moment parameters, such as scale parameters and correlation parameters, from the EM algorithm. This is because in the EM algorithm assumes that the full model is available, in which borrowing information across margins is essential. The existence of any partially identifiable parameters makes the sharing of information impossible. So, the results from both the EM algorithm and the multiple imputation method appear strongly biased and cannot be trusted in real world applications.

The estimates of partially identifiable parameter γ_{12} provided by the EM algorithm and the multiple imputation method are incorrect with excessively large biases. In contrast, our method provides an estimated range $(-0.7543, 0.9943) \subset (-1, 1)$ as well as associated standard errors for both lower and upper bound estimates. The data information of the third margin does help us to reach a narrower parameter range than $(-1, 1)$.

3.6.2 Integration of Four Data Sources

In the second simulation experiment, we aim to examine the performance of the proposed method to deal with missing data when multiple data sources are combined. We consider integrate 4 data sets, each of which is generated from a tri-variate linear regression model with normally distributed errors, similar to the model used in the first simulation experiment in Section 3.6.1. Simulation setup is given as follows. For the first data set, three marginal outcomes are fully observed; for the second data set, the first margin is not measured (fully missing), but the second and the third margins are fully observed; for the third group, only the first margin is fully observed, but the second and the third margins are not measured (fully missing); for the fourth data set, 30% of subjects have missing values at random, according to the following missing mechanism,

$$P(R_0 = 0|X_1, Y_1) = \begin{cases} 0.3, \text{ if } X_1 = 1; \\ 0.42, \text{ if } X_1 = 0, \text{ and } Y_1 > \mu_1; \\ 0.18, \text{ if } X_1 = 0, \text{ and } Y_1 < \mu_1, \end{cases}$$

where R_0 is a missing indicator for a subject, and when $R_0 = 0$, one or two margins are randomly missing with equal probability.

In each data set the 3-dimensional regression model is specified in the same form as that given in the first simulation study in Section 3.6.1. The sample size is for each data source is fixed at 100, so the total sample size of the integrated data is 400. We run 1000 replicates to draw summary statistics. We compare the results obtained from the complete-case composite likelihood with the peeling algorithm with those obtained from the gold standard using the full data, from the EM algorithm, from the multiple Imputation method, and from the hot-deck imputation method. The two imputation methods are implemented under the default settings of R packages *MI* and $HotDeckImputation$, respectively. Similar to the first simulation experiment in Section 3.6.1 our comparison on the two types of standard errors are also included.

As shown in Table 3.2, both estimates and standard errors obtained from complete-case composite likelihood method and the EM algorithm are all close to the true values. This is because in this data integration, the misaligned missingness in one data source disappears when the multiple data sets are combined. Also, the standard errors from the complete-case composite likelihood and the EM algorithm appear to be similar. This is due to the fact that under MAR both the MLE and the complete-case composite likelihood estimation are fully efficient (refer to Theorem III.5). For the multiple imputations methods, noticeable differences appear between the empirical standard errors and the average model-based standard errors for marginal regression parameters remain in this simulation study. Overall, the complete-case composite likelihood and the EM algorithm outperform the two imputation methods, judged by both bias and standard errors.

3.6.3 Estimation with Two Partially Identifiable Parameters

The third simulation study aims to provide additional numerical evidence on the proposed complete-case composite likelihood estimation when there are two partially identifiable correlation parameters. This is a more complicated scenario than the first simulation that contains only one partially identifiable parameter. Here we consider a five-dimensional linear regression model with normally distributed errors, where two pairs of marginal outcomes are subject to misaligned missing. We adopt the same univariate regression model with two covariates as that specified in the first simulation study, and the correlation matrix $\mathbf{\Gamma}$ for the 5-dimensional correlated errors is specified as follows:

$$\begin{pmatrix} 1 & 0.6 & 0.5 & 0.4 & 0.3 \\ 0.6 & 1 & 0.5 & 0.45 & 0.3 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.4 & 0.45 & 0.5 & 1 & 0.6 \\ 0.3 & 0.3 & 0.5 & 0.6 & 1 \end{pmatrix}.$$

The misaligned missingness between the first two margins is generated by the probability model given in Section 3.6.1, and the misaligned missingness for the last two margins is generated according to the following probability model: for the fourth margin, we set

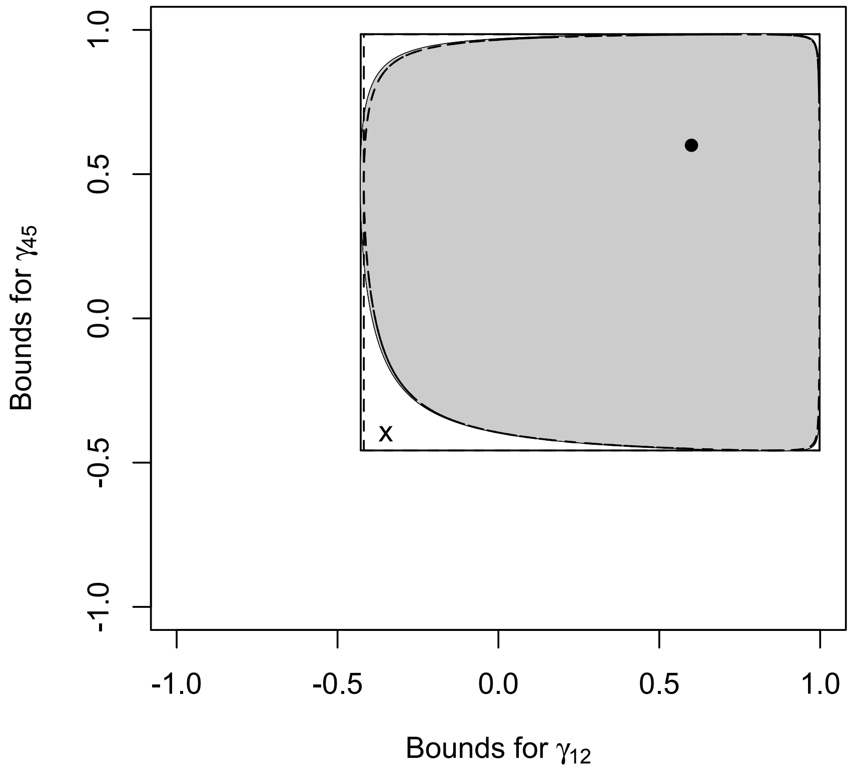
$$P(R_4 = 0|X_2, Y_4) = \begin{cases} 0.4, \text{ if } X_2 > 0; \\ 0.48, \text{ if } X_2 < 0, \text{ and } Y_4 > \mu_4; \\ 0.32, \text{ if } X_2 < 0, \text{ and } Y_4 < \mu_4; \end{cases}$$

and for the last margin, set $R_5 = 1 - R_4$. The above missing data model gives rise on average 40% missingness in the fourth variable and 70% missingness in the fifth variable. In this setting, both correlation parameters γ_{12} and γ_{45} are not point identifiable.

The sample size is fixed at 200, and we run 1000 replicates to draw summary statistics. We compare the complete-case composite likelihood to the gold standard of the maximum likelihood estimation using the full data in terms of both estimation bias and standard error.

Table 3.3 indicates that the estimates of point identifiable parameters from the complete-case composite likelihood method are close to the true values, with slightly larger standard errors than those given by the golden standard. This is not a surprise because the complete-case composite likelihood estimation utilizes the reduced

Figure 3.2: Bounds for Partial Identifiable Parameters



sample based only on the set of observed data. Also, a high agreement between the empirical standard error and the average model-based standard error is evident across all parameters. For the two partially identifiable parameters, their respective upper and lower bounds for γ_{12} and γ_{45} are calculated by the method discussed in Section 3.3.4 and displayed in Figure 3.2. In this figure the solid boundaries of the rectangle represent their individual upper and lower bounds for γ_{12} and γ_{45}, respectively, and the dashed lines are the true the one-dimensional parameter range. The black dot represents the true values of these two correlation parameters. The gray area shows the joint range of the two parameters, which is slightly smaller than the rectangle area because of some internal constraint between γ_{12} and γ_{45}. In other words, with a parameter value indicated by x in the figure, the correlation matrix $\mathbf{\Gamma}$

is not necessarily positive definite.

3.7 PROMIS Data Analysis

PROMIS data was introduced in Section 3.2 as a motivating example. Here, we present the relevant detail concerning the joint regression analysis of three QoL scores domains (pain, fatigue, and anxiety) where pain and fatigue are misaligned missing. Out of 224 subjects, 108 subjects have measurements of pain but no measurements of fatigue, while the other 116 subjects have measurements of fatigue but no measurements of pain. In addition, two subjects have neither measurements of pain nor measurements of fatigue. Interestingly, measurements of anxiety have been fully recorded on all 224 individuals with no missing data.

We first conduct the complete-case univariate analysis of each QoL domain score (y_1 = Pain, y_2 = Fatigue, y_3 = Anxiety) on covariates of age (x_1), gender (x_2), edema (x_3), race (white, x_4; black, x_5; and other as reference). Next, we consider a three-dimensional Gaussian copula regression model to jointly analyze the three domain scores. Since these QoL measurements are all highly positively skewed, we adopt gamma distribution for the marginal outcomes. In this case, it is natural to use rank-based correlation Kendall's tau in matrix Γ, and the corresponding standard errors of the estimated tau correlations are obtained by the delta method using the transformation relationship between Pearson correlation and Kendall's tau (Ding and Song (2014)). The Kendall's tau is estimated as 0.368 between fatigue and anxiety and 0.569 between pain and anxiety, while the correlation between fatigue and pain is not point-identifiable.

In each mean marginal model, the mean of gamma distribution takes a log-linear model of the following form: $\mu_{ik} = \mathrm{E}(y_{ik}|\mathbf{X}_i) = \exp(\mathbf{X}_i^T\boldsymbol{\beta}_k), k = 1, 2, 3,$ where

$\mathbf{X}_i = (1, \mathbf{x}_i^T)^T$ and regression vector $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \cdots, \beta_{k5})', k = 1, 2, 3$ is a 6-dimensional unknown regression vector. Moreover, the shape parameter of the k-th gamma margin is $\frac{1}{\sigma_k^2}$.

Table 3.4 includes results obtained from the traditional estimation methods. In the application of the EM algorithm, we run the algorithm twice with two different initial values for the correlation parameters, which produces two very different estimation results, especially the estimates of scale parameters and correlation parameters. This suggests the EM algorithm does not converge. Also, we run the multiple imputations twice, and also obtain two sets of estimation results. Although most of the estimates are similar between the two multiple imputations, there does exist a significant difference in the point estimation for the partially identifiable correlation parameter γ_{12}. Using the most reliable results obtained from our complete-case composite likelihood method, we can conclude that edema is significantly associated with two QoL domains of pain and anxiety but is not associated with QoL domain fatigue. Also, black children tend to have higher QoL scores in pain and fatigue in comparison to children of other race. Both age and gender are not important factors for QoL in all three domains.

3.8 Concluding Remarks

In this paper we develop a complete-case composite likelihood approach to handle regression analysis with misaligned missing data pattern. As shown in Theorem III.1, this method indeed is also applicable to a general estimation framework with MAR missing data mechanism. Regression analysis using the copula regression model is useful to deal with nonlinear and non-normal univariate mean regression model. This paper presents a meticulous treatment on missing data in the Gaussian copula regression model. An advantage of this proposed method is that it takes the observed information into the formulation of likelihood function and more importantly only point identifiable parameters are involved in the proposed likelihood, so estimation of point identifiable parameters can proceed with no influence from unidentifiable parameters. In addition, we propose to estimate bounds of a parameter range for partially identifiable parameters.

We use the location-scale family as the class of marginal models in this paper. As a matter of fact, as pointed out by Song et al. (2009b), the marginal regression model in the framework of Gaussian copula regression models can be rather arbitrary; for example, one may use the class of generalized linear models in the margins, which was done by Song et al. (2009b) in the absence of missing data.

One limitation of this method is that when there are three or more partially identifiable parameters in the estimation, the multivariate delta's method for the derivation of the asymptotic covariance for the bound estimates become tedious. Some further exploration may be of interest.

Table 3.1: Summary of simulation results, including average point estimates, empirical standard errors (ESE) and average model-based standard errors (AMSE), under the misaligned missing data pattern for 3-dimensional linear regression model.

		Full Data		Complete-Case CL		EM algorithm		Multiple Imputation	
Response	Ture Value	Estimate	ESE/AMSE	Estimate	ESE/AMSE	Estimate	ESE/AMSE	Estimate	ESE/AMSE
β_{10}	0	-0.0014	0.0709 / 0.0706	-0.0634	0.0826 / 0.0822	-0.0663	0.0840 / 0.0822	-0.0879	0.1037 / 0.0708
β_{11}	1	1.0063	0.1425 / 0.1412	1.1255	0.1717 / 0.1643	1.1319	0.1740 / 0.1644	1.1342	0.1884 / 0.1415
β_{12}	1	0.9978	0.0735 / 0.0704	0.9999	0.0880 / 0.0817	0.9999	0.0890 / 0.0818	0.9998	0.0962 / 0.0710
σ_1	1	0.9936	0.0506 / 0.0489	0.9855	0.0619 / 0.0581	0.9228	0.0833 / 0.0507	0.9958	0.0691 / -
β_{20}	0	-0.0046	0.0716 / 0.0703	0.0593	0.1247 / 0.1236	0.0683	0.1247 / 0.1237	0.1107	0.2156 / 0.0716
β_{21}	1	1.0031	0.1431 / 0.1405	0.8780	0.2536 / 0.2469	0.8606	0.2550 / 0.2471	0.8584	0.2786 / 0.1431
β_{22}	1	0.9977	0.0706 / 0.0699	0.9998	0.1298 / 0.1227	0.9995	0.1309 / 0.1228	1.0003	0.1408 / 0.0718
σ_2	1	0.9889	0.0496 / 0.0490	0.9644	0.0935 / 0.0860	0.7237	0.2152 / 0.0598	1.0067	0.1209 / -
β_{20}	0	-0.0022	0.0745 / 0.0704	-0.0025	0.0778 / 0.0739	-0.0030	0.0783 / 0.0740	-0.0015	0.0821 / 0.0710
β_{31}	1	1.0056	0.1421 / 0.1407	1.0063	0.1471 / 0.1473	1.0070	0.1476 / 0.1476	1.0098	0.1523 / 0.1419
β_{32}	1	1.0026	0.0719 / 0.0700	1.0019	0.0744 / 0.0734	1.0017	0.0752 / 0.0736	1.0025	0.0788 / 0.0712
σ_3	1	0.9906	0.0499 / 0.0489	0.9904	0.0512 / 0.0514	0.9640	0.0535 / 0.0479	0.9984	0.0540 / -
γ_{12}	0.5	0.4977	0.0543 / 0.0527	-	- / -	-0.3538	0.2194 / -	0.0962	0.4182 / -
γ_{12}^U	0.9943	-	- / -	0.9812	0.0252 / 0.0235	-	- / -	-	- / -
γ_{12}^L	-0.7543	-	- / -	-0.7491	0.1040 / 0.0975	-	- / -	-	- / -
γ_{13}	0.4	0.3946	0.0618 / 0.0590	0.3964	0.0762 / 0.0731	0.4736	0.1045 / 0.0550	0.3953	0.0865 / -
γ_{23}	0.3	0.2952	0.0651 / 0.0638	0.2924	0.1319 / 0.1209	0.3171	0.1394 / 0.0875	0.2831	0.1699 / -

Table 3.2: Summary of simulation results in the integration of four data sets, including average point estimates, empirical standard errors (ESE) and average model-based standard errors (AMSE), under various missing data patterns over four different data sources, each of which is generated from a three-dimensional linear regression model

		Full Data		Complete Case CL		MLE/EM		Multiple Imputations		Hot-Deck Imputation	
Ture Value		Est.	ESE/AMSE	Est.	ESE/AMSE	Est.	ESE/AMSE	Est.	ESE/AMSE	Est.	ESE
β_0	0	0.0012	0.0511 / 0.0500	0.0009	0.0584 / 0.0569	0.0008	0.0602 / 0.0569	0.0006	0.0687 / 0.0502	0.0008	0.0602
β_1	1	1.0010	0.1000 / 0.1000	1.0038	0.1130 / 0.1137	1.0039	0.1177 / 0.1138	1.0055	0.1200 / 0.1004	1.0039	0.1177
β_2	1	1.0028	0.0489 / 0.0500	1.0012	0.0565 / 0.0567	1.0015	0.0588 / 0.0567	0.9999	0.0608 / 0.0503	1.0015	0.0588
σ	1	0.9976	0.0353 / 0.0351	0.9969	0.0412 / 0.0410	1.0019	0.0419 / 0.0413	1.0014	0.0463 / -	1.0019	0.0419
β_0	0	0.0017	0.0492 / 0.0498	0.0023	0.0573 / 0.0572	0.0021	0.0589 / 0.0572	0.0020	0.0675 / 0.0499	0.0021	0.0589
β_1	1	0.9998	0.0965 / 0.0997	1.0011	0.1111 / 0.1143	0.9998	0.1154 / 0.1144	0.9987	0.1222 / 0.0999	0.9998	0.1154
β_2	1	1.0017	0.0488 / 0.0498	1.0005	0.0573 / 0.0571	1.0004	0.0600 / 0.0572	1.0009	0.0624 / 0.0500	1.0004	0.0600
σ	1	0.9943	0.0352 / 0.0350	0.9934	0.0414 / 0.0411	0.9979	0.0418 / 0.0414	0.9966	0.0460 / -	0.9980	0.0417
β_0	0	0.0010	0.0529 / 0.0499	0.0006	0.0602 / 0.0578	0.0004	0.0615 / 0.0579	0.0015	0.0707 / 0.0500	0.0010	0.0716
β_1	1	0.9980	0.0972 / 0.0998	0.9968	0.1150 / 0.1156	0.9958	0.1159 / 0.1157	0.9954	0.1242 / 0.1001	0.9979	0.1258
β_2	1	0.9998	0.0494 / 0.0497	0.9993	0.0567 / 0.0576	0.9991	0.0576 / 0.0576	0.9980	0.0634 / 0.0501	0.9988	0.0610
σ	1	0.9957	0.0347 / 0.0350	0.9934	0.0407 / 0.0412	0.9983	0.0414 / 0.0416	0.9983	0.0457 / -	0.9956	0.0474
γ_{12}	0.5	0.4994	0.0364 / 0.0373	0.5018	0.0528 / 0.0529	0.5093	0.0534 / 0.0518	0.4984	0.0643 / -	0.4974	0.0602
γ_{13}	0.4	0.3989	0.0406 / 0.0418	0.4009	0.0606 / 0.0594	0.4070	0.0608 / 0.0589	0.3994	0.0746 / -	0.4002	0.0684
γ_{23}	0.3	0.3005	0.0449 / 0.0452	0.2995	0.0548 / 0.0539	0.3070	0.0577 / 0.0537	0.3002	0.0640 / -	0.3014	0.0636

Table 3.3: Summary of simulation results in a 5-dimensional linear model, including average point estimates, empirical standard errors (ESE) and average model-based standard errors (AMSE), under two pairs of misaligned missingness between the first and second margins and between the fourth and fifth margins, respectively.

		Full Data		Complete-Case CL	
Parameter	True Value	Estimate	ESE/AMSE	Estimate	ESE/AMSE
β_{10}	0	-0.0007	0.0726 / 0.0703	-0.0524	0.0830 / 0.0805
β_{11}	1	0.9985	0.1439 / 0.1406	1.0998	0.1625 / 0.1610
β_{12}	1	1.0027	0.0730 / 0.0699	0.9991	0.0836 / 0.0802
σ_1	1	0.9896	0.0490 / 0.0490	0.9822	0.0596 / 0.0580
β_{20}	0	-0.0004	0.0729 / 0.0704	0.0535	0.1269 / 0.1147
β_{21}	1	1.0006	0.1408 / 0.1408	0.8882	0.2467 / 0.2307
β_{22}	1	0.9997	0.0749 / 0.0702	1.0029	0.1269 / 0.1138
σ_2	1	0.9906	0.0513 / 0.0489	0.9652	0.0933 / 0.0849
β_{30}	0	-0.0025	0.0709 / 0.0702	-0.0022	0.0737 / 0.0727
β_{31}	1	1.0054	0.1376 / 0.1405	1.0059	0.1433 / 0.1452
β_{32}	1	0.9979	0.0727 / 0.0701	0.9974	0.0756 / 0.0726
σ_3	1	0.9888	0.0480 / 0.0489	0.9866	0.0509 / 0.0511
β_{40}	0	-0.0040	0.0695 / 0.0704	-0.0044	0.0856 / 0.0856
β_{41}	1	1.0008	0.1416 / 0.1408	0.9975	0.1714 / 0.1714
β_{42}	1	0.9995	0.0732 / 0.0701	0.9989	0.0876 / 0.0853
σ_4	1	0.9907	0.0498 / 0.0491	0.9861	0.0658 / 0.0617
β_{50}	0	-0.0006	0.0701 / 0.0705	-0.0019	0.1063 / 0.1029
β_{51}	1	0.9948	0.1380 / 0.1409	0.9886	0.2124 / 0.2052
β_{52}	1	1.0004	0.0718 / 0.0701	0.9978	0.1120 / 0.1020
σ_5	1	0.9916	0.0505 / 0.0490	0.9787	0.0795 / 0.0748
γ_{12}	0.6	0.5967	0.0471 / 0.0450	-	- / -
γ_{12}^U	0.9982	-	- / -	0.9984	0.0461 / -
γ_{12}^L	-0.4182	-	- / -	-0.4279	0.1321 / -
γ_{13}	0.5	0.4988	0.0534 / 0.0526	0.4989	0.0636 / 0.0643
γ_{14}	0.4	0.3979	0.0602 / 0.0588	0.4001	0.0894 / 0.0852
γ_{15}	0.3	0.3019	0.0645 / 0.0635	0.3041	0.1243 / 0.1110
γ_{23}	0.5	0.4957	0.0526 / 0.0527	0.4912	0.1068 / 0.0981
γ_{24}	0.45	0.4480	0.0564 / 0.0560	0.4438	0.1341 / 0.1213
γ_{25}	0.3	0.3010	0.0659 / 0.0635	0.2984	0.1775 / 0.1623
γ_{34}	0.5	0.4978	0.0546 / 0.0526	0.4979	0.0732 / 0.0687
γ_{35}	0.5	0.5014	0.0533 / 0.0523	0.5025	0.0896 / 0.0828
γ_{45}	0.6	0.6003	0.0450 / 0.0448	-	- / -
γ_{45}^U	0.9846	-	- / -	0.9853	0.0594 / -
γ_{45}^L	-0.4579	-	- / -	-0.4578	0.1212 / -

Table 3.4: Estimation of correlation and marginal regression parameters for quality of life study obtained by complete-case univariate analysis, EM algorithm with two initial values of correlation parameters, two runs multiple imputations, and one method of complete-case composite likelihood.

		Univariate		EM algorithm(Initial γ=0/0.5)				Two Multiple Imputations				Our Method	
Variable		Est.	AMSE	Est.	AMSE	Est.	AMSE	Est.	AMSE	Est.	AMSE	Est.	AMSE
Y_1	Int.	3.4961	0.1159	3.5046	0.1133	3.5219	0.0892	3.5600	0.1400	3.5041	0.1224	3.5441	0.1024
	X_1	0.0156	0.0069	0.0145	0.0067	0.0133	0.0053	0.0113	0.0083	0.0151	0.0076	0.0121	0.0061
	X_2	0.0127	0.0419	0.0052	0.0404	0.0068	0.0320	0.0153	0.0535	0.0232	0.0436	0.0158	0.0399
	X_3	0.1581	0.0551	0.1723	0.0540	0.1672	0.0426	0.1823	0.0604	0.1686	0.0673	0.1614	0.0375
	X_4	0.0796	0.0521	0.0705	0.0512	0.0586	0.0403	0.0529	0.0546	0.0571	0.0575	0.0615	0.0447
	X_5	0.1512	0.0605	0.1451	0.0591	0.1483	0.0468	0.1544	0.0685	0.1523	0.0696	0.1597	0.0518
	σ	0.2012	0.0136	0.2080	0.0146	0.1620	0.0081	0.2074	0.0156	0.1993	0.0174	0.1997	0.0095
Y_2	Int.	3.7393	0.1324	3.7389	0.1146	3.6435	0.0869	3.6858	0.1398	3.6736	0.1463	3.6207	0.1103
	X_1	0.0009	0.0077	-0.0001	0.0068	0.0043	0.0051	0.0065	0.0080	0.0060	0.0082	0.0088	0.0064
	X_2	0.0012	0.0442	-0.0015	0.0388	0.0009	0.0294	-0.0113	0.0461	-0.0077	0.0475	-0.0028	0.0392
	X_3	0.1292	0.0718	0.1334	0.0609	0.1056	0.0464	0.1281	0.0760	0.1285	0.0775	0.0965	0.0615
	X_4	0.0277	0.0720	0.0173	0.0606	0.0603	0.0461	0.0408	0.0755	0.0481	0.0768	0.0664	0.0550
	X_5	0.1468	0.0819	0.1364	0.0692	0.1549	0.0526	0.1231	0.0867	0.1397	0.0902	0.1428	0.0632
	σ	0.2281	0.0148	0.2375	0.0150	0.1704	0.0074	0.2351	0.0195	0.2317	0.0204	0.2358	0.0110
Y_3	Int.	3.7147	0.0882	3.7507	0.0887	3.7070	0.0796	3.7147	0.0882	3.7147	0.0882	3.7105	0.0944
	X_1	0.0001	0.0052	-0.0029	0.0053	-0.0004	0.0047	0.0001	0.0052	0.0001	0.0052	0.0004	0.0052
	X_2	0.0388	0.0312	0.0370	0.0311	0.0372	0.0281	0.0388	0.0312	0.0388	0.0312	0.0389	0.0323
	X_3	0.1085	0.0455	0.1202	0.0457	0.1086	0.0410	0.1085	0.0455	0.1085	0.0455	0.1071	0.0435
	X_4	0.0106	0.0437	-0.0077	0.0440	0.0090	0.0394	0.0106	0.0437	0.0106	0.0437	0.0096	0.0439
	X_5	0.0982	0.0506	0.0971	0.0509	0.0994	0.0456	0.0982	0.0506	0.0982	0.0506	0.0988	0.0510
	σ	0.2221	0.0104	0.2243	0.0107	0.2018	0.0085	0.2221	0.0104	0.2221	0.0104	0.2220	0.0084
Γ	γ_{12}	-	-	0.2902	-	0.4448	-	0.0464	-	0.1913	-	-	-
	γ_{12}^U	0.9507	-	-	-	-	-	-	-	-	-	0.9264	0.0406
	γ_{12}^L	-0.0989	-	-	-	-	-	-	-	-	-	-0.2376	0.1098
	γ_{13}	0.5465	-	0.4547	0.0746	0.3819	0.0623	0.4505	-	0.4528	-	0.4551	0.0813
	γ_{23}	0.7793	-	0.7485	0.0403	0.6221	0.0474	0.7563	-	0.7361	-	0.7568	0.0406
τ	τ_{12}	-	-	0.1875	-	0.2934	-	0.0295	-	0.1225	-	-	-
	τ_{12}^U	0.7992	-	-	-	-	-	-	-	-	-	0.7543	0.1078
	τ_{12}^L	-0.0631	-	-	-	-	-	-	-	-	-	-0.1527	0.1130
	τ_{13}	0.3681	-	0.3005	0.0837	0.2494	0.0675	0.2975	-	0.2992	-	0.3008	0.0913
	τ_{23}	0.5688	-	0.5385	0.0608	0.4275	0.0605	0.5459	-	0.5266	-	0.5465	0.0622

CHAPTER IV

Multilevel Gaussian Copula Regression Model

4.1 Summary

Motivated by an electroencephalography (EEG) data collected from 128 electrodes on the scalps of 91 9-months-old infants, Project 3 concerns the regression analysis of multilevel correlated data. Arguably, multilevel correlated data are pervasive in practice, which are routinely modeled by the hierarchical modeling system often utilizing random effects. We develop an alternative approach based on a class of multi-dimensional parametric regression models in the framework of Gaussian copulas, in which implementation of the maximum likelihood estimation is established. The proposed model enjoys great flexibility; in the aspect of regression model, it can accommodate continuous outcomes, discrete outcomes or outcomes of mixed types, while in the aspect of dependence, it can allow temporal (e.g. AR), spatial (e.g. Matérn), clustered (e.g. exchangeable), or a mixture of these dependence structures. Parameters in the proposed model have marginal interpretation, which is absent in the hierarchical model when outcomes of interest are non-normal (e.g. binary or ordinal). The EM algorithm introduced in Chapter II with the peeling procedure provides a fast and stable iterative procedure for parameter estimation in this chapter. The proposed model and algorithm are assessed by simulation studies, and

further illustrated by the analysis of EEG data for an adverse effect of prenatal iron deficiency on infant's visual recognition memory.

4.2 Introduction

Multilevel data, also known as hierarchical data, clustered data, and nested data, are a common type of data structure in temporal-spatial analysis, or when subjects are grouped by some specific characters, as a generalization of regression model with parameters that vary at more than one level. For example, Aitkin and Longford (1986) designed a two-level model for educational data, in which students are clustered in schools. Random effects model, also known as variance components model, is one of the most popular methods to establish multilevel models.

Random effects model was introduced by Laird and Ware (1982), where the "fixed" and "random" effects to respectively refer to the population-average and subject-specific effects. Related theories and applications of random effects model in data analysis can be found in Verbeke et al. (2010); Liang and Zeger (1986); Zeger et al. (1988); Zeger and Liang (1986), among others.

Under the traditional random effects model, repeated measurements of subjects within certain cluster, (e.g., students in a certain school in a study) are collected, where the cluster factor is modeled as a random effect. Motivated from our collaborative study on infant's visual recognition memory, where multilevel data are measured from a multivariate setting, we consider a Gaussian copula multilevel regression model as an alternative to the random effects model.

The multilevel Gaussian copula framework focuses on multilevel correlations, which allows covariates to be incl.ded in the marginal regression model as well as interactions within and between levels. This chapter will be organized as follows.

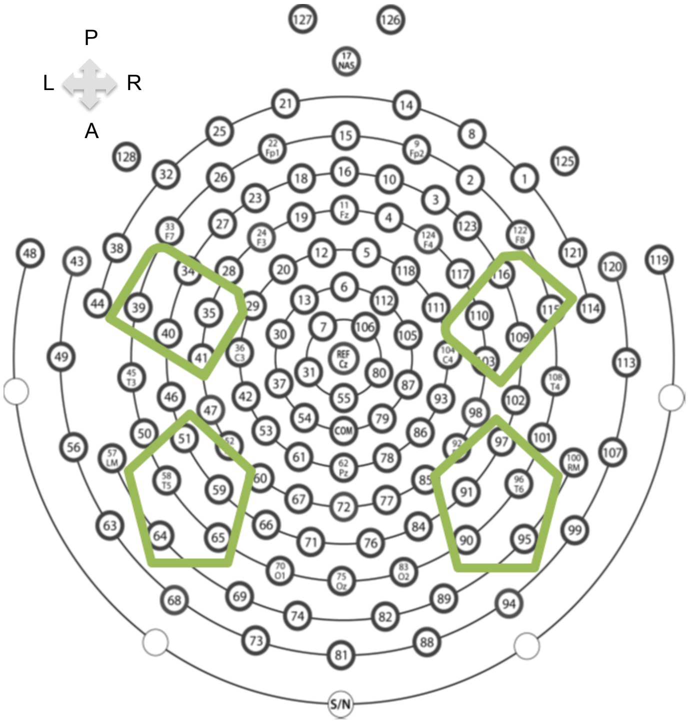
Section 4.3 presents the background and exploratory analysis of EEG data, which motivates the development of the proposed model. Section 4.4 describes the multilevel Gaussian copula model, with specific useful examples including ordinal and mixed margins. Section 4.5 presents the peeling algorithm. Section 4.6 presents simulation studies, and the EEG data analysis is included in Section 4.7. Section 4.8 provides some concluding remarks.

4.3 EEG Data

Infant's visual recognition memory is an important marker of child development in early age. According to Barker et al. (1989)'s theory of development origins, it is hypothesized that mother's exposure to toxicants and nutritional deficiency may affect her child's growth and development, such as neural functional development. Subjects recruited into a collaborative project at University of Michigan, Center for Human Growth and Development, are 9-months old infants without prenatal or acute or chronic illness. To address the scientific hypothesis, this study aims to evaluate whether or not, and if so, how, prenatal iron deficiency affects visual recognition memory for infants. Refer to some of important related work in de Haan et al. (2003). Event-related potential (ERP) is a widely used measure of a specific sensory, cognitive, or motor event. In this study, infant's memory capability reflecting the activity of the brain is measured over a period of 1700 milliseconds using electroencephalograph (EEG) net of 128-channel sensors on the scalp (Reynolds et al. (2011)). Figure 4.1 shows the layout the 128-channel EEG sensor net. The REF node is placed on the central top of the scalp.

The data collection occurs along with a sequence of pictorial stimuli at two time points: when an infant sees his/her mother's picture and when he or she sees a

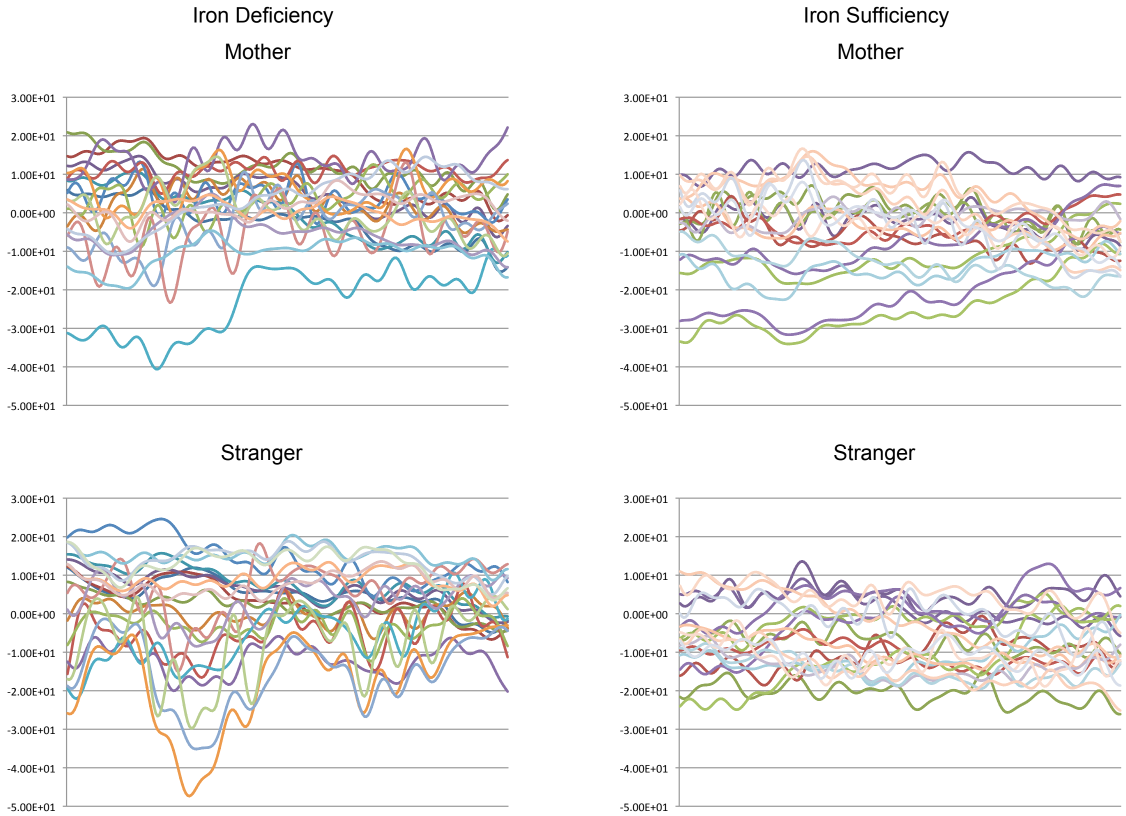
Figure 4.1: Contour of the 128-Channel EEG Sensor Net (L: left; R: right; A: anterior; P: posterior)



stranger's picture. At each time point, an event-related potential (ERP) of interest, late slow wave (LSW), defined as an average magnitude of EEG time series occurring during the last 500 milliseconds, is extracted from the standard data processing. LSW is widely used as a primary outcome of visual recognition memory. In total, there are 91 children with fully observed data. According to our collaborator, there are 20 out of 128 electrodes are of particular interest, with 5 electrodes in each of the four subregions. These four subregions are left-posterior (L-P, subregion 1), right-posterior (subregion 2), left-anterior (L-A, subregions 3), and right-anterior (R-A, subregion 4), outlined with polygons in Figure 4.1.

Figure 4.2 shows examples of the LSW related time series cross-classified by iron status and pictorial stimulus. Figure 4.2 shows that the infants with iron sufficiency appear to have more stable curves over the time window of 500 milliseconds, and for the infants with prenatal iron deficiency contribute more volatile curves when stranger's picture is presented than those when mother's picture is presented.

Figure 4.2: LSW related time series cross-classified by iron status and stimulus.



4.3.1 Data Exploration

Through data processing, including an average of 250 time series data points on each node for each infant with different stimuli, we end up with a vector of 40 outcomes. The first 20 outcomes are LSW measurements on 20 electrodes under the stimulus of mother's picture, and the rest 20 are LSW measurements from the stimulus of stranger's picture. These two vectors of LSW data are collected at two time pints on each infant, so they are serially correlated. We refer to this as the first level of correlation. In each vector of 20 measurements, they are collected from 4 spatially correlated subregions of the highlighted polygons in Figure 4.1. This gives rise to the second level correlation. Moreover, 5 electrodes in each subregion are also clustered and correlated. This may be regarded as the third level correlation.

In Figure 4.3, densities of LSW measurements are plotted over stimulus level where five densities in each panel corresponding to 5 electrodes in each subregion. Figure 4.3 indicates that LSW outcome is approximately normally distributed at each electrode, and this is the data evidence that we utilize to specify normal distribution as marginal model in the analysis.

Figure 4.4 and Figure 4.5 is the scatter plot of LSE measurements in each subregion stratified by pictorial stimuli, and Table 4.1 shows corresponding correlation matrix of each subregion when infants are presented different pictorial stimuli. Based on findings in Figure 4.4, Figure 4.5, and Table 4.1, we choose exchangeable matrix to model the within cluster correlation.

Furthermore, using the average LSW measurements in each subregion, we explore how iron status or stimulus affects region-level visual recognition memory. As shown in Figure 4.6, it seems that the median LSW difference between two iron conditions in subregion 3 is more evident when a stranger's picture is presented to the infants.

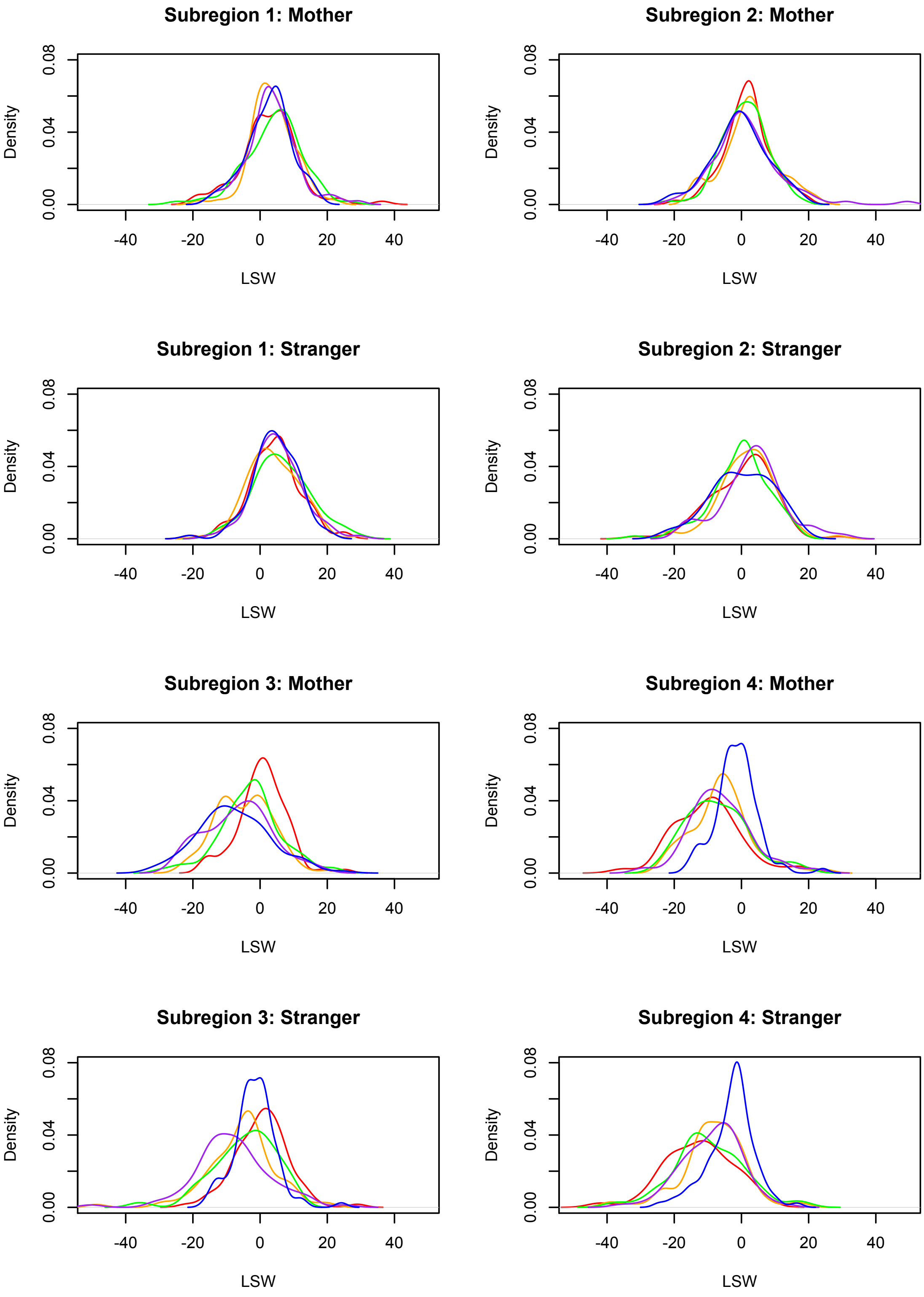
In addition, some collected in the study. They are mothers' age at birth of her baby, gestational age (in weeks), cord blood Pb levels (in ug/dL), first born child or not, baby gender (boy=1, girl=0), and delivery type (vaginal=1, C-section=0).

Figure 4.7 displays boxplots of subject-specific LSW measurements over two stimuli between girls and boys. Once again, such differences in subregion 3 is slightly bigger than the other three subregions.

4.3.2 Mixed Effects Model

As a part of data exploration, we utilized mixed effects model to analyze the EEG data. For the mixed effects model, the covariates mentioned above are included as fixed effects factors. Subregions and pictorial stimuli are considered two completely

Figure 4.3: Density of LSW measurements at each electrodes over stimulus level.



Subregion 1: Mother

Subregion 2: Mother

Subregion 1: Stranger

Subregion 2: Stranger

Subregion 3: Mother

Subregion 4: Mother

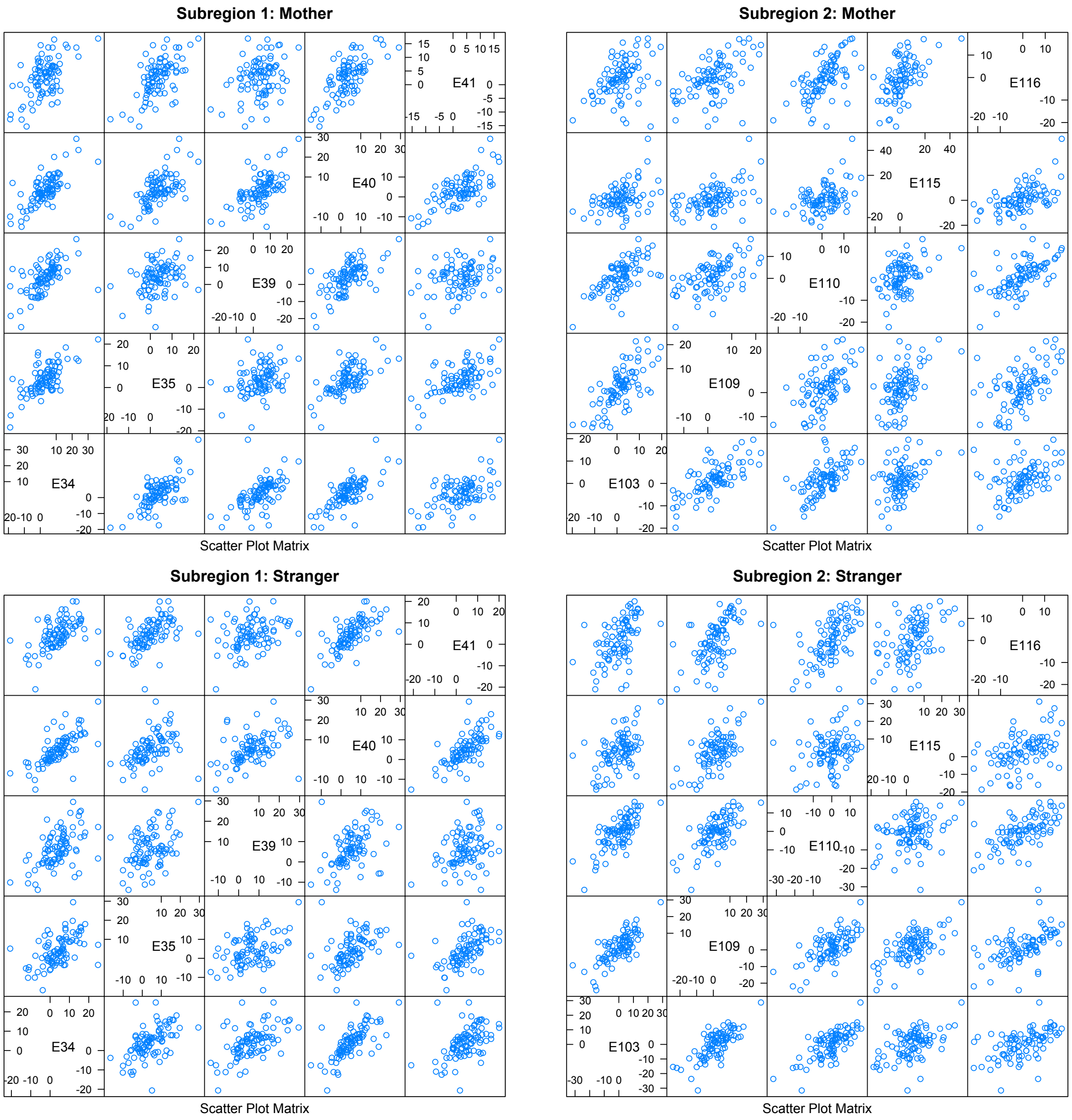
Subregion 3: Stranger

Subregion 4: Stranger

Table 4.1: Summary of Pearson Correlation of LSW Measurements in Each Subregion Stratified by Pictorial Stimuli

		Subregion 1						Subregion 2				
Mother		E34	E35	E39	E40	E41		E103	E109	E110	E115	E116
	E34	1	0.6833	0.5724	0.7432	0.5721	E103	1	0.7059	0.6604	0.4261	0.4188
	E35	0.6833	1	0.3290	0.6513	0.6117	E109	0.7059	1	0.5273	0.5086	0.5270
	E39	0.5724	0.3290	1	0.6262	0.3190	E110	0.6604	0.5273	1	0.4409	0.6163
	E40	0.7432	0.6513	0.6262	1	0.7052	E115	0.4261	0.5086	0.4409	1	0.5437
	E41	0.5721	0.6117	0.3190	0.7052	1	E116	0.4188	0.5270	0.6163	0.5437	1
Stranger		E34	E35	E39	E40	E41		E103	E109	E110	E115	E116
	E34	1	0.5263	0.5475	0.7210	0.4923	E103	1	0.7491	0.7439	0.5139	0.5317
	E35	0.5263	1	0.3990	0.4692	0.5483	E109	0.7491	1	0.6760	0.6495	0.5710
	E39	0.5475	0.3990	1	0.4454	0.3267	E110	0.7439	0.6760	1	0.2979	0.5719
	E40	0.7210	0.4692	0.4454	1	0.6950	E115	0.5139	0.6495	0.2979	1	0.4470
	E41	0.4923	0.5483	0.3267	0.6950	1	E116	0.5317	0.5710	0.5719	0.4470	1
		Subregion 3						Subregion 4				
Mother		E51	E58	E59	E64	E65		E90	E91	E95	E96	E97
	E51	1	0.4940	0.5171	-0.0365	0.2222	E90	1	0.5687	0.6146	0.7236	0.3407
	E58	0.4940	1	0.5291	0.4327	0.6345	E91	0.5687	1	0.4541	0.6411	0.5490
	E59	0.5171	0.5291	1	0.1201	0.6480	E95	0.6146	0.4541	1	0.5865	0.3513
	E64	-0.0365	0.4327	0.1201	1	0.5357	E96	0.7236	0.6411	0.5865	1	0.5929
	E65	0.2222	0.6345	0.6480	0.5357	1	E97	0.3407	0.5490	0.3513	0.5929	1
Stranger		E51	E58	E59	E64	E65		E90	E91	E95	E96	E97
	E51	1	0.5446	0.6537	0.2326	0.3347	E90	1	0.5364	0.6108	0.7477	0.3118
	E58	0.5446	1	0.6244	0.5706	0.7139	E91	0.5364	1	0.1408	0.4133	0.5377
	E59	0.6537	0.6244	1	0.2760	0.5979	E95	0.6108	0.1408	1	0.6813	0.2077
	E64	0.2326	0.5706	0.2760	1	0.7070	E96	0.7477	0.4133	0.6813	1	0.3905
	E65	0.3347	0.7139	0.5979	0.7070	1	E97	0.3118	0.5377	0.2077	0.3905	1

Figure 4.4: Scatter Plot of LSW Measurements in Each Subregion Stratified by Pictorial Stimuli.

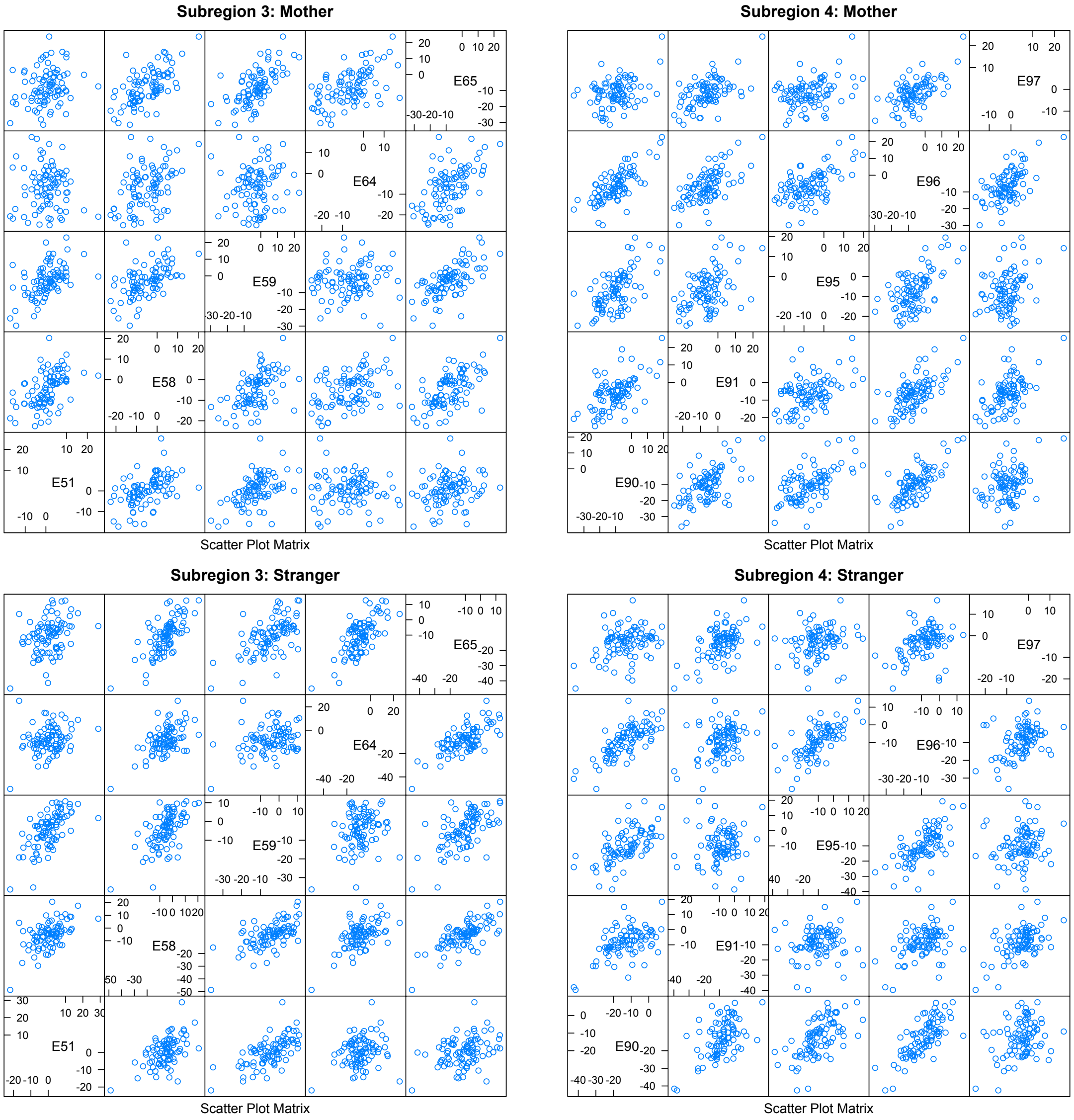


crossed random effects factors nested to each infant, and node is nested random effects factor in each subregion. The model may be written as follows,

$$LSW_{itjk} = X_i^T\beta + b_{1,it} + b_{2,itj} + b_{3,ijk} + \epsilon_{itjk}, \tag{4.1}$$

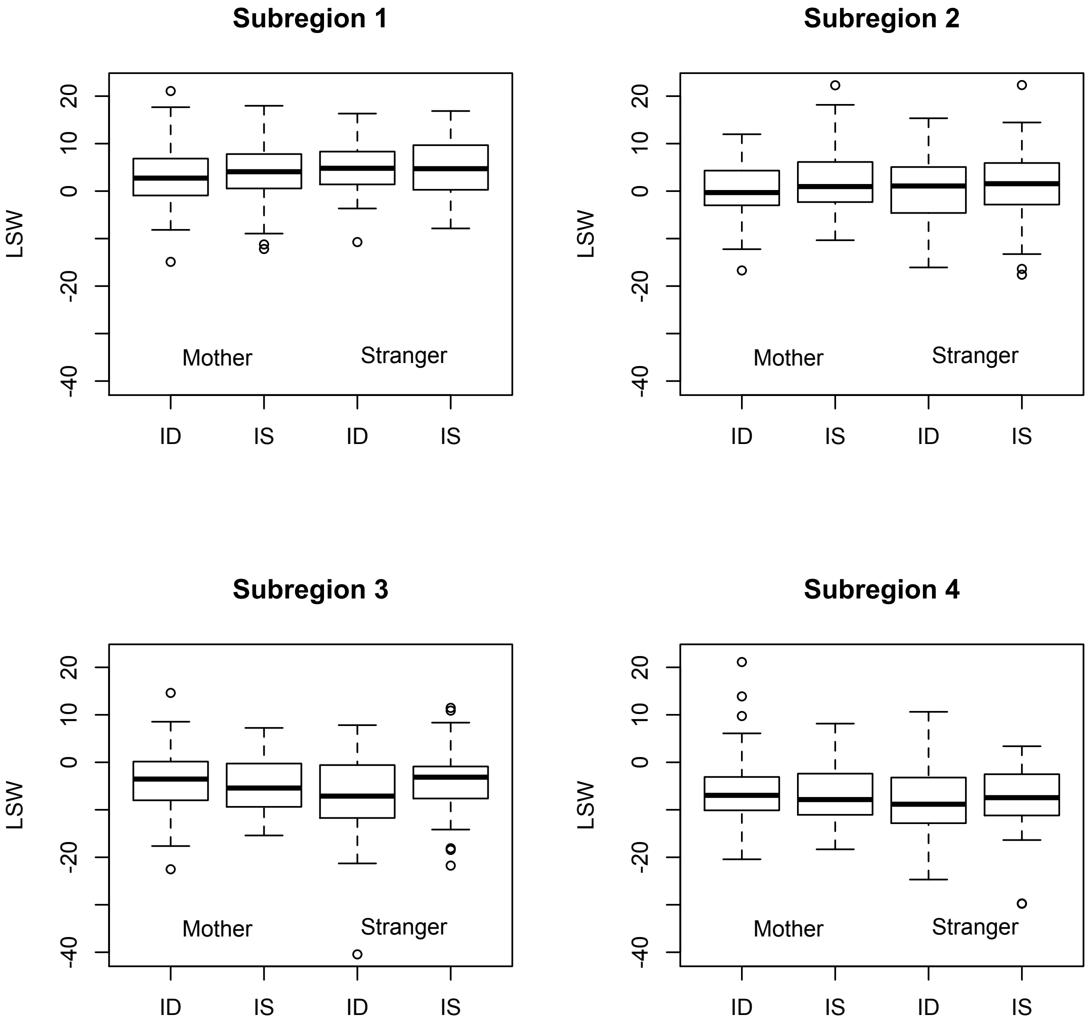
where $t = 1, 2$ for mother and stranger's pictorial stimuli, $i = 1, \cdots, 91$ for 91 infants, $j = 1, \cdots, 4$ for 4 subregions, and $k = 1, \cdots, 5$ for 5 nodes within each subregion. $X_i, i = 1, \cdots, 91$ are the fixed effects factors. $b_{1,it}$ is a random effect that introduces

Figure 4.5: Scatter Plot of LSW Measurements in Each Subregion Stratified by Pictorial Stimuli.



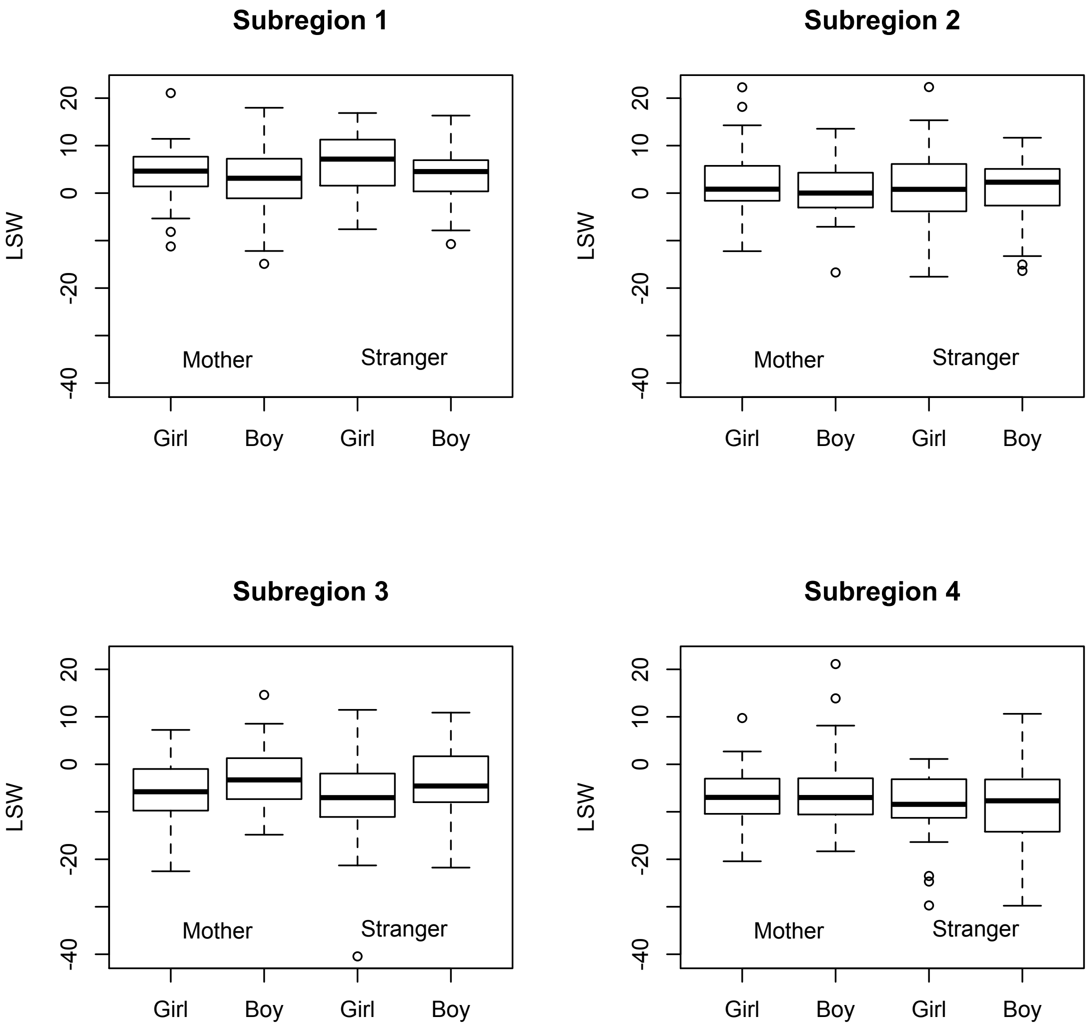
an exchangeable correlation among 20 nodes, because it is a common random variable shared by 4 (subregions) × 5 (nodes in each subregion) LSWs, with the distribution $N(0, \sigma^2_{b_1})$. $b_{2,itj}$ introduces an exchangeable correlation among 5 nodes, because it is a common random variable shared by LSWs from 5 nodes in subregion j, with the distribution $N(0, \sigma^2_{b_2})$. $b_{3,ijk}$ introduces an exchangeable correlation between two time points, because it is a common random variable shared by LSW between two time points, with the distribution $N(0, \sigma^2_{b_3})$.

Figure 4.6: Boxplots of region-specific LSW cross iron status and stimulus.



Applying *lmer* function in R package "lme4" by default settings under REML, Table 4.2 provides maximum likelihood estimates of the fixed effects from the mixed effects model. For subregion dummy variables, subregion1 is defined as reference. We notice that mother's age and subregions are the only factor that has significant effect on LSW. However, none of the effect of iron sufficiency or interaction term between iron sufficiency and subregions is significant. Moreover, according to the defined dummy variables, iron sufficiency is the interaction effect of iron sufficiency

Figure 4.7: Boxplots of region-specific LSW cross iron status and stimulus.


Subregion 1

Subregion 2

Subregion 3

Subregion 4

and subregion1, which is also not significant.

Table 4.3 provides the restricted maximum likelihood estimates of the variance components from the mixed effects model. It is evident that the correlation estimators are different from the preliminary results in Table 4.1.

Since we are interested in the effect of iron deficiency on LSW, especially in certain subregions, and in the tempo and spatial correlations between LSW. The mixed effects model is not able to provide such results in an easy way, and fails to

Table 4.2: Maximum Likelihood Estimates of the Fixed Effects

Parameter	Estimate	Std.Err	t.value
Intercept	0.7712	9.2935	0.083
Mother's Age	0.2030	0.0776	2.616
Gestation Length	-0.0350	0.2204	-0.159
Pb Level	-0.2073	0.1630	-1.272
First Born	0.0513	0.5618	0.091
Baby Gender	-0.0663	0.4168	-0.159
Delivery Type	0.0834	0.4957	0.168
Iron Sufficiency	-0.0924	1.0008	-0.092
Subregion 2	-10.0819	1.1807	-8.539
Subregion 3	-3.9500	0.9797	-4.032
Subregion 4	-11.1144	1.2324	-9.019
Subregion 2*Iron Sufficiency	1.4786	1.8271	0.809
Subregion 3*Iron Sufficiency	1.6838	1.5161	1.111
Subregion 4*Iron Sufficiency	0.3843	1.9071	0.202

Table 4.3: Restricted Maximum Likelihood Estimates of the Variance Components

Groups	Name	Variance	Std.Dev.	Corr			
Subregion / Infant	Node 1	25.9395	5.0931				
	Node 2	15.3446	3.9172	-0.86			
	Node 3	9.0081	3.0013	-0.79	0.67		
	Node 4	26.3337	5.1316	-0.7	0.96	0.53	
	Node 5	53.8977	7.3415	-0.87	0.97	0.82	0.92
Infant	Subregion 1	10.2222	3.1972				
	Subregion 2	50.6216	7.1149	-0.71			
	Subregion 3	27.6058	5.2541	-0.55	0.07		
	Subregion 4	57.2341	7.5653	-1	0.69	0.49	
Pictorial Stimulus	Mother	0.1532	0.3914				
Residual		49.324	7.0231				

detect the significance of iron deficiency's effect. Moreover, the mixed effects model is also not as easy as a marginal model to specify the correlation structure.

Here we develop a more straightforward method that combines the marginal models and dependence model, with fewer parameters.

4.4 Model

Motivated by the EEG data structure, we propose to develop a Gaussian copula model for multilevel correlated data, an alternative method to analyze the EEG data.

We here first assume that data are fully observed with no missing values, and then later extend the proposed methodology by allowing missing values in responses through the utility of EM algorithm. In the EEG data, for each subject, let $Y = (y_1, \cdots, y_{40})'$ be a 40-dimensional random vector of LSW measurements from 20 of electrodes under mother's and stranger's stimuli. As pointed above, a three-level correlation among 40 LSW outcomes: the first corresponds to the serial dependence between two stimuli; the second one is a spatial correlation of four subregions, and the third one is the within-cluster correlation in each subregion. Furthermore, in the proposed multilevel Gaussian copula regression model, the data are allowed to be either continuous or discrete outcomes.

4.4.1 Location-scale Family Marginal Model

Location-scale family are assumed for marginal distributions, which are given by $f_j(y_j|\theta_j), j = 1, \cdots, d$, where $\theta = (\theta_1, \cdots, \theta_d)'$ and $\theta_j = (\mu_j, \sigma_j), j = 1, \cdots, d$, with μ_j as the marginal location parameter and σ_j as the sacel parameter. Note that the marginal parameters are associated with the marginal density function, $f_j(y_j|\theta_j), j = 1, \cdots, d$. Denote the marginal cumulative distribution function by $u_j = F_j(y_j|\theta_j)$.

4.4.2 Gaussian Copula

As discussed in Section 2.3.2, Y is assumed to follows the d-dimensional distribution generated by a Gaussian copula (Song (2007)), whose density function is given by

$$f(Y|\theta, \Gamma) = c(u|\Gamma) \prod_{j=1}^{d} f_j(y_j|\theta_j), u = (u_1, u_2, \cdots, u_d)' \in [0, 1]^d, \tag{4.2}$$

and the joint density of a Gaussian copula function $c(\cdot|\Gamma)$ takes the form:

$$c(u|\Gamma) = |\Gamma|^{-\frac{1}{2}} \exp\left\{\frac{1}{2}Q(u)^T(\mathrm{I} - \Gamma^{-1})Q(u)\right\}, u \in [0,1]^d, \tag{4.3}$$

where $\Gamma = [\gamma_{j_1 j_2}]_{d\times d}$ is the Pearson correlation matrix of $Q(u) = (q_1(u_1), \cdots, q_d(u_d))'$, and $q_j = q_j(u_j) = \Phi^{-1}(u_j), j = 1, \cdots, d$ is the j^{th} marginal normal quantile, where Φ is CDF of the standard normal distribution.

The Gaussian copula model may be extended to embrace multilevel correlation via Kronecker product of multiple correlation matrices. Suppose there are L types of correlation matrices in the multilevel correlated data, denoted by $\Gamma_1, \cdots, \Gamma_L$, respectively, with the corresponding dimensions $d_1, \cdots, d_L$, and $d_1 d_2 \cdots d_L = d$ is the total dimension. Then, the correlation matrix of the d-dimensional vector of outcomes Y may be modeled by $\Gamma = \Gamma_1 \otimes \cdots \otimes \Gamma_L$ with dimension d, where a Kronecker product of two matrix $A = (a_{j_1 j_2})_{d_1\times d_1}$ and $B = (b_{j_1 j_2})_{d_2\times d_2}$ is

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1d_1}B \\ \vdots & \ddots & \vdots \\ a_{d_1 1}B & \cdots & a_{d_1 d_1}B \end{bmatrix}, \tag{4.4}$$

and the Kronecker product is associative that for three matrix A, B, C, $A\otimes B\otimes C = (A \otimes B) \otimes C = A \otimes (B \otimes C)$.

Consequently, the determinant of Γ is

$$|\Gamma| = \prod_{l=1}^{L} |\Gamma_l|^{d/d_l},$$

and the inverse of Γ (or the precision matrix) is

$$\Gamma^{-1} = \Gamma_1^{-1} \otimes \cdots \otimes \Gamma_L^{-1}.$$

In this chapter, we only consider parametric margins. We present two examples to illustrate our proposed approach. In an example of skewed marginal distributions, a gamma distribution is used. In another example of a discrete marginal distributions, a multiple logistic regression model is considered for ordinal data.

4.4.3 Example of Marginal Model: Gamma Margin

In this chapter, we only consider parametric margins. In an example of skewed marginal distributions, a gamma distribution is used. To include covariates in the mean marginal model, let $X_i = (1, x_i^T)^T$, $i = 1, \cdots, n$. For the j^{th} gamma distributed margin, the log-linear model is imposed on the mean parameter, $\mu_{ij} = \mathrm{E}(y_{ij}|X_i) = \exp(X_i^T \beta_j), j = 1, \cdots, d$, where $\beta_j = (\beta_{j0}, \beta_{j1}, \cdots, \beta_{jp})^T$ is a $(p+1)$-element unknown regression vector. Moreover, for the j^{th} gamma distributed margin, the shape parameter is $\dfrac{1}{\sigma_j^2}$, and the rate parameter is $\dfrac{1}{\sigma_j^2 \exp(X_i^T \beta_j)}$. The CDF of marginal distribution model is $u_{ij} = F_j(y_j|\theta_j), j = 1, \cdots, d$.

4.5 Maximum Likelihood Estimation

Our goal is to obtain the maximum likelihood estimation for the model parameter (θ, Γ) with data of multilevel correlation. To deal with a large number of parameters in the estimation, we invoke the peeling procedure developed in Chapter II, which has been shown as an effective numerical optimization engine to obtain the MLE in the multi-dimensional copula regression model. Steps of this optimization procedure are discussed in detail and some illustrative examples are provided in the subsequent subsections.

In the presence of missing data, EM algorithm will be applied; refer to the details in Section 2.4.1, where the M-step is based on the peeling algorithm. When the data is fully observed with no missing values, the log likelihood function can be greatly simplified, and the peeling algorithm works effectively to obtain parameter estimates.

4.5.1 Log Likelihood Function

The log likelihood function given by

$$\lambda(\theta,\Gamma)=\sum_{i=1}^{n}\lambda_i(\theta,\Gamma), \tag{4.5}$$

where the log likelihood of the i^{th} subject is given as follows. Consider the case where the data are fully observed, so the E-step is not required. For the ease of exposition, the subscription of the i^{th} subject, $i=1,\cdots,n$ is omitted.

$$\begin{aligned}
\lambda_i(\theta,\Gamma) &= \ln\{c(u|\Gamma)\}+\sum_{j=1}^{d}f_j(y_j|\theta_j)\\
&= -\frac{1}{2}\ln|\Gamma|-\frac{1}{2}Q(u)^T(\Gamma^{-1}-\mathrm{I})Q(u)+\sum_{j=1}^{d}f_j(y_j|\theta_j)\\
&= -\frac{d}{2}\sum_{l=1}^{L}\frac{1}{d_l}\ln|\Gamma_l|-\frac{1}{2}\mathrm{vec}(\Gamma_1^{-1}\otimes\cdots\otimes\Gamma_L^{-1}-\mathrm{I})^T\mathrm{vec}\left\{Q(u)^TQ(u)\right\}\\
&\quad +\sum_{j=1}^{d}f_j(y_j|\theta_j) \qquad (4.6)\\
&= -\frac{d}{2}\sum_{l=1}^{L}\frac{1}{d_l}\ln|\Gamma_l|+\frac{1}{2}\sum_{j=1}^{d}(1-A_{jj})q_j^2-\frac{1}{2}\sum_{j_2\neq j_1}^{d}A_{j_1j_2}q_{j_1}q_{j_2}+\sum_{j=1}^{d}f_j(y_j|\theta_j) \qquad (4.7)
\end{aligned}$$

where $A=[A_{j_1j_2}]_{d\times d}$ is the inverse matrix of Γ, and $A=\Gamma_1^{-1}\otimes\cdots\otimes\Gamma_L^{-1}=A_1\otimes\cdots\otimes A_L$. The peeling algorithm for optimization is used to obtain MLE $(\hat{\theta},\hat{\Gamma})=\arg\max_{(\theta,\Gamma)}\lambda(\theta,\Gamma)$. The *vec* is a function that forces to transform matrix into a vector column by column.

4.5.2 Peeling Algorithm

The peeling algorithm allows us to iteratively solve the score equation, $\sum_{i=1}^{n}\nabla_{\theta,\Gamma}\lambda_i(\theta,\Gamma)=0$, in which it updates parameter values by maximizing equation (4.5) sequentially with respect to individual parameters components of θ and Γ. By updating low dimensional parameters in each iteration, a kind of "profile" log likelihood function is used with much simpler expressions..

Step P-1: Updating Marginal Parameters

For a specific marginal parameter $\theta_j, j = 1, \cdots, d$, we rewrite the parts involved with the j^{th} margin in equation (4.7) as a "profile" log likelihood function, at the $(t+1)^{th}$ iteration,

$$\begin{aligned} & \lambda_{ij}\left(\theta_1^{(t+1)}, \cdots, \theta_{j-1}^{(t+1)}, \theta_j, \theta_{j+1}^{(t)}, \cdots, \theta_d^{(t)}, \Gamma^{(t)}\right) \\ = & \frac{1}{2}(1 - A_{jj}^{(t)})q_{ij}^2 - \left\{A[j, -j]^{(t)} q_{i,-j}^{(t)}\right\} q_{ij} + f_j(y_{ij}|\theta_j), \end{aligned} \tag{4.8}$$

where $q_{i,-j}^{(t)} = (q_{i1}^{(t+1)}, \cdots, q_{i,j-1}^{(t+1)}, q_{i,j+1}^{(t)}, \cdots, q_{id}^{(t)})^T$ is a subvector of $q_i^{(t)}$ with the first j elements already updated at the $(t+1)^{th}$ iteration. Then, the update of θ_j is obtained by maximizing the following profile log likelihood function:

$$\theta_j^{(t+1)} = \arg\max_{\theta_j} \sum_{i=1}^{n} \lambda_{ij}\left(\theta_1^{(t+1)}, \cdots, \theta_{j-1}^{(t+1)}, \theta_j, \theta_{j+1}^{(t)}, \cdots, \theta_d^{(t)}, \Gamma^{(t)}\right),$$

This optimization is carried out numerically by a quasi-Newton optimization routine available in R function *nlm*, and this step is computationally fast as the optimization involves only a set of low-dimensional parameters θ_j at one time. Consequently, we obtain updates of the other quantiles: $u_{ij}^{(t+1)} = F_j(y_{ij}|\theta_j^{(t+1)})$ and $q_{ij}^{(t+1)} = \Phi^{-1}(u_{ij}^{(t+1)})$, for $j = 1, \cdots, d$, and $i = 1, \cdots, n$.

Step P-2: Updating Correlation Parameters

For the correlation parameters in Γ, we rewrite the part involving the correlation matrix in equation (4.6) as follows,

$$\begin{aligned} \lambda_{i,\Gamma}(\theta^{(t+1)}, \Gamma) &= -\frac{d}{2}\sum_{l=1}^{L} \frac{1}{d_l} \ln|\Gamma_l| \\ & -\frac{1}{2}\mathrm{vec}(\Gamma_1^{-1} \otimes \cdots \otimes \Gamma_L^{-1} - \mathrm{I})^T \mathrm{vec}\left\{Q_i^{(t+1)}(u)^T Q_i^{(t+1)}(u)\right\}, \end{aligned} \tag{4.9}$$

where $Q_i^{(t+1)}(u) = (q_{i1}^{(t+1)}, \cdots, q_{id}^{(t+1)})^T$. Then, the update of Γ is obtained by sequentially maximizing the profile log likelihood function of the form $\sum_{i=1}^{n} \lambda_i(\theta^{(t+1)}, \Gamma)$.

This optimization can be done numerically by applying R function *optim* (Nelder & Mead, 1965). For correlation matrix $\Gamma_l, l = 1, \cdots, L$, when it is the exchangeable or the first-order auto-regressive correlation, its determinant and inverse matrix have closed-form expressions in that only one correlation parameter is involved in optimization. The case of Matérn correlation matrix contains two parameters. For an unstructured correlation matrix, there are $\frac{1}{2} d_l \times (d_l - 1)$ parameters involved. In the latter two cases, updating the correlation matrix Γ_l is done on the basis of the entire matrix, so the related optimization depends largely on the dimension of a correlation matrix. With the R function *optim*, the computing works reasonably fast.

Initialization

For marginal parameters, the initial values are obtained by running marginal regression under the independence working correlation. To generate initial values of correlation parameters, Step P-2 is applied with the given initial estimates of marginal parameters.

4.5.3 Statistical Inference

For the proposed multilevel copula regression model, Fisher information can be calculated to provide the asymptotic variance and covariance for the MLE. In the presence of missing data, the Louis' Formula discussed in Chapter II Section 2.4.3 is applied to calculate the asymptotic variance and covariance of the MLE. In the framework of maximum likelihood estimation, the MLE given for the proposed models follows the classical large-sample properties under some regularity conditions.

4.6 Simulation Study

We conduct simulation experiments to evaluate the performance of the peeling algorithm. In the first, third and fourth experiments, we consider 2-level correlation models with 2-dimensional correlation at level-1 and 3-dimensional correlation at level 2. In the second simulation we consider a 4-level correlation model. Different types of correlation matrices Γ at each level may take various correlation structures, including first-order autoregressive correlation (AR-1), exchangeable, unstructured, Matérn, and wave correlations.

Point estimates and empirical standard errors obtained from the peeling algorithm are provided, together with empirical 95% confidence intervals. In each simulation experiment, the sample size is fixed at 200, while 1000 replicates are run to draw summary statistics.

4.6.1 Multilevel Model with Normal Margins I

First we examine the multilevel model for normally distributed margins with two levels of correlation. The first level is the wave correlation with dimension $d_1 = 3$, and the second level is the AR(1) correlation with dimension $d_2 = 2$. The total dimension is $d = d_1 \times d_2 = 6$. The wave correlation is a spatial correlation function with parameter $\phi > 0$ of the form $\rho(h) = \dfrac{\phi}{h} \sin\left(\dfrac{h}{\phi}\right)$, where h is the distance between two locations. This wave correlation function allows both positive and negative correlations.

The simulation setup is given as follows. We include $p = 2$ covariates $X_1 \sim \text{Bin}(0, 1) - 0.5$, and $X_2 \sim \Gamma(2, 1) - 2$ in the marginal linear model with $\mu_k = X^T \beta_k, k = 1, \cdots, d = 6$. The wave correlation function at the first level is specified with $\phi = 0.5$, and the AR-1 correlation at the second level is specified by the parameter $\gamma = 0.5$. To generate marginal outcomes, errors are assumed follow to a 6-variate normal $N_6(0, \Gamma)$ with the standard normal margin $N(0, 1)$ and correlation matrix $\Gamma = \Gamma_1 \otimes \Gamma_2$, where Γ_1 is a 3-dimensional wave correlation matrix, and Γ_2 is a 2-dimensional AR-1 correlation matrix.

We compare the results obtained from the multilevel Gaussian copula regression model with those obtained from the univariate analysis that ignores correlations. Two types of standard errors are reported: the first type is the empirical standard error in the two methods, and the other type is the average of 1000 model-based standard errors calculated from Fisher Information.

As shown in Table 4.4, the estimates of the marginal estimates and correlation parameters obtained from both two methods are close to their true values. This is because both models are assumed to be correctly specified, so these two methods provide consistent estimates. The standard errors from the multilevel Gaussian copula regression model are, with no surprise, smaller, when the correlation is used in the estimation, especially for the variance parameters and the correlation parameters.

4.6.2 Multilevel Model with Normal Margins II

In the second simulation experiment, we examine the multilevel Gaussian copula regression model with a four level correlation. The first level correlation matrix Γ_1 is set as a 6 (d_1)-dimensional Matérn class correlation with spatial correlation parameter $\alpha = 1$ and shape parameter $\nu = 1$. The second level correlation matrix Γ_2 is set as a 3 (d_2)-dimensional exchangeable correlation with parameter $\gamma_2 = 0.5$. The third level correlation matrix Γ_3 is set as a 2 (d_3)-dimensional AR-1 correlation with parameter $\gamma_3 = 0.5$. The fourth level correlation matrix Γ_4 is set as a 3 (d_4)-dimensional unstructured correlation with parameters $(0.6, 0.5, 0.4)$. The resulting total dimension is $d = \prod_{l=1}^{4} d_l = 108$, and the resulting correlation matrix is $\Gamma = \Gamma_1 \otimes \cdots \otimes \Gamma_4$.

We also include $p = 2$ covariates $X_1 \sim \text{Bin}(0,1) - 0.5$, and $X_2 \sim \Gamma(2,1) - 2$ in the marginal linear models with $\mu_k = X^T\beta_k, k = 1, \cdots, d = 108$. To generate correlated outcomes, errors are generated from a 108-variate normal $N_{108}(0, \Gamma)$ with

Table 4.4: Summary of simulation results from the analysis of correlated normal outcomes under 2-level correlation: the wave correlation and AR-1 correlation. The multilevel Gaussian copula regression model is compared with the univariate analysis, including average point estimates, empirical standard errors (ESE) and average model-based standard errors (AMSE) over 1000 running of simulations.

				Univariate		Multilevel Model	
Level 1	Level 2	Parameter	True Value	Estimate	ESE/AMSE	Estimate	ESE/AMSE
1	1	β_0	0	-0.0011	0.1404 / 0.1428	-0.0011	0.1404 / 0.1416
		β_1	1	0.9984	0.1352 / 0.1426	0.9984	0.1352 / 0.1413
		β_2	1	1.0008	0.0502 / 0.0509	1.0008	0.0502 / 0.0504
		σ	1	1.0030	0.0510 / -	0.9940	0.0471 / 0.0463
	2	β_0	0	0.0022	0.1400 / 0.1425	0.0022	0.1400 / 0.1414
		β_1	1	0.9970	0.1398 / 0.1423	0.9970	0.1398 / 0.1411
		β_2	1	1.0003	0.0502 / 0.0507	1.0003	0.0502 / 0.0503
		σ	1	1.0010	0.0488 / -	0.9930	0.0447 / 0.0457
2	1	β_0	0	0.0001	0.1445 / 0.1424	0.0001	0.1445 / 0.1413
		β_1	1	0.9994	0.1422 / 0.1421	0.9994	0.1422 / 0.1410
		β_2	1	1.0011	0.0508 / 0.0507	1.0011	0.0508 / 0.0503
		σ	1	0.9999	0.0502 / -	0.9923	0.0463 / 0.0462
	2	β_0	0	0.0020	0.1443 / 0.1424	0.0020	0.1443 / 0.1415
		β_1	1	1.0063	0.1406 / 0.1422	1.0063	0.1406 / 0.1412
		β_2	1	0.9990	0.0508 / 0.0507	0.9990	0.0508 / 0.0504
		σ	1	1.0002	0.0501 / -	0.9935	0.0463 / 0.0463
3	1	β_0	0	0.0049	0.1460 / 0.1426	0.0049	0.1460 / 0.1417
		β_1	1	1.0012	0.1450 / 0.1423	1.0012	0.1450 / 0.1415
		β_2	1	0.9985	0.0511 / 0.0508	0.9985	0.0511 / 0.0505
		σ	1	1.0013	0.0505 / -	0.9953	0.0473 / 0.0459
	2	β_0	0	0.0052	0.1462 / 0.1427	0.0052	0.1462 / 0.1417
		β_1	1	1.0017	0.1450 / 0.1424	1.0017	0.1450 / 0.1414
		β_2	1	0.9970	0.0498 / 0.0508	0.9970	0.0498 / 0.0504
		σ	1	1.0018	0.0493 / -	0.9949	0.0458 / 0.0464
		ϕ	0.5	0.5042	0.0369 / -	0.4922	0.0237 / 0.0241
		γ	0.5	0.5047	0.0284 / -	0.5018	0.0283 / 0.0281

the standard normal marginal $N(0,1)$ and a 108×108-dimensional correlation matrix Γ.

We compare the results obtained from the multilevel Gaussian copula regression model with those obtained from the univariate analysis that ignores the correlation. The empirical standard error in the two methods are provided.

Table 4.5: Summary of simulation results in four-level correlated normally distributed margins with Matérn correlation, exchangeable correlation, AR-1 correlation, and unstructured correlation obtained from the multilevel Gaussian copula regression model, including average point estimates and empirical standard errors (ESE).

		Multilevel Model	
Parameter	True Value	Estimate	ESE
α	1	0.9937	0.0159
γ_R	0.5	0.5031	0.0101
γ_T	0.5	0.5027	0.0108
$\gamma_{M,12}$	0.6	0.6028	0.0109
$\gamma_{M,13}$	0.5	0.5034	0.0125
$\gamma_{M,23}$	0.4	0.4025	0.0149

The simulation results for the correlation parameters are summarized in Table 4.5. There are some findings during the estimation worth mentioning. We have 438 parameters in total to estimate in this simulation, but it only takes a few minutes for the algorithm to converge. However, since the multilevel Gaussian copula regression model is a unified framework, the model-based standard errors of parameters are calculated in one Hessian matrix, whose dimension is 438 as well in this simulation. This is the time-consuming part in this estimation procedure, which needs further research.

4.7 Analysis of EEG Data

We now present the analysis of the EEG data introduced in Section 4.3 using the proposed multilevel Gaussian copula regression model and compare the results with those obtained from the univariate analysis that ignores the correlation. The

40-dimensional vector of LSW measures are collected from 91 infants, and associated with three levels of correlations. The first level of correlation is modeled by a 2×2 correlation matrix of mother's and stranger's pictorial stimuli. The second level of correlation is modeled by a 4×4 spatial correlation matrix generated by two 2×2 correlation matrices by the Kronecker product, of two matrices corresponding to two paired subregions of left / right hemispheres and anterior / posterior, respectively. The third level of correlation is modeled by a 5×5 within-cluster exchangeable correlation matrix corresponding to 5 electrodes within each subregion.

Covariates included in the marginal regression model are iron sufficiency (1 for iron sufficiency, and 0 for iron deficiency), mother's age at birth delivery, gestation age in weeks, cord blood Pb levels in ug/dL, whether first born (1 for yes, and 0 for no), baby gender (1 for boy, and 0 for girl), and delivery type (1 for vaginal, and 0 for C section). The fact that the estimates for marginal regression parameters are close the multilevel Gaussian copula regression model and univariate analysis suggested that the correlation structure had little impact on the point estimation. Similar numerical evidence was also reported in the GEE model and the random effect model. However, the estimated standard errors from the multilevel Gaussian copula regression model are smaller due to the multivariate analysis approach.

Table 4.6 summarized the effect of iron status on LSW measurements collected from each node and each stimulus, including estimates, model-based standard errors, and z-statistics (i.e., parameter estimate divided by the model-based standard errors). Figure 4.8 shows the z-statistics of iron status across both mother's and stranger's pictorial stimuli. From Figure 4.8, we can see that iron deficiency significantly affects the LSW within the left anterior subregion when infants saw stranger's picture. Infants with iron deficiency had lower LSW than those with iron sufficiency

Table 4.6: Summary of iron status effect on LSW measurements collected from each node and each stimulus, including estimate, model-based standard errors, and z-statistics.

Level-1	Level2-1	Level2-2	Level 3	Estimate	Std.Err	z-statistic
Mother	Left	Posterior	Node 34	2.7243	1.7187	1.5851
			Node 35	1.3733	1.2828	1.0706
			Node 39	-0.6972	1.8685	-0.3731
			Node 40	2.0734	1.4304	1.4495
			Node 41	0.6665	1.3351	0.4992
		Anterior	Node 51	0.2540	1.7351	0.1464
			Node 58	-0.3897	1.7472	-0.2230
			Node 59	1.7669	1.9977	0.8845
			Node 64	-1.0353	2.4419	-0.4240
			Node 65	0.3563	2.3176	0.1537
	Right	Posterior	Node 103	1.3820	1.5178	0.9105
			Node 109	0.9564	1.6768	0.5704
			Node 110	1.1341	1.4421	0.7864
			Node 115	2.3445	2.3447	0.9999
			Node 116	2.8214	1.7432	1.6185
		Anterior	Node 90	-2.7382	1.9958	-1.3720
			Node 91	-0.2981	1.7567	-0.1697
			Node 95	-0.4620	2.0373	-0.2268
			Node 96	-2.2881	1.8876	-1.2122
			Node 97	1.4454	1.2902	1.1202
Stranger	Left	Posterior	Node 34	-0.7327	1.6211	-0.4520
			Node 35	-1.3869	1.6082	-0.8624
			Node 39	-3.5700	1.7909	-1.9934
			Node 40	0.5667	1.5324	0.3698
			Node 41	-1.4991	1.4743	-1.0168
		Anterior	Node 51	3.4765	1.6861	2.0619
			Node 58	6.0610	2.0404	2.9705
			Node 59	5.4746	1.8125	3.0205
			Node 64	2.1499	2.5233	0.8520
			Node 65	5.8331	2.3060	2.5295
	Right	Posterior	Node 103	0.9051	1.8425	0.4912
			Node 109	2.8546	1.6340	1.7470
			Node 110	1.3625	1.7832	0.7641
			Node 115	0.8898	2.0579	0.4324
			Node 116	-0.9979	1.9939	-0.5005
		Anterior	Node 90	0.5283	2.0553	0.2570
			Node 91	-1.0468	2.0419	-0.5127
			Node 95	2.1114	2.4072	0.8771
			Node 96	-2.2506	1.8455	-1.2195
			Node 97	1.6682	1.5559	1.0722

at the stimulus of strangers picture. Moreover, we tested the significance of iron sufficiency's effect on LSW measurement in Subregion 3, and the corresponding null hypothesis test is $H_0:\ \beta_{3,1,IS} = \beta_{3,2,IS} = \beta_{3,3,IS} = \beta_{3,4,IS} = \beta_{3,5,IS} = 0$. The statistic of Wald test is equal to 15.0848, which follows a chi-square distribution with degree of freedom equal to 5 under null hypothesis, and the p value is 0.01. Thus, we reject the null hypothesis, and we believe iron sufficiency has signifiant effect on LSW measurement in Subregion 3.

Table 4.7 shows a summary of analysis results for the correlation parameters. Our findings include that: (i) LSW measurements between the left and right hemisphere are not significantly correlated; (ii) LSW measurements on the electrodes between the anterior and posterior regions are significantly negatively correlated.

Table 4.7: Summary of estimation results for the correlation parameters.

Correlation Parameter	Estimate	Std.Err
Level 1	0.2764	0.0219
Level 2-1	-0.0120	0.0237
Level 2-2	-0.0776	0.0241
Level 1	0.5155	0.0175

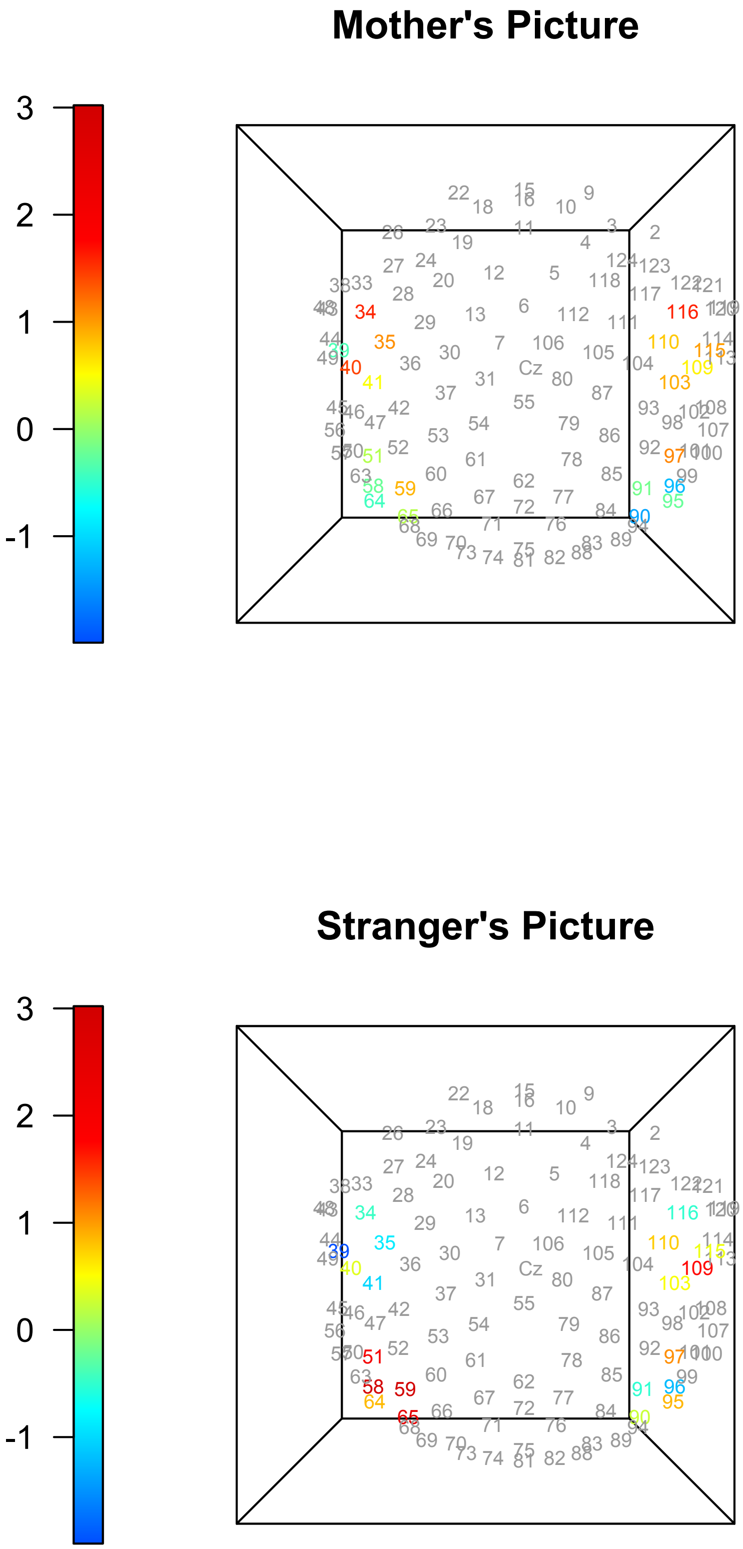
Though it only took a few minutes for the peeling algorithm to converge for the point estimation, it was very time-consuming for the estimation of standard errors. This is because the dimension of Hessian matrix with is as high as 364.

4.8 Conclusions & Discussion

This chapter presents a Gaussian copula framework for multilevel rank-based correlations that may be estimated by the proposed approach. The peeling algorithm is developed and implemented to estimate both marginal parameters and correlation parameters. The proposed methodology allows to adjust for covariates via the marginal regression models. The proposed peeling procedure to facilitate the computation of MLE.

All numerical examples have shown that the Gaussian copula multilevel model performed well. On one hand, this proposed approach can be used to analyze longitu-

Figure 4.8: Z-statistics of Iron Status Effect on Mother's and Stranger's Pictorial Stimuli



Mother's Picture

Stranger's Picture

dinal, or spatial, or spatio-temporal data; on the other hand, it provides a framework that to organizes correlated data by multiple hierarchical layers, each it including a low dimensional dependence structure. Consequently, the resulting correlation parameters are more interpretable and more easily to be to estimated via the proposed peeling algorithm.

One limitation worth mentioning is the computation burden on the estimation of standard errors. Because all the model parameters in the proposed multilevel Gaussian copula regression model is related within a unified framework, the asymptotic covariance matrix is generally not a block diagonal matrix. Then the required computation is carried out with a large dimension. As proposed model being a class of parametric models, model diagnosis is the most critical component in the application of the proposed model. Residual analysis is useful to detect any potential violation of model assumptions such as Gaussian copula, marginal location-scale family, and correlation structures.

In data exploration analysis, we found some outliers, which may lead to some confusing results. We plan to communicate with our collaborators to decide the criteria of including data. Moreover, we noticed some heavy tails of densities of LSW measurements in Figure 4.3, we plan to consider empirical location-scale family to improve the analysis.

CHAPTER V

Discussions and Future Works

In this dissertation, we developed two methods to deal with missing data problems. In particular misaligned missing data pattern has been treated systematically in Chapters II and III. In addition, a multilevel copula regression model is proposed to deal with complex correlation data in Chapter IV. Here we have some discussions on future work.

One of the big challenges we met in our work concerns computational burden. In Chapter II, although both EM algorithm and Gaussian copula have been well studied in the existing literature, our work improved the estimation procedure by simplifying the multiple integrals into many one-dimension integrals. Therefore, the related computation was fast. However, in Chapter III, the complete-case composite likelihood was computationally challenging. This is because in our application of the peeling algorithm for parameters estimation, at each iteration, the inverse of each subject-specific correlation matrix based on observed data requires computational effort. Thus, when dimension of data is large, the related calculation procedure may be time consuming. In addition, for the estimation of the asymptotic covariance, calculating both sensitivity matrix and variability matrix are necessary. The computational efficiency on this calculation depends on the dimension of the parameters.

Consequently, a large dimensional vector of parameters requires substantially more computational power.

The same computational challenge remains in Chapter IV, where estimation of the Hessian matrix is also required for the estimation of asymptotic covariances. To speed up the peeling algorithm for parameter estimation, more computing memory is necessary to store some matrices essential for the estimation procedure, whose dimensions are proportional to both sample size and dimension of parameters under estimation. Hence, improving computation speed of the peeling further effort of algorithm needs further effort. Furthermore, with more time, we will consider using C++ instead of R for implementation, which we believe will be much faster.

Future work for multilevel Gaussian copula regression model: one is to deal with missing data. Despite the approaches developed in Chapters II and III for data analysis with missing data, we need to address the computational challenge related to large dimensional data.

APPENDIX

APPENDIX A

The Likelihood Orthogonal for Complete-Case Composite Likelihood

A.1 Theorem III.2: Unbiasedness

Proof. For a subject, the expectation of the complete-case composite score function is,

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\eta}_0}\Psi(\mathbf{y},\mathbf{R};\boldsymbol{\eta}) &= \int \Psi(\mathbf{y},\mathbf{R};\boldsymbol{\eta})f(\mathbf{y};\boldsymbol{\xi}_0)d\mathbf{y} \\
&= \int \sum_{s\in S} w_s \frac{\partial}{\partial\boldsymbol{\eta}} \ln f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)f(\mathbf{y};\boldsymbol{\xi}_0)d\mathbf{y} \\
&= \sum_{s\in S} w_s \int \frac{\partial}{\partial\boldsymbol{\eta}} \ln f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)f_s(\mathbf{y}_s|\boldsymbol{\xi}_{0,s})d\mathbf{y}_s \\
&= \sum_{s\in S} w_s \int \frac{\dot{f}_s(\mathbf{y}_s|\boldsymbol{\xi}_s)}{f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)} f_s(\mathbf{y}_s|\boldsymbol{\xi}_{0,s})d\mathbf{y}_s.
\end{aligned}$$

When $\boldsymbol{\eta} = \boldsymbol{\eta}_0$, the above equation above may be rewritten as,

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\eta}_0}\Psi(\mathbf{y},\mathbf{R};\boldsymbol{\eta})|_{\boldsymbol{\eta}_0} &= \sum_{s\in S} w_s \int \frac{\partial}{\partial\boldsymbol{\eta}} f_s(\mathbf{y}_s|\boldsymbol{\xi}_{0,s})d\mathbf{y}_s \\
&= \sum_{s\in S} w_s \frac{\partial}{\partial\boldsymbol{\eta}} \int f_s(\mathbf{y}_s|\boldsymbol{\xi}_{0,s})d\mathbf{y}_s \\
&= \mathbf{0}.
\end{aligned}$$

For the entire sample, $\mathrm{E}_{\boldsymbol{\eta}_0}\Psi(\mathbf{Y},\mathbf{R};\boldsymbol{\eta})|_{\boldsymbol{\eta}_0} = \sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{\eta}_0}\Psi(\mathbf{y}_i,\mathbf{R}_i;\boldsymbol{\eta})|_{\boldsymbol{\eta}_0}$ □

A.2 Theorem III.5: Composite Barlett Identity

Proof. For a subject, the variability matrix is given by,

$$\begin{aligned}\mathbf{J}(\boldsymbol{\eta}) &= \mathrm{var}(\Psi(\mathbf{y},\mathbf{R};\boldsymbol{\eta})) \\ &= \mathrm{var}\left(\frac{\partial}{\partial\boldsymbol{\eta}}\sum_{s\in S} w_s \ln f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right).\end{aligned}$$

Because $w_s^2 = w_s$ for all $s \in S$, and $w_{s_1}w_{s_2} = 0$, when $s_1 \neq s_2 \in S$,

$$\begin{aligned}\mathbf{J}(\boldsymbol{\eta}) &= \sum_{s\in S} w_s \mathrm{var}\left(\frac{\partial}{\partial\boldsymbol{\eta}}\ln f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right), \\ &= \sum_{s\in S} w_s \mathrm{E}\left\{\left(\frac{\partial}{\partial\boldsymbol{\eta}}\ln f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right)\left(\frac{\partial}{\partial\boldsymbol{\eta}}\ln f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right)^T\right\},\end{aligned}$$

which is because of the unbiasedness.

For a subject, the sensitivity matrix may be rewritten by,

$$\begin{aligned}\mathbf{H}(\boldsymbol{\eta}) &= -\sum_{s\in S} w_s \mathrm{E}\left\{\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}\ln f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right\} \\ &= -\sum_{s\in S} w_s \mathrm{E}\left\{\frac{\partial}{\partial\boldsymbol{\eta}}\left(\frac{1}{f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)}\frac{\partial}{\partial\boldsymbol{\eta}^T}\ln f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right)\right\} \\ &= \sum_{s\in S} w_s \mathrm{E}\left\{\left(\frac{\partial}{\partial\boldsymbol{\eta}}\ln f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right)\left(\frac{\partial}{\partial\boldsymbol{\eta}}\ln f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right)^T\right\}, \\ &\quad -\sum_{s\in S} w_s \mathrm{E}\left\{\frac{1}{f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)}\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right\},\end{aligned}$$

To prove $\mathbf{H}(\boldsymbol{\eta}) = \mathbf{J}(\boldsymbol{\eta})$, we prove $\sum_{s\in S} w_s \mathrm{E}\left\{\frac{1}{f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)}\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right\} = 0$ as follows,

$$\begin{aligned}
& \sum_{s\in S} w_s \mathrm{E}\left\{\frac{1}{f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)}\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right\} \\
= & \sum_{s\in S} w_s \int \frac{1}{f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)}\left\{\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right\} f(\mathbf{y}|\boldsymbol{\xi})d\mathbf{y} \\
= & \sum_{s\in S} w_s \int \frac{1}{f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)}\left\{\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)\right\} f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)d\mathbf{y}_s \\
= & \sum_{s\in S} w_s \frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}\int f_s(\mathbf{y}_s|\boldsymbol{\xi}_s)d\mathbf{y}_s \\
= & \sum_{s\in S} w_s \frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}1. \\
= & \ 0.
\end{aligned}$$

Therefore, $\mathbf{H}(\boldsymbol{\eta}) = \mathbf{J}(\boldsymbol{\eta})$. □

A.3 Corollary III.6: Uniqueness

Proof. From equation (3.17), for a subject,

$$\begin{aligned}
\Psi(\mathbf{y},\mathbf{R};\boldsymbol{\eta}) &= \frac{\partial}{\partial(\boldsymbol{\theta},\mathbf{A})}\mathrm{ln}L_c(\boldsymbol{\eta}|\mathbf{y},\mathbf{R}) = \frac{\partial}{\partial(\boldsymbol{\theta},\mathbf{A})}\sum_{s\in S} w_s \ln f_s(\mathbf{y}_s|\boldsymbol{\xi}_s) \\
&= \sum_{s\in S} w_s \frac{\partial}{\partial(\boldsymbol{\theta},\tilde{\mathbf{A}}_s)} \ln f_s(\mathbf{y}_s|\boldsymbol{\xi}_s).
\end{aligned}$$

In the equation above, we denote $\mathbf{A}_s = \mathbf{\Gamma}_s^{-1}$, and $\tilde{\mathbf{A}}_s$ is a $d \times d$ matrix, where if $k_1, k_2 \in s$, and $\mathbf{\Gamma}_{s,j_1j_2} = \mathbf{\Gamma}_{k_1k_2}$, then $\tilde{\mathbf{A}}_{s,k_1k_2} = (\mathbf{\Gamma}_s^{-1})_{j_1j_2}$, else $\tilde{\mathbf{A}}_{s,k_1k_2} = 0$. Therefore,

$$
\begin{aligned}
& \mathrm{E}_{\boldsymbol{\eta}_0} \Psi(\mathbf{y}, \mathbf{R}; \boldsymbol{\eta}) \\
= \; & \int \Psi(\mathbf{y}, \mathbf{R}; \boldsymbol{\eta}) f(\mathbf{y}; \boldsymbol{\xi}_0) d\mathbf{y} \\
= \; & \int \sum_{s \in S} w_s \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \ln f_s(\mathbf{y}_s | \boldsymbol{\xi}_s) f(\mathbf{y}; \boldsymbol{\xi}_0) d\mathbf{y} \\
= \; & \sum_{s \in S} w_s \int \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \ln f_s(\mathbf{y}_s | \boldsymbol{\xi}_s) f_s(\mathbf{y}_s | \boldsymbol{\xi}_{0,s}) d\mathbf{y}_s \\
= \; & \sum_{s \in S} w_s \int \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \left\{ \sum_{k \in s} \ln f_k(y_k | \boldsymbol{\theta}_k) + \ln c(\mathbf{u}_s | \mathbf{\Gamma}_s) \right\} f(\mathbf{y}_s; \boldsymbol{\xi}_{0,s}) d\mathbf{y}_s \\
= \; & \sum_{s \in S} w_s \sum_{k \in s} \int \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \ln f_k(y_k | \boldsymbol{\theta}_k) f_k(y_k | \boldsymbol{\theta}_{0,k}) dy_k \\
& + \sum_{s \in S} w_s \int \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \left\{ -\frac{1}{2} \ln |\mathbf{\Gamma}_s| + \frac{1}{2} \mathbf{q}_s^T (\mathrm{I} - \mathbf{\Gamma}_s^{-1}) \mathbf{q}_s \right\} \varphi(\mathbf{q}_s | \mathbf{\Gamma}_{0,s}) d\mathbf{q}_s \\
= \; & \sum_{s \in S} w_s \sum_{k \in s} \int \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \ln f_k(y_k | \boldsymbol{\theta}_k) f_k(y_k | \boldsymbol{\theta}_{0,k}) dy_k \\
& - \frac{1}{2} \sum_{s \in S} w_s \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \ln |\mathbf{\Gamma}_s| \\
& + \sum_{s \in S} w_s \int \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \left\{ \frac{1}{2} \sum_{k \in s} (1 - \tilde{\mathbf{A}}_{s,kk}) q_k^2 - \sum_{k_1 < k_2 \in s} \tilde{\mathbf{A}}_{s,k_1k_2} q_{k_1} q_{k_2} \right\} \varphi(\mathbf{q}_s | \mathbf{\Gamma}_{0,s}) d\mathbf{q}_s \\
= \; & \sum_{s \in S} w_s \sum_{k \in s} \int \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \ln f_k(y_k | \boldsymbol{\theta}_k) f_k(y_k | \boldsymbol{\theta}_{0,k}) dy_k \\
& - \frac{1}{2} \sum_{s \in S} w_s \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \ln |\mathbf{\Gamma}_s| + \frac{1}{2} \sum_{s \in S} w_s \sum_{k \in s} \int \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} (1 - \tilde{\mathbf{A}}_{s,kk}) q_k^2 \varphi(q_k) dq_k \\
& - \sum_{s \in S} w_s \sum_{k_1 < k_2 \in s} \int \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \tilde{\mathbf{A}}_{s,k_1k_2} q_{k_1} q_{k_2} \varphi(q_{k_1}, q_{k_2} | \mathbf{\Gamma}_{0,k_1k_2}) dq_{k_1} dq_{k_2} \\
= \; & \sum_{s \in S} w_s \sum_{k \in s} \int \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \ln f_k(y_k | \boldsymbol{\theta}_k) f_k(y_k | \boldsymbol{\theta}_{0,k}) dy_k \\
& + \sum_{s \in S} w_s \frac{\partial}{\partial(\boldsymbol{\theta}, \tilde{\mathbf{A}}_s)} \left\{ -\frac{1}{2} \ln |\mathbf{\Gamma}_s| + \frac{1}{2} \sum_{k \in s} (1 - \tilde{\mathbf{A}}_{s,kk}) - \sum_{k_1 < k_2 \in s} \mathbf{\Gamma}_{0,k_1k_2} \tilde{\mathbf{A}}_{s,k_1k_2} \right\}.
\end{aligned}
$$

There are two terms in equation above, the former of which is a linear combination of the score functions of marginal models with no correlation parameters involved, and the latter is partial derivatives of a function of $\mathbf{\Gamma}$. Thus, it suffices to prove that the second term has a unique zero at $\mathbf{\Gamma}_0$. Since for unstructured correlation matrix,

$$\frac{\partial}{\partial \tilde{\mathbf{A}}_{s,k_1k_2}} \left\{ -\frac{1}{2} \ln |\mathbf{\Gamma}_s| + \frac{1}{2} \sum_{k \in s} (1 - \tilde{\mathbf{A}}_{s,kk}) - \sum_{k_1 < k_2 \in s} \mathbf{\Gamma}_{0,k_1k_2} \tilde{\mathbf{A}}_{s,k_1k_2} \right\} = \mathbf{\Gamma}_{k_1k_2} - \mathbf{\Gamma}_{0,k_1k_2},$$

which has a unique zero at $\mathbf{\Gamma}_{0,k_1k_2}$, and

$$\frac{\partial}{\partial \tilde{\mathbf{A}}_{s,kk}} \left\{ -\frac{1}{2} \ln |\mathbf{\Gamma}_s| + \frac{1}{2} \sum_{k \in s} (1 - \tilde{\mathbf{A}}_{s,kk}) - \sum_{k_1 < k_2 \in s} \mathbf{\Gamma}_{0,k_1k_2} \tilde{\mathbf{A}}_{s,k_1k_2} \right\} = 0.$$

For an exchangeable correlation matrix with parameter γ, for $d_s = \dim(s) \geq 2$,

$$\frac{\partial}{\partial \gamma} \left\{ -\frac{1}{2} \ln |\mathbf{\Gamma}_s| + \frac{1}{2} \sum_{k \in s} (1 - \tilde{\mathbf{A}}_{s,kk}) - \sum_{k_1 < k_2 \in s} \mathbf{\Gamma}_{0,k_1k_2} \tilde{\mathbf{A}}_{s,k_1k_2} \right\} = \frac{(d_s - 1) d_s (1 + (d_s - 1)\gamma^2)}{2(1-\gamma)^2 (1 + (d_s - 1)\gamma)^2} (-\gamma + \gamma_0),$$

which has a unique 0 at γ_0,

For an AR-1 correlation matrix with parameter γ, for $d_s = \dim(s) \geq 2$,

$$\frac{\partial}{\partial \gamma} \left\{ -\frac{1}{2} \ln |\mathbf{\Gamma}_s| + \frac{1}{2} \sum_{k \in s} (1 - \tilde{\mathbf{A}}_{s,kk}) - \sum_{k_1 < k_2 \in s} \mathbf{\Gamma}_{0,k_1k_2} \tilde{\mathbf{A}}_{s,k_1k_2} \right\} = \frac{(d_s - 1)(\gamma^2 + 1)}{(1-\gamma^2)^2} (-\gamma + \gamma_0),$$

which has a unique 0 at γ_0. Therefore, we know the second term has a unique zero at $\mathbf{\Gamma}_0$ for unstructured, exchangeable, and AR-1 correlation matrices. □

BIBLIOGRAPHY

Bibliography

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Number 55. Courier Dover Publications, 1972.

E. F. Acar, C. Genest, and J. NešLehová. Beyond simplified pair-copula constructions. *Journal of Multivariate Analysis*, 110:74–90, 2012.

M. Aitkin and N. Longford. Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General)*, pages 1–43, 1986.

R. R. Andridge and R. J. Little. A review of hot deck imputation for survey nonresponse. *International Statistical Review*, 78(1):40–64, 2010.

T. Ané and C. Kharoubi. Dependence structure and risk measure*. *The journal of business*, 76(3):411–438, 2003.

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

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

D. J. Barker, C. Osmond, P. Winter, B. Margetts, and S. Simmonds. Weight in infancy and death from ischaemic heart disease. *The Lancet*, 334(8663):577–580, 1989.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

J. Besag. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, pages 616–618, 1977.

C. Czado. Pair-copula constructions of multivariate copulas. In *Copula theory and its applications*, pages 93–109. Springer, 2010.

M. de Haan, M. H. Johnson, and H. Halit. Development of face-sensitive event-related potentials during infancy: a review. *International Journal of Psychophysiology*, 51 (1):45–58, 2003.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

W. Ding and P. X.-K. Song. Em algorithm in gaussian copula with missing data. *Submitted*, 2014.

Y. Fan and D. Zhu. Partial identification and confidence sets for functionals of the joint distribution of potential outcomes. Technical report, Working paper, 2009.

P. Fearnhead and P. Donnelly. Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):657–680, 2002.

J. Fries, B. Bruce, and D. Cella. The promise of promis: using item response theory to improve assessment of patient-reported outcomes. *Clinical and experimental rheumatology*, 23(5):S53, 2005.

X. Gao and P. X.-K. Song. Composite likelihood em algorithm with applications to multivariate hidden markov model. *Statistica Sinica*, 21:165–185, 2011.

C. Genest, J. Quesada Molina, and J. Rodríguez Lallena. De l'impossibilité de construire des lois à marges multidimensionnelles données à partir de copules. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, 320(6):723–726, 1995.

C. Genest, J. Nešlehová, and N. Ben Ghorbal. Estimators based on kendall's tau in multivariate copula models. *Australian & New Zealand Journal of Statistics*, 53 (2):157–177, 2011.

D. S. Gipson, D. T. Selewski, S. F. Massengill, L. Wickman, K. L. Messer, E. Herreshoff, C. Bowers, M. E. Ferris, J. D. Mahan, L. A. Greenbaum, et al. Gaining the promis perspective from children with nephrotic syndrome: a midwest pediatric nephrology consortium study. *Health Qual Life Outcomes*, 11(3), 2013.

C. Glasbey. Non-linear autoregressive time series with multivariate gaussian mixtures as marginal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2):143–154, 2001.

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

J. W. Graham, A. E. Olchowski, and T. D. Gilreath. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213, 2007.

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

N. L. Hjort and C. Varin. Ml, pl, ql in markov chain models. *Scandinavian Journal of Statistics*, 35(1):64–82, 2008.

N. L. Hjort, H. Omre, M. Frisén, F. Godtliebsen, J. Helgeland, J. Møller, E. B. V. Jensen, M. Rudemo, and H. Stryhn. Topics in spatial statistics [with discussion, comments and rejoinder]. *Scandinavian Journal of Statistics*, pages 289–357, 1994.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260): 663–685, 1952.

H. Joe. *Multivariate models and multivariate dependence concepts*, volume 73. CRC Press, 1997.

H. Joe, H. Li, and A. K. Nikoloulopoulos. Tail dependence functions and vine copulas. *Journal of Multivariate Analysis*, 101(1):252–270, 2010.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

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

B. G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1): 221–39, 1988.

R. J. Little and D. B. Rubin. Statistical analysis with missing data. *Statistical analysis with missing data, 2nd ed., by RJA Little and DB Rubin. Wiley series in probability and stistics. New York, NY: Wiley, 2002*, 1, 2002.

T. A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233, 1982.

C. F. Manski. *Partial identification of probability distributions*. Springer, 2003.

G. Masarotto, C. Varin, et al. Gaussian copula marginal regression. *Electronic Journal of Statistics*, 6:1517–1549, 2012.

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management: concepts, techniques, and tools*. Princeton university press, 2010.

X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

G. Molenberghs and G. Verbeke. Models for discrete longitudinal data. *Springer Series in Statistics*, 2005.

G. Molenberghs, M. G. Kenward, G. Verbeke, and T. Birhanu. Pseudo-likelihood estimation for incomplete data. *Statistica Sinica*, 21(1):187, 2011.

J. A. Nelder and R. Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.

E. Parner. A composite likelihood approach to multivariate survival data. *Scandinavian Journal of Statistics*, 28(2):295–302, 2001.

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

T. E. Raghunathan. What do we do with missing data? some options for analysis of incomplete data. *Annu. Rev. Public Health*, 25:99–117, 2004.

D. Renard, G. Molenberghs, and H. Geys. A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis*, 44 (4):649–667, 2004.

G. D. Reynolds, M. W. Guy, and D. Zhang. Neural correlates of individual differences in infant visual attention and recognition memory. *Infancy*, 16(4):368–391, 2011.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

J. Segers, R. Van den Akker, B. Werker, et al. Semiparametric gaussian copula models: Geometry and efficient rank-based estimation. In *18th European Young Statisticians Meeting*, page 6, 2013.

M. Sklar. *Fonctions de répartition à n dimensions et leurs marges.* Université Paris 8, 1959.

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

P. X.-K. Song. *Correlated data analysis: modeling, analytics, and applications.* Springer, 2007.

P. X.-K. Song, Y. Fan, and J. D. Kalbfleisch. Maximization by parts in likelihood inference. *Journal of the American Statistical Association*, 100(472):1145–1158, 2005.

P. X.-K. Song, Z. Jiang, E. Park, and A. Qu. Quadratic inference functions in marginal models for longitudinal data. *Statistics in medicine*, 28(29):3683–3696, 2009a.

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

M. L. Stein, Z. Chi, and L. J. Welty. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296, 2004.

C. Varin and P. Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528, 2005.

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

G. Verbeke, G. Molenberghs, and D. Rizopoulos. Random effects models for longitudinal data. In *Longitudinal research with latent variables*, pages 37–96. Springer, 2010.

S. L. Zeger and K.-Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130, 1986.

S. L. Zeger, K.-Y. Liang, and P. S. Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.