Longitudinal Data Analysis with Composite Likelihood Methods

by

Haocheng Li

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics - Biostatistics

Waterloo, Ontario, Canada, 2012

© Haocheng Li 2012

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Longitudinal data arise commonly in many fields including public health studies and survey sampling. Valid inference methods for longitudinal data are of great importance in scientific researches. In longitudinal studies, data collection are often designed to follow all the interested information on individuals at scheduled times. The analysis in longitudinal studies usually focuses on how the data change over time and how they are associated with certain risk factors or covariates. Various statistical models and methods have been developed over the past few decades. However, these methods could become invalid when data possess additional features.

First of all, incompleteness of data presents considerable complications to standard modeling and inference methods. Although we hope each individual completes all of the scheduled measurements without any absence, missing observations occur commonly in longitudinal studies. It has been documented that biased results could arise if such a feature is not properly accounted for in the analysis. There has been a large body of methods in the literature on handling missingness arising either from response components or covariate variables, but relatively little attention has been directed to addressing missingness in both response and covariate variables simultaneously. Important reasons for the sparsity of the research on this topic may be attributed to substantially increased complexity of modeling and computational difficulties.

In Chapter 2 and Chapter 3 of the thesis, I develop methods to handle incomplete longitudinal data using the pairwise likelihood formulation. The proposed methods can handle longitudinal data with missing observations in both response and covariate variables. A unified framework is invoked to accommodate various types of missing data patterns. The performance of the proposed methods is carefully assessed under a variety of circumstances. In particular, issues on efficiency and robustness are investigated. Longitudinal survey data from the National Population Health Study are analyzed with the proposed methods.

The other difficulty in longitudinal data is model selection. Incorporating a large number of irrelevant covariates to the model may result in computation, interpretation and prediction difficulties, thus selecting parsimonious models are typically desirable. In particular, the penalized likelihood method is commonly employed for this purpose. However, when we apply the penalized likelihood approach in longitudinal studies, it may involve high dimensional integrals which are computationally expensive.

We propose an alternative method using the composite likelihood formulation. Formulation of composite likelihood requires only a partial structure of the correlated data such as marginal or pairwise distributions. This strategy shows modeling tractability and computational cheapness in model selection. Therefore, in Chapter 4 of this thesis, I propose a composite likelihood approach with penalized function to handle the model selection issue. In practice, we often face the model selection problem not only from choosing proper covariates for regression predictor, but also from the component of random effects. Furthermore, the specification of random effects distribution could be crucial to maintain the validity of statistical inference. Thus, the discussion on selecting both covariates and random effects as well as misspecification of random effects are also included in Chapter 4.

Chapter 5 of this thesis mainly addresses the joint features of missingness and model selection. I propose a specific composite likelihood method to handle this issue. A typical advantage of the approach is that the inference procedure does not involve explicit missing process assumptions and nuisance parameters estimation.

Acknowledgements

For helpful advices and suggestions on how to make this thesis possible and better, I wish to express my gratitude to my supervisor, Professor Grace Y. Yi. Her academic knowledge and personal encouragement light up the path for me.

I thank my thesis committee members: Professors Joel A. Dubin, Richard J. Cook, Naisyin Wang (University of Michigan), and Janice Husted.

I wish to acknowledge Statistics Canada for making available to me the National Population Health Survey Data, which is an important statistical analysis example in this thesis. I would like to thank David Binder, Georgia Robert and Ivan Carrillo-Garcia for their support during my internship at Statistics Canada from September 2009 to December 2009.

I would also like to thank Baojiang Chen, Xiaoqin Xiong, Xu Wang and Ker-Ai Lee for their kind help.

Finally, I would like to thank the faculty members, administrative staffs and my colleagues. I never walked alone during my studies.

Dedication

To my parents;
who constantly supported me with understanding and encouragement.
To my grandfather;
who wished me to become a good scholar.

Contents

List of Tables **xxi**

List of Figures **xxiv**

1 Introduction **1**

1.1 Overview 1

1.2 Modeling Strategies 3

1.2.1 Generalized Linear Models 3

1.2.2 Generalized Linear Mixed Models 4

1.2.3 Generalized Estimating Equations 5

1.3 Composite Likelihood 6

1.3.1 Formulation of Composite Likelihood 7

1.3.2 Statistical Properties 8

1.3.3 Computational Issue 10

1.4 Model Selection 11

1.4.1 Best Subset Selection 11

1.4.2 Penalized Likelihood 12

1.4.3 Model Selection for Longitudinal Data 17

1.5 Missing Data in Longitudinal Studies 18

1.5.1 Missing Data Mechanism 18

1.5.2 Likelihood-Based Methods 19

1.5.3 GEE-Based Methods 22

1.5.4 Other Methods 23

1.5.5 Nonidentifiability Issue 23

1.6 Model Misspecification 26

1.7 Motivating Example: The National Population Health Survey Data 28

1.7.1 Background 28

1.7.2 Missing Data 28

1.7.3 A Subsample from NPHS 30

1.8 Outline of Thesis 31

2 A Pairwise Likelihood Approach for Longitudinal Data with Missing Observations in Both Response and Covariates 35

2.1 Introduction 35

2.2 Notations and Model Setups 37

2.2.1 The Response Process 37

2.2.2 The Covariate Process 38

2.2.3 Missing Data Process 38

2.3 Estimation and Inference 39

2.3.1 Marginal and Pairwise Likelihoods 39

2.3.2 Inference Procedures 40

2.4 Numerical Studies 43

2.4.1 Empirical Assessment of the Proposed Methods 43

2.4.2 Application to the NPHS Data 46

2.5 Efficiency Assessment 50

2.6 Sensitivity Analysis for Model Misspecification 55

3 Analysis of Longitudinal Binary Data with Missing Response and Covariates 63

3.1 Introduction 63

3.2 Model Formulation 64

3.2.1 Response Process 64

3.2.2 Covariate Process 66

3.2.3 Missing Data Process 67

3.3 Estimation and Inference 68

3.3.1 Marginal and Pairwise Likelihoods 68

3.3.2 Inference Procedures 69

3.4 The NPHS Data Sample 71

3.5 Empirical Studies 74

3.5.1 Performance of the Proposed Methods 74

3.5.2 Sensitivity Analysis 78

3.5.3 Efficiency Assessment 79

4 Simultaneous Methods of Variable Selection and Estimation for Longitudinal Data Arising in Clusters 87

4.1 Introduction 87

4.2 Model Setup 89

4.2.1 Generalized Linear Mixed Models 90

4.2.2 Generalized Linear Mixed Pairwise Models 91

4.2.3 Pairwise Likelihoods 93

4.3 Methodology: Selecting Fixed Effects 94

4.4 Methodology: Selecting Both Fixed and Random Effects 96

4.4.1 Review of Cholesky Decomposition 96

4.4.2 The Algorithm 98

4.5 Asymptotic Results 100

4.6 Numerical Studies 102

4.6.1 Simulation for Selecting Fixed Effects 102

4.6.2 Simulation for Both Fixed and Random Effects 107

4.6.3 Data Analysis 111

4.7 Model Selection under Misspecified Models 114

4.7.1 Misspecified Models 114

4.7.2 Numerical Studies 116

5 Variable Selection via Composite Likelihood for Incomplete Longitudinal Data Arising in Clusters 155

5.1 Introduction 155

5.2 Model Formulation 157

5.2.1 Generalized Linear Mixed Models 157

5.2.2 Missing Data Models 158

5.2.3 Composite Likelihood 159

5.2.4 Extensions 161

5.3 Selecting Fixed Effects Using the Composite Likelihood 161

5.4 Application 163

5.5 Simulation Studies of the Proposed Methods 166

5.5.1 Measure of Marginal Model Error 166

5.5.2 Linear Mixed Model 167

5.5.3 Poisson Mixed Model 168

6 Discussion and Future Research **183**

6.1 Composite Likelihood Analysis for Incomplete Longitudinal Data 183

6.2 Variable Selection via Composite Likelihood for Analysis of Longitudinal Data Arising in Clusters 185

6.3 Variable Selection via Composite Likelihood for Incomplete Longitudinal Data arising in Clusters 186

References **189**

List of Tables

1.1 The initial values and likelihood estimates for Model 1 and Model 2, respectively. 26

1.2 Missing data rates for health status and household income variables in the NPHS data (%) 30

1.3 Missing data patterns for health status and household income variables in the NPHS data 31

2.1 Simulation results for incomplete longitudinal data with missingness in both continuous response and covariate under pairwise likelihood (PL) and two-stage (TS) estimation algorithms 45

2.2 Sample data from the NPHS 47

2.3 Analysis of the NPHS data using the pairwise likelihood, two-stage estimation approach and naive method: Response models 48

2.4 Analysis of the NPHS data using the pairwise likelihood and two-stage estimation approach: Covariate and missing-data models 49

2.5 Efficiency comparison of the $\boldsymbol{\beta}$ estimators under various scenarios of unknown nuisance parameters 53

3.1 Missing data patterns for the HUI and INC variables in the NPHS data (%) 72

3.2 Analysis of the NPHS data using the pairwise likelihood (PL), two-stage estimation (TS) and naive GEE (NGEE) methods: response models 73

3.3 Analysis of the NPHS data using the pairwise likelihood (PL) and two-stage estimation (TS) approaches: results for parameters associated with the covariate and missing data processes 74

3.4 Simulation results for longitudinal binary data with missingness in both response and covariate variables 77

4.1 Analysis results for the NPHS data: entries represent the estimates and standard errors (in brackets) 113

4.2 The parameter values for the true model and misspecified model for the simulation study 118

4.3 Simulation results for the fixed effects selection under linear mixed model: model selection 132

4.4 Simulation results for the fixed effects selection under linear mixed model: estimation of selected regression coefficients $\boldsymbol{\beta}$ 133

4.5 Simulation results for the fixed effects selection under linear mixed model: estimation of random effects parameters 134

4.6 Simulation results for the fixed effects selection under logistic mixed model: model selection 135

4.7 Simulation results for the fixed effects selection under logistic mixed model: estimation of regression coefficients $\boldsymbol{\beta}$ 136

4.8 Simulation results for the fixed effects selection under logistic mixed model: estimation of random effects parameters 137

4.9 Simulation results for the fixed effects selection under Poisson mixed model: model selection 138

4.10 Simulation results for the fixed effects selection under Poisson mixed model: estimation of regression coefficients $\boldsymbol{\beta}$ 139

4.11 Simulation results for the fixed effects selection under Poisson mixed model: estimation of random effects parameters 140

4.12 Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMMs: model selection in the linear mixed model 141

4.13 Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMMs: estimation of selected regression coefficients $\boldsymbol{\beta}$ in the linear mixed model 142

4.14 Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMM: estimation of random effects parameters in the linear mixed model 143

4.15 Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMPMs: model selection in the linear mixed model 144

4.16 Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMPMs: estimation of selected regression coefficients $\boldsymbol{\beta}$ in the linear mixed model 145

4.17 Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMPMs: estimation of random effects parameters in the linear mixed model 146

4.18 Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMMs: model selection in the Poisson mixed model 147

4.19 Simulation results for doubly selection fixed and random effects under the case that data are generated by GLMMs: estimation of selected regression coefficients $\boldsymbol{\beta}$ in the Poisson mixed model 148

4.20 Simulation results for doubly selection fixed and random effects under the case that data are generated by GLMMs: estimation of random effects parameters in the Poisson mixed model 149

4.21 Simulation results for doubly selection fixed and random effects under the case that data are generated by GLMPMs: model selection in the Poisson mixed model 150

4.22 Simulation results for doubly selection fixed and random effects under the case that data are generated by GLMPMs: estimation of selected regression coefficients $\boldsymbol{\beta}$ in the Poisson mixed model 151

4.23 Simulation results for doubly selection fixed and random effects under the case that data are generated by GLMPMs: estimation of random effects parameters in the Poisson mixed model 152

4.24 Simulation results for the fixed effects selection under misspecified linear mixed model: model selection precision rate (%) for each variable 153

4.25 Simulation results for the fixed effects selection under misspecified linear mixed model: regression coefficient estimation for β_1 154

5.1 Missing data proportions for HUI and INC variables in the NPHS data (%) 164

5.2 Analysis results for the NPHS data: entries represent the estimates and standard errors (in brackets) 166

5.3 Simulation results for the incomplete data via the linear mixed model: model selection 179

5.4 Simulation results for the incomplete data via the Poisson mixed model: model selection 180

5.5 Simulation results for the incomplete data via linear mixed model: model estimation on regression coefficients 181

5.6 Simulation results for the incomplete data via Poisson mixed model: model estimation on regression coefficients 182

List of Figures

1.1 *Comparison between the log likelihood function (logL) and the penalized log likelihood function (PlogL) with SCAD penalty. - - -: logL function; ——: PlogL function. The estimates from logL and PlogL are labeled as $\hat{\beta}_{logL}$ and $\hat{\beta}_{PlogL}$, respectively.* 15

1.2 *Comparison between the log likelihood function (logL) and the penalized log likelihood function (PlogL) with LASSO penalty. - - -: logL function; ——: PlogL function. The estimates from logL and PlogL are labeled as $\hat{\beta}_{logL}$ and $\hat{\beta}_{PlogL}$, respectively.* 16

1.3 *The profile likelihood values with α_1 to be fixed at a set of values for Model 1 and Model 2, respectively.* 25

2.1 *Relative efficiency with respect to common correlation coefficient ρ. $R_{ML:TS}(\beta_0)$: ━ ━ ━ ; $R_{ML:PL}(\beta_0)$: – – – ; $R_{ML:TS}(\beta_1)$: ━━━ ; $R_{ML:PL}(\beta_1)$: ———.* 54

2.2 *Relative efficiency with respect to common correlation coefficient ρ. $R_{PL:TS}(\beta_0)$: ━ ━ ━ ; $R_{PL:TS}(\beta_1)$: ━━━.* 55

2.3 *Asymptotic relative bias for regression coefficients β_0 and β_1 for PL and TS methods when the association structures for the response, covariate and missing processes are all misspecified. The models for estimation involves common correlation coefficient. However, the true correlation matrix for response process has the form in (2.13).* 57

2.4 *Asymptotic percent relative bias for regression coefficients* β_0 *and* β_1 *when response models are misspecified. The model for estimation is* $\mu^y_{ij} = \beta_0 + \beta_1 X_{ij}$*, while true models are:* $\mu^y_{ij} = \beta_0 + \beta_1 X_{ij} + \kappa \cdot j$ *for mean model (1) and* $\mu^y_{ij} = \beta_0 + \beta_1 X_{ij} + \kappa \cdot X_{ij} \cdot j$ *for mean model (2), respectively. PL method:* - - -*; TS method:* ———*.* 58

2.5 *Asymptotic relative bias for regression coefficients* β_0 *and* β_1 *when the missing data process is misspecified. The model for estimation is specified in Section 2.4.1, while the true model is* $\eta^{Ry}_{ij} = \lambda^y_0 + \lambda^y_1 X_{ij} + \lambda^y_2 R^x_{ij} + \kappa Y_{ij}$, $\eta^{Rx}_{ij} = \lambda^x_0 + \lambda^x_1 X_{ij} + \kappa Y_{ij}$*. PL method:* - - -*; TS method:* ———*.* 59

3.1 *Asymptotic relative bias for regression coefficients* β_0 *and* β_1 *when the marginal structures in missing data process are misspecified. The model for estimation is specified in Section 3.5.1, while the true model is* $\eta^{Ry}_{ij} = \lambda^y_0 + \lambda^y_1 y_{ij} + \lambda^y_2 R^x_{ij} + \kappa x_{ij}$, $\eta^{Rx}_{ij} = \lambda^x_0 + \lambda^x_1 y_{ij} + \kappa x_{ij}$*. PL method:* - - -*; TS method:* ———*.* 79

3.2 *Relative efficiency of estimators for* β_1*. The TS method:* ━━━*; the PL method:* ———*.* 80

Chapter 1

Introduction

1.1 Overview

Longitudinal data arise commonly in many fields including clinical trials and health research. Data are typically collected by following up individuals over a period of time. Statistical methods for longitudinal analysis have been quickly developed over the past few decades (e.g. Laird and Ware, 1982; Diggle et al., 2002). For dealing with different research interests in longitudinal data, three classes of models are commonly employed in applications. The three classes of models are mixed effects models, marginal models, and transition models (Fitzmaurice et al., 2004).

Mixed effects models are desirable when research focuses on the response for an individual rather than for the entire population. Early studies of this area involves linear mixed models for repeated measurements proposed by Harville (1977) and Patterson and Thompson (1971), who develop the restricted maximum likelihood (REML) to modify the loss of degree of freedom issue arising in the estimation for the variance of components. Laird and Ware (1982) propose estimation method for linear mixed effects models using the EM algorithm (Dempster et al., 1977) and the empirical Bayes method. Extensions that accommodate both linear mixed effects models (LMM) and generalized linear models

(GLM) (McCullagh and Nelder, 1989) are generalized linear mixed effects models (GLMM) (Breslow and Clayton, 1993; Stiratelli et al., 1984), which have been widely used for various settings.

Marginal models are commonly used in population studies. A typical estimation method for marginal models is the so-called generalized estimating equations (GEE) approach (Liang and Zeger, 1986). Early theoretical discussions on estimating functions include Godambe (1960, 1976) and Godambe and Thompson (1984). Liang and Zeger (1986), Zeger and Liang (1986) and Zeger et al. (1988) introduce the idea of estimating functions into the setting of longitudinal studies. The GEE method does not require specification of the full joint distribution for the longitudinal data, but only the marginal structure. In its implementation, a working correlation matrix is called in if the true association structure for longitudinal data is not modeled. Consistent estimates of parameters in the marginal structure can be obtained, provided the mean structure is correctly specified. An extension of the GEE method, named GEE2, is discussed by Prentice (1988) and Zhao and Prentice (1990) among others. The GEE2 approach facilitates estimation of association parameters.

Transitional models (Molenberghs and Verbeke, 2005) focus on modeling the dependence of individual's response on its history, together with covariates. Therefore, it is convenient if the research interest lies in the influence of previous outcomes on the current response. Frequently, transition models are formulated in conjunction with certain Markov conditions, which restrict the dependence of the current response to a limited number of past observations.

Longitudinal Data Arising in Clusters

In many situations, longitudinal data arise in clusters. A typical case is sociological survey studies that involve communities, families or schools with repeated assessments of individual members over time. For example, Payment et al. (1991) conduct a randomized intervention trial based on 606 households. The study measures the health outcomes of interest for each household member over a 15-month period. Cameron et al. (1999) study the social influences on smoking prevention by following 100 elementary schools from grades 6 to 8.

There are many potential goals when analyzing longitudinal data arising in clusters. For example, Roy and Lin (2002) discuss the EM algorithm to handle outcomes with nonignorable dropouts and missing covariates. Yi and Cook (2002) propose a weighted GEE approach to handle longitudinal data arising in clusters with missingness. Fieuws and Verbeke (2006) discuss a pairwise fitting strategy under the framework of mixed models.

1.2 Modeling Strategies

In this section, we introduce basic notations and symbols. Suppose that there are n subjects with m visits. Let Y_{ij} denote the response for subject i at visit j, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Take $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im})^T$, $i = 1, 2, \ldots, n$. Let $X_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$ be the $p \times 1$ covariate vector for subject i in visit j, and $X_i = (X_{i1}^T, X_{i2}^T, \ldots, X_{im}^T)^T$. The interest of longitudinal studies usually lies in understanding the relationship between response Y_i and covariates X_i. In particular, we let $f(Y_i|X_i; \theta)$ denote the conditional probability density or mass function of Y_i given X_i, where parameter θ takes values in a parameter space Θ.

1.2.1 Generalized Linear Models

Specification of $f(Y_i|X_i; \theta)$ often involves modeling the marginal distribution $f(Y_{ij}|X_i; \theta)$ for which generalized linear models (GLM) family can be introduced with

$$f(Y_{ij}|X_i; \theta) = \exp\Big[\big\{Y_{ij}\tau_{ij} - b(\tau_{ij})\big\}/a(\phi) + c(Y_{ij}; \phi)\Big], \tag{1.1}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are some specific functions, ϕ is a scale parameter and τ_{ij} is the canonical parameter with $E(Y_{ij}|X_i) = b'(\tau_{ij})$ and $Var(Y_{ij}|X_i) = a(\phi)b''(\tau_{ij})$. We further assume that the marginal distribution of Y_{ij} depends only on the covariate vector for subject

i at time j (Pepe and Anderson, 1994), and thus $f(Y_{ij}|X_i;\theta) = f(Y_{ij}|X_{ij};\theta)$. Furthermore, a regression model can be introduced as

$$h\big\{E(Y_{ij}|X_i)\big\} = X_{ij}^T\boldsymbol{\beta},$$

where h is a differentiable monotone link function, and $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients. Particularly, the canonical link function h satisfies $\tau_{ij} = h\big\{E(Y_{ij}|X_i)\big\} = X_{ij}^T\boldsymbol{\beta}$.

If the Y_{ij} are assumed to be independent for all $j = 1, \ldots, m$, given covariates X_i, the $f(Y_i|X_i;\theta)$ is then given by

$$f(Y_i|X_i;\theta) = \prod_{j=1}^{m} f(Y_{ij}|X_i;\theta).$$

However, this assumption is normally invalid for longitudinal settings. Therefore, various types of joint distributions are proposed to feature different association structures of longitudinal data. For instance, multivariate normal distributions are commonly employed to handle continuous data, and multivariate probit models (e.g. Ashford and Sowden, 1970; Ochi and Prentice, 1984) are used for binary outcomes. Although there are some available multivariate distributions, directly modeling the joint distribution of $f(Y_i|X_i;\theta)$ for individual applications still remains to be a daunting task if not impossible (Lindsay et al., 2011).

1.2.2 Generalized Linear Mixed Models

Generalized linear mixed models (GLMMs) are used to handle associated observations by adding random effects and further assuming independence for the Y_{ij} $(j = 1, \ldots, m)$ given covariates and random effects. Denote u_i to be the vector for random effects. Let $Z_{ij} = (Z_{ij1}, \ldots, Z_{ijq})^T$ be the $q \times 1$ random effects covariate vector for subject i at visit j, and $Z_i = (Z_{i1}^T, Z_{i2}^T, \ldots, Z_{im}^T)^T$. Z_i is most typically a subset of columns of X_i. Given random effects u_i and covariates X_i and Z_i, the conditional distribution of Y_{ij} is given by

$$f(Y_{ij}|X_i, Z_i, u_i;\theta) = \exp\Big[\{Y_{ij}\tau_{ij} - b(\tau_{ij})\}/a(\phi) + c(Y_{ij};\phi)\Big], \tag{1.2}$$

where with similar notations to (1.1), $\mathrm{E}(Y_{ij}|X_i, Z_i, u_i) = b'(\tau_{ij})$, $\mathrm{Var}(Y_{ij}|X_i, Z_i, u_i) = a(\phi)b''(\tau_{ij})$ and the regression model is specified as

$$h\{\mathrm{E}(Y_{ij}|X_i, Z_i, u_i; \theta)\} = X_{ij}^T\boldsymbol{\beta} + Z_{ij}^T u_i,$$

in which, again, $f(Y_{ij}|X_i, Z_i, u_i; \theta) = f(Y_{ij}|X_{ij}, Z_{ij}, u_i; \theta)$ is assumed.

As a result, the joint distribution of $f(Y_i|X_i; \theta)$ is obtained by integrating out the unobservable random effects u_i:

$$f(Y_i|X_i, Z_i) = \int \left\{ \prod_{j=1}^{m} f(Y_{ij}|X_i, Z_i, u_i) \right\} f(u_i) du_i, \tag{1.3}$$

where $f(u_i)$ is the joint distribution for random effects.

The integrals in (1.3) can be intractable as there are generally no closed forms in GLMM settings. To deal with this issue, many algorithms are developed to approximate the integrals, such as Gauss-Hermite quadrature (Longford, 1994), Laplacian approximation, adaptive Gauss-Hermite quadrature (Pinheiro and Bates, 1995), penalized quasi-likelihood (Breslow and Clayton, 1993), marginal quasi-likelihood (Goldstein, 2002), Monte Carlo Newton-Raphson and Monte Carlo EM (McCulloch, 1997; Booth and Hobert, 1999).

1.2.3 Generalized Estimating Equations

Generalized estimating equations (GEE) methods circumvent the direct modeling on $f(Y_i|X_i; \theta)$ by basing inference on appropriately "combining" marginal distribution elements of Y_i. For simplicity, we rewrite the notations in (1.1) with $E(Y_{ij}|X_i) = \mu_{ij}$ and $Var(Y_{ij}|X_i) = v_{ij}$. Take $\mu_i = (\mu_{i1}, \ldots, \mu_{im})^T$, and $\theta = (\boldsymbol{\beta}^T, \xi^T)^T$, where ξ represents all parameters other than $\boldsymbol{\beta}$. Define

$$U_i(\boldsymbol{\beta}, \xi) = D_i V_i^{-1}(Y_i - \mu_i), \tag{1.4}$$

where $D_i = \partial\mu_i^T / \partial\boldsymbol{\beta}$, $V_i = B_i^{1/2} R_i(\xi) B_i^{1/2}$, $B_i = \mathrm{diag}(v_{i1}, \ldots, v_{im})$, and $R_i(\xi)$ is a working correlation matrix for Y_i. The GEE approach estimates $\boldsymbol{\beta}$ by solving

$$\sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \xi) = \sum_{i=1}^{n} D_i V_i^{-1}(Y_i - \mu_i) = \mathbf{0},$$

where the correlation parameters ξ are treated as nuisance. Parameters ξ can be estimated via the method of moments given $\boldsymbol{\beta}$ (Liang and Zeger, 1986). An advantage of the GEE approach is that the estimator of regression coefficients $\boldsymbol{\beta}$ is robust even if the correlation structure $R_i(\xi)$ is misspecified.

1.3 Composite Likelihood

Composite likelihood, initiated by Besag (1975, 1977) and further developed by Lindsay (1988), Arnold and Strauss (1991) and Cox and Reid (2004), provides a useful inference alternative in place of the full likelihood based inference. Instead of specifying the full distribution, we only need to specify some partial structures of $f(Y_i|X_i;\theta)$ in the composite likelihood formulation. The composite likelihood method can ease issues related to complex modeling. Moreover, inference results based on the composite likelihood formulations are robust in the sense that association structures higher than those used in the formulation can be misspecified. These advantages become more obvious when the dimension of Y_i increases.

Efficiency loss is the typical price that the composite likelihood method pays as opposed to the likelihood approach. Geys et al. (1997, 1998) confirm that the composite likelihood estimators are less efficient than maximum likelihood. Kuk (2007) claims that the pairwise likelihood inference can be inefficient and a hybrid pairwise likelihood method is proposed to augment efficiency. Simulation studies by Troxel et al. (1998) empirically demonstrate that inference based on the marginal likelihood method is less efficient than that of the full likelihood method.

Many applications of composite likelihood methods can be found in a variety of settings. To name some, for example, Heagerty and Lele (1998), Curriero and Lele (1999) and Varin et al. (2005) discuss the composite likelihood estimation approach for binary spatial data analysis, while Fearnhead and Donnelly (2002) use the composite likelihood idea to handle genetic data. Hanfelt (2004) takes the composite conditional likelihood approach for sparse clustered data. Chatelain et al. (2008) study pairwise likelihood estimation for multivariate mixed Poisson models. Multilevel probit models are discussed with the composite likelihood method by Kuk and Nott (2000). Renard et al. (2002), Zhao and Joe (2005) and Joe and Lee (2009) conduct pairwise likelihood inferences for analyzing correlated binary data. Yi et al. (2011b) and He and Yi (2011) employ the composite likelihood method to handle clustered binary data with missing observations. Wei et al. (1989) use marginal distribution to handle multivariate incomplete failure time data. Parner (2001) uses composite likelihood to analyze familial survival data. Gao and Song (2011) propose the composite likelihood EM algorithm and apply it to handle multivariate hidden Markov models. Detailed discussion and review on the composite likelihood method can be found in Lindsay et al. (2011), Varin (2008) and Varin et al. (2011).

1.3.1 Formulation of Composite Likelihood

With longitudinal response $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im})^T$, we consider the composite likelihood formulation following Lindsay et al. (2011):

$$C(\theta) = \prod_{k=1}^{N_{cl}} \left\{ L(S_k; \theta) \right\}^{w_k}, \tag{1.5}$$

where N_{cl} is the number of factors in $C(\theta)$, each $L(S_k; \theta)$ is a user-selected sub-likelihood generated from $f(Y_i; \theta)$ with S_k being a conditional or marginal set of variables, and w_k is a certain weight.

For example, if S_k consists of a single or paired response component, the log composite likelihood contributed from subject i is given by

$$\ell_c(Y_i; \theta) = \sum_{j<j'} \ell_c(Y_{ij}, Y_{ij'}; \theta) = \sum_{j<j'} \left\{ B_{ijj'} \ell(Y_{ij}, Y_{ij'}; \theta) - B_{ij} \ell(Y_{ij}; \theta) - B_{ij'} \ell(Y_{ij'}; \theta) \right\}, \tag{1.6}$$

where $B_{ijj'}$, B_{ij} and $B_{ij'}$ are scalar weights. When all $B_{ijj'} = 1$ and $B_{ij} = B_{ij'} = 0$, equation (1.6) results in all-pairwise marginal log likelihood (APW), obtained by considering $\prod_{j<j'} f(Y_{ij}, Y_{ij'}; \theta)$. When all $B_{ijj'} = 2$ and $B_{ij} = B_{ij'} = 1$, equation (1.6) gives

all-pairwise conditional log likelihood (APC), obtained by considering $\prod_{j\neq j'} f(Y_{ij}|Y_{ij'};\theta)$. Pairwise marginal or pairwise conditional likelihood are perhaps the most widely used formulations. In our following discussions, we will focus on such forms.

1.3.2 Statistical Properties

Consistency

Under regularity conditions, equation (1.6) provides consistent estimators of θ, since all elements in the right side of the equation have zero expectation, $E\left\{\partial \log f(Y_{ij}, Y_{ij'};\theta)/\partial\theta\right\} = \mathbf{0}$. That is,

$$\begin{aligned} & E\Big\{\frac{\partial \log f(Y_{ij}, Y_{ij'};\theta)}{\partial\theta}\Big\} \\ = & \int \frac{\partial \log f(Y_{ij}, Y_{ij'};\theta)}{\partial\theta} f(Y_{i1}, Y_{i2}, \ldots, Y_{im};\theta) dY_{i1} dY_{i2}\cdots dY_{im} \\ = & \int \frac{\partial \log f(Y_{ij}, Y_{ij'};\theta)}{\partial\theta} f(Y_{ij}, Y_{ij'};\theta) dY_{ij} dY_{ij'} = \mathbf{0}. \end{aligned}$$

Efficiency

Next, we consider possible efficiency loss in contrast to the full likelihood method. Let $S(\theta) = \sum_{i=1}^n \partial \log f(Y_i;\theta)/\partial\theta$ be the score function obtained from the full likelihood, and $H(\theta) = \sum_{i=1}^n \sum_{j<j'} \partial \ell_c(Y_{ij}, Y_{ij'};\theta)/\partial\theta$ be the composite score function, respectively. The Godambe information matrix (Godambe, 1991) is then given by

$$I_H(\theta) = E\{\partial H(\theta)/\partial\theta\}^T \Big[E\{H(\theta)H^T(\theta)\}\Big]^{-1} E\{\partial H(\theta)/\partial\theta\},$$

and

$$I_S(\theta) = E\{S(\theta)S^T(\theta)\},$$

for the composite and full likelihood, respectively. If θ is a scalar, Lindsay (1988) indicates that

$$I_H(\theta) = \frac{\mathrm{Cov}^2\{H(\theta), S(\theta)\}}{\mathrm{Var}(H(\theta))} = \rho^2_{H(\theta),S(\theta)} I_S(\theta) \tag{1.7}$$

where $\rho_{H(\theta),S(\theta)}$ denotes the linear correlation coefficient between $H(\theta)$ and $S(\theta)$. Therefore, it implies that compared to the full likelihood, the composite likelihood method may incur efficiency loss.

To further explain the efficiency loss issue, we propose a general framework to portray the relationship between the full likelihood and the composite likelihood derived from equation (1.6):

$$\log f(Y_i;\theta) = k\left\{\sum_{j<j'} \ell_c(Y_{ij}, Y_{ij'};\theta) + \sum_{j<j'} \tilde{\ell}_{ijj'}(\theta)\right\}, \tag{1.8}$$

where $k = 1/\{\sum_{j<j'} B_{ijj'}\}$ $(\sum_{j<j'} B_{ijj'} \neq 0)$ and $\tilde{\ell}_{ijj'}(\theta)$ has

$$B_{ij}\ell(Y_{ij};\theta) + B_{ij'}\ell(Y_{ij'};\theta) + B_{ijj'}\log f(Y_i|Y_{ij}, Y_{ij'};\theta).$$

It can be seen from equation (1.8) that composite likelihood can be viewed as a partial "section" from full likelihood with a term (i.e. $\tilde{\ell}_{ijj'}(\theta)$) removed.

Let $\tilde{H}(\theta) = \sum_{i=1}^{n}\sum_{j<j'} \partial\tilde{\ell}_{ijj'}(\theta)/\partial\theta$. Suppose we still assume θ to be scalar, and apply

the argument in equation (1.7), we can obtain

$$
\begin{aligned}
& I_H(\theta) \\
= \; & \frac{\operatorname{cov}^2(S(\theta), H(\theta))}{\operatorname{Var}(H(\theta))} \\
= \; & \frac{\operatorname{cov}^2(k(H(\theta) + \tilde{H}(\theta)), H(\theta))}{\operatorname{Var}(H(\theta))} \\
= \; & \frac{k^2 \Big\{ \operatorname{Var}(H(\theta))^2 + 2\operatorname{cov}(H(\theta), \tilde{H}(\theta))\operatorname{Var}(H(\theta)) + \operatorname{cov}^2(H(\theta), \tilde{H}(\theta)) \Big\}}{\operatorname{Var}(H(\theta))} \\
= \; & k^2 \Big\{ \operatorname{Var}(H(\theta)) + 2\operatorname{cov}(H(\theta), \tilde{H}(\theta)) + \frac{\operatorname{cov}^2(H(\theta), \tilde{H}(\theta))}{\operatorname{Var}(H(\theta))} \Big\} \\
= \; & k^2 \Big\{ \operatorname{Var}(H(\theta)) + 2\operatorname{cov}(H(\theta), \tilde{H}(\theta)) + \operatorname{Var}(\tilde{H}(\theta)) - \operatorname{Var}(\tilde{H}(\theta)) \\
& + \frac{\operatorname{cov}^2(H(\theta), \tilde{H}(\theta))}{\operatorname{Var}(H(\theta))\operatorname{Var}(\tilde{H}(\theta))} \operatorname{Var}(\tilde{H}(\theta)) \Big\} \\
= \; & k^2 \Big\{ \operatorname{Var}(H(\theta)) + 2\operatorname{cov}(H(\theta), \tilde{H}(\theta)) + \operatorname{Var}(\tilde{H}(\theta)) - \operatorname{Var}(\tilde{H}(\theta))(1 - \rho^2_{H(\theta), \tilde{H}(\theta)}) \Big\} \\
= \; & k^2 \Big\{ \operatorname{Var}(H(\theta) + \tilde{H}(\theta)) - \operatorname{Var}(\tilde{H}(\theta))(1 - \rho^2_{H(\theta), \tilde{H}(\theta)}) \Big\} \\
= \; & I_S(\theta) - k^2 \operatorname{Var}(\tilde{H}(\theta))(1 - \rho^2_{H(\theta), \tilde{H}(\theta)}).
\end{aligned}
$$

Thus, we can have an intuitive idea that the information loss of composite likelihood depends on both the variance of the “removed” term and the correlation between the composite likelihood score function and the “removed” term.

1.3.3 Computational Issue

The lower-dimension modeling strategy in composite likelihoods leads to computation cheapness in many studies. In particular, it reduces the dimensions of integrals in many scenarios. For example, GLMM models with crossed random effects often involve high-dimensional intractable integrals. Bellio and Varin (2005) propose pairwise likelihood approach to reduce 20-dimensional integrals to 3-dimensional integrals in the analysis of salamander mating data. Troxel et al. (1998) use the implementation of marginal likelihood to reduce high-dimensional integrals for longitudinal data analysis. Moreover, Fieuws and Verbeke (2006) argue that computation can become difficult as the dimension of the random-effects vector increases, even in the case of linear mixed models where the integrals can be calculated analytically. They introduce a pairwise modeling strategy to circumvent this problem.

Parzen et al. (2007) and Lindsay et al. (2011) discuss that the calculation of the likelihood functions for all pairs can be computational expensive. If the composite likelihood functions include all bivariate distributions, the number of pairs could also increase fast as the data dimension increases. However, this issue of composite likelihood could be handled with parallel computing facilities (Almasi and Gottlieb, 1989), in which many simpler calculations are carried out simultaneously under the computer architecture with multicore processors. Thus, the composite likelihood is promising in many applications with parallel computing resources.

1.4 Model Selection

Model selection is an important topic in statistical inference. When more than one model is possible to fit the data, we are interested in selection of the one that fits data the best or nearly the best. To achieve this goal, many approaches are developed. Below we discuss several strategies of model selection.

1.4.1 Best Subset Selection

A large family of model selection methods is based on the best subset selection. Normally, the best subset selection first conducts likelihood estimation for all possible candidate models, and then calculates a measure corresponding to a certain criterion for each model. The candidate model with minimum (or maximum) criterion value would be preferred.

Denote $\ell(Y;\theta)$ to be the log full likelihood function. Some well-known information criteria involve the Akaike information criterion (AIC) (Akaike, 1973)

$$\text{AIC} = -2\ell(Y;\theta) + 2k,$$

where k is the dimension of θ, and Bayesian information criterion (BIC) (Schwarz, 1978)

$$\text{BIC} = -2\ell(Y;\theta) + k\log n.$$

Further studies in this area include Konishi et al. (2004) for applying the BIC criterion to the choice of smoothing parameters and the adaptive model selection approach proposed by Shen and Ye (2002).

Note that the AIC/BIC methods can only be applied when a full likelihood function is available. Varin and Vidoni (2005) discuss a composite likelihood Akaike information criterion (cAIC) with

$$\text{cAIC} = -2\ell_c(Y;\theta) + 2 \times \text{df}(\theta),$$

where $\ell_c(Y;\theta)$ is the log composite likelihood function and the effective number of degrees of freedom $\text{df}(\theta)$ is defined as

$$\text{df}(\theta) = \text{tr}\big\{J(\theta)H^{-1}(\theta)\big\}.$$

Here $J(\theta) = \sum_{i=1}^{n}\{\partial\ell_c(Y_i;\theta)/\partial\theta\}\{\partial\ell_c(Y_i;\theta)/\partial\theta\}^T$ and $H(\theta) = -\partial^2\ell_c(Y;\theta)/\partial\theta\partial\theta^T$. Gao and Song (2010) propose a composite likelihood Bayesian information criteria (cBIC) with

$$\text{cBIC} = -2\ell_c(Y;\theta) + \log(n) \times \text{df}(\theta),$$

1.4.2 Penalized Likelihood

Although the best subset selection is widely used in statistical inference, Fan and Li (2001, 2006) argue that these selection procedures ignore stochastic errors inherited in the stages of variable selections, and their computational time increases exponentially with the parameter dimensionality. To overcome this problem, many techniques involving simultaneous estimation and variable selection are developed. These include the bridge regression (Frank and Friedman, 1993), the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996, 2011), smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), least angle regression (LARS) (Efron et al., 2004), elastic net (Zou and Hastie, 2005), adaptive LASSO (Zou, 2006), minimax concave (MCP) penalty (Zhang, 2010) and LASSO regression with the strong heredity constraint (Choi et al., 2010). Other studies of model selections such as single index methods can be found in Naik and Tsai (2001), Kong and Xia (2007), etc.

Fan and Li (2001, 2004, 2006) propose a unified penalized likelihood framework that extends these approaches to generalized linear models (GLM). Similar extensions can also be found in Park and Hastie (2007). For variable selection, a penalized likelihood can be written as

$$\ell_{pen}(Y;\theta) = \ell(Y;\theta) - n\sum_{s=1}^{p} p_\lambda(|\beta_s|), \tag{1.9}$$

where $p_\lambda(|\beta_s|)$ is a penalty function for the s-th element in $\boldsymbol{\beta}$. Various penalty functions can be implemented. For instance, the SCAD penalty (Fan and Li, 2001) is

$$p_\lambda(|\beta_s|) = \lambda\int_0^{|\beta_s|} \min\{1, \frac{(a\lambda - x)_+}{(a-1)\lambda}\}dx,$$

and LASSO penalty (Tibshirani, 1996) is taken as

$$p_\lambda(|\beta_s|) = \lambda|\beta_s|,$$

for some $a > 2$ and $\lambda > 0$.

According to the above examples, it can be seen that the variable selection can be achieved by introducing penalized functions. The influence of the penalty can be simply described as “pressing down except zero”, which leads to a function that is much easier to have extreme value at zero. To further illustrate this, we consider a toy example with only one observation y and one parameter β for regression model $y = \beta + \epsilon$, where $\epsilon \sim N(0, 1)$. Then the log likelihood function is logL $= -\log(\sqrt{2\pi}) - (y - \beta)^2/2$ and the penalized log likelihood function is PlogL $= -\log(\sqrt{2\pi}) - (y - \beta)^2/2 - p_\lambda(|\beta|)$. Here we set $a = 3.7$, $\lambda = 0.5$ and $y = 0, -0.5, 0.5, 2$, and plot both SCAD and LASSO functions against different values of β. The likelihood estimates are obtained by maximizing likelihood functions with respect to β. Let $\hat{\beta}_{logL}$ and $\hat{\beta}_{PlogL}$ denote the estimates from logL and PlogL likelihood functions, respectively. Figures 1.1 and 1.2 display the comparison between logL and PlogL with SCAD and LASSO penalties, respectively. It can be seen that the SCAD and LASSO penalties "press down" the likelihood functions except for the points with $\beta = 0$. Therefore, comparing with $\hat{\beta}_{logL}$, $\hat{\beta}_{PlogL}$ is more likely to have $\hat{\beta}_{PlogL} = 0$.

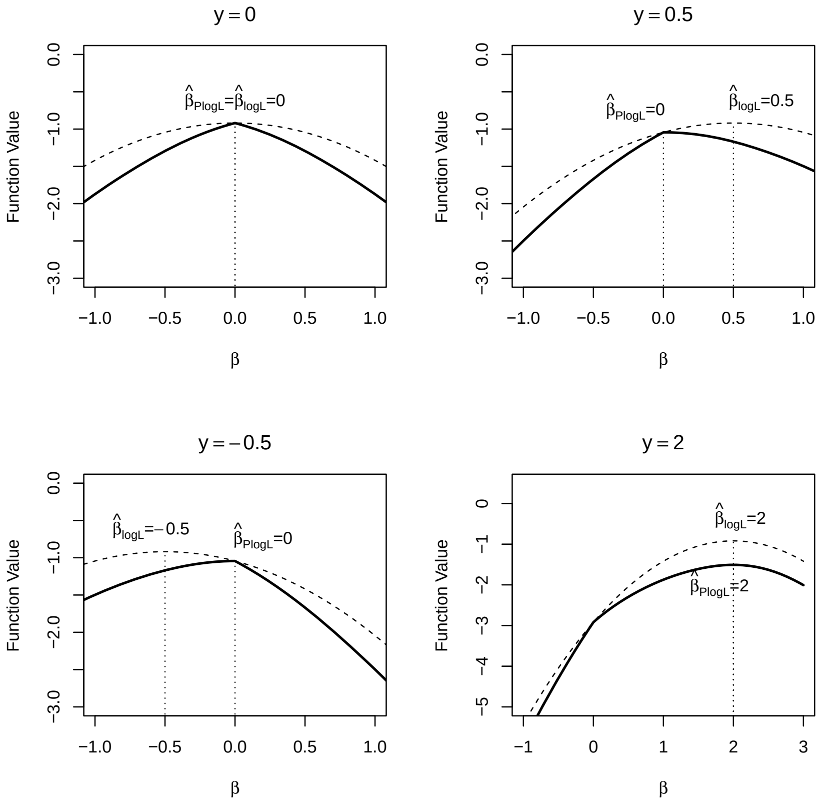

Figure 1.1: *Comparison between the log likelihood function (logL) and the penalized log likelihood function (PlogL) with SCAD penalty. - - -: logL function; ——: PlogL function. The estimates from logL and PlogL are labeled as $\hat{\beta}_{logL}$ and $\hat{\beta}_{PlogL}$, respectively.*

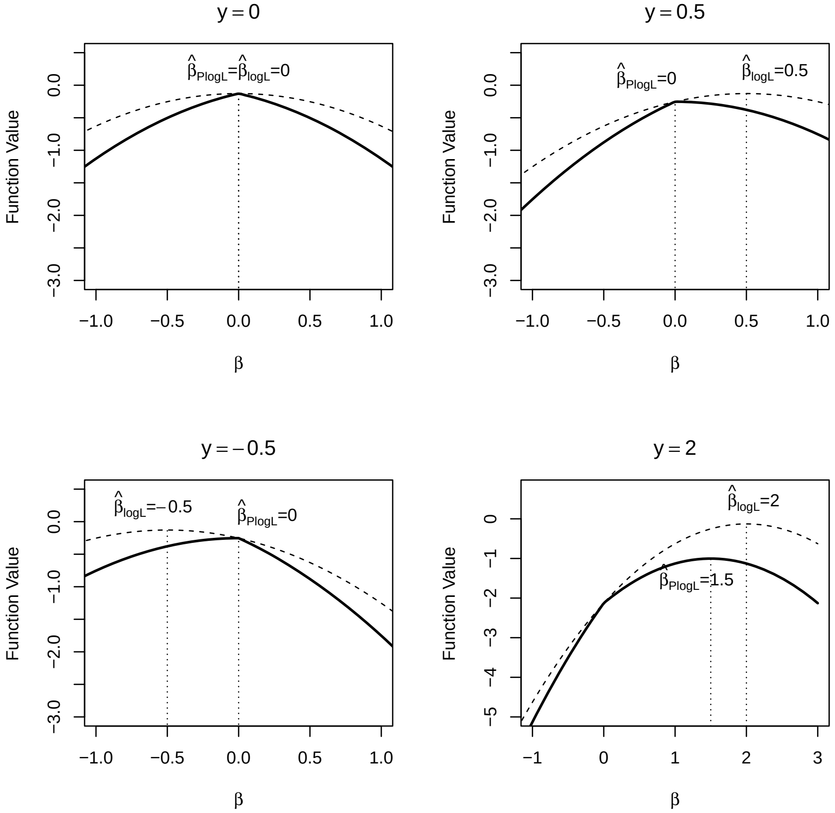

Figure 1.2: *Comparison between the log likelihood function (logL) and the penalized log likelihood function (PlogL) with LASSO penalty. - - -: logL function; ——: PlogL function. The estimates from logL and PlogL are labeled as $\hat{\beta}_{logL}$ and $\hat{\beta}_{PlogL}$, respectively.*

Now we discuss the implementation procedure. In principle, model selection and estimation results can be obtained by maximizing the penalized log likelihood function in (1.9). However, the penalty functions such as SCAD and LASSO are singular at the origin, and they do not have continuous second order derivatives. Following Fan and Li (2001), a local approximation approach is available to approximate the penalty term by a quadratic function. This approach is based on the fact that when β_s is close to the true value β_{s0}, we have

$$\big\{p_\lambda(|\beta_s|)\big\}' = p'_\lambda(|\beta_s|)\mathrm{sgn}(\beta_s) \approx \{p'_\lambda(|\beta_{s0}|)/|\beta_{s0}|\}\beta_s,$$

for $\beta_{s0} \neq 0$. Then we have

$$\big\{p_\lambda(|\beta_s|)\big\}'' \approx \Big[\{p'_\lambda(|\beta_{s0}|)/|\beta_{s0}|\}\beta_s\Big]' = p'_\lambda(|\beta_{s0}|)/|\beta_{s0}|,$$

which leads to the Newton-Raphson algorithm that can be used in searching for the estimates.

1.4.3 Model Selection for Longitudinal Data

Recently, many researchers extend model selection methods to longitudinal data analysis. To deal with a correlated dataset, Liu et al. (1999) propose a generalized cross-validation selection method based on the Predicted Residual Sum of Squares (PRESS). Pauler (1998) proposes a BIC method for choosing fixed effects in normal linear mixed models, and Weiss et al. (1997) conduct fixed effects selection in random effects models using Bayesian approaches. Pinheiro and Bates (2000) discuss the use of likelihood ratio tests, AIC and BIC for selecting fixed effects and random effects under mixed effect models.

Much recent work focuses on the model selection on both fixed and random effects in longitudinal data studies. Yafune et al. (2005) discuss an extended information criterion and Vaida and Blanchard (2005) discuss a conditional Akaike information criterion, respectively. Moreover, Smith and Kohn (2002), Chen and Dunson (2003) and Kinney and Dunson (2007) propose Bayesian approaches for fixed and random effects selections. Under the penalized likelihood framework, Bondell et al. (2010) discuss the penalized joint likelihood method with an adaptive penalty for the selection and estimation of both fixed and random effects, and Ibrahim et al. (2010) propose a method for a general class of mixed effects models using maximum penalized likelihood estimation along with SCAD and adaptive LASSO penalty functions.

Furthermore, semiparametric models (Diggle et al., 2002) are widely adopted to analyze longitudinal data with parametric fixed effects to represent covariate effects and a smooth function to model the time effects. Fan and Li (2004) propose model selection and estimation procedures for regression covariates with semi-parametric models. Ni et al. (2010) discuss a double-penalized likelihood approach, where two types of penalties are jointly imposed on the ordinary log-likelihood: the roughness penalty on the nonparametric baseline function and a nonconcave shrinkage penalty on linear coefficients to accommodate model sparsity. Other work related to longitudinal model selection includes penalized GEE approaches discussed by Fu (2003), Johnson et al. (2008) and Tong et al. (2009).

1.5 Missing Data in Longitudinal Studies

Suppose we fit a dataset with model $f(Y_i|X_i;\theta)$ and the observations involve incomplete response. Let $R_i = (R_{i1}, \ldots, R_{im})^T$ be the corresponding missing data indicator vector, where $R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ if Y_{ij} is missing.

Monotone missing data patterns occur if a subject misses one assessment, returning to the study is impossible. That is, $R_{ij} = 0$ implies $R_{ij'} = 0$ whenever $j' > j$. Monotone missingness is also phrased as *drop-out*. Otherwise, missing data patterns are called *non-monotone*. That is, a subject may miss one assessment, but returning to the study is still possible, this is also referred to as intermittent missingness.

For ease of exposition, sometimes we write $Y_i = (Y_i^{obs}, Y_i^{mis})$, where Y_i^{obs} and Y_i^{mis} represent subvectors consisting of observed and unobserved components of Y_i, respectively. Either Y_i^{obs} or Y_i^{mis} can be null, depending on whether or not Y_{ij} $(j = 1, \ldots, m)$ is observed.

1.5.1 Missing Data Mechanism

Early work on dealing with missing data involves complete-case/available-data analysis (Kim and Curry, 1977) and naive imputation missing values (Buck, 1960). Recent work is generally based on the framework discussed by Rubin (1976) and Little and Rubin (2002). Missing data mechanism is often classified into three classes: *missing completely at random (MCAR)*, *missing at random (MAR)* and *missing not at random (MNAR)*. MCAR features the situation where the missing data probability is independent of the variables subject to missingness, given covariates

$$P(R_i|Y_i, X_i) = P(R_i|X_i).$$

MAR says that given covariates, the missing data probability may depend on the variables prone to missingness, but only depend on the observed variables:

$$P(R_i|Y_i, X_i) = P(R_i|Y_i^{obs}, X_i).$$

MNAR facilitates the most general situation for which the missing data probability can depend on the unobserved data, even conditional on covariates:

$$P(R_i|Y_i, X_i) = P(R_i|Y_i^{obs}, Y_i^{mis}, X_i).$$

1.5.2 Likelihood-Based Methods

Likelihood approaches for incomplete longitudinal data are developed by constructing the joint distribution of response variable Y_i and the missing data indicators R_i, given the covariates X_i. Three classes of likelihood-based models are commonly applied. One is based on the so-called *selection models* (Little and Rubin, 2002) with the joint distribution of Y_i and R_i factorized as

$$f(R_i, Y_i|X_i; \theta, \alpha) = f(R_i|Y_i, X_i; \alpha) f(Y_i|X_i; \theta), \tag{1.10}$$

where the distribution of R_i given response and covariates involves parameter α, which is assumed to be functionally independent of θ, the parameter vector associated with the response model. Another approach is *pattern-mixture models* (Little, 1995; Thijs et al., 2002), in which the factorization of the joint distribution is

$$f(R_i, Y_i|X_i; \delta, \gamma) = f(Y_i|R_i, X_i; \delta)f(R_i|X_i; \gamma),$$

where the distribution of Y_i is modeled conditionally on both covariates and missing data indicators, and parameters δ and γ are often assumed to be distinct.

Furthermore, *shared-parameter models* (Wu and Carroll, 1988; Albert and Follmann, 2003) assume that Y_i and R_i are conditionally independent, given a random variable ξ_i, therefore, we can write

$$f(R_i, Y_i|X_i; \delta, \gamma) = \int f(Y_i|X_i, \xi_i; \delta)f(R_i|X_i, \xi_i; \gamma)f(\xi_i)\, d\xi_i,$$

where $f(\xi_i)$ is the density function for the random variable ξ_i.

When the research interest focuses on the model of $f(Y_i|X_i; \theta)$, it is often natural to use selection models, where the response model does not include any missing indicators. In this thesis, we limit the discussion mainly to selection models. In particular, inference can be achieved using the observed likelihood

$$L_i(Y_i^{obs}, R_i|X_i; \theta, \alpha) = \int f(Y_i^{obs}, Y_i^{mis}|X_i; \theta)f(R_i|Y_i^{obs}, Y_i^{mis}, X_i; \alpha)dY_i^{mis}. \qquad (1.11)$$

When the missing mechanism is MAR (or MCAR), equation (1.11) becomes

$$\begin{aligned} L_i(Y_i, R_i|X_i; \theta, \alpha) &= \int f(Y_i^{obs}, Y_i^{mis}|X_i; \theta)f(R_i|Y_i^{obs}, Y_i^{mis}, X_i; \alpha)dY_i^{mis} \\ &= \int f(Y_i^{obs}, Y_i^{mis}|X_i; \theta)f(R_i|Y_i^{obs}, X_i; \alpha)dY_i^{mis} \\ &= f(R_i|Y_i^{obs}, X_i; \alpha) \cdot \int f(Y_i^{obs}, Y_i^{mis}|X_i; \theta)dY_i^{mis} \\ &= f(R_i|Y_i^{obs}, X_i; \alpha) \cdot f(Y_i^{obs}|X_i; \theta). \end{aligned}$$

Since we also assume α and θ to be functionally independent, inference about θ can directly be conducted based on the model $f(Y_i^{obs}|X_i; \theta)$ for the observed data only and the missing data model can be ignored. When the missing mechanism is MNAR, the integrals in the likelihood function (1.11) are often intractable.

Maximization of the observed likelihood can be implemented using the Newton-Raphson algorithm. However, the Newton-Raphson algorithm could be sensitive to the initial values. An alternative approach for handling missing data is to use the so-called *Expectation Maximization (EM)* algorithm (Dempster et al., 1977). To be specific, the EM algorithm involves the E and M steps. In the E-step, we evaluate the conditional expectation of the complete data log likelihood:

$$Q_i(\theta, \alpha|\theta^{(t)}, \alpha^{(t)}) = E\big\{ \log L_i(Y_i, R_i|X_i; \theta, \alpha)|Y_i^{obs}, X_i, R_i; \theta^{(t)}, \alpha^{(t)}\big\} \qquad (1.12)$$

where $\theta^{(t)}$ and $\alpha^{(t)}$ denote the parameters' value estimated from the previous tth iteration, and $L_i(Y_i, R_i|X_i; \theta, \alpha)$ is the complete data likelihood contributed from subject i, which is determined by (1.10). The M-step maximizes $Q_i(\theta, \alpha|\theta^{(t)}, \alpha^{(t)})$ with respect to parameters θ and α, and the maximizer is taken as $\theta^{(t+1)}$ and $\alpha^{(t+1)}$. The EM algorithm iterates the E and M steps until $(\theta^{(t+1)}, \alpha^{(t+1)})$ reaches convergence.

We comment on the numerical performance of the Newton-Raphson and EM methods. Although directly maximizing observed likelihood functions via Newton-Raphson can reach the estimation purpose, the maximization might be very sensitive to starting values. Poor starting values can lead to the failure of convergence. The EM algorithm is relatively more stable but is subject to slow convergence. Often, a combined the Newton-Raphson and the EM approach is used, where the algorithm starts with the EM, then the Newton-Raphson is used for speed after a certain number of iterations.

In implementing the E-step, commonly, the integrals in equation (1.12) have no analytical form. A typical method to handle this is the so-called MC-EM algorithm (Ibrahim et al., 2001), which approximates the intractable expectation form. Namely, for a sufficiently large M_i, generate M_i samples of Y_i^{mis} from the conditional distribution

$$f(Y_i^{mis}|Y_i^{obs}, X_i, R_i; \theta^{(t)}, \alpha^{(t)}),$$

and approximate the Q_i function in equation (1.12) by

$$\hat{Q}_i(\theta, \alpha|\theta^{(t)}, \alpha^{(t)}) = \frac{1}{M_i} \sum_{w=1}^{M_i} \log L_i(Y_{i,w}^{mis}, Y_i^{obs}, R_i|X_i; \theta, \alpha),$$

where $Y_{i,w}^{mis}$ is the wth sample of Y_i^{mis}. The M step then maximizes $\hat{Q}_i(\theta, \alpha|\theta^{(t)}, \alpha^{(t)})$ with respect to parameters θ and α.

1.5.3 GEE-Based Methods

GEE analysis based on (1.4) is valid when the missing data mechanism is MCAR. When data are MAR or MNAR, GEE approaches may result in biased estimators (Fitzmaurice et al., 1995). Robins et al. (1995), and Rotnitzky et al. (1998) developed a modified approach using the inverse probability weighted generalized estimating equations (IPWGEE) to handle incomplete data with MAR.

Let $\theta = (\boldsymbol{\beta}^T, \xi^T)^T$, where ξ represents all parameters other than $\boldsymbol{\beta}$ in the response models. Take α to be the parameters corresponding to missingness probabilities. The IPWGEE are formulated with (1.4) modified as:

$$\sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \xi, \alpha) = \sum_{i=1}^{n} D_i V_i^{-1} \Delta_i(\alpha)(Y_i - \mu_i),$$

where $\Delta_i(\alpha)$ is a diagonal weight matrix with $\Delta_i(\alpha) = \text{diag}\big\{I(R_{ij} = 1)/\pi_{ij}(\alpha)\big\}$, $j = 1, 2, \ldots, m)$, and $\pi_{ij}(\alpha) = P(R_{ij} = 1|Y_i, X_i; \alpha)$.

Much recent work provides various extensions of the IPWGEE methods. For example, Yi and Cook (2002) propose a modified IPWGEE approach to handle incomplete longitudinal data arising in clusters. Cook et al. (2004) compare IPWGEE with the imputation method using the last observation carried forward (LOCF) strategy. Carpenter and Kenward (2006) discuss a doubly robust estimation strategy based on IPWGEE. Chen et al. (2010) introduce an IPWGEE approach to handle longitudinal datasets with missingness in both response and covariates. Yi et al. (2012) propose a functional generalized method of moments method to handle missing data and measurement error simultaneously.

1.5.4 Other Methods

Besides likelihood and GEE based methods, many other approaches are developed to deal with incomplete longitudinal data as well. For example, missing data can be handled via the Bayesian approach. This approach involves specifying the distribution of variables subject to missingness and the prior distribution of parameters, and then uses the posterior distribution to obtain estimates. Related studies include Press and Scott (1976), Ibrahim et al. (2002) and Daniels and Hogan (2008).

Alternatively, multiple imputation is another useful method to handle missing data. It first creates multiple "complete" datasets by imputing certain values into missing blanks, then individually analyzes each "complete" dataset, and finally combines the results into final estimates. Multiple imputation is discussed by many authors, including Glynn et al. (1993), Schafer and Olsen (1998) and Schafer and Yucel (2002). A comparative review for the four classes approaches is provided by Ibrahim et al. (2005). Some specific applications of multiple imputation for incomplete data are studied by Landrum and Becker (2001).

1.5.5 Nonidentifiability Issue

When we handle the missingness in *missing not at random (MNAR)*, nonidentifiability in missing data process could be an issue due to the lack of information on the unobserved variable components. As discussed by many authors, such as Ibrahim et al. (2005) and Yi et al. (2011a), it is often difficult to analytically check whether or not the models are identifiable. When this concern arises, a viable way is to carry out sensitivity analyses to assess how inference results may change by altering the models and parameter values for the missing data processes.

Fitzmaurice et al. (1996) illustrate that there still exists identifiable models even the missing mechanism is MNAR. Under MNAR, Ibrahim et al. (2005) suggest that the EM algorithm can be applied to numerically distinguish identifiable/nonidentifiable models. For nonidentifiable models, the EM algorithm may diverge.

To further demonstrate the nonidentifiability issues with missingness, we consider an example involving two models. Suppose the binary response variable Y_i are independent for all $i = 1, \ldots, n$. Let $n = 1000$. Denote $R_i = 1$ if Y_i is observed, and $R_i = 0$ otherwise. We assume that the missing data probability depends on unobserved response variable Y_i, which leads the missing data mechanism to be MNAR. The observed likelihood defined in (1.11) becomes

$$\begin{aligned} L_i &= \prod_{i=1}^{n} \Big[\big\{ P(R_i = 1|Y_i) P(Y_i) \big\}^{R_i} \\ &\quad \times \big\{ P(R_i = 0|Y_i = 1) P(Y_i = 1) + P(R_i = 0|Y_i = 0) P(Y_i = 0) \big\}^{1-R_i} \Big]. \end{aligned} \tag{1.13}$$

We introduce two models as follows.

Model 1 Let $P(Y_i = 1) = p$, $P(R_i = 1|Y_i) = \text{expit}(\alpha_0 + \alpha_1 Y_i)$, where $\text{expit}(t) = \exp(t)/(1+\exp(t))$. The likelihood function (1.13) has a parameter set (p, α_0, α_1). One dataset is generated with $p = 0.2$, $\exp(\alpha_0) = 0.2$ and $\exp(\alpha_1) = 0.5$.

Model 2 Let $P(Y_i = 1|X_i) = \text{expit}(\beta_0 + \beta_1 X_i)$, where X_i is a completely observed binary variable with $P(X_i = 1) = 0.5$. $P(R_i = 1|Y_i, X_i) = \text{expit}(\alpha_0 + \alpha_1 Y_i)$, which follows the one in Model 1. The likelihood function has parameter set $(\beta_0, \beta_1, \alpha_0, \alpha_1)$. One dataset is generated with $\exp(\beta_0) = 1.5$, $\exp(\beta_1) = 0.5$, $\exp(\alpha_0) = 0.2$ and $\exp(\alpha_1) = 0.5$.

According to Fitzmaurice et al. (1996), the parameters (p, α_0, α_1) or $(\beta_0, \beta_1, \alpha_0, \alpha_1)$ are not statistically identifiable if there exists parameters $(p^*, \alpha_0^*, \alpha_1^*) \neq (p, \alpha_0, \alpha_1)$ or $(\beta_0^*, \beta_1^*, \alpha_0^*, \alpha_1^*) \neq (\beta_0, \beta_1, \alpha_0, \alpha_1)$, such that

$$L_i(p, \alpha_0, \alpha_1) = L_i(p^*, \alpha_0^*, \alpha_1^*),$$

or

$$L_i(\beta_0, \beta_1, \alpha_0, \alpha_1) = L_i(\beta_0^*, \beta_1^*, \alpha_0^*, \alpha_1^*).$$

To evaluate the model identifiability, we fix α_1 at a set of values in the likelihood function (1.13). Given fixed α_1, the likelihood function (1.13) is maximized with respect to (p, α_0) and $(\beta_0, \beta_1, \alpha_0)$ for Model 1 and Model 2, respectively. Thus, we obtain the profile likelihoods for two models. Figure 1.3 displays the values of the maximized profile likelihoods given various of α_1. It can be observed that the profile likelihood for Model 1 is flat which implies that Model 1 is nonidentifiable. Because we have $L_i(p, \alpha_0, \alpha_1) = L_i(p^*, \alpha_0^*, \alpha_1^*)$ with $\alpha_1 \neq \alpha_1^*$. On the other hand, the profile likelihood for Model 2 is a curve, which suggests that Model 2 could be identifiable.

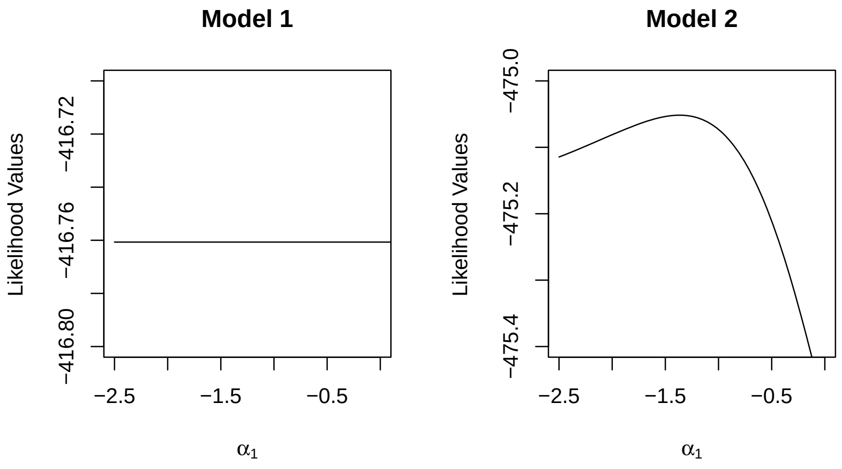

Figure 1.3: *The profile likelihood values with α_1 to be fixed at a set of values for Model 1 and Model 2, respectively.*

We set different initial values and maximize the likelihood function (1.13) for Model 1 and Model 2, respectively. Table 1.1 displays the corresponding likelihood estimates when the maximization algorithm converges. Model 1 results in different estimates from various initial values, while the estimates from Model 2 are stable.

Table 1.1: The initial values and likelihood estimates for Model 1 and Model 2, respectively.

Model 1							
Initial values				Likelihood estimates			
	p	α_1	α_0		p	α_0	α_1
			-1		0.380	-2.045	-0.016
	1	-1	0		0.509	-1.777	-0.606
			1		0.537	-1.709	-0.732
Model 2							
Initial values				Likelihood estimates			
β_0	β_1	α_1	α_0	β_0	β_1	α_0	α_1
			-1	0.247	-0.582	-1.812	-0.576
1	1	1	0	0.247	-0.582	-1.812	-0.576
			1	0.247	-0.582	-1.812	-0.576

Therefore, identifiability issues may arise when the data records are missing with MNAR mechanism. It may not be identifiable for some models, but can be identifiable for others. In practice, setting a grid of initial values can be helpful in checking model identifiability. With diverse starting values, nonidentifiable likelihoods may lead to different maximized results. This agrees with the discussion in Glonek (1999). On the other hand, the identifiable models would be stable with various of initial values.

1.6 Model Misspecification

Let $g(y)$ be the "true" joint density function for independent random vectors Y_i, $i = 1, \ldots, n$. Suppose a working density function $f(y;\theta) = \prod_{i=1}^{n} f(y_i;\theta)$ with $\theta \in \Theta$ is used for estimation of θ, where y_i is the realizations of Y_i. The validity of the statistical inference requires correct model specification to some extent. White (1982) investigates the impact of model misspecification on estimation of the parameter θ. Under certain regularity conditions, if we apply a misspecified model to fit a dataset, then the resultant estimator, denoted by $\hat{\theta}^*$, for the parameter θ would converge in probability to a limit, say θ^*, which may differ from the true parameter value θ_0. If the working density function is correctly specified in a sense that the class of $\Big\{f(y;\theta) : \theta \in \Theta\Big\}$ contains $g(y)$, i.e., there exists $\theta_0 \in \Theta$ such that $f(y;\theta_0) = g(y)$, then the working estimator $\hat{\theta} = \arg\max_{\theta\in\Theta} n^{-1} \log f(y;\theta)$ is consistent for the "true" parameter θ_0.

Yi and Reid (2010) extend White's results from the maximum likelihood framework to the framework of estimating equations. Suppose our inference is based on a biased working estimating function $h(y;\theta)$, which means $E_\theta\big\{h(Y;\theta)\big\} \neq 0$. Assume that the equation

$$H_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} h(y_i;\theta) = 0$$

has a root $\hat{\theta} \in \Theta$ for any given random sample $y_1, \ldots, y_n$, then Yi and Reid (2010) show that under regularity conditions, $\hat{\theta}$ is consistent to a limit, say θ^*, where θ^* is determined by

$$E_\theta\big\{h(Y;\theta^*)\big\} = 0.$$

This result can be used in the study of the misspecification issue related to composite likelihood. Specifically, let $\ell_c(y;\theta)$ be a log composite likelihood function formulated from a model which could be misspecified. Then under certain regularity conditions, the limit θ^* is the solution of

$$E_Y\left\{\frac{\partial \ell_c(y;\theta^*)}{\partial\theta}\right\} = 0, \tag{1.14}$$

where the expectation is taken under the true joint distribution for the Y variable with parameter θ. In most situations, equation (1.14) does not have an analytically closed solution. Hence the relationship between θ^* and θ is frequently evaluated via numerical assessment.

1.7 Motivating Example: The National Population Health Survey Data

1.7.1 Background

The National Population Health Survey (NPHS) collects health information and related socio-demographic information by following a group of Canadian household residents for 10 cycles. The survey is conducted every second year from 1994/1995 and has completed eight cycles: Cycle 1 (1994/1995), Cycle 2 (1996/1997), Cycle 3 (1998/1999), Cycle 4 (2000/2001), Cycle 5 (2002/2003), Cycle 6 (2004/2005), Cycle 7 (2006/2007) and Cycle 8 (2008/2009). The questions for the NPHS include many aspects of in-depth health information such as health status, use of health services, chronic conditions and activity restrictions. Moreover, social background questions, including age, sex, education, income level and marital status, are contained in the questionnaire.

1.7.2 Missing Data

The NPHS started with a sample of 17276 individuals spreading out in the ten provinces across Canada. Each individual is asked to complete a questionnaire in every two years. Although we hope the survey would successfully collect complete health records for 17276 members in all cycles, the NPHS data are subject to information incompletion due to many reasons. Three main possible cases are non-tracing, refusal or unknown to question items, and death.

Non-tracing denotes the situation that interviewers failed to reach the respondents. To deal with non-tracing issue, many approaches were introduced into the survey. For example, workload restriction on maximum interviewees is set for reducing overburden cases; interviewers are trained to apply several survey skills (e.g., making calls or visits at various times of the day, making an appointment to call back or come back if previous time is not convenient); and the survey also attempted to track individuals who moved within Canada or to United States. Despite those efforts, there were still a few non-tracing cases in each cycle and the non-tracing rate in all 17276 members slightly increased with each cycle from 1.7% in Cycle 2 to 5.4% in Cycle 7.

Refusal or unknown to question items leads to another source of information loss. Respondents would refuse to participate in the survey because of privacy, time schedule arrangement or other concerns. The NPHS made efforts to persuade all members to continue the study. For example, a persuasive letter would be sent to respondents if they decided to quit the survey; senior interviewers or other experienced interviewers would try to follow refusals to convince them to rejoin the survey. Though many strategies were applied, refusal rate in survey sample increased from 3.1% in Cycle 1 to 13.2% in Cycle 7. Besides the situation that respondents refused to attend the survey, respondents might attend the survey but refuse to report some question items. A typical example in the NPHS data is that respondents may finish other questions but refuse to report their income status. Moreover, for some questions, a respondent may not find a proper result and then just report as unknown, which also results in an incomplete record.

Until Cycle 7, there are 2032 (11.76%) members who died before the end of the NPHS. Death causes longitudinal health information to be cut off at a specific cycle. However, different from previous situations, where the related health information is existent but unobserved, death leads to another source of information loss that may not be well handled by general approaches. For example, if a respondent was dead at a particular cycle, we may not record variables such as Body Mass Index (BMI), alcohol or tobacco consumption.

To handle longitudinal data with death, one option is to build joint models to postulate both longitudinal records and death information (Diehr and Patrick, 2003; Dufouil et al., 2004; Kurland and Heagerty, 2005; Harel et al., 2007). However, such discussions are beyond the scope of this thesis. Here, we only focus on the case that missing data arise from non-tracing and refusal-to-answer settings.

1.7.3 A Subsample from NPHS

The analysis of the NPHS data focuses on modeling the influence of income, age, education and marital status on population health. The data we select contains 6 cycles' observations (from Cycle 1 to Cycle 6), including 1349 males with age between 50-70 at Cycle 1, and less than 80 at Cycle 6. All the deceased are excluded from our sample data. Missingness occurs in two variables: health status and household income.

Health status is measured by the Health Utilities Index Mark (HUI) from eight attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain and discomfort (Feeny et al., 2002). Household income (INC) is measured by provincial level of household income which ranges from 1 to 10, where 1 denotes household income ranks at decile 1 in the related province, while 10 denotes highest 10 percent of household income.

In our sample data, 36.69% individuals have missing observations in the HUI variable and 52.93% have missing observations in the INC. Only 43.21% of the members have complete observations for both the HUI and the INC in 6 cycles. Table 1.2 shows the missing data rate of both variables, and Table 1.3 displays various missing data patterns.

Table 1.2: Missing data rates for health status and household income variables in the NPHS data (%)

Cycle	1	2	3	4	5	6
HUI	5.3	8.8	11.9	16.8	22.3	25.6
INC	8.7	13.2	17.1	24.0	29.0	33.4

Table 1.3: Missing data patterns for health status and household income variables in the NPHS data

Percentage	HUI						INC					
in Observation	1	2	3	4	5	6	1	2	3	4	5	6
43.2%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4.2%	✓	×	×	×	×	×	✓	×	×	×	×	×
...												
2%	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓
1%	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	×	✓
1%	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

✓ Observed; × Missing

1.8 Outline of Thesis

This thesis develops various inference strategies for longitudinal data using the composite likelihood framework. We particularly address features on missing observations and model selections. Issues of consistency and efficiency are investigated. The remaining chapters are organized in the following structure.

Chapter 2

In Chapter 2, analysis methods using the composite likelihood framework are explored for incomplete longitudinal continuous data. Incomplete data can involve non-monotone missingness for both response and covariates with MNAR mechanisms. In particular, we compare a two-stage estimation strategy and a pairwise method. Simulation studies show that both methods lead to consistent estimators. Issues of efficiency and robustness are carefully investigated. Longitudinal survey data from the National Population Health Study are analyzed with the proposed methods.

Chapter 3

Chapter 3 discusses analysis methods using the composite likelihood for incomplete longitudinal binary data. This chapter parallels Chapter 2 in structures, but considers probit models that are useful for binary data. Again, both response and covariates may be missing with a MNAR mechanism. We explore a two-stage estimation strategy and a pairwise likelihood method. Simulation studies show that both methods result in consistent estimators. Efficiency and robustness are investigated as well. Longitudinal survey data from the National Population Health Study are analyzed with the proposed methods.

Chapter 4

In Chapter 4, we address issues on model selection using the composite likelihood method for more complex data: longitudinal data arising in clusters. We propose a flexible modeling framework to account for complex association structures. In particular, we discuss two forms of composite likelihood function: *all pairwise marginal likelihood* (APW) and *all pairwise conditional likelihood* (APC). The SCAD penalty is applied in the composite likelihood functions, and the related oracle properties are established. Simulations demonstrate that the proposed method gives consistent estimators and is able to select important variables. The composite likelihood EM algorithm and the model misspecification issues are explored in detail.

Chapter 5

Chapter 5 extends the development in Chapter 4 to accommodate the situation that response or covariates are subject to missingness. Conditional likelihood functions are constructed to accommodate missingness effects. Preliminary simulation results demonstrate that the proposed approach outperforms the naive estimation method.

Chapter 6

Chapter 6 summarizes overall results and outlines some further possible extensions of the proposed methods.

Chapter 2

A Pairwise Likelihood Approach for Longitudinal Data with Missing Observations in Both Response and Covariates

2.1 Introduction

Longitudinal data arise commonly in fields including clinical trials and health research. Longitudinal studies are often designed to collect information on individuals at scheduled times, but missing observations occur frequently. Incompleteness of data presents considerable challenges in standard analysis methods, especially when both response and covariate variables incur missingness. A large body of methods have been developed with the primary focus being on either the missingness in response or the missingness in covariates (e.g. Diggle and Kenward, 1994; Little, 1995; Ibrahim et al., 1999, 2001). Research on missingness in both response and covariates is relatively limited, although several authors have developed methods for certain situations.

Under different model assumptions, Shardell and Miller (2008), Chen et al. (2008), Stubbendick and Ibrahim (2003) and Stubbendick and Ibrahim (2006) develop likelihood-based approaches, while Chen et al. (2010) propose a marginal method using the inverse probability weighted generalized estimating equation. Although likelihood-based methods are efficient in estimation of parameters, they require full distributional assumptions, which makes the results sensitive to model misspecification. On the other hand, Chen et al. (2010) relax modeling assumption for the response process by assuming only the marginal structure. The method is mainly developed to handle data that are missing at random.

It is desirable to develop methods that are robust yet flexible to handle various types of missingness in both response and covariate measurements. The purpose of this manuscript is to describe a general approach based on the pairwise likelihood formulation (Lindsay, 1988; Cox and Reid, 2004; Lindsay et al., 2011) to handle longitudinal data with incomplete response and covariates. A unified framework is invoked to accommodate various types of missing data patterns. In particular, our methods can accommodate the existing work as a special case. For instance, Troxel et al. (1998), Parzen et al. (2007) and Troxel et al. (2010) propose marginal and pairwise likelihood methods respectively to deal with missing data when the missingness occurs only in response. Parzen et al. (2006) propose a marginal modeling approach that is suitable for the simultaneous missingness in response and covariates. Our method is flexible to handle the situation when the response and covariates are missing not necessarily simultaneously.

The reminder of the chapter is organized as follows. Section 2.2 introduces the notations and model setups. Inference methods are presented in Section 2.3. In Section 2.4, we report on numerical assessment of the performance of the proposed methods, together with an application to the data arising from the longitudinal National Population Health Survey (NPHS). To further evaluate the performance of the proposed methods, we study the relative efficiency and robustness to model the misspecification in Sections 2.5 and 2.6, respectively.

2.2 Notations and Model Setups

Suppose that there are n subjects and m follow-up occasions. Let Y_{ij} and X_{ij} be the response variable and a covariate vector for subject i at occasion j, respectively, $i = 1, 2, \ldots, n; j = 1, 2, \ldots, m$. Both Y_{ij} and X_{ij} are subject to missingness. Let Z_{ij} be a vector of covariates that can be observed completely. Here we start with the case that X_{ij} is a scalar. Extensions to accommodate multiple covariates X_{ij} are discussed in Chapter 6. Denote $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im})^T$, $X_i = (X_{i1}, X_{i2} \ldots, X_{im})^T$ and $Z_i = (Z_{i1}^T, Z_{i2}^T, \ldots, Z_{im}^T)^T$.

If interest lies in understanding the complete relationship between response Y_i and covariates (X_i, Z_i), one may invoke a full distribution of $f(Y_i|X_i, Z_i; \theta)$ with parameter θ varying in a space Θ. Then inference objective would focus on estimation of parameter θ. In practice, it may be difficult to specify a proper distribution form f, especially when the dimensions of Y_i and covariates (X_i, Z_i) are large. Often, instead of working on the full distribution structure, our interest centers on a partial structure of f such as lower order distributions for some components of Y_i, for example, marginal or pairwise distributions. This strategy has a number of advantages, including transparent interpretation, modeling tractability and lower computational cost. In the chapter we confine our attention to explore pairwise modeling strategies in the context with missing observations.

2.2.1 The Response Process

For $j < k$, let $f(Y_{ij}, Y_{ik}|X_i, Z_i; \boldsymbol{\beta}, \sigma_y^2, \boldsymbol{\psi}^y)$ be the probability density or mass function for paired responses Y_{ij} and Y_{ik}, where $\boldsymbol{\beta}$, σ_y^2 and $\boldsymbol{\psi}^y$ are parameters associated with marginal mean, variance and association measures, respectively. Assume that $f(Y_{ij}, Y_{ik}|X_i, Z_i; \boldsymbol{\beta}, \sigma_y^2, \boldsymbol{\psi}^y)$ is a bivariate normal density function. That is, conditional on (X_i, Z_i),

$$(Y_{ij}, Y_{ik}) \sim N_2((\mu_{ij}^y, \mu_{ik}^y)^T, \boldsymbol{\Sigma}_{ijk}(\sigma_y^2, \boldsymbol{\psi}_{jk}^y)),$$

where $N_2(\cdot, \cdot)$ denotes a bivariate normal distribution with mean and covariance matrix indicated by the arguments, and $\boldsymbol{\Sigma}_{ijk}(\sigma_y^2, \boldsymbol{\psi}_{jk}^y)$ is a 2×2 covariance matrix with diagonal elements σ_y^2 and correlation coefficient $\boldsymbol{\psi}_{jk}^y$. Commonly, a regression model is postulated to reflect the dependence of marginal mean μ_{ij}^y on the covariates at occasion j. For instance, consider $\mu_{ij}^y = X_{ij}\beta_x + Z_{ij}^T\boldsymbol{\beta_z}$, where $\boldsymbol{\beta} = (\beta_x, \boldsymbol{\beta}_z^T)^T$ is a $(q+1) \times 1$ vector of regression parameters linking covariates and response.

2.2.2 The Covariate Process

For $j < k$, let $f(X_{ij}, X_{ik}|Z_i; \boldsymbol{\alpha}, \sigma_x^2, \boldsymbol{\psi}^x)$ be the probability density or mass function for paired covariates X_{ij} and X_{ik}, where $\boldsymbol{\alpha}$, σ_x^2 and $\boldsymbol{\psi}^x$ are parameters corresponding to marginal mean, variance and association measures, respectively. Analogous to the modeling of the response variable, we assume that, condition on Z_i,

$$(X_{ij}, X_{ik}) \sim N_2((\mu_{ij}^x, \mu_{ik}^x)^T, \boldsymbol{\Sigma}_{ijk}(\sigma_x^2, \boldsymbol{\psi}_{jk}^x)),$$

where μ_{ij}^x and σ_x^2 are the marginal mean and variance of X_{ij}, respectively, and $\boldsymbol{\psi}_{jk}^x$ is the correlation coefficient between X_{ij} and X_{ik}. Furthermore, we feature marginal mean μ_{ij}^x by a regression model, such as $\mu_{ij}^x = Z_{ij}^T\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is a vector of regression coefficients.

2.2.3 Missing Data Process

Define $R_{ij}^y = 1$ if Y_{ij} is observed, and $R_{ij}^y = 0$ otherwise. $R_{ij}^x = 1$ if X_{ij} is observed, and $R_{ij}^x = 0$ otherwise. Let $R_i^y = (R_{i1}^y, R_{i2}^y, \ldots, R_{im}^y)^T$ and $R_i^x = (R_{i1}^x, R_{i2}^x, \ldots, R_{im}^x)^T$. Write $Y_i = (Y_i^{obsT}, Y_i^{misT})^T$, and $X_i = (X_i^{obsT}, X_i^{misT})^T$ to distinguish the observed and unobserved components of Y_i and X_i, respectively. For ease of exposition, we put $Y_{ij} = (Y_{ij}^{obs}, Y_{ij}^{mis})$, where either Y_{ij}^{obs} and Y_{ij}^{mis} can be null, depending on whether or not Y_{ij} is observed. Similarly, write $X_{ij} = (X_{ij}^{obs}, X_{ij}^{mis})$.

For the missing data process, we follow the same lines to postulate pairwise models. In particular, we model $P(R_{ij}^y = 1, R_{ik}^y = 1|Y_i, X_i, Z_i, R_{ij}^x, R_{ik}^x)$ and $P(R_{ij}^x = 1, R_{ik}^x = 1|Y_i, X_i, Z_i)$ for $j < k$. As a result, the distribution $P(R_{ij}^y = 1, R_{ik}^y = 1, R_{ij}^x = 1, R_{ik}^x =$

$1|Y_i, X_i, Z_i)$ is uniquely determined. A common assumption (e.g., Troxel et al. (1998)) is made:

$$\begin{aligned} & P(R_{ij}^y = 1, R_{ik}^y = 1, R_{ij}^x = 1, R_{ik}^x = 1|Y_i, X_i, Z_i) \\ = \quad & P(R_{ij}^y = 1, R_{ik}^y = 1, R_{ij}^x = 1, R_{ik}^x = 1|\ Y_{ij}, Y_{ik}, X_{ij}, X_{ik}, Z_{ij}, Z_{ik}). \end{aligned}$$

We employ a pairwise probit model to postulate (R_{ij}^y, R_{ik}^y) or (R_{ij}^x, R_{ik}^x). Specifically, assume there are latent variables $(\tilde{R}_{ij}^y, \tilde{R}_{ik}^y)^T$ that follow a bivariate normal distribution $N_2((0,0)^T, \mathbf{\Sigma}_{ijk}(1, \boldsymbol{\rho}_{jk}^y))$; then $\tilde{R}_{ij}^y$ determines the binary variable R_{ij}^y according to $R_{ij}^y = I(\tilde{R}_{ij}^y \leq \eta_{ij}^{Ry}), \quad j = 1, \cdots, m$, where $I(\cdot)$ is the indicator function and η_{ij}^{Ry} is the linear predictor for R_{ij}^y. Such a modeling scheme has been constantly used for binary data analysis. See Ashford and Sowden (1970), Joe (1997), Renard et al. (2002) and Chaganty and Joe (2004), for details. More explicitly, the pairwise model can be written as $P(R_{ij}^y = 1, R_{ik}^y = 1|Y_i, X_i, Z_i, R_{ij}^x, R_{ik}^x) = \Phi_2((\eta_{ij}^{Ry}, \eta_{ik}^{Ry})^T, \mathbf{\Sigma}_{ijk}(1, \boldsymbol{\rho}_{jk}^y))$, and $P(R_{ij}^x = 1, R_{ik}^x = 1|Y_i, X_i, Z_i) = \Phi_2((\eta_{ij}^{Rx}, \eta_{ik}^{Rx})^T, \mathbf{\Sigma}_{ijk}(1, \boldsymbol{\rho}_{jk}^x))$, where $\Phi_2(\mathbf{u}, \mathbf{v})$ is the bivariate cumulative distribution function for the $N_2((0,0)^T, \mathbf{v})$ evaluated at $\mathbf{u} = (u_1, u_2)$.

Furthermore, regression models are employed to facilitate the dependence of each conditional probability on associated variables. To be specific, we have $\eta_{ij}^{Ry} = \boldsymbol{\lambda}^{yT} \boldsymbol{\xi}_{ij}^y$ and $\eta_{ij}^{Rx} = \boldsymbol{\lambda}^{xT} \boldsymbol{\xi}_{ij}^x$. $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^{yT}, \boldsymbol{\lambda}^{xT})^T$ are missing process related regression parameters. $\boldsymbol{\xi}_{ij}^y$ and $\boldsymbol{\xi}_{ij}^x$ are subsets of $\{Y_{ij}, X_{ij}, Z_{ij}, R_{ij}^x\}$ and $\{Y_{ij}, X_{ij}, Z_{ij}\}$, respectively. Varying choices of these subsets can feature different types of dependence among missing data indicators.

2.3 Estimation and Inference

2.3.1 Marginal and Pairwise Likelihoods

Let $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\lambda}^T, \sigma_y^2, \sigma_x^2)^T$ be the parameters associated with the marginal structure, and $\boldsymbol{\delta} = (\boldsymbol{\psi}^{yT}, \boldsymbol{\psi}^{xT}, \boldsymbol{\rho}^{yT}, \boldsymbol{\rho}^{xT})^T$ be the set of parameters which governs the association structure in the pairwise models. Write $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \boldsymbol{\delta}^T)^T$. Let

$$\begin{aligned}
\mathcal{L}_{C1,i}(\boldsymbol{\gamma}) &= \prod_{j=1}^{m} f(Y_{ij}^{obs}, X_{ij}^{obs}, R_{ij}^{y}, R_{ij}^{x} | Z_{ij}) \\
&= \prod_{j=1}^{m} \int \int f(Y_{ij}|X_{ij}, Z_{ij}) f(X_{ij}|Z_{ij}) f(R_{ij}^{y}, R_{ij}^{x}|Y_{ij}, X_{ij}, Z_{ij}) dY_{ij}^{mis} dX_{ij}^{mis},
\end{aligned}$$

be the observed likelihood for subject i with an independence structure temporarily assumed for repeated measurements, and

$$\begin{aligned}
L_{C2,i}(\boldsymbol{\theta}) &= \prod_{j<k} \Big\{ \int \cdots \int f(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) f(X_{ij}, X_{ik}|Z_{ij}, Z_{ik}) \\
&\quad \times f(R_{ij}^{y}, R_{ij}^{x}, R_{ik}^{y}, R_{ik}^{x}|Y_{ij}, Y_{ik}, X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) dY_{ij}^{mis} dX_{ij}^{mis} dY_{ik}^{mis} dX_{ik}^{mis} \Big\}.
\end{aligned}$$

be the observed pairwise likelihood for subject i. Then the marginal likelihood and pairwise likelihood are respectively given by

$$\mathcal{L}_{C1}(\boldsymbol{\gamma}) = \prod_{i=1}^{n} \mathcal{L}_{C1,i}(\boldsymbol{\gamma}), \tag{2.1}$$

$$\mathcal{L}_{C2}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{L}_{C2,i}(\boldsymbol{\theta}). \tag{2.2}$$

Provided mild regularity conditions, solving the pseudo-score functions $\partial \text{log} \mathcal{L}_{C1}(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma} = \mathbf{0}$ and $\partial \text{log} \mathcal{L}_{C2}(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$ results in consistent estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, respectively. A proof is sketched in supplementary material.

2.3.2 Inference Procedures

We now employ two algorithms for estimation of $\boldsymbol{\theta}$. Approach 1 involves direct maximization of the pairwise likelihood (2.2) (labeled as PL). An alternative method is a two-stage approach (labeled as TS) which first maximizes marginal likelihood (2.1) to obtain the estimator of $\boldsymbol{\gamma}$, and then maximizes pairwise likelihood (2.2), resulting in the estimator of $\boldsymbol{\delta}$.

Compared to the PL, although some efficiency loss may incur in the TS, an obvious advantage is the substantial gain in the ease of computation due to the fact that the dimension of integrals in marginal likelihood is a lot smaller than that in the pairwise likelihood.

Let $S_{1i}(\boldsymbol{\gamma}) = \partial\log\mathcal{L}_{C1,i}(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^T$, $S_{2i}(\boldsymbol{\gamma}) = \partial\log\mathcal{L}_{C2,i}(\boldsymbol{\theta})/\partial\boldsymbol{\gamma}^T$, and $S_{2i}(\boldsymbol{\delta}) = \partial\log\mathcal{L}_{C2,i}(\boldsymbol{\theta})/\partial\boldsymbol{\delta}^T$. Define $H_i = (S_{1i}(\boldsymbol{\gamma})^T, S_{2i}(\boldsymbol{\delta})^T)^T$, and $S_{2i}(\boldsymbol{\theta}) = (S_{2i}(\boldsymbol{\gamma})^T, S_{2i}(\boldsymbol{\delta})^T)^T$.

Pairwise Likelihoods (PL) Inference

We employ the Newton-Raphson algorithm to maximize the pairwise likelihood function (2.2). The pairwise likelihood (PL) estimators are denoted by $\boldsymbol{\theta}_{PL} = (\boldsymbol{\gamma}_{PL}^T, \boldsymbol{\delta}_{PL}^T)^T$. We update the estimates by the iterative equation

$$\begin{pmatrix} \boldsymbol{\gamma}_{PL}^{(t+1)} \\ \boldsymbol{\delta}_{PL}^{(t+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma}_{PL}^{(t)} \\ \boldsymbol{\delta}_{PL}^{(t)} \end{pmatrix} - \Big\{ \sum_{i=1}^{n} D_i^{(t)} \Big\}^{-1} \cdot \sum_{i=1}^{n} S_{2i}(\boldsymbol{\gamma}_{PL}^{(t)}, \boldsymbol{\delta}_{PL}^{(t)}), \tag{2.3}$$

where

$$D_i^{(t)} = \begin{pmatrix} \partial S_{2i}(\boldsymbol{\gamma}_{PL}^{(t)})/\partial\boldsymbol{\gamma}^T & \partial S_{2i}(\boldsymbol{\gamma}_{PL}^{(t)})/\partial\boldsymbol{\delta}^T \\ \partial S_{2i}(\boldsymbol{\delta}_{PL}^{(t)})/\partial\boldsymbol{\gamma}^T & \partial S_{2i}(\boldsymbol{\delta}_{PL}^{(t)})/\partial\boldsymbol{\delta}^T \end{pmatrix},$$

and $t = 0, 1, \ldots,$ until $(\boldsymbol{\gamma}_{PL}^{T(t+1)}, \boldsymbol{\delta}_{PL}^{T(t+1)})^T$ converges to the solution $(\hat{\boldsymbol{\gamma}}_{PL}^T, \hat{\boldsymbol{\delta}}_{PL}^T)^T$.

Under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta})$ has an asymptotic normal distribution with mean $\mathbf{0}$ and covariance matrix $\{E(D_i)\}^{-1}E\{S_{2i}(\boldsymbol{\theta})S_{2i}(\boldsymbol{\theta})^T\}\{E(D_i)\}^{-1T}$. In particular, for primarily interesting parameter $\boldsymbol{\beta}$, we need to establish the asymptotic distribution of its estimator $\hat{\boldsymbol{\beta}}_{PL}$. Rewrite $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\upsilon}^T)^T$, and $S_{2i}(\boldsymbol{\theta}) = (S_{2i}(\boldsymbol{\beta})^T, S_{2i}(\boldsymbol{\upsilon})^T)^T$. Define

$$J^* = E\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\} - E\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\upsilon}^T\} \cdot E^{-1}\{\partial S_{2i}(\boldsymbol{\upsilon})/\partial\boldsymbol{\upsilon}^T\} \cdot E^T\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\upsilon}^T\},$$

and

$$\begin{aligned} K^* \;=\; & E\{S_{2i}(\boldsymbol{\beta}) \cdot S_{2i}(\boldsymbol{\beta})^T\} - E\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\upsilon}^T\} \cdot E^{-1}\{\partial S_{2i}(\boldsymbol{\upsilon})/\partial\boldsymbol{\upsilon}^T\} \cdot E\{S_{2i}(\boldsymbol{\upsilon})S_{2i}(\boldsymbol{\beta})^T\} \\ & -\Big[E\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\upsilon}^T\} \cdot E^{-1}\{\partial S_{2i}(\boldsymbol{\upsilon})/\partial\boldsymbol{\upsilon}^T\} \cdot E\{S_{2i}(\boldsymbol{\upsilon})S_{2i}(\boldsymbol{\beta})^T\}\Big]^T \\ & +E\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\upsilon}^T\} \cdot E^{-1}\{\partial S_{2i}(\boldsymbol{\upsilon})/\partial\boldsymbol{\upsilon}^T\} \cdot E\{S_{2i}(\boldsymbol{\upsilon})S_{2i}(\boldsymbol{\upsilon})^T\} \\ & \cdot E^{-1}\{\partial S_{2i}(\boldsymbol{\upsilon})/\partial\boldsymbol{\upsilon}^T\} \cdot E^T\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\upsilon}^T\}. \end{aligned}$$

Then $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{PL}-\boldsymbol{\beta}\right)$ has an asymptotic normal distribution with mean $\mathbf{0}$ and covariance matrix $J^{*-1}K^{*}\{J^{*-1}\}^{T}$. The proof is outlined in supplementary material.

Two-Stage Inference

Under the two-stage inference scheme, an estimate, denoted $\hat{\boldsymbol{\gamma}}_{TS}$, of $\boldsymbol{\gamma}$ is first obtained as the maximizer of the marginal likelihood $\mathcal{L}_{C1}(\boldsymbol{\gamma})$. With this $\hat{\boldsymbol{\gamma}}_{TS}$, we then maximize the pairwise likelihood $\mathcal{L}_{C2}(\hat{\boldsymbol{\gamma}}_{TS},\boldsymbol{\delta})$, with respect to $\boldsymbol{\delta}$, and the maximizer $\hat{\boldsymbol{\delta}}_{TS}$ is taken as the estimate of $\boldsymbol{\delta}$. To be specific, the two-stage procedure can be realized using the iterative equation

$$\boldsymbol{\gamma}_{TS}^{(t+1)}=\boldsymbol{\gamma}_{TS}^{(t)}-\Big\{\sum_{i=1}^{n}\partial S_{1i}(\boldsymbol{\gamma}_{TS}^{(t)})/\partial\boldsymbol{\gamma}^{T}\Big\}^{-1}\cdot\sum_{i=1}^{n}S_{1i}(\boldsymbol{\gamma}_{TS}^{(t)}),\quad t=1,2,\ldots$$

until convergence. Similarly, update the estimate of $\boldsymbol{\delta}$ using the iterative equation

$$\boldsymbol{\delta}_{TS}^{(t+1)}=\boldsymbol{\delta}_{TS}^{(t)}-\Big\{\sum_{i=1}^{n}\partial S_{2i}(\hat{\boldsymbol{\gamma}}_{TS},\boldsymbol{\delta}_{TS}^{(t)})/\partial\boldsymbol{\delta}^{T}\Big\}^{-1}\cdot\sum_{i=1}^{n}S_{2i}(\hat{\boldsymbol{\gamma}}_{TS},\boldsymbol{\delta}_{TS}^{(t)}),\quad t=1,2,\ldots$$

until convergence.

An alternative to obtain the estimator $\hat{\boldsymbol{\theta}}_{TS}=(\hat{\boldsymbol{\gamma}}_{TS}^{T},\hat{\boldsymbol{\delta}}_{TS}^{T})^{T}$ is to employ the joint iterative equation to update the estimate:

$$\begin{pmatrix}\boldsymbol{\gamma}_{TS}^{(t+1)}\\ \boldsymbol{\delta}_{TS}^{(t+1)}\end{pmatrix}=\begin{pmatrix}\boldsymbol{\gamma}_{TS}^{(t)}\\ \boldsymbol{\delta}_{TS}^{(t)}\end{pmatrix}\ -\ \Big\{\sum_{i=1}^{n}D_{i}^{*(t)}\Big\}^{-1}\cdot\sum_{i=1}^{n}H_{i}(\boldsymbol{\gamma}_{TS}^{(t)},\boldsymbol{\delta}_{TS}^{(t)}),\qquad(2.4)$$

where

$$D_{i}^{*(t)}=\begin{pmatrix}\partial S_{1i}(\boldsymbol{\gamma}_{TS}^{(t)})/\partial\boldsymbol{\gamma}^{T} & \mathbf{0}\\ \partial S_{2i}(\boldsymbol{\delta}_{TS}^{(t)})/\partial\boldsymbol{\gamma}^{T} & \partial S_{2i}(\boldsymbol{\delta}_{TS}^{(t)})/\partial\boldsymbol{\delta}^{T}\end{pmatrix}.$$

At each iteration, the update obtained from (2.4) may differ from that obtained from the two-stage algorithm. However, the updated values from these two procedures converge to the same limit under mild regularity conditions (Newey and McFadden, 1994).

While the two-stage algorithm provides an easy way for estimation, the algorithm based on (2.4) is more convenient to establish the asymptotic distribution of the estimator. Under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}}_{TS} - \boldsymbol{\theta})$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix $\{E(D_i^*)\}^{-1}E\{H_iH_i^T\}\{E(D_i^*)\}^{-1T}$. Rewrite $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \boldsymbol{v}^{*T})^T$, and $S_{1i}(\boldsymbol{\gamma}) = (S_{1i}(\boldsymbol{\beta})^T, S_{1i}(\boldsymbol{v}^*)^T)^T$.

Define

$$J^{**} = E\{\partial S_{1i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\} - E\{\partial S_{1i}(\boldsymbol{\beta})/\partial\boldsymbol{v}^{*T}\} \cdot E^{-1}\{\partial S_{1i}(\boldsymbol{v}^*)/\partial\boldsymbol{v}^{*T}\} \cdot E^T\{\partial S_{1i}(\boldsymbol{\beta})/\partial\boldsymbol{v}^{*T}\},$$

and

$$\begin{aligned} K^{**} \ = \ & E\{S_{1i}(\boldsymbol{\beta}) \cdot S_{1i}(\boldsymbol{\beta})^T\} - E\{\partial S_{1i}(\boldsymbol{\beta})/\partial\boldsymbol{v}^{*T}\} \cdot E^{-1}\{\partial S_{1i}(\boldsymbol{v}^*)/\partial\boldsymbol{v}^{*T}\} \cdot E\{S_{1i}(\boldsymbol{v}^*)S_{1i}(\boldsymbol{\beta})^T\} \\ & -\Big[E\{\partial S_{1i}(\boldsymbol{\beta})/\partial\boldsymbol{v}^{*T}\} \cdot E^{-1}\{\partial S_{1i}(\boldsymbol{v}^*)/\partial\boldsymbol{v}^{*T}\} \cdot E\{S_{1i}(\boldsymbol{v}^*)S_{1i}(\boldsymbol{\beta})^T\}\Big]^T \\ & +\Big[E\{\partial S_{1i}(\boldsymbol{\beta})/\partial\boldsymbol{v}^{*T}\} \cdot E^{-1}\{\partial S_{1i}(\boldsymbol{v}^*)/\partial\boldsymbol{v}^{*T}\} \cdot E\{S_{1i}(\boldsymbol{v}^*)S_{1i}(\boldsymbol{v}^*)^T\} \\ & \cdot E^{-1}\{\partial S_{1i}(\boldsymbol{v}^*)/\partial\boldsymbol{v}^{*T}\} \cdot E^T\{\partial S_{1i}(\boldsymbol{\beta})/\partial\boldsymbol{v}^{*T}\}\Big]. \end{aligned}$$

Then $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{TS} - \boldsymbol{\beta}\right)$ has an asymptotic normal distribution with mean $\mathbf{0}$ and covariance matrix $J^{**-1}K^{**}\{J^{**-1}\}^T$. The proof is outlined in the supplementary material.

2.4 Numerical Studies

2.4.1 Empirical Assessment of the Proposed Methods

In this section, we assess the empirical performance of the proposed methods through a simulation study. One hundred and 500 simulations are run for the PL and TS methods, respectively. We consider a setting with $m = 3$ and $n = 150$, and simulate longitudinal continuous responses from a normal model with $\mu_{ij}^y = \beta_0 + \beta_1 X_{ij}$, where X_{ij} is a time-dependent continuous covariate generated from a normal distribution with $\mu_{ij}^x = \alpha_0$. Set $\beta_0 = -2$, $\beta_1 = 2$ and $\alpha_0 = 1$. The association among responses is specified as exchangeable with $\sigma_y^2 = 1$ and correlation coefficient ψ^y, specified as 0.5. The association among covariate components is specified as exchangeable with $\sigma_x^2 = 1$ and $\psi^x = 0.5$.

For the response and covariate missingness process, we take

$$\eta_{ij}^{Ry} = \lambda_0^y + \lambda_1^y X_{ij} + \lambda_2^y R_{ij}^x, \quad \text{and}$$

$$\eta_{ij}^{Rx} = \lambda_0^x + \lambda_1^x X_{ij}.$$

The true values for the regression parameters of missing data processes are set to be $\lambda_0^y = \lambda_0^x = 1.5$, $\lambda_1^y = \lambda_1^x = -1$, and $\lambda_2^y = 0.5$. For the joint distribution of the response and covariate missing processes, we consider

$$P(R_{i1}^y = 1, R_{i2}^y = 1, R_{i3}^y = 1 | Y_i, X_i, R_{i1}^x, R_{i2}^x, R_{i3}^x) = \Phi_3((\eta_{i1}^{Ry}, \eta_{i2}^{Ry}, \eta_{i3}^{Ry})^T, \Sigma_i(1, \boldsymbol{\rho}_{123}^y)),$$

and $P(R_{i1}^x = 1, R_{i2}^x = 1, R_{i3}^x = 1 | Y_i, X_i) = \Phi_3((\eta_{i1}^{Rx}, \eta_{i2}^{Rx}, \eta_{i3}^{Rx})^T, \Sigma_i(1, \boldsymbol{\rho}_{123}^x))$, respectively, where $\Phi_3(\mathbf{u}^*, \mathbf{v}^*)$ is the cumulative distribution function for the $N_3((0,0,0)^T, \mathbf{v}^*)$ evaluated at $\mathbf{u}^* = (u_1^*, u_2^*, u_3^*)$. We take $\Sigma_i(1, \boldsymbol{\rho}_{123}^y)$ and $\Sigma_i(1, \boldsymbol{\rho}_{123}^x)$ to have exchangeable association forms with correlation coefficients ρ^y, ρ^x, respectively. The true values are set as $\rho^y = \rho^x = 0.5$.

The results are reported in Table 2.1, where the bias is the percent relative bias, ASE and ESE are the average of model-based standard errors and empirical standard errors, respectively, and CP% represents the empirical coverage probability for the 95% confidence intervals. The table shows that our PL and TS approaches both yield small bias and satisfactory coverage probability for the response parameters in both the mean and association structures. As expected, the PL approach results in smaller ASE and ESE for parameter β_1 than the TS method, which confirms the PL approach is more efficient than the TS method. A good agreement between ASE and ESE indicates that variance estimates for the corresponding estimators are valid. In covariate and missing processes, it can be seen that the biases are negligible and ASE/ESE are similar for most of the parameters, which implies two approaches also provide reasonable inference on covariate and missing models.

Table 2.1: Simulation results for incomplete longitudinal data with missingness in both continuous response and covariate under pairwise likelihood (PL) and two-stage (TS) estimation algorithms

		Response				Covariate			Response Missingness				Covariate Missingness		
		β_0	β_1	ψ^y	σ_y^2	α_0	ψ^x	σ_x^2	λ_0^y	λ_1^y	λ_2^y	ρ^y	λ_0^x	λ_1^x	ρ^x
PL†	Bias%*	-0.024	-0.223	-0.288	0.570	0.238	-0.471	0.209	6.227	4.804	-3.454	0.138	-0.170	0.253	-0.061
	ASE**	0.089	0.075	0.065	0.102	0.075	0.054	0.105	0.326	0.160	0.211	0.121	0.190	0.122	0.109
	ESE***	0.096	0.081	0.065	0.108	0.070	0.048	0.112	0.324	0.157	0.207	0.149	0.216	0.145	0.100
	CP%	93.0	91.0	94.0	93.0	96.0	98.0	94.0	98.0	94.0	96.0	91.0	91.0	91.0	98.0
TS	Bias%	0.027	0.036	-2.161	-1.228	0.377	-1.790	0.127	4.548	3.659	-2.579	-2.667	1.092	1.375	0.824
	ASE	0.092	0.085	0.064	0.100	0.078	0.054	0.111	0.342	0.164	0.229	0.121	0.207	0.132	0.108
	ESE	0.096	0.088	0.067	0.099	0.072	0.053	0.111	0.353	0.171	0.225	0.125	0.225	0.146	0.110
	CP%	93.4	93.0	91.6	93.8	96.8	95.2	95.4	93.4	93.6	96.0	94.2	95.0	92.0	94.8

† PL and TS denote pairwise likelihood and two-stage inference procedures. There are 100 and 500 simulation runs for PL and TS, respectively.

$*$ Relative bias defined by $(\hat{\beta} - \beta_{true})/\beta_{true} \times 100$.

$**$ ASE is the average standard error for r times simulations, which is defined by $r^{-1}\sum_{i=1}^{r}\sqrt{\widehat{Var}(\hat{\beta}^i)}$, where $\sqrt{\widehat{Var}(\hat{\beta}^i)}$ is the standard error estimates in ith simulation result.

$***$ ESE is the empirical standard error for r times simulation, which is defined by $\{(r-1)^{-1}\sum_{i=1}^{r}(\hat{\beta}^i - \bar{\hat{\beta}})^2\}^{1/2}$, where $\hat{\beta}^i$ is the ith simulation result, and $\bar{\hat{\beta}} = r^{-1}\sum_{i=1}^{r}\hat{\beta}^i$.

2.4.2 Application to the NPHS Data

The National Population Health Survey (NPHS) is a longitudinal study that collects health information and related socio-demographic information by following a group of Canadian household residents. The questions for the NPHS include many aspects of in-depth health information such as health status, use of health services, chronic conditions and activity restrictions. Moreover, social background questions, including age, sex and income level, are contained in the questionnaire. A research interest focuses on modeling the influence of income on population health. The data we analyze here contain 3 cycles' observations (from Cycle 4 to Cycle 6), including $n = 300$ males with age between 50-70 at Cycle 1, and less than 80 at Cycle 6. All the deceased subjects are excluded from the analysis.

Health status is measured by the Health Utilities Index (HUI) Mark after zero-mean normalization with observed average 0.85 and standard deviation 0.21. The higher HUI score indicates better health. The covariate prone to missingness is household income (INC), which is measured by provincial level of household income with zero-mean normalization with observed average 5.27 and standard deviation 2.88. The other covariate, denoted by CYCLE is cycle number with values $-1, 0, 1$ that correspond to Cycle $4, 5$ and 6, respectively.

In the data analyzed here, 21.3% individuals have missing observations in HUI variable and 35.7% have missing observations in INC. Only 62.3% of the members have complete observations for both HUI and INC in all 3 cycles. The missingness proportions in HUI from Cycle 4 to Cycle 6 are 2.7%, 11.0% and 17.7%, respectively, while the missingness proportions in INC from Cycle 4 to Cycle 6 are 9.3%, 17.3%, 27.3%, respectively. Table 5.1 displays a sample data subset.

Table 2.2: Sample data from the NPHS

ID	HUI			INC		
	4	5	6	4	5	6
1	0.577	0.577	0.577	1.645	1.645	1.645
2	0.577	·	0.256	-0.440	·	-0.788
3	0.134	-0.582	-0.314	0.950	1.297	·
4	0.704	0.704	0.256	-1.135	·	-0.093
5	·	-0.945	0.577	-1.135	-1.483	-1.135
6	·	0.577	·	-0.788	-0.440	·
7	0.704	·	·	-0.788	·	·

· represents missing observations

Let HUI_{ij}, INC_{ij} and CYCLE_{ij} be the normalized Health Utility Index score, normalized income level, and cycle numbers for individual i at Cycle j. Let $R_{ij} = (R^y_{ij}, R^x_{ij})$ represent the missing indicator where $R^y_{ij} = 1$ denotes subject i's HUI is observed at Cycle j, and $R^y_{ij} = 0$ otherwise. Similarly, $R^x_{ij} = 1$ means that subject i's INC is observed at Cycle j and $R^x_{ij} = 0$ otherwise.

We assume that HUI and INC follow marginal models

$$\text{HUI}_{ij} = \beta_0 + \beta_1 \text{INC}_{ij} + \beta_2 \text{CYCLE}_{ij} + \varepsilon^y_{ij}, \tag{2.5}$$

$$\text{INC}_{ij} = \alpha_0 + \alpha_1 \text{CYCLE}_{ij} + \varepsilon^x_{ij}, \tag{2.6}$$

respectively, where $\varepsilon^y_{ij} \sim N(0, \sigma^2_y)$, $\varepsilon^x_{ij} \sim N(0, \sigma^2_x)$.

The missing data processes are specified as

$$\eta^{Ry}_{ij} = \lambda^y_0 + \lambda^y_1 \text{HUI}_{ij} + \lambda^y_2 \text{INC}_{ij} + \lambda^y_3 R^x_{ij} + \lambda^y_4 \text{CYCLE}_{ij}, \tag{2.7}$$

$$\eta^{Rx}_{ij} = \lambda^x_0 + \lambda^x_1 \text{HUI}_{ij} + \lambda^x_2 \text{INC}_{ij} + \lambda^x_3 \text{CYCLE}_{ij}. \tag{2.8}$$

We further assume an AR(1) association structure for each process with corresponding association parameters ψ^y, ψ^x, ρ^y and ρ^x for HUI, INC, R^y and R^x, respectively.

With models (2.5)-(2.8), we analyze the data using the PL and TS methods, and report the results in Table 2.3 and Table 2.4.

Table 2.3: Analysis of the NPHS data using the pairwise likelihood, two-stage estimation approach and naive method: Response models

		PL[‡]			TS		
Parameter		Estimate	SE	p-value	Estimate	S.E.	p-value
INTERC.	(β_0)	-0.045	0.053	0.393	-0.040	0.064	0.530
INC	(β_1)	0.219	0.042	< 0.001	0.231	0.045	< 0.001
CYCLE	(β_2)	-0.041	0.027	0.125	-0.029	0.035	0.405
Variance	(σ_y^2)	0.957	0.122	< 0.001	0.938	0.120	< 0.001
Association	(ψ^y)	0.677	0.046	< 0.001	0.667	0.045	< 0.001

[‡] PL and TS respectively denote the pairwise likelihood and two-stage inference procedures, respectively.

Table 2.4: Analysis of the NPHS data using the pairwise likelihood and two-stage estimation approach: Covariate and missing-data models

		PL			TS		
Parameter		Estimate	SE	p-value	Estimate	SE	p-value
		Model for INC					
Intercept	α_0	0.001	0.025	0.969	0.141	0.133	0.288
CYCLE	α_1	-0.107	0.025	< 0.001	-0.067	0.048	0.165
Variance in HUI	σ_x^2	1.015	0.048	< 0.001	1.066	0.106	< 0.001
Association	ψ^x	0.835	0.022	< 0.001	0.832	0.023	< 0.001
		Response Missing Model					
Intercept	λ_0^y	0.074	0.129	0.568	-0.144	0.224	0.519
HUI	λ_1^y	0.045	0.088	0.606	-0.166	0.121	0.170
INC	λ_2^y	0.058	0.114	0.610	0.226	0.158	0.152
R_{ij}^x	λ_3^y	2.166	0.170	< 0.001	2.475	0.295	< 0.001
CYCLE	λ_4^y	-0.254	0.099	0.010	-0.258	0.104	0.013
Association	ρ^y	0.636	0.118	< 0.001	0.624	0.130	< 0.001
		Covariate Missing Model					
Intercept	λ_0^x	0.971	0.064	< 0.001	1.105	0.257	< 0.001
HUI	λ_1^x	0.145	0.053	0.007	0.189	0.146	0.196
INC	λ_2^x	0.029	0.105	0.782	-0.429	0.328	0.192
CYCLE	λ_3^x	-0.343	0.059	< 0.001	-0.413	0.085	< 0.001
Association	ρ^x	0.570	0.059	< 0.001	0.595	0.068	< 0.001

For the response model in Table 2.3, PL and TS approaches reveal that the cycle time is not statistically significant, whereas income has a significant positive effect on health index. People are more likely to have better health if they have higher income. Moreover, it can be seen that the PL method yields smaller standard errors than the TS approach, which agrees with the finding in the previous subsection. For the model of household income in Table 2.4, the PL method indicates as the survey cycle increases, the income would significantly decrease, while the TS approach reveals an insignificant temporal effect on income.

For the missing probability in Table 2.4, both PL and TS show insignificance of HUI and INC in the response missing data model, and only PL suggests a significant positive effect of HUI in the covariate missing-data model. Moreover, the significance of λ_4^y and λ_3^x suggests that the missing rate for both response and covariate increases as the longitudinal research cycle increases. Estimation of λ_3^y indicates an association between missingness of the response and of the covariate.

2.5 Efficiency Assessment

To fully understand the performance of the proposed methods, in this section we assess the efficiency of the PL and TS algorithms. To this end, we invoke estimating function theory. Suppose $U(\boldsymbol{\theta}) = \sum_{i=1}^{n} U_i(\boldsymbol{\theta})$ are estimating functions for parameter $\boldsymbol{\theta}$, where $E[U_i(\boldsymbol{\theta})] = \mathbf{0}$, then under regularity conditions, the solution, say $\hat{\boldsymbol{\theta}}$, to $U(\boldsymbol{\theta}) = \mathbf{0}$ has an asymptotic normal distribution

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow_D N(\mathbf{0}, I^{-1}(\boldsymbol{\theta})), \tag{2.9}$$

where $I(\boldsymbol{\theta})$ is the Godambe information matrix (Godambe, 1991) defined as

$$I(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\partial \mathbf{U}_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^T]^T E_{\boldsymbol{\theta}}[\mathbf{U}_i(\boldsymbol{\theta})\mathbf{U}_i(\boldsymbol{\theta})^T]^{-1}(\boldsymbol{\theta}) E_{\boldsymbol{\theta}}[\partial \mathbf{U}_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^T].$$

The Godambe information matrix or its inverse provides us a basis to evaluate efficiency of estimators obtained from different methods or from distinct conditions. In particular, we are interested in two scenarios concerning the marginal response parameter $\boldsymbol{\beta}$.

In the first case, we are interested in comparing the efficiency of estimators of $\boldsymbol{\beta}$ that are obtained when nuisance parameters are known or estimated. This study would provide insight into variability induced by an additional estimation procedure for nuisance parameters. Following the notations in Section 3.3, if nuisance parameter $\boldsymbol{\nu}$ or $\boldsymbol{\nu}^*$ is unknown, then the estimation of $\boldsymbol{\beta}$ can proceed by solving $(\partial/\partial\boldsymbol{\theta})\log\mathcal{L}_{C2}(\boldsymbol{\theta}) = \mathbf{0}$ for the PL approach and $(\partial/\partial\boldsymbol{\gamma})\log\mathcal{L}_{C1}(\boldsymbol{\gamma}) = \mathbf{0}$ for the TS approach. Let $\hat{\boldsymbol{\theta}}_{PL} = (\hat{\boldsymbol{\beta}}_{PL}^T, \hat{\boldsymbol{\nu}}_{PL}^T)^T$ and $\hat{\boldsymbol{\gamma}}_{TS} = (\hat{\boldsymbol{\beta}}_{TS}^T, \hat{\boldsymbol{\nu}}_{TS}^{*T})^T$ be the result estimators for PL and TS approaches, respectively. Then its asymptotic covariance is determined by (2.9), yielding the asymptotic covariance $I_{PL}^{-1}(\boldsymbol{\beta})$ for $\hat{\boldsymbol{\beta}}_{PL}$:

$$\begin{aligned} I_{PL}(\boldsymbol{\beta}) \quad = \quad & E\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\} \cdot \mathbf{D}_1 \cdot E^T\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\} \\ & -E\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\} \cdot E^{-1}\{S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\beta})^T\} \cdot E\{S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\nu})^T\} \cdot \mathbf{D}_2 \cdot E^T\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\nu}^T\} \\ & -\left[E\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\} \cdot E^{-1}\{S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\beta})^T\} \cdot E\{S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\nu})^T\} \cdot \mathbf{D}_2 \cdot E^T\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\nu}^T\}\right]^T \\ & +E\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\nu}^T\} \cdot \mathbf{D}_2 \cdot E^T\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\nu}^T\} \end{aligned} \qquad (2.10)$$

where $\mathbf{D}_1 = \Big[E\{S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\beta})^T\} - E\{S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\nu})^T\} \cdot E^{-1}\{S_{2i}(\boldsymbol{\nu})S_{2i}(\boldsymbol{\nu})^T\} \cdot E\{S_{2i}(\boldsymbol{\nu})S_{2i}(\boldsymbol{\beta})^T\}\Big]^{-1}$, and $\mathbf{D}_2 = \Big[E\{S_{2i}(\boldsymbol{\nu})S_{2i}(\boldsymbol{\nu})^T\} - E\{S_{2i}(\boldsymbol{\nu})S_{2i}(\boldsymbol{\beta})^T\} \cdot E^{-1}\{S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\beta})^T\} \cdot E\{S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\nu})^T\}\Big]^{-1}$. Moreover, for TS method, we can obtain $I_{TS}(\boldsymbol{\beta})$ by respectively replacing $S_{2i}(\boldsymbol{\beta})$, $S_{2i}(\boldsymbol{\nu})$ and $\boldsymbol{\nu}$ into $S_{1i}(\boldsymbol{\beta})$, $S_{1i}(\boldsymbol{\nu}^*)$ and $\boldsymbol{\nu}^*$ in (2.10).

On the other hand, if nuisance parameter $\boldsymbol{\nu}$ is known, the estimation of the $\boldsymbol{\beta}$ parameter can proceed by solving $(\partial/\partial\boldsymbol{\beta})\log\mathcal{L}_{C2}(\boldsymbol{\beta}) = \mathbf{0}$ for the PL approach and $(\partial/\partial\boldsymbol{\beta})\log\mathcal{L}_{C1}(\boldsymbol{\beta}) = \mathbf{0}$ for the TS approach, respectively. The resulting estimator, denoted by $\tilde{\boldsymbol{\beta}}_{PL}$ and $\tilde{\boldsymbol{\beta}}_{TS}$ have the asymptotic covariance $\tilde{I}_{PL}^{-1}(\boldsymbol{\beta})$ given by

$$\tilde{I}_{PL}(\boldsymbol{\beta}) = E[\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T] \cdot \{E[S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\beta})^T]\}^{-1} \cdot E^T[\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T], \qquad (2.11)$$

while $\tilde{I}_{TS}^{-1}(\boldsymbol{\beta})$ can be obtained by replacing $S_{2i}(\boldsymbol{\beta})$ into $S_{1i}(\boldsymbol{\beta})$.

To compare the efficiency of the PL estimators $\hat{\boldsymbol{\beta}}_{PL}$ and $\tilde{\boldsymbol{\beta}}_{PL}$, one needs only to compare $I_{PL}(\boldsymbol{\beta})$ and $\tilde{I}_{PL}(\boldsymbol{\beta})$. Similarly, comparison of $I_{TS}(\boldsymbol{\beta})$ and $\tilde{I}_{TS}(\boldsymbol{\beta})$ indicates the efficiency of the TS estimators $\hat{\boldsymbol{\beta}}_{TS}$ and $\tilde{\boldsymbol{\beta}}_{TS}$. The difference in (2.10) and (2.11) quantify the amount of additional variation induced in estimating parameter $\boldsymbol{\nu}$ that would be contained in the asymptotic covariance matrix of the estimator for $\boldsymbol{\beta}$ if $\boldsymbol{\nu}$ were unknown. It is a common conception that $\tilde{\boldsymbol{\beta}}_{PL}$ and $\tilde{\boldsymbol{\beta}}_{TS}$ are more efficient than $\hat{\boldsymbol{\beta}}_{PL}$ and $\hat{\boldsymbol{\beta}}_{TS}$, respectively. However, this is not obviously perceived from (2.10) and (2.11). In principle, the differences of (2.10) and (2.11) depend on the model structures as well as the true value of relevant parameters, agreeing with the discussion in Henmi and Eguchi (2004). To illustrate this, we conduct a numerical study here.

To be specific, we consider the two scenarios. Scenario I assumes the same missing data model as in Section 2.4.1, while in scenario II, we specify the missing data process as $\eta_{ij}^{Ry} = 1.5 - 0.5y_{ij} - 0.5R_{ij}^{x}$ and $\eta_{ij}^{Rx} = 1.5 - y_{ij}$. Let $\text{avar}(\hat{\beta}_j^X)$ denote the asymptotic variance of estimator $\hat{\beta}_j^X$ for parameter β_j $(j = 0, 1)$, obtained from the X method, where X refers to either the PL or TS method. Table 2.5 displays the relative efficiency of the estimators for $\boldsymbol{\beta}$ parameters that is defined as the ratio $R_s^X(\beta_j) = \text{avar}_s(\hat{\beta}_j^X)/\text{avar}(\tilde{\beta}_j^X)$ for $j = 0, 1$, where $\text{avar}_s(\hat{\beta}_j^X)$ and $\text{avar}(\tilde{\beta}_j^X)$ are jth diagonal element of $I_{s,X}^{-1}(\boldsymbol{\beta})$ and $\tilde{I}_X^{-1}(\boldsymbol{\beta})$, respectively, and $I_{s,X}(\boldsymbol{\beta})$ is similar to $I_X(\boldsymbol{\beta})$ in (2.10) under the assumption some or all nuisance parameters are unknown. All the entries for PL and TS are no bigger than 1, suggesting that the involvement of unknown nuisance parameters in the estimation would reduce the efficiency for $\boldsymbol{\beta}$ estimators. The more unknown nuisance parameters are involved, the larger efficiency loss tend to occur for both PL and TS. Furthermore, the efficiency loss depends on the model form as well. Under scenario I, the efficiency loss is less striking. But scenario II leads to more substantial efficiency deduction which can be as high as nearly 20% for two methods. It is also interesting to report that the efficiency loss induced from unknown association parameters is null for TS and very small for PL, which is at nearly 1.5% in Scenario II.

Table 2.5: Efficiency comparison of the $\boldsymbol{\beta}$ estimators under various scenarios of unknown nuisance parameters

Scenario 1[†]

s	Nuisance Para.				$R_s^{PL}(\beta_0)$	$R_s^{TS}(\beta_0)$	$R_s^{PL}(\beta_1)$	$R_s^{TS}(\beta_1)$
	σ_y^2	$\boldsymbol{\alpha}$	$\boldsymbol{\lambda}$	$\boldsymbol{\delta}$				
1	$\times$[*]	$\surd$	$\surd$	$\surd$	0.999	1.000	0.998	1.000
2	$\surd$	$\times$	$\surd$	$\surd$	1.000	0.999	0.989	0.993
3	$\surd$	$\surd$	$\times$	$\surd$	0.992	0.988	0.963	0.957
4	$\times$	$\times$	$\surd$	$\surd$	1.000	0.999	0.987	0.993
5	$\times$	$\surd$	$\times$	$\surd$	0.990	0.988	0.962	0.955
6	$\surd$	$\times$	$\times$	$\surd$	0.990	0.979	0.942	0.928
7	$\times$	$\times$	$\times$	$\surd$	0.989	0.979	0.940	0.927
8	$\surd$	$\surd$	$\surd$	$\times$	0.999	1.000	0.996	1.000
9	$\times$	$\times$	$\times$	$\times$	0.988	0.979	0.939	0.927

Scenario 2

s	Nuisance Para.				$R_s^{PL}(\beta_0)$	$R_s^{TS}(\beta_0)$	$R_s^{PL}(\beta_1)$	$R_s^{TS}(\beta_1)$
	σ_y^2	$\boldsymbol{\alpha}$	$\boldsymbol{\lambda}$	$\boldsymbol{\delta}$				
1	$\times$[*]	$\surd$	$\surd$	$\surd$	0.985	0.994	1.000	0.990
2	$\surd$	$\times$	$\surd$	$\surd$	1.000	0.999	0.881	0.856
3	$\surd$	$\surd$	$\times$	$\surd$	0.994	0.988	0.960	0.955
4	$\times$	$\times$	$\surd$	$\surd$	0.985	0.994	0.867	0.818
5	$\times$	$\surd$	$\times$	$\surd$	0.975	0.978	0.959	0.937
6	$\surd$	$\times$	$\times$	$\surd$	0.993	0.987	0.854	0.824
7	$\times$	$\times$	$\times$	$\surd$	0.975	0.978	0.832	0.773
8	$\surd$	$\surd$	$\surd$	$\times$	0.997	1.000	0.984	1.000
9	$\times$	$\times$	$\times$	$\times$	0.973	0.978	0.830	0.773

[†] Scenario 1 follows identical settings in continuous variable simulation study in Section 2.4.1. Scenario 2 involves analogous settings in response and covariate processes, but the missing process has $\eta_{ij}^{Ry} = 1.5 - 0.5y_{ij} - 0.5R_{ij}^{x}$, $\eta_{ij}^{Rx} = 1.5 - y_{ij}$.

[*] $\times$ and $\surd$ indicate the corresponding nuisance parameter is unknown or known,respectively.

Next, we are interested in assessing efficiency for estimators obtained from different methods. Again, we consider the model settings in Section 2.4.1. To highlight comparisons on the $\boldsymbol{\beta}$ parameter, we assume all nuisance parameters are known for simplicity. For the TS method, avar$(\tilde{\beta}_j^{TS})$ is the diagonal element of $[E\{\partial S_{1i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\}]^{-1}\cdot$

$E\{S_{1i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\beta})^T\} \cdot [E\{\partial S_{1i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\}]^{-1T}$; for the PL method $\text{avar}(\tilde{\beta}_j^{PL})$ is the diagonal element of $[E\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\}]^{-1} \cdot E\{S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\beta})^T\} \cdot [E\{\partial S_{2i}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\}]^{-1T}$; while for the ML approach, $\text{avar}(\tilde{\beta}_j^{ML})$ is obtained from the diagonal element of $[E\{S_i^F(\boldsymbol{\beta})S_i^F(\boldsymbol{\beta})^T\}]^{-1}$, where $S_i^F(\boldsymbol{\beta})$ is the score function of $\boldsymbol{\beta}$ from the fully specified likelihood function. Let $R_{ML:TS}(\beta_j) = \text{avar}(\tilde{\beta}_j^{ML})/\text{avar}(\tilde{\beta}_j^{TS})$, $R_{ML:PL}(\beta_j) = \text{avar}(\tilde{\beta}_j^{ML})/\text{avar}(\tilde{\beta}_j^{PL})$, and $R_{PL:TS}(\beta_j) = \text{avar}(\tilde{\beta}_j^{PL})/\text{avar}(\tilde{\beta}_j^{TS})$ $(j = 0, 1)$ be the relative efficiency for corresponding estimators. We consider the case with a common exchangeable correlation coefficient $\rho = \psi^y = \psi^x = \rho^y = \rho^x$.

We evaluate the relative efficiency of the PL and TS estimators with respect to the ML estimator and display the result in Figure 2.1. As expected, both the PL and TS methods incur efficiency loss. As the correlation becomes stronger, the loss of efficiency increases. When the measurements are uncorrelated, the PL, TS and ML methods produce the same asymptotic variance. In addition, the efficiency loss in using the PL method is less striking than that incurred by using the TS method. It is noted that efficiency loss associated with intercept β_0 is less profound than that for the covariate effect β_1. To better visualize the relative performance of the PL and TS methods, we show the relative efficiency $R_{PL:TS}(\beta_j)$ $(j = 0, 1)$ in Figure 2.2 as well.

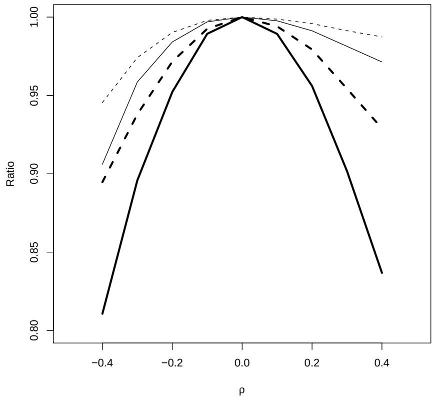

Figure 2.1: *Relative efficiency with respect to common correlation coefficient ρ. $R_{ML:TS}(\beta_0)$: ▬ ▬ ▬ ; $R_{ML:PL}(\beta_0)$: − − − ; $R_{ML:TS}(\beta_1)$: ━━━ ; $R_{ML:PL}(\beta_1)$: ——.*

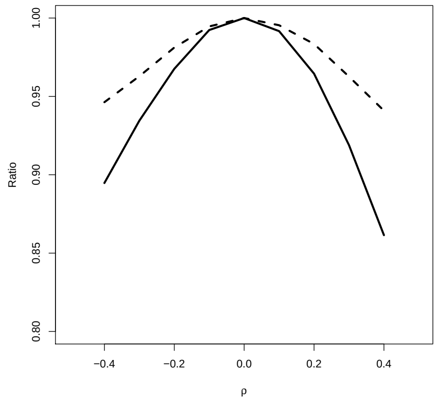

Figure 2.2: *Relative efficiency with respect to common correlation coefficient ρ. $R_{PL:TS}(\beta_0)$: – – – ; $R_{PL:TS}(\beta_1)$: ——.*

2.6 Sensitivity Analysis for Model Misspecification

The validity of the proposed method requires the correct model specification, and this involves modeling of the response, covariate and missing data processes. Now we investigate the impact of model misspecification on the estimation of the parameter $\boldsymbol{\theta}$.

If we apply a misspecified model to fit data, then the resultant estimator, denoted by $\hat{\boldsymbol{\theta}}^*$, for the parameter $\boldsymbol{\theta}$ would converge in probability to a limit, say $\boldsymbol{\theta}^*$, which may differ from the true parameter value $\boldsymbol{\theta}$. Specifically, let $\mathcal{L}^*(\boldsymbol{\theta}^*)$ be the marginal or pairwise likelihood function formulated from a misspecified model. Then according to the result in Yi and Reid (2010), under certain regularity conditions, the limit $\boldsymbol{\theta}^*$ is the solution of

$$E_{(Y,X,R)}\left\{\frac{\partial \log \mathcal{L}^*(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*}\right\} = \mathbf{0}, \tag{2.12}$$

where the expectation is taken under the true joint distribution for Y, X and R variables. In most situations, equation (2.12) does not have an analytically closed solution. Hence the relationship between $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$ is frequently evaluated via numerical assessment. Now we undertake numerical studies by assuming the similar settings described in Section 2.4.1, and focus the discussion primarily on parameter $\boldsymbol{\beta}$.

Firstly, to compare the robustness of the PL method relative to the TS method, we first consider the case that all the marginal models including response, covariate and missing processes are correctly specified, but the association structures are misspecified. The true correlation matrix for the response process

$$\begin{pmatrix} 1 & \psi^y + \kappa & \psi^y - \kappa \\ \psi^y + \kappa & 1 & \psi^y + \kappa \\ \psi^y - \kappa & \psi^y + \kappa & 1 \end{pmatrix},$$

is used to generate the data, but a misspecified correlation structure with common correlation coefficient

$$\begin{pmatrix} 1 & \psi^y & \psi^y \\ \psi^y & 1 & \psi^y \\ \psi^y & \psi^y & 1 \end{pmatrix},$$

is used to fit the data. Moreover, the covariate and missing processes are misspecified by a common correlation coefficient but the true correlation matrix follows same form as the response process.

In Figure 2.3 we display the relative biases defined as $(100 \times (\beta^* - \beta)/\beta)$. It is seen that for both PL and TS methods, the asymptotic relative biases for β_0 and β_1 are negligible, showing that both approaches are robust to the misspecification of association structures under current model settings.

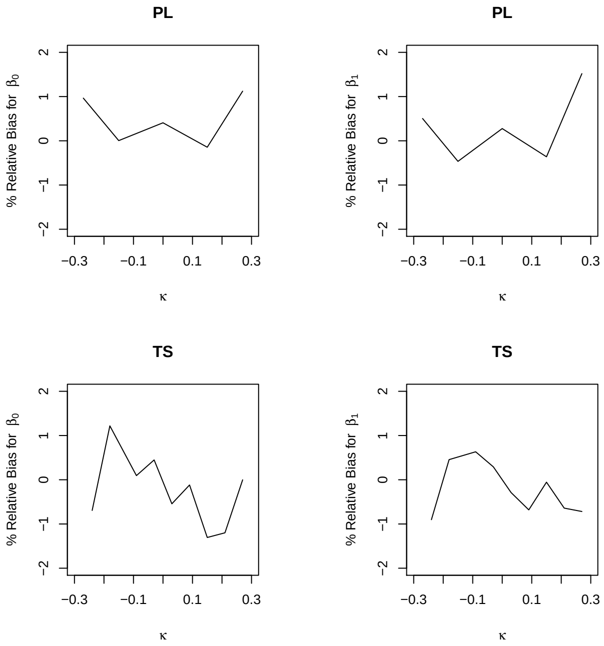

Figure 2.3: *Asymptotic relative bias for regression coefficients β_0 and β_1 for PL and TS methods when the association structures for the response, covariate and missing processes are all misspecified. The models for estimation involves common correlation coefficient. However, the true correlation matrix for response process has the form in (2.13).*

In the reminder of this section, we focus the assessment on the misspecification of some marginal models. First, we consider the case that the marginal mean model for the response process is misspecified but other processes are modeled correctly. In particular, we generate data from the following two means models along with other models described in Section 2.4.1: (1) $\mu_{ij}^y = \beta_0 + \beta_1 X_{ij} + \kappa \cdot j$; and (2) $\mu_{ij}^y = \beta_0 + \beta_1 X_{ij} + \kappa \cdot x_{ij} \cdot j$. Regardless of the true model, we always fit the data with the model in Section 2.4.1 where the mean is specified as $\mu_{ij}^y = \beta_0 + \beta_1 X_{ij}$. Figure 2.4 displays the asymptotic percent relative bias against varying degrees of κ. It is observed that when a specific term in response process is ignored, the bias would occur. As expected, the stronger influence of the omitting term on response, the larger the relative bias. Moreover, the PL and TS methods result in same bias patterns.

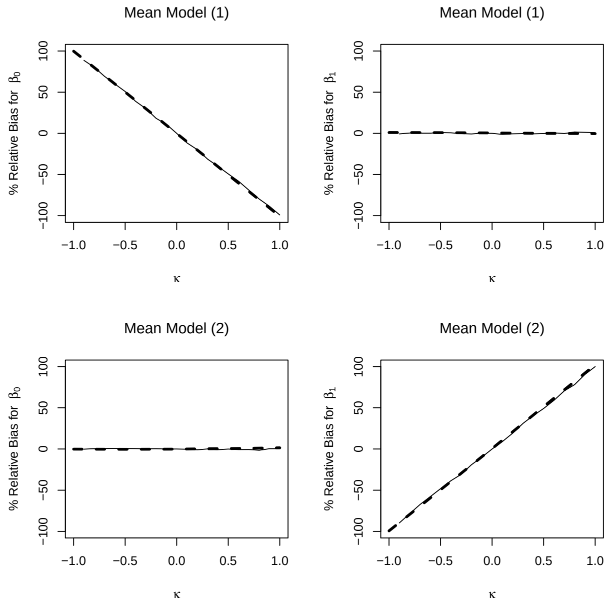

Figure 2.4: *Asymptotic percent relative bias for regression coefficients β_0 and β_1 when response models are misspecified. The model for estimation is $\mu_{ij}^y = \beta_0 + \beta_1 X_{ij}$, while true models are: $\mu_{ij}^y = \beta_0 + \beta_1 X_{ij} + \kappa \cdot j$ for mean model (1) and $\mu_{ij}^y = \beta_0 + \beta_1 X_{ij} + \kappa \cdot X_{ij} \cdot j$ for mean model (2), respectively. PL method: - - -; TS method: ——.*

Finally, we evaluate the impact of misspecifying the missing processes while the response and covariate process are retained being correctly specified. True models of the missing processes given by $\eta_{ij}^{Ry} = \lambda_0^y + \lambda_1^y X_{ij} + \lambda_2^y R_{ij}^x + \kappa Y_{ij}$ and $\eta_{ij}^{Rx} = \lambda_0^x + \lambda_1^x X_{ij} + \kappa Y_{ij}$, are particularly considered. But we fit data with models described in Section 2.4.1, where in particular, the missing data models are $\eta_{ij}^{Ry} = \lambda_0^y + \lambda_1^y X_{ij} + \lambda_2^y R_{ij}^x$, and $\eta_{ij}^{Rx} = \lambda_0^x + \lambda_1^x X_{ij}$. In Figure 2.5, we display the asymptotic relative biases for β_0 and β_1. Again, various patterns of inflating biases are observed, and the PL and TS methods follow similar pattern.

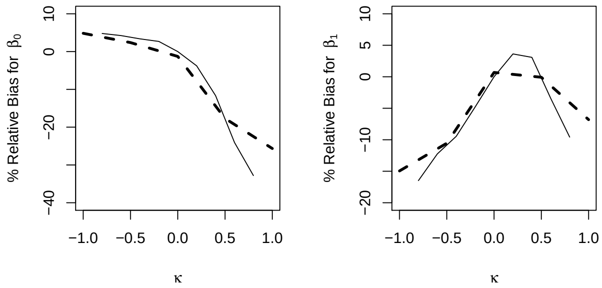

Figure 2.5: *Asymptotic relative bias for regression coefficients β_0 and β_1 when the missing data process is misspecified. The model for estimation is specified in Section 2.4.1, while the true model is $\eta_{ij}^{Ry} = \lambda_0^y + \lambda_1^y X_{ij} + \lambda_2^y R_{ij}^x + \kappa Y_{ij}$, $\eta_{ij}^{Rx} = \lambda_0^x + \lambda_1^x X_{ij} + \kappa Y_{ij}$. PL method: - - -; TS method: ———.*

Appendix A: Proof of Unbiasedness of Estimating Functions

To show unbiasedness of estimating functions, it suffices to show that

$$E_{Y_i, X_i, R_i^y, R_i^x | Z_i} \left[\sum_{i=1}^{n} \frac{\partial \log \mathcal{L}_{C1,i}(\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}} \right] = \mathbf{0}.$$

The proof of $E_{Y_i, X_i, R_i^y, R_i^x | Z_i} \left[\sum_{i=1}^{n} \partial \log \mathcal{L}_{C2,i}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \right] = \mathbf{0}$ follows analogously. Let

$$K_{1,ij} = f(Y_{ij}|X_{ij}, Z_{ij}) f(X_{ij}|Z_{ij}) P(R_{ij}^y = 1, R_{ij}^x = 1 | Y_{ij}, X_{ij}, Z_{ij}),$$

$$K_{2,ij} = \int \left\{ f(Y_{ij}|X_{ij}, Z_{ij}) f(X_{ij}|Z_{ij}) P(R_{ij}^y = 0, R_{ij}^x = 1 | Y_{ij}, X_{ij}, Z_{ij}) \right\} dY_{ij},$$

$$K_{3,ij} = \int \left\{ f(Y_{ij}|X_{ij}, Z_{ij}) f(X_{ij}|Z_{ij}) P(R_{ij}^y = 1, R_{ij}^x = 0 | Y_{ij}, X_{ij}, Z_{ij}) \right\} dX_{ij},$$

and

$$K_{4,ij} = \iint \left\{ f(Y_{ij}|X_{ij}, Z_{ij}) f(X_{ij}|Z_{ij}) P(R_{ij}^y = 0, R_{ij}^x = 0 | Y_{ij}, X_{ij}, Z_{ij}) \right\} dY_{ij} dX_{ij},$$

then we write

$$\begin{aligned}\log \mathcal{L}_{C1}(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\sum_{j=1}^{m}\Big\{R_{ij}^{y}R_{ij}^{x}\log K_{1,ij} + (1-R_{ij}^{y})R_{ij}^{x}\log K_{2,ij} \\ &\quad +R_{ij}^{y}(1-R_{ij}^{x})\log K_{3,ij} + (1-R_{ij}^{y})(1-R_{ij}^{x})\log K_{4,ij}\Big\}. \qquad (2.13)\end{aligned}$$

By the distinctness of the parameters in different processes, we have

$$\begin{aligned}&E_{(Y_i,X_i,R_i^y,R_i^x|Z_i)}\Big(R_{ij}^{y}R_{ij}^{x}\frac{\partial \mathrm{log} K_{1,ij}}{\partial \boldsymbol{\beta}}\Big) \\ =\ & E_{(Y_i,X_i|Z_i)}\Big\{E_{(R_i^y,R_i^x|Y_i,X_i,Z_i)}\Big(R_{ij}^{y}R_{ij}^{x}\frac{\partial \mathrm{log} K_{1,ij}}{\partial \boldsymbol{\beta}}\Big)\Big\} \\ =\ & E_{(Y_i,X_i|Z_i)}\Big\{P(R_{ij}^{y}=1, R_{ij}^{x}=1|Y_{ij},X_{ij},Z_{ij})\frac{\partial \mathrm{log} f(Y_{ij}|X_{ij},Z_{ij})}{\partial \boldsymbol{\beta}}\Big\}.\end{aligned}$$

Note that

$$K_{2,ij} = f(X_{ij}|Z_{ij})\cdot E_{(Y_i|X_i,Z_i)}\left\{P(R_{ij}^{y}=0, R_{ij}^{x}=1|Y_{ij},X_{ij},Z_{ij})\right\},$$

then for the second term in (2.13), we have

$$\begin{aligned}&E_{(Y_i,X_i,R_i^y,R_i^x|Z_i)}\Big\{(1-R_{ij}^{y})R_{ij}^{x}\frac{\partial \mathrm{log} K_{2,ij}}{\partial \boldsymbol{\beta}}\Big\} \\ =\ & E_{(X_i|Z_i)}\left[E_{(Y_i|X_i,Z_i)}\Big\{E_{(R_i^y,R_i^x|Y_i,X_i,Z_i)}\Big((1-R_{ij}^{y})R_{ij}^{x}\frac{\partial \mathrm{log} K_{2,ij}}{\partial \boldsymbol{\beta}}\Big)\Big\}\right] \\ =\ & E_{(X_i|Z_i)}\left[E_{(Y_i|X_i,Z_i)}\Big\{P(R_{ij}^{y}=0, R_{ij}^{x}=1|Y_{ij},X_{ij},Z_{ij})\Big(\frac{\partial \mathrm{log} K_{2,ij}}{\partial \boldsymbol{\beta}}\Big)\Big\}\right] \\ =\ & E_{(X_i|Z_i)}\left[\Big\{E_{(Y_i|X_i,Z_i)}\{P(R_{ij}^{y}=0, R_{ij}^{x}=1|Y_{ij},X_{ij},Z_{ij})\}\Big\}\times\frac{1}{K_{2,ij}}\times\frac{\partial K_{2,ij}}{\partial \boldsymbol{\beta}}\right] \\ =\ & E_{(X_i|Z_i)}\left\{\frac{1}{f(X_{ij}|Z_{ij})}\cdot\frac{\partial K_{2,ij}}{\partial \boldsymbol{\beta}}\right\}. \qquad (2.14)\end{aligned}$$

By the distinctness of the parameters in different processes, we have

$$\begin{aligned}
\frac{\partial K_{2,ij}}{\partial \boldsymbol{\beta}} &= \int \left\{ \frac{\partial f(Y_{ij}|X_{ij}, Z_{ij})}{\partial \boldsymbol{\beta}} f(X_{ij}|Z_{ij}) P(R_{ij}^y = 0, R_{ij}^x = 1|Y_{ij}, X_{ij}, Z_{ij}) \right\} dY_{ij} \\
&= \int \left\{ \frac{\partial \log f(Y_{ij}|X_{ij}, Z_{ij})}{\partial \boldsymbol{\beta}} f(Y_{ij}|X_{ij}, Z_{ij}) f(X_{ij}|Z_{ij}) P(R_{ij}^y = 0, R_{ij}^x = 1|Y_{ij}, X_{ij}, Z_{ij}) \right\} dY_{ij} \\
&= f(X_{ij}|Z_{ij}) \cdot E_{(Y_i|X_i,Z_i)} \left\{ \frac{\partial \log f(Y_{ij}|X_{ij}, Z_{ij})}{\partial \boldsymbol{\beta}} P(R_{ij}^y = 0, R_{ij}^x = 1|Y_{ij}, X_{ij}, Z_{ij}) \right\},
\end{aligned}$$

therefore, (2.14) becomes

$$E_{(Y_i,X_i|Z_i)}\Big\{ P(R_{ij}^y = 0, R_{ij}^x = 1|Y_{ij}, X_{ij}, Z_{ij}) \frac{\partial \log f(Y_{ij}|X_{ij}, Z_{ij})}{\partial \boldsymbol{\beta}} \Big\}.$$

Analogously, for the third and fourth terms in (2.13), we obtain

$$E\Big(\frac{\partial \log K_{3,ij}}{\partial \boldsymbol{\beta}}\Big) = E_{(Y_i,X_i|Z_i)}\Big\{ P(R_{ij}^y = 1, R_{ij}^x = 0|Y_{ij}, X_{ij}, Z_{ij}) \frac{\partial \log f(Y_{ij}|X_{ij}, Z_{ij})}{\partial \boldsymbol{\beta}} \Big\},$$

and

$$E\Big(\frac{\partial \log K_{4,ij}}{\partial \boldsymbol{\beta}}\Big) = E_{(Y_i,X_i|Z_i)}\Big\{ P(R_{ij}^y = 0, R_{ij}^x = 0|Y_{ij}, X_{ij}, Z_{ij}) \frac{\partial \log f(Y_{ij}|X_{ij}, Z_{ij})}{\partial \boldsymbol{\beta}} \Big\},$$

where the expectation "E" is evaluated with respect to the conditional distribution of (Y_i, X_i, R_i^y, R_i^x) given Z_i.

Then combining these results leads to

$$E\left\{ \sum_{i=1}^{n} \partial \log \mathcal{L}_{C1,i}(\boldsymbol{\gamma})/\partial \boldsymbol{\beta} \right\} = \sum_{i=1}^{n} \sum_{j=1}^{m} E_{(Y_i,X_i|Z_i)} \left\{ \partial \log f(Y_{ij}|X_{ij}, Z_{ij})/\partial \boldsymbol{\beta} \right\} = \mathbf{0}.$$

Appendix B: Asymptotic Distribution for $\hat{\boldsymbol{\beta}}_{PL}$

We sketch the proof of the asymptotic distribution for $\hat{\boldsymbol{\beta}}_{PL}$ and the asymptotic distribution for $\hat{\boldsymbol{\beta}}_{TS}$ follows similarly. Appendix A shows that $E\{S_{2i}(\boldsymbol{\theta})\} = \mathbf{0}$. Apply estimating function theory leads to the asymptotic distribution

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta}) \to_D N(\mathbf{0}, \{E(D_i)\}^{-1} E\{S_{2i}(\boldsymbol{\theta}) S_{2i}(\boldsymbol{\theta})^T\} \{E(D_i)\}^{-1T}). \qquad (2.15)$$

Rewrite $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\nu}^T)^T$, and $S_{2i}(\boldsymbol{\theta}) = (S_{2i}(\boldsymbol{\beta})^T, S_{2i}(\boldsymbol{\nu})^T)^T$, yielding

$$E(D_i) = E\begin{pmatrix} \partial S_{2i}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^T & \partial S_{2i}(\boldsymbol{\beta})/\partial \boldsymbol{\nu}^T \\ \partial S_{2i}(\boldsymbol{\nu})/\partial \boldsymbol{\beta}^T & \partial S_{2i}(\boldsymbol{\nu})/\partial \boldsymbol{\nu}^T \end{pmatrix},$$

and

$$E\{S_{2i}(\boldsymbol{\theta})S_{2i}(\boldsymbol{\theta})^T\} = E\begin{pmatrix} S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\beta})^T & S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\nu})^T \\ S_{2i}(\boldsymbol{\nu})S_{2i}(\boldsymbol{\beta})^T & S_{2i}(\boldsymbol{\nu})S_{2i}(\boldsymbol{\nu})^T \end{pmatrix}.$$

Using (2.15), we obtain the asymptotic covariance matrix for $\sqrt{n}(\hat{\boldsymbol{\beta}}_{PL} - \boldsymbol{\beta})$ using the left-upper block matrix from

$$E^{-1}\begin{pmatrix} \frac{\partial S_{2i}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} & \frac{\partial S_{2i}(\boldsymbol{\beta})}{\partial \boldsymbol{\nu}^T} \\ \frac{\partial S_{2i}(\boldsymbol{\nu})}{\partial \boldsymbol{\beta}^T} & \frac{\partial S_{2i}(\boldsymbol{\nu})}{\partial \boldsymbol{\nu}^T} \end{pmatrix} E\begin{pmatrix} S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\beta})^T & S_{2i}(\boldsymbol{\beta})S_{2i}(\boldsymbol{\nu})^T \\ S_{2i}(\boldsymbol{\nu})S_{2i}(\boldsymbol{\beta})^T & S_{2i}(\boldsymbol{\nu})S_{2i}(\boldsymbol{\nu})^T \end{pmatrix} E^{-1}\begin{pmatrix} \frac{\partial S_{2i}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} & \frac{\partial S_{2i}(\boldsymbol{\beta})}{\partial \boldsymbol{\nu}^T} \\ \frac{\partial S_{2i}(\boldsymbol{\nu})}{\partial \boldsymbol{\beta}^T} & \frac{\partial S_{2i}(\boldsymbol{\nu})}{\partial \boldsymbol{\nu}^T} \end{pmatrix}^T.$$

Chapter 3

Analysis of Longitudinal Binary Data with Missing Response and Covariates

3.1 Introduction

To provide a complement of Chapter 2 which focuses on continuous responses, we address the analysis of longitudinal binary data with the composite likelihood method. The remainder of this chapter is organized as follows. Section 3.2 introduces notations and the model setups. Inference methods are presented in Section 3.3. In Section 3.4, we analyze the National Population Health Survey (NPHS) data with the proposed methods. To evaluate the performance of the proposed methods, we conduct various empirical studies and report the results in Section 3.5. Concluding remarks are given in Section 3.6.

3.2 Model Formulation

3.2.1 Response Process

Suppose that there are n subjects and m assessment times. Let Y_{ij} be the binary response variable, and X_{ij} be a covariate vector for subject i at occasion j. Both Y_{ij} and X_{ij} are subject to missingness. For ease of exposition, here we consider the case that X_{ij} is a scalar. An extension to multiple covariates X_{ij} is discussed in Chapter 6. Let Z_{ij} be a vector of covariates that have complete observations. Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im})^T$, $X_i = (X_{i1}, X_{i2} \ldots, X_{im})^T$ and $Z_i = (Z_{i1}^T, Z_{i2}^T, \ldots, Z_{im}^T)^T$.

To model the relationship between the response and the covariates, one may attempt to fully specify a distributional form for $P(Y_i = y_i | X_i, Z_i)$, where y_i is a binary vector. However, this could be difficult in many situations, especially when the dimension m is large. Moreover, fully modeling a multivariate distribution can introduce considerable computation cost (e.g. Ochi and Prentice (1984)). To protect against model misspecification and ease computation, we consider a pairwise modeling strategy. First, we introduce some notations. For a given 2×2 correlation matrix

$$\mathbf{v} = \begin{pmatrix} 1 & v_{12} \\ v_{12} & 1 \end{pmatrix},$$

let $\phi_2(\mathbf{z}; \mathbf{v})$ be the probability density function for a bivariate normal distribution, given by

$$\phi_2(\mathbf{z}; \mathbf{v}) = (2\pi)^{-1} |\mathbf{v}|^{-1/2} \exp\big(- \frac{1}{2} \mathbf{z}^T \mathbf{v}^{-1} \mathbf{z} \big),$$

where $\mathbf{z} = (z_1, z_2)^T$. For $u = (u_1, u_2)^T$, let $\Phi_2(\mathbf{u}; \mathbf{v})$ denote the corresponding bivariate cumulative distribution function:

$$\Phi_2(\mathbf{u}; \mathbf{v}) = \int_{-\infty}^{u_2} \int_{-\infty}^{u_1} \phi_2(\mathbf{z}; \mathbf{v}) \, \mathrm{d}z_1 \mathrm{d}z_2.$$

Now we consider bivariate probit models for paired responses Y_{ij} and Y_{ik}, $j < k$, $i = 1, \cdots, n$. Namely, we set

$$P(Y_{ij} = 1, Y_{ik} = 1 \mid X_i, Z_i) = \Phi_2\Big(\boldsymbol{\eta}_{ijk}^y ; \boldsymbol{\Sigma}(\psi_{ijk}^y) \Big), \tag{3.1}$$

where $\boldsymbol{\eta}_{ijk}^{y} = (\eta_{ij}^{y}, \eta_{ik}^{y})^T$ is the linear predictors, and $\boldsymbol{\Sigma}(\psi_{ijk}^{y})$ is a 2×2 covariance matrix with diagonal elements 1 and correlation coefficient ψ_{ijk}^{y}. Requiring the diagonal elements of $\boldsymbol{\Sigma}(\psi_{ijk}^{y})$ to be 1 is to ensure model identifiability (e.g., Renard et al. (2002) and Roy and Banerjee (2009)).

To make modeling more parsimonious, we further consider a regression model to reflect the dependence of η_{ij}^{y} on the covariates at occasion j :

$$\eta_{ij}^{y} = X_{ij}\beta_x + Z_{ij}^{T}\boldsymbol{\beta}_z,$$

where $\boldsymbol{\beta} = (\beta_x, \boldsymbol{\beta}_z^T)^T$ is a vector of regression parameters linking the covariates and the response. With this step, it is immediate that

$$P(Y_{ij} = 1|X_i, Z_i) = \Phi_1(\eta_{ij}^{y}), \tag{3.2}$$

where $\Phi_1(u_1)$ represents the cumulative distribution function for the standard normal distribution $N(0,1)$, i.e., $\Phi_1(u_1) = \int_{-\infty}^{u_1} \phi_1(z_1)\,\mathrm{d}z_1$ with $\phi_1(z_1) = (2\pi)^{-1/2}\exp\big(-z_1^2/2\big)$.

Analogous to a transformation discussed in (Hawkins, 1989), we model correlation coefficient ψ_{ijk}^{y} with a regression model

$$\log\left(\frac{1+\psi_{ijk}^{y}}{1-\psi_{ijk}^{y}}\right) = h^y(\boldsymbol{\psi}^y; \mathbf{w}_{ijk}^{y}), \tag{3.3}$$

where $\boldsymbol{\psi}^y$ is the vector of regression coefficients, $\mathbf{w}_{ijk}^{y}$ is a vector of covariates, and h^y is a known function that takes values over the entire real number line. For instance, setting $h^y(\boldsymbol{\psi}^y; \mathbf{w}_{ijk}^{y})$ to be a scalar $\boldsymbol{\psi}^y$ leads to an exchangeable correlation structure, while taking

$$h^y(\boldsymbol{\psi}^y; \mathbf{w}_{ijk}^{y}) = \log\left(\frac{1+\Big[\{\exp(\boldsymbol{\psi}^y)-1\}/\{\exp(\boldsymbol{\psi}^y)+1\}\Big]^{|j-k|}}{1-\Big[\{\exp(\boldsymbol{\psi}^y)-1\}/\{\exp(\boldsymbol{\psi}^y)+1\}\Big]^{|j-k|}}\right)$$

results in an AR(1) correlation structure. An obvious advantage of (3.3) is to provide a reparameterization for correlation coefficient ψ_{ijk}^{y}, so there is no need to impose any constraints on parameter $\boldsymbol{\psi}^y$. Moreover, (3.3) enables us to describe complex dependence of association structures on covariates by specifying different forms of the $h^y(\boldsymbol{\psi}^y; \mathbf{w}_{ijk}^{y})$ function, such as a linear function $h^y(\boldsymbol{\psi}^y; \mathbf{w}_{ijk}^{y}) = (\mathbf{w}_{ijk}^{y})^T\boldsymbol{\psi}^y$.

3.2.2 Covariate Process

By analogy, to postulate the covariate process, we do not attempt to specify the full distribution with probability density (or mass) function $f(X_i|Z_i)$ (or $P(X_i = \mathbf{x}_i|Z_i)$). Instead, we focus on specifying a pairwise distribution to gain protection from misspecification of higher order structures. If X_{ij} is binary, we consider for $j < k$,

$$P(X_{ij} = 1, X_{ik} = 1 \mid Z_i) = \Phi_2\Big(\boldsymbol{\eta}^x_{ijk}; \boldsymbol{\Sigma}(\psi^x_{ijk})\Big), \tag{3.4}$$

where $\boldsymbol{\eta}^x_{ijk} = (\eta^x_{ij}, \eta^x_{ik})^T$, and a regression model is applied to reflect the dependence of η^x_{ij} on the covariates Z_i

$$\eta^x_{ij} = Z^T_{ij}\boldsymbol{\alpha}, \tag{3.5}$$

with $\boldsymbol{\alpha}$ being the vector of regression parameters. It is immediate that

$$P(X_{ij} = 1 \mid Z_i) = \Phi_1(\eta^x_{ij}). \tag{3.6}$$

Following the same spirit in response process, correlation coefficient ψ^x_{ijk} is modeled as

$$\log\left(\frac{1 + \psi^x_{ijk}}{1 - \psi^x_{ijk}}\right) = h^x(\boldsymbol{\psi}^x; \mathbf{w}^x_{ijk}), \tag{3.7}$$

where $\boldsymbol{\psi}^x$ is the vector of regression coefficients, $\mathbf{w}^x_{ijk}$ is a vector of covariates, and h^x is a specified function.

If X_{ij} is continuous, a bivariate normal distribution can be an option to postulate paired variables $X_{ijk} = (X_{ij}, X_{ik})^T$. That is, conditional on Z_i, assume X_{ijk} has the probability density function

$$\begin{aligned} & f_2(x_{ijk}; \boldsymbol{\mu}^x_{ijk}, \boldsymbol{\Sigma}(\sigma^2_x, \psi^x_{ijk})) \\ = \;& (2\pi)^{-1}|\boldsymbol{\Sigma}(\sigma^2_x, \psi^x_{ijk})|^{-1/2} \exp\Big\{-\frac{1}{2}(x_{ijk} - \boldsymbol{\mu}^x_{ijk})^T\boldsymbol{\Sigma}(\sigma^2_x, \psi^x_{ijk})^{-1}(x_{ijk} - \boldsymbol{\mu}^x_{ijk})\Big\}, \end{aligned} \tag{3.8}$$

where $x_{ijk} = (x_{ij}, x_{ik})^T$, $\boldsymbol{\mu}^x_{ijk} = (\mu^x_{ij}, \mu^x_{ik})^T$, and $\boldsymbol{\Sigma}(\sigma^2_x, \psi^x_{ijk})$ is a 2×2 covariance matrix with diagonal elements being σ^2_x and the correlation coefficient being ψ^x_{ijk}. It is noted that μ^x_{ij}

and σ_x^2 are the conditional marginal mean and variance of X_{ij} given Z_i, respectively. By analogy, μ_{ij}^x and ψ_{ijk}^x may be respectively modulated as (3.5) and (3.7). More generally, a bivariate skew normal distribution can be employed to model non-normal X_{ijk} for greater flexibility. Properties of this type of distributions are discussed by Azzalini and Capitanio (1999).

3.2.3 Missing Data Process

Let $R_{ij}^y = 1$ if Y_{ij} is observed and 0 otherwise. Let $R_{ij}^x = 1$ if X_{ij} is observed and 0 otherwise. Denote $R_i^y = (R_{i1}^y, R_{i2}^y, \ldots, R_{im}^y)^T$ and $R_i^x = (R_{i1}^x, R_{i2}^x, \ldots, R_{im}^x)^T$. For the missing data process, we follow the same lines to postulate pairwise models. In particular, we only model $P(R_{ij}^y = 1, R_{ik}^y = 1 | Y_i, X_i, Z_i, R_{ij}^x, R_{ik}^x)$ and $P(R_{ij}^x = 1, R_{ik}^x = 1 | Y_i, X_i, Z_i)$, which uniquely determine the distribution $P(R_{ij}^y = 1, R_{ik}^y = 1, R_{ij}^x = 1, R_{ik}^x = 1 | Y_i, X_i, Z_i)$. Specifically, for $j < k$, the pairwise model is specified as

$$P(R_{ij}^y = 1, R_{ik}^y = 1 | Y_i, X_i, Z_i, R_{ij}^x, R_{ik}^x) = \Phi_2(\boldsymbol{\eta}_{ijk}^{Ry}; \boldsymbol{\Sigma}(\rho_{ijk}^y)),$$

and

$$P(R_{ij}^x = 1, R_{ik}^x = 1 | Y_i, X_i, Z_i) = \Phi_2(\boldsymbol{\eta}_{ijk}^{Rx}; \boldsymbol{\Sigma}(\rho_{ijk}^x)), \tag{3.9}$$

where $\boldsymbol{\eta}_{ijk}^{Ry} = (\eta_{ij}^{Ry}, \eta_{ik}^{Ry})^T$, $\boldsymbol{\eta}_{ijk}^{Rx} = (\eta_{ij}^{Rx}, \eta_{ik}^{Rx})^T$, and the regression models

$$\eta_{ij}^{Ry} = \boldsymbol{\lambda}^{y^T} \boldsymbol{\xi}_{ij}^y$$

and

$$\eta_{ij}^{Rx} = \boldsymbol{\lambda}^{x^T} \boldsymbol{\xi}_{ij}^x$$

can be introduced to reflect the dependence of $(\eta_{ij}^{Ry}, \eta_{ij}^{Rx})$ on response and covariate variables, respectively. $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^{y^T}, \boldsymbol{\lambda}^{x^T})^T$ are the regression parameters related to the missing data process, and $\boldsymbol{\xi}_{ij}^y$ and $\boldsymbol{\xi}_{ij}^x$ are subsets of $\{Y_{ij}, X_{ij}, Z_{ij}, R_{ij}^x\}$ and $\{Y_{ij}, X_{ij}, Z_{ij}\}$, respectively. Similarly, the correlation coefficients ρ_{ijk}^y and ρ_{ijk}^x can be modeled as

$$\log\left(\frac{1 + \rho_{ijk}^y}{1 - \rho_{ijk}^y}\right) = h^{Ry}(\boldsymbol{\rho}^y; \mathbf{w}_{ijk}^{Ry}),$$

and

$$\log\left(\frac{1+\rho^x_{ijk}}{1-\rho^x_{ijk}}\right) = h^{Rx}(\boldsymbol{\rho}^x; \mathbf{w}^{Rx}_{ijk}),$$

respectively, where $\boldsymbol{\rho}^y$ and $\boldsymbol{\rho}^x$ are the vectors of regression coefficients, $\mathbf{w}^{Ry}_{ijk}$ and $\mathbf{w}^{Rx}_{ijk}$ are subsets of $\{Y_{ij}, X_{ij}, Z_{ij}, R^x_{ij}\}$ and $\{Y_{ij}, X_{ij}, Z_{ij}\}$, respectively, and h^{Ry} and h^{Rx} are given functions. It is immediate that

$$P(R^y_{ij}=1, R^x_{ij}=1|Y_i, X_i, Z_i) = \sum_{r^y_{ik}=0}^{1}\sum_{r^x_{ik}=0}^{1} P(R^y_{ij}=1, R^y_{ik}=r^y_{ik}, R^x_{ij}=1, R^x_{ik}=r^x_{ik}|Y_i, X_i, Z_i). \tag{3.10}$$

3.3 Estimation and Inference

3.3.1 Marginal and Pairwise Likelihoods

Let $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\lambda}^T)^T$ be the parameters associated with the marginal structure, and $\boldsymbol{\delta} = (\boldsymbol{\psi}^{y^T}, \boldsymbol{\psi}^{x^T}, \boldsymbol{\rho}^{y^T}, \boldsymbol{\rho}^{x^T})^T$ be the set of parameters which governs the association structure in the pairwise models. For ease of exposition, we only consider the case with binary variable X_{ij} for the covariate process. With a continuous X_{ij}, modifications in the exposition are immediate by changing the probability mass function to the probability density function and replacing the corresponding summation with an integral. We put $y_{ij} = (y^{obs}_{ij}, y^{mis}_{ij})$, where either y^{obs}_{ij} and y^{mis}_{ij} can be null, depending on whether or not y_{ij} is observed. Similarly, write $x_{ij} = (x^{obs}_{ij}, x^{mis}_{ij})$.

First, we temporarily assume an independence structure among the response components, and denote

$$\begin{aligned}\mathcal{L}_{C1,i}(\boldsymbol{\gamma}) \;=\; & \prod_{j=1}^{m}\Big\{\sum_{y^{mis}_{ij}=0}^{1}\sum_{x^{mis}_{ij}=0}^{1} P(Y_{ij}=y_{ij}|X_{ij}, Z_{ij})P(X_{ij}=x_{ij}|Z_{ij}) \\ & \times P(R^y_{ij}=r^y_{ij}, R^x_{ij}=r^x_{ij}|Y_{ij}, X_{ij}, Z_{ij})\Big\}\end{aligned}$$

as the observed likelihood contributed by subject i, where the probability mass (or density) functions are determined by (3.2), (3.6), and (3.10). For $j < k$, let

$$\begin{aligned}\mathcal{L}_{C2,i}(\boldsymbol{\gamma}, \boldsymbol{\delta}) \;=\; & \prod_{j<k} \Big\{ \sum_{y_{ij}^{mis}, y_{ik}^{mis}} \sum_{x_{ij}^{mis}, x_{ik}^{mis}} P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik} | X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) \\ & \times P(X_{ij} = x_{ij}, X_{ik} = x_{ik} | Z_{ij}, Z_{ik}) \\ & \times P(R_{ij}^{y} = r_{ij}^{y}, R_{ij}^{x} = r_{ij}^{x}, R_{ik}^{y} = r_{ik}^{y}, R_{ik}^{x} = r_{ik}^{x} | Y_{ij}, Y_{ik}, X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) \Big\}\end{aligned}$$

be the observed pairwise likelihood contributed from subject i, where the probability mass (or density) functions are determined by (3.1), (3.4), and (3.9). Then the marginal likelihood and pairwise likelihood are respectively given by

$$\mathcal{L}_{C1}(\boldsymbol{\gamma}) = \prod_{i=1}^{n} \mathcal{L}_{C1,i}(\boldsymbol{\gamma}), \tag{3.11}$$

and

$$\mathcal{L}_{C2}(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \prod_{i=1}^{n} \mathcal{L}_{C2,i}(\boldsymbol{\gamma}, \boldsymbol{\delta}). \tag{3.12}$$

3.3.2 Inference Procedures

Let $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \boldsymbol{\delta}^T)^T$, $\mathbf{S}_{1i,\gamma} = \partial \log \mathcal{L}_{C1,i}(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}$, $\mathbf{S}_{2i,\gamma} = \partial \log \mathcal{L}_{C2,i}(\boldsymbol{\gamma}, \boldsymbol{\delta})/\partial \boldsymbol{\gamma}$, and $\mathbf{S}_{2i,\boldsymbol{\delta}} = \partial \log \mathcal{L}_{C2,i}(\boldsymbol{\gamma}, \boldsymbol{\delta})/\partial \boldsymbol{\delta}$. Define $\mathbf{H}_i = (\mathbf{S}_{1i,\gamma}^T, \mathbf{S}_{2i,\boldsymbol{\delta}}^T)^T$, and $\mathbf{S}_{2i,\boldsymbol{\theta}} = (\mathbf{S}_{2i,\gamma}^T, \mathbf{S}_{2i,\boldsymbol{\delta}}^T)^T$. We employ two approaches for the estimation of $\boldsymbol{\theta}$: the pairwise likelihood (PL) approach and the two-stage (TS) estimation.

Pairwise Likelihoods

Estimation of $\boldsymbol{\theta}$ can be carried out using the Newton-Raphson algorithm. Let $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\gamma}^{T(t)}, \boldsymbol{\delta}^{T(t)})^T$ denote the estimate at the tth iteration. We update the estimates of $\boldsymbol{\theta}$ by the iterative equation

$$\begin{pmatrix} \boldsymbol{\gamma}^{(t+1)} \\ \boldsymbol{\delta}^{(t+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma}^{(t)} \\ \boldsymbol{\delta}^{(t)} \end{pmatrix} \;-\; \Big\{ \sum_{i=1}^{n} \mathbf{D}_i(\boldsymbol{\theta}^{(t)}) \Big\}^{-1} \cdot \Big\{ \sum_{i=1}^{n} \mathbf{S}_{2i,\boldsymbol{\theta}}(\boldsymbol{\theta}^{(t)}) \Big\}, \qquad t = 0, 1, \ldots$$

until convergence, where $\mathbf{D}_i = \partial \mathbf{S}_{2i,\boldsymbol{\theta}} / \partial \boldsymbol{\theta}^T$. Let $\hat{\boldsymbol{\theta}}_{PL} = (\hat{\boldsymbol{\gamma}}_{PL}^T, \hat{\boldsymbol{\delta}}_{PL}^T)^T$ denote the convergence value. Using estimating function theory, it can be shown that under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta})$ has an asymptotic normal distribution with mean $\mathbf{0}$ and covariance matrix $\{E(\mathbf{D}_i)\}^{-1} E\{\mathbf{S}_{2i,\boldsymbol{\theta}} \mathbf{S}_{2i,\boldsymbol{\theta}}^T\} \{E(\mathbf{D}_i)\}^{-1T}$.

Two-Stage Algorithm

For the ease of computation, we describe a two-stage estimation algorithm. In the first stage, we estimate the marginal parameter $\boldsymbol{\gamma}$ based on $\mathbf{S}_{1i,\gamma}$ using the iteration equation

$$\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)} - \Big\{ \sum_{i=1}^{n} \partial \mathbf{S}_{1i,\gamma}(\boldsymbol{\gamma}^{(t)}) / \partial \boldsymbol{\gamma}^T \Big\}^{-1} \cdot \Big\{ \sum_{i=1}^{n} \mathbf{S}_{1i,\gamma}(\boldsymbol{\gamma}^{(t)}) \Big\}, \quad t = 1, 2, \ldots$$

where $\boldsymbol{\gamma}^{(t)}$ represents the estimate of $\boldsymbol{\gamma}$ at the tth iteration. Let $\hat{\boldsymbol{\gamma}}_{TS}$ denote the estimate of $\boldsymbol{\gamma}$ at convergence.

In the second stage, we use $\mathbf{S}_{2i,\boldsymbol{\delta}}$ to estimate the association parameter $\boldsymbol{\delta}$ by fixing $\boldsymbol{\gamma}$ to be $\hat{\boldsymbol{\gamma}}_{TS}$. Specifically, we update the estimate of $\boldsymbol{\delta}$ by the iteration equation:

$$\boldsymbol{\delta}^{(t+1)} = \boldsymbol{\delta}^{(t)} - \Big\{ \sum_{i=1}^{n} \partial \mathbf{S}_{2i,\boldsymbol{\delta}}(\hat{\boldsymbol{\gamma}}_{TS}, \boldsymbol{\delta}^{(t)}) / \partial \boldsymbol{\delta}^T \Big\}^{-1} \cdot \Big\{ \sum_{i=1}^{n} \mathbf{S}_{2i,\boldsymbol{\delta}}(\hat{\boldsymbol{\gamma}}_{TS}, \boldsymbol{\delta}^{(t)}) \Big\}, \quad t = 1, 2, \ldots,$$

where $\boldsymbol{\delta}^{(t)}$ represents the estimate of $\boldsymbol{\delta}$ at the tth iteration. Let $\hat{\boldsymbol{\delta}}_{TS}$ denote the estimate of $\boldsymbol{\delta}$ at convergence, and let $\hat{\boldsymbol{\theta}}_{TS} = (\hat{\boldsymbol{\gamma}}_{TS}^T, \hat{\boldsymbol{\delta}}_{TS}^T)^T$. Under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}}_{TS} - \boldsymbol{\theta})$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix $\{E(\mathbf{D}_i^*)\}^{-1} E\{\mathbf{H}_i \mathbf{H}_i^T\} \big[\{E(\mathbf{D}_i^*)\}^{-1}\big]^T$, where

$$\mathbf{D}_i^* = \begin{pmatrix} \partial \mathbf{S}_{1i,\gamma} / \partial \boldsymbol{\gamma}^T & \mathbf{0} \\ \partial \mathbf{S}_{2i,\boldsymbol{\delta}} / \partial \boldsymbol{\gamma}^T & \partial \mathbf{S}_{2i,\boldsymbol{\delta}} / \partial \boldsymbol{\delta}^T \end{pmatrix}.$$

The proof is sketched in the Appendix.

3.4 The NPHS Data Sample

We apply the proposed methods to analyze the NPHS data of 1394 males who were assessed for 6 cycles. At Cycle 1, the individuals' age ranged between 50 and 70. At Cycle 6 all the subjects were under age 80. All the deceased subjects were excluded from the analysis. The response of interest is the indicator of normal Health Utilities Index (HUI) Mark versus abnormal Health Utilities Index Mark measured at each cycle, where 0.89 was taken as a threshold value. Meanwhile, covariate measurements describing participants' provincial level of household income (INC) were also taken. The income covariate was obtained by classifying the provincial level income (ranging from 1-10) as high or low, where 5 is a cutoff point. One objective of the study was to investigate how an individual's health status was associated with his/her income, and whether or not there was a temporal effect on health. Let the binary response variable Y_{ij} equal to 1 if the ith individual has HUI score higher than 0.89 at time j, and 0 otherwise; let X_{ij} equal to 1 if the ith individual has INC higher than 5 at time j, and 0 otherwise.

In the data set, only 43.2% of the individuals have complete observations for both response and covariate in all the 6 cycles. The response missingness proportions for the 6 cycles are 5.3%, 8.9%, 11.9%, 16.8%, 22.3%, and 25.6%, respectively, while the INC covariate missingness proportions are 8.7%, 13.2%, 17.1%, 24.0%, 29.0% and 33.4%, respectively. Various types of missingness patterns are present. A sample of missingness patterns is displayed in Table 5.1.

Table 3.1: Missing data patterns for the HUI and INC variables in the NPHS data (%)

Percentage	HUI						INC					
	1	2	3	4	5	6	1	2	3	4	5	6
43.2%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4.2%	✓	×	×	×	×	×	✓	×	×	×	×	×
...												
2%	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓
1%	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	×	✓
1%	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

✓ Observed; × Missing

We assume that response and covariate processes followed marginal structures

$$\eta_{ij}^{y} = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij}, \tag{3.13}$$

and

$$\eta_{ij}^{x} = \alpha_0 + \alpha_1 Z_{ij}, \tag{3.14}$$

respectively, where $Z_{ij} = j$ is set to be $-2, -1, 0, 1, 2, 3$, corresponding to Cycle 1 to Cycle 6.

The marginal structures of missing data processes were specified as

$$\eta_{ij}^{Ry} = \lambda_0^y + \lambda_1^y Y_{ij} + \lambda_2^y X_{ij} + \lambda_3^y R_{ij}^x + \lambda_4^y Z_{ij}, \tag{3.15}$$

and

$$\eta_{ij}^{Rx} = \lambda_0^x + \lambda_1^x Y_{ij} + \lambda_2^x X_{ij} + \lambda_3^x Z_{ij}. \tag{3.16}$$

To complete pairwise modeling, we used an AR(1) correlation structure for paired variables at times j and k for the response, covariate and missing data processes. Thus, for the models described in Section 3.2, we had the association parameters ψ^y, ψ^x, ρ^y and ρ^x for response, covariate and missing models, respectively.

We analyzed the data using the PL and TS methods. As a comparison, we employed a naive approach that is often used by analysts to handle data with missing observations. That is, we applied the generalized estimating equations (GEE) method to the complete data only, and denoted this method by NGEE. The correlation structure for the NGEE method was set to be unstructured. Tables 3.2 and 3.3 record the analysis results. For the response model in Table 3.2, the PL and TS approaches suggest that income has a significant positive effect on health index. People are more likely to have a better health status if they have higher income. There is no evidence of temporal effects on health status. The analysis results suggest a positive pairwise correlation among outcome measurements. The naive GEE approach indicates the same nature of findings.

Table 3.2: Analysis of the NPHS data using the pairwise likelihood (PL), two-stage estimation (TS) and naive GEE (NGEE) methods: response models

		PL			TS			NGEE		
Parameter		Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	(β_0)	0.351	0.039	< 0.001	0.328	0.037	< 0.001	0.405	0.032	< 0.001
INC	(β_1)	0.355	0.041	< 0.001	0.410	0.057	< 0.001	0.238	0.037	< 0.001
Cycle	(β_2)	-0.014	0.010	0.148	-0.012	0.011	0.282	-0.016	0.009	0.086
Association	(ψ^y)	1.873	0.065	< 0.001	1.811	0.073	< 0.001	-	-	-

For the covariate model of household income in Table 3.3, the PL and TS methods indicate different estimate results. The PL method suggests a negative temporal effect on the income, whereas the TS approach does not find a significant temporal effect on the income. For the missing data processes, although the PL and TS approaches produce estimates with different magnitudes, they suggest similar nature of the estimates.

Table 3.3: Analysis of the NPHS data using the pairwise likelihood (PL) and two-stage estimation (TS) approaches: results for parameters associated with the covariate and missing data processes

		PL			TS		
Parameter		Estimate	SE	p-value	Estimate	SE	p-value
Covariate (INC) Model							
Intercept	(α_0)	-0.023	0.044	0.599	0.291	0.051	< 0.001
Cycle	(α_1)	-0.122	0.011	< 0.001	-0.020	0.029	0.481
Association	(ψ^x)	2.626	0.087	< 0.001	2.094	0.173	< 0.001
Response Missingness Model							
Intercept	(λ_0^y)	-0.318	0.108	0.003	-0.131	0.160	0.411
HUI	(λ_1^y)	0.110	0.158	0.486	-0.277	0.169	0.101
INC	(λ_2^y)	0.075	0.085	0.377	0.055	0.098	0.579
Cycle	(λ_3^y)	-0.071	0.016	< 0.001	-0.072	0.016	< 0.001
R_{ij}^x	(λ_4^y)	2.228	0.061	< 0.001	2.371	0.075	< 0.001
Association	(ρ^y)	1.626	0.140	< 0.001	1.587	0.109	< 0.001
Covariate Missingness Model							
Intercept	(λ_0^x)	0.755	0.085	< 0.001	2.609	2.262	0.249
HUI	(λ_1^x)	-0.014	0.071	0.838	0.085	0.112	0.447
INC	(λ_2^x)	0.439	0.208	0.035	-2.102	2.355	0.372
Cycle	(λ_3^x)	-0.165	0.011	< 0.001	-0.243	0.022	< 0.001
Association	(ρ^x)	1.977	0.086	< 0.001	2.379	0.095	< 0.001

3.5 Empirical Studies

3.5.1 Performance of the Proposed Methods

In this section, we assess the empirical performance of the proposed methods through simulation studies. Five hundred simulations are run for the parameter configuration considered. We take a setting with $m = 3$ and $n = 500$, and simulate longitudinal binary responses from the joint distribution $P(Y_{i1} = 1, Y_{i2} = 1, Y_{i3} = 1 \mid X_i) = \Phi_3((\eta_{i1}^y, \eta_{i2}^y, \eta_{i3}^y)^T; \Sigma)$, where Φ_3 is the cumulative distribution function for a trivariate normal distribution that is defined similarly to Φ_2 in Section 3.2, and Σ is a correlation matrix with an exchangeable correlation coefficient ψ^y:

$$\Sigma = \begin{pmatrix} 1 & \psi^y & \psi^y \\ \psi^y & 1 & \psi^y \\ \psi^y & \psi^y & 1 \end{pmatrix}.$$

The regression model linking η_{ij}^y with covariate is specified as

$$\eta_{ij}^y = \beta_0 + \beta_1 X_{ij},$$

where we set $\beta_0 = -0.5$, $\beta_1 = 1$ and $\psi^y = 0.9$.

Analogously, missingness-prone binary covariates X_{ij} are generated from $P(X_{i1} = 1, X_{i2} = 1, X_{i3} = 1) = \Phi_3((\eta_{i1}^x, \eta_{i2}^x, \eta_{i3}^x)^T; \Sigma^x)$, where we set

$$\eta_{ij}^x = \alpha_0,$$

and Σ^x takes the same form as Σ, except that ψ^y is replaced by ψ^x. We take $\alpha_0 = 0.25$ and $\psi^x = 0.5$.

The response missingness process is generated similarly using $P(R_{i1}^y = 1, R_{i2}^y = 1, R_{i3}^y = 1 \mid Y_i, X_i, R_i^x) = \Phi_3((\eta_{i1}^{Ry}, \eta_{i2}^{Ry}, \eta_{i3}^{Ry})^T; \Sigma^{Ry})$, where we specify

$$\eta_{ij}^{Ry} = \lambda_0^y + \lambda_1^y Y_{ij} + \lambda_2^y R_{ij}^x,$$

and Σ^{Ry} takes the same form as Σ except that ψ^y is replaced by ρ^y. For the covariate missingness process, we generate R_{ij}^x using the distribution $P(R_{i1}^x = 1, R_{i2}^x = 1, R_{i3}^x = 1 \mid Y_i, X_i) = \Phi_3((\eta_{i1}^{Rx}, \eta_{i2}^{Rx}, \eta_{i3}^{Rx})^T; \Sigma^{Rx})$, where the marginal regression model is

$$\eta_{ij}^{Rx} = \lambda_0^x + \lambda_1^x Y_{ij},$$

and Σ^{Rx} takes the same form as Σ except that ψ^y is replaced by ρ^x. The true values for the regression parameters in the missing-data processes are set to be $\lambda_0^y = \lambda_0^x = -0.5$, $\lambda_1^y = \lambda_1^x = 1.5$, $\lambda_2^y = -0.5$ and $\rho^y = \rho^x = 0.5$.

We assess the performance of the PL and the TS approaches in contrast to the naive method NGEE, described in Section 3.4. In the NGEE approach, all incomplete observations are ignored and only the complete data are used for the estimation. We report the results in Table 3.4, where "bias" represents the percent relative bias, "ASE" and "ESE" are the average of model-based and empirical standard errors, respectively, and CP% represents the empirical coverage probability for the 95% confidence intervals. The results show that our PL and TS approaches yield small biases and satisfactory coverage probabilities for both the mean and the association parameters. ASE and ESE agree reasonably well for the PL and the TS methods, suggesting the consistency of variance estimates. The NGEE method, on the other hand, yields remarkably biased results.

Table 3.4: Simulation results for longitudinal binary data with missingness in both response and covariate variables

		Response			Covariate		Response Missingness				Covariate Missingness		
Method[†]		β_0	β_1	ψ^y	α_0	ψ^x	λ_0^y	λ_1^y	λ_2^x	ρ^y	λ_0^x	λ_1^x	ρ^x
PL	Bias%[*]	-1.026	0.331	2.682	1.265	-0.104	0.063	0.914	1.324	-1.043	0.766	0.518	-1.437
	ASE[**]	0.095	0.155	0.484	0.064	0.163	0.068	0.212	0.135	0.138	0.096	0.118	0.150
	ESE[***]	0.095	0.155	0.467	0.065	0.160	0.071	0.221	0.139	0.141	0.091	0.117	0.147
	CP%	95.0	95.8	93.6	94.0	94.4	93.4	93.6	93.8	95.4	95.4	93.4	95.0
TS	Bias%	0.328	-0.326	0.072	2.535	1.083	0.010	4.110	2.663	-1.338	-1.608	-0.657	-0.778
	ASE	0.118	0.245	0.575	0.084	0.164	0.071	0.422	0.193	0.142	0.160	0.182	0.149
	ESE	0.114	0.228	0.624	0.081	0.162	0.073	0.364	0.185	0.152	0.139	0.164	0.150
	CP%	96.5	95.9	95.7	94.5	95.1	93.9	95.3	96.9	96.3	97.3	95.5	95.5
NGEE	Bias%	-299.485	-20.772	-	-	-	-	-	-	-	-	-	-
	ASE	0.146	0.187	-	-	-	-	-	-	-	-	-	-
	ESE	0.157	0.190	-	-	-	-	-	-	-	-	-	-
	CP%	0	80.2	-	-	-	-	-	-	-	-	-	-

[†] "TS" represents the two-stage inference procedures, "PL" denotes the pairwise likelihood, and "NGEE" is the naive GEE method.

[*] Relative bias is defined as $100 \times (\hat{\beta} - \beta_{true})/\beta_{true}$.

[**] ASE is the average standard error for 500 simulations, defined as $500^{-1} \sum_{i=1}^{500} \sqrt{\widehat{Var}(\hat{\beta}^i)}$, where $\sqrt{\widehat{Var}(\hat{\beta}^i)}$ is the standard error estimates in the ith simulation.

[***] ESE is the empirical standard error for 500 simulations, defined as $\left\{(500-1)^{-1} \sum_{i=1}^{500} (\hat{\beta}^i - \bar{\hat{\beta}})^2\right\}^{1/2}$, where $\hat{\beta}^i$ is the ith simulation result, and $\bar{\hat{\beta}} = 500^{-1} \sum_{i=1}^{500} \hat{\beta}^i$.

3.5.2 Sensitivity Analysis

Now we evaluate the sensitivity of our methods. In particular, we consider the case that the marginal structures for missing data processes are misspecified, while the association structures of the missing models are correctly specified. The response and covariate processes are retained to be correctly specified, as described in Section 3.5.1. To be specific, we generate the missing data indicators from the model with the true marginal structures $\eta_{ij}^{Ry} = \lambda_0^y + \lambda_1^y y_{ij} + \lambda_2^y R_{ij}^x + \kappa x_{ij}$, and $\eta_{ij}^{Rx} = \lambda_0^x + \lambda_1^x y_{ij} + \kappa x_{ij}$, but we fit data with models described in Section 3.5.1 with $\eta_{ij}^{Ry} = \lambda_0^y + \lambda_1^y y_{ij} + \lambda_2^y R_{ij}^x$ and $\eta_{ij}^{Rx} = \lambda_0^x + \lambda_1^x y_{ij}$.

Under model misspecification, the resultant estimator for the parameter $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}^*$, would converge in probability to a limit $\boldsymbol{\theta}^*$, say. This limit $\boldsymbol{\theta}^*$ is, under certain regularity conditions, the solution of

$$E_{(Y,X,R^y,R^x|Z)}\left\{\frac{\partial \log \mathcal{L}^*(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*}\right\} = \mathbf{0}, \tag{3.17}$$

where the expectation is taken under the true joint distribution for the (Y, X, R^y, R^x) variables given Z, and $\mathcal{L}^*(\boldsymbol{\theta}^*)$ is the marginal or pairwise likelihood function formulated from the misspecified model (Yi and Reid, 2010).

In our analysis here, (3.17) does not have a closed form solution. We use numerical approximations to display the asymptotic relative biases, defined as $(100 \times (\beta^* - \beta)/\beta)$, against varying degrees of κ. The results are shown in Figure 3.1. It is seen that when a specific term in missing data process is ignored, the bias may occur. As expected, the stronger influence of the omitting term on the missing process model, the larger the relative bias. While the PL and TS methods show similar trends in bias, the PL method tends to produce smaller bias than the TS method.

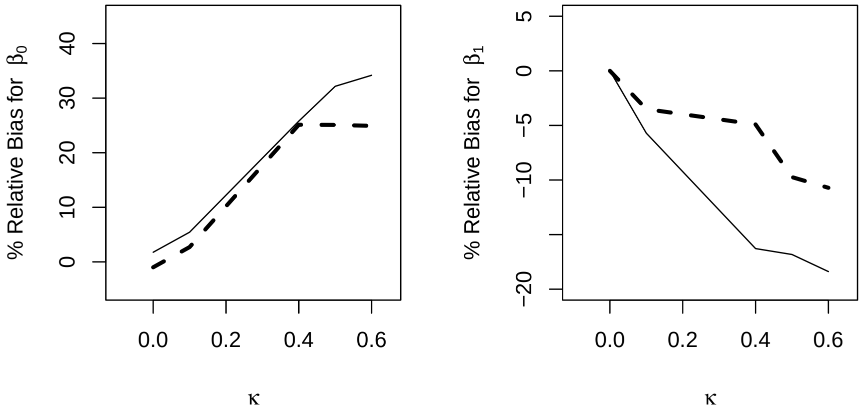

Figure 3.1: *Asymptotic relative bias for regression coefficients β_0 and β_1 when the marginal structures in missing data process are misspecified. The model for estimation is specified in Section 3.5.1, while the true model is $\eta_{ij}^{Ry} = \lambda_0^y + \lambda_1^y y_{ij} + \lambda_2^y R_{ij}^x + \kappa x_{ij}$, $\eta_{ij}^{Rx} = \lambda_0^x + \lambda_1^x y_{ij} + \kappa x_{ij}$. PL method: − − −; TS method: ———.*

3.5.3 Efficiency Assessment

We are also interested in assessing the efficiency of the estimators obtained from the PL and the TS methods. This assessment is carried out as opposed to the maximum likelihood (ML) method. We consider the model setting in Section 3.5.1, but set a common exchangeable correlation coefficient $\rho = \psi^y = \psi^x = \rho^y = \rho^x$. To highlight comparison on the $\boldsymbol{\beta}$ parameter, we assume all other nuisance parameters are known for simplicity.

Let $\text{avar}(\hat{\beta}_1^{PL})$ denote the asymptotic variance for the estimator of β_1 obtained from the PL method. It is calculated by similar approaches in Section 3.3, with all nuisance parameters set to be fixed. Analogously, we obtain the asymptotic variance $\text{avar}(\hat{\beta}_1^{TS})$ for the estimator of β_1 obtained from the TS approach. Let $\text{avar}(\hat{\beta}_1^{ML})$ denote the asymptotic variance of the estimator for β_1 obtained from the maximum likelihood method, i.e., obtained from the diagonal element of $\left[E\{\mathbf{S}_i^F(\boldsymbol{\beta})\mathbf{S}_i^F(\boldsymbol{\beta})^T\}\right]^{-1}$ evaluated at the maximum likelihood estimate, where $\mathbf{S}_i^F(\boldsymbol{\beta})$ is the score function of $\boldsymbol{\beta}$ from the fully specified likelihood function. Then, the relative efficiency of the PL estimator with respect to the ML estimator is given by avar($\hat{\beta}_1^{ML}$)/avar($\hat{\beta}_1^{PL}$), and the relative efficiency of the TS estimator against the ML estimator is given by avar($\hat{\beta}_1^{ML}$)/avar($\hat{\beta}_1^{TS}$).

Figure 3.2 shows that the PL and TS methods incur different degrees of efficiency loss. When the measurements are uncorrelated (i.e. $\rho = 0$), the PL, TS and ML methods produce the same asymptotic variance, as shown by the peak of the curves. As the correlation becomes stronger, the efficiency loss increases. It is seen that the efficiency loss in using the PL method is less striking than that incurred by using the TS method.

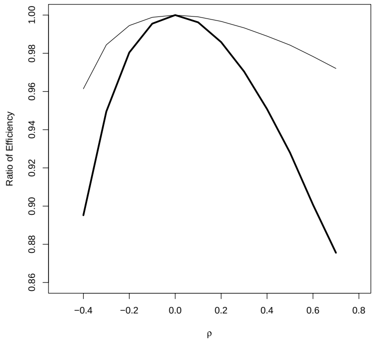

Figure 3.2: *Relative efficiency of estimators for β_1. The TS method:* ━━━*; the PL method:* ———*.*

Appendix

To show the asymptotic distribution of our two-stage approach, we proceed with two steps. First, we show $E(\mathbf{H}_i) = 0$, and then we derive the asymptotic distribution.

The proof of $E(\mathbf{H}_i) = 0$

To show $E(\mathbf{H}_i) = 0$, it suffices to show that $E\left\{\sum_{i=1}^n \partial \log \mathcal{L}_{C1,i}(\boldsymbol{\gamma})/\partial \boldsymbol{\beta}\right\} = \mathbf{0}$. The proof for other elements in $\mathbf{H}_i$ follows analogously. Let

$$K_{1,ij} = f(Y_{ij}|X_{ij}, Z_{ij}) f(X_{ij}|Z_{ij}) P(R_{ij}^y = 1, R_{ij}^x = 1|Y_{ij}, X_{ij}, Z_{ij}),$$

$$K_{2,ij} = \sum_{Y_{ij}=0}^{1} \Big\{ f(Y_{ij}|X_{ij}, Z_{ij}) f(X_{ij}|Z_{ij}) P(R_{ij}^y = 0, R_{ij}^x = 1|Y_{ij}, X_{ij}, Z_{ij}) \Big\},$$

$$K_{3,ij} = \sum_{X_{ij}=0}^{1} \Big\{ f(Y_{ij}|X_{ij}, Z_{ij}) f(X_{ij}|Z_{ij}) P(R_{ij}^y = 1, R_{ij}^x = 0|Y_{ij}, X_{ij}, Z_{ij}) \Big\},$$

and

$$K_{4,ij} = \sum_{Y_{ij}=0}^{1} \sum_{X_{ij}=0}^{1} \Big\{ f(Y_{ij}|X_{ij}, Z_{ij}) f(X_{ij}|Z_{ij}) P(R_{ij}^y = 0, R_{ij}^x = 0|Y_{ij}, X_{ij}, Z_{ij}) \Big\},$$

then we write

$$\begin{aligned} \log \mathcal{L}_{C1}(\boldsymbol{\gamma}) \quad = \quad & \sum_{i=1}^{n} \sum_{j=1}^{m} \Bigg\{ R_{ij}^y R_{ij}^x \log K_{1,ij} + (1 - R_{ij}^y) R_{ij}^x \log K_{2,ij} \\ & + R_{ij}^y (1 - R_{ij}^x) \log K_{3,ij} + (1 - R_{ij}^y)(1 - R_{ij}^x) \log K_{4,ij} \Bigg\}. \end{aligned} \tag{3.18}$$

By the distinctness of the parameters in different processes, we have

$$\begin{aligned} & E_{(Y_i, X_i, R_i^y, R_i^x | Z_i)} \left(R_{ij}^y R_{ij}^x \frac{\partial \text{log} K_{1,ij}}{\partial \boldsymbol{\beta}} \right) \\ = \quad & E_{(Y_i, X_i | Z_i)} \left\{ E_{(R_i^y, R_i^x | Y_i, X_i, Z_i)} \left(R_{ij}^y R_{ij}^x \frac{\partial \text{log} K_{1,ij}}{\partial \boldsymbol{\beta}} \right) \right\} \\ = \quad & E_{(Y_i, X_i | Z_i)} \left\{ P(R_{ij}^y = 1, R_{ij}^x = 1 | Y_{ij}, X_{ij}, Z_{ij}) \frac{\partial \text{log} f(Y_{ij} | X_{ij}, Z_{ij})}{\partial \boldsymbol{\beta}} \right\}. \end{aligned}$$

Note that

$$K_{2,ij} = f(X_{ij}|Z_{ij}) \cdot E_{(Y_i|X_i, Z_i)} \left\{ P(R_{ij}^y = 0, R_{ij}^x = 1|Y_{ij}, X_{ij}, Z_{ij}) \right\},$$

then for the second term in (3.18), we have

$$\begin{aligned}
& E_{(Y_i,X_i,R_i^y,R_i^x|Z_i)}\Big\{(1-R_{ij}^y)R_{ij}^x\frac{\partial \text{log} K_{2,ij}}{\partial \boldsymbol{\beta}}\Big\} \\
= \;& E_{(X_i|Z_i)}\left[E_{(Y_i|X_i,Z_i)}\left\{E_{(R_i^y,R_i^x|Y_i,X_i,Z_i)}\Big((1-R_{ij}^y)R_{ij}^x\frac{\partial \text{log} K_{2,ij}}{\partial \boldsymbol{\beta}}\Big)\right\}\right] \\
= \;& E_{(X_i|Z_i)}\left[E_{(Y_i|X_i,Z_i)}\left\{P(R_{ij}^y=0,R_{ij}^x=1|Y_{ij},X_{ij},Z_{ij})\Big(\frac{\partial \text{log} K_{2,ij}}{\partial \boldsymbol{\beta}}\Big)\right\}\right] \\
= \;& E_{(X_i|Z_i)}\left[\left\{E_{(Y_i|X_i,Z_i)}\{P(R_{ij}^y=0,R_{ij}^x=1|Y_{ij},X_{ij},Z_{ij})\}\right\}\times\frac{1}{K_{2,ij}}\times\frac{\partial K_{2,ij}}{\partial \boldsymbol{\beta}}\right] \\
= \;& E_{(X_i|Z_i)}\left\{\frac{1}{f(X_{ij}|Z_{ij})}\cdot\frac{\partial K_{2,ij}}{\partial \boldsymbol{\beta}}\right\}. \qquad (3.19)
\end{aligned}$$

By the distinctness of the parameters in different processes, we have

$$\begin{aligned}
\frac{\partial K_{2,ij}}{\partial \boldsymbol{\beta}} &= \sum_{Y_{ij}=0}^{1}\left\{\frac{\partial f(Y_{ij}|X_{ij},Z_{ij})}{\partial \boldsymbol{\beta}}f(X_{ij}|Z_{ij})P(R_{ij}^y=0,R_{ij}^x=1|Y_{ij},X_{ij},Z_{ij})\right\} \\
&= \sum_{Y_{ij}=0}^{1}\left\{\frac{\partial \log f(Y_{ij}|X_{ij},Z_{ij})}{\partial \boldsymbol{\beta}}f(Y_{ij}|X_{ij},Z_{ij})f(X_{ij}|Z_{ij})P(R_{ij}^y=0,R_{ij}^x=1|Y_{ij},X_{ij},Z_{ij})\right\} \\
&= f(X_{ij}|Z_{ij})\cdot E_{(Y_i|X_i,Z_i)}\left\{\frac{\partial \log f(Y_{ij}|X_{ij},Z_{ij})}{\partial \boldsymbol{\beta}}P(R_{ij}^y=0,R_{ij}^x=1|Y_{ij},X_{ij},Z_{ij})\right\},
\end{aligned}$$

therefore, (3.19) becomes

$$E_{(Y_i,X_i|Z_i)}\Big\{P(R_{ij}^y=0,R_{ij}^x=1|Y_{ij},X_{ij},Z_{ij})\frac{\partial \log f(Y_{ij}|X_{ij},Z_{ij})}{\partial \boldsymbol{\beta}}\Big\}.$$

Analogously, for the third and fourth terms in (3.18), we obtain

$$E\Big(\frac{\partial \log K_{3,ij}}{\partial \boldsymbol{\beta}}\Big) = E_{(Y_i,X_i|Z_i)}\Big\{P(R_{ij}^y=1,R_{ij}^x=0|Y_{ij},X_{ij},Z_{ij})\frac{\partial \log f(Y_{ij}|X_{ij},Z_{ij})}{\partial \boldsymbol{\beta}}\Big\},$$

and

$$E\Big(\frac{\partial \log K_{4,ij}}{\partial \boldsymbol{\beta}}\Big) = E_{(Y_i,X_i|Z_i)}\Big\{P(R_{ij}^y=0,R_{ij}^x=0|Y_{ij},X_{ij},Z_{ij})\frac{\partial \log f(Y_{ij}|X_{ij},Z_{ij})}{\partial \boldsymbol{\beta}}\Big\},$$

where the expectation "E" is evaluated with respect to the conditional distribution of (Y_i, X_i, R_i^y, R_i^x) given Z_i.

Then combining these results leads to

$$E\left\{\sum_{i=1}^{n} \partial \log \mathcal{L}_{C1,i}(\boldsymbol{\gamma})/\partial\boldsymbol{\beta}\right\} = \sum_{i=1}^{n}\sum_{j=1}^{m} E_{(Y_i,X_i|Z_i)}\left\{\partial \log f(Y_{ij}|X_{ij}, Z_{ij})/\partial\boldsymbol{\beta}\right\} = \mathbf{0}.$$

Asymptotic Distribution

An alternative to obtain the estimator $\hat{\boldsymbol{\theta}}_{TS} = (\hat{\boldsymbol{\gamma}}_{TS}^T, \hat{\boldsymbol{\delta}}_{TS}^T)^T$ is to employ the joint iterative equation to update the estimate:

$$\begin{pmatrix} \boldsymbol{\gamma}^{(t+1)} \\ \boldsymbol{\delta}^{(t+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma}^{(t)} \\ \boldsymbol{\delta}^{(t)} \end{pmatrix} \quad - \quad \Big\{\sum_{i=1}^{n} \mathbf{D}_i^*(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)})\Big\}^{-1} \cdot \sum_{i=1}^{n} \Big\{\mathbf{H}_i(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\delta}^{(t)})\Big\}, \quad (3.20)$$

At each iteration, the update obtained from (3.20) may differ from that obtained from the two-stage algorithm. However, updated values from these two procedures converge to the same limit under mild regularity conditions (Prentice, 1988). When the algorithm in (3.20) reaches convergence, the $n^{-1}\sum_{i=1}^{n} \mathbf{H}_i(\hat{\boldsymbol{\theta}}_{TS}) = \mathbf{0}$ condition will be satisfied. Then the mean-value theorem gives

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{H}_i(\boldsymbol{\theta}) + \Big\{\frac{1}{n}\sum_{i=1}^{n} \mathbf{D}_i^*(\tilde{\boldsymbol{\theta}})\Big\}(\hat{\boldsymbol{\theta}}_{TS} - \boldsymbol{\theta}) = \mathbf{0}, \quad (3.21)$$

where $\tilde{\boldsymbol{\theta}}$ is a value "between" the true value $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_{TS}$.

Multiplying $\sqrt{n}$ and solving for $\sqrt{n}(\hat{\boldsymbol{\theta}}_{TS} - \boldsymbol{\theta})$ gives

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{TS} - \boldsymbol{\theta}) = -\Big\{\frac{1}{n}\sum_{i=1}^{n} \mathbf{D}_i^*(\tilde{\boldsymbol{\theta}})\Big\}^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{H}_i(\boldsymbol{\theta}). \quad (3.22)$$

Under regularity conditions, the property $E(\mathbf{H}_i) = 0$ ensures that $\hat{\boldsymbol{\theta}}_{TS} \to_p \boldsymbol{\theta}$. Because $\tilde{\boldsymbol{\theta}}$ lies between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_{TS}$, it will also be consistent to $\boldsymbol{\theta}$. Then the first term in (3.22) is consistent to $[E(\mathbf{D}_i^*)]^{-1}$ if matrix $E(\mathbf{D}_i^*)$ is nonsingular. On the other hand, the central limit theorem implies that the second term in (3.22) has the limiting distribution $N(\mathbf{0}, E\{\mathbf{H}_i\mathbf{H}_i^T\})$. Therefore, it follows from the Slutzky theorem that the asymptotic distribution for $\sqrt{n}(\hat{\boldsymbol{\theta}}_{TS} - \boldsymbol{\theta})$ is a normal distribution with mean $\mathbf{0}$ and covariance matrix $\{E(\mathbf{D}_i^*)\}^{-1}E\{\mathbf{H}_i\mathbf{H}_i^T\}\{E(\mathbf{D}_i^*)\}^{-1T}$.

Some Computation Details

Here we present some derivatives that are used in the implementation of our methods. Let

$$\begin{aligned} A_1 &= \sum_{y_{ij}^{mis}, y_{ik}^{mis}} \sum_{x_{ij}^{mis}, x_{ik}^{mis}} \Big\{ P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik} | X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) P(X_{ij} = x_{ij}, X_{ik} = x_{ik} | Z_{ij}, Z_{ik}) \\ & \quad \times P(R_{ij}^y = r_{ij}^y, R_{ij}^x = r_{ij}^x, R_{ik}^y = r_{ik}^y, R_{ik}^x = r_{ik}^x | Y_{ij}, Y_{ik}, X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) \Big\}, \end{aligned}$$

and

$$\begin{aligned} A_2 &= \log P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik} | X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) + \log P(X_{ij} = x_{ij}, X_{ik} = x_{ik} | Z_{ij}, Z_{ik}) \\ & \quad + \log P(R_{ij}^y = 1, R_{ij}^x = 1, R_{ik}^y = 1, R_{ik}^x = 1 | Y_{ij}, Y_{ik}, X_{ij}, X_{ik}, Z_{ij}, Z_{ik}), \end{aligned}$$

then $\log \mathcal{L}_{C2,i}(\boldsymbol{\theta})$ in (3.12) can be rewritten as

$$\log \mathcal{L}_{C2,i}(\boldsymbol{\theta}) = \sum_{j<k} I(r_{ij}^y + r_{ij}^x + r_{ik}^y + r_{ik}^x < 4) \log A_1 + I(r_{ij}^y + r_{ij}^x + r_{ik}^y + r_{ik}^x = 4) A_2,$$

leading to the score function

$$\frac{\partial \log \mathcal{L}_{C2,i}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{j<k} I(r_{ij}^y + r_{ij}^x + r_{ik}^y + r_{ik}^x < 4) \frac{1}{A_1} \frac{\partial A_1}{\partial \boldsymbol{\theta}} + I(r_{ij}^y + r_{ij}^x + r_{ik}^y + r_{ik}^x = 4) \frac{\partial A_2}{\partial \boldsymbol{\theta}},$$

where

$$\begin{aligned}
\frac{\partial A_1}{\partial \boldsymbol{\theta}} &= \sum_{y_{ij}^{mis}, y_{ik}^{mis}} \sum_{x_{ij}^{mis}, x_{ik}^{mis}} \Bigg[\frac{\partial}{\partial \boldsymbol{\theta}} \Big\{ P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik} | X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) \Big\} \\
&\quad \cdot P(X_{ij} = x_{ij}, X_{ik} = x_{ik} | Z_{ij}, Z_{ik}) \\
&\quad \cdot P(R_{ij}^y = r_{ij}^y, R_{ij}^x = r_{ij}^x, R_{ik}^y = r_{ik}^y, R_{ik}^x = r_{ik}^x | Y_{ij}, Y_{ik}, X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) \\
&\quad + P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik} | X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) \cdot \frac{\partial}{\partial \boldsymbol{\theta}} \Big\{ P(X_{ij} = x_{ij}, X_{ik} = x_{ik} | Z_{ij}, Z_{ik}) \Big\} \\
&\quad \cdot P(R_{ij}^y = r_{ij}^y, R_{ij}^x = r_{ij}^x, R_{ik}^y = r_{ik}^y, R_{ik}^x = r_{ik}^x | Y_{ij}, Y_{ik}, X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) \\
&\quad + P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik} | X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) \cdot P(X_{ij} = x_{ij}, X_{ik} = x_{ik} | Z_{ij}, Z_{ik}) \\
&\quad \cdot \frac{\partial}{\partial \boldsymbol{\theta}} \Big\{ P(R_{ij}^y = r_{ij}^y, R_{ij}^x = r_{ij}^x, R_{ik}^y = r_{ik}^y, R_{ik}^x = r_{ik}^x | Y_{ij}, Y_{ik}, X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) \Big\} \Bigg],
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial A_2}{\partial \boldsymbol{\theta}} &= \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \Big\{ P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik} | X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) \Big\}}{P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik} | X_{ij}, X_{ik}, Z_{ij}, Z_{ik})} + \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \Big\{ P(X_{ij} = x_{ij}, X_{ik} = x_{ik} | Z_{ij}, Z_{ik}) \Big\}}{P(X_{ij} = x_{ij}, X_{ik} = x_{ik} | Z_{ij}, Z_{ik})} \\
&\quad + \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \Big\{ P(R_{ij}^y = r_{ij}^y, R_{ij}^x = r_{ij}^x, R_{ik}^y = r_{ik}^y, R_{ik}^x = r_{ik}^x | Y_{ij}, Y_{ik}, X_{ij}, X_{ik}, Z_{ij}, Z_{ik}) \Big\}}{P(R_{ij}^y = r_{ij}^y, R_{ij}^x = r_{ij}^x, R_{ik}^y = r_{ik}^y, R_{ik}^x = r_{ik}^x | Y_{ij}, Y_{ik}, X_{ij}, X_{ik}, Z_{ij}, Z_{ik})}.
\end{aligned}$$

Similarly, we can work out the second derivatives:

$$\begin{aligned}
\frac{\partial^2 \log \mathcal{L}_{C2,i}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} &= \sum_{j<k} I(r_{ij}^y + r_{ij}^x + r_{ik}^y + r_{ik}^x < 4) \Bigg[-\frac{1}{A_1^2} \frac{\partial A_1}{\partial \boldsymbol{\theta}} \Big\{ \frac{\partial A_1}{\partial \boldsymbol{\theta}} \Big\}^T + \frac{1}{A_1} \frac{\partial A_1^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Bigg] \\
&\quad + I(r_{ij}^y + r_{ij}^x + r_{ik}^y + r_{ik}^x = 4) \Bigg(\frac{\partial A_2^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Bigg).
\end{aligned}$$

Chapter 4

Simultaneous Methods of Variable Selection and Estimation for Longitudinal Data Arising in Clusters

4.1 Introduction

In longitudinal studies, datasets can involve a large number of covariates. However, not all of them are relevant to explain the response variable. Properly selecting variables to build a feasible model is important for valid inference.

Many studies on variable selection methods focus on the analysis of univariate data. The methods include the best subset selection (Akaike, 1973; Schwarz, 1978), stepwise selection (Yan and Su, 2009), and shrinkage methods (Frank and Friedman, 1993; Tibshirani, 1996, 2011). However, relatively limited work has been done for longitudinal data arising in clusters. Fan and Li (2001) propose a variable selection approach by imposing the smoothly clipped absolute deviation (SCAD) penalty on log likelihood for generalized linear models on independent data. Fan and Li (2004) discuss a variable selection method based on semiparametric model for longitudinal data. However, their methods ignore the correlation in longitudinal data. Ni et al. (2010) study the model selection methods for both covariates and semiparametric components under linear mixed effects models with a double penalty strategy. Bondell et al. (2010) and Ibrahim et al. (2010) discuss double penalty ideas for the selection of both covariates and random effects via the EM algorithm.

A challenge on handling longitudinal data, or even longitudinal data arising in clusters, comes from substantially increased modeling complexity and computational difficulty. With clusters present in longitudinal studies, the likelihood functions become cumbersome. Fieuws and Verbeke (2006) argue that for longitudinal clustered data under random effects models, computation will become difficult as the dimension of the random-effects vector is often high, even in the case of linear mixed models where the integrals may be calculated analytically. Thus, an obvious paradox for longitudinal data arising in clusters is that although likelihood methods are straightforward to be formulated with penalty functions accommodated for variable section, the complexity in modeling and the intensity in computing seriously prevent universal use of such methods.

It is desirable to develop methods that preserve advantages of existing methods and overcome their shortcomings. The purpose of this chapter is to describe a general variable selection approach based on the pairwise likelihood formulation (Lindsay, 1988; Arnold and Strauss, 1991; Cox and Reid, 2004; Lindsay et al., 2011) to handle longitudinal clustered data. Pairwise likelihood functions focus only on partial structures of data, and often enjoy transparent interpretation, modeling tractability and computational cheapness. Furthermore, as opposed to the full likelihood method, the pairwise likelihood formulation is robust in the sense that association structures higher than those used in the formulation are left unspecified. Two specific types of pairwise likelihood, *all-pairwise marginal likelihood* (APW) and *all-pairwise conditional likelihood* (APC), are introduced in this chapter. The SCAD penalty is used for variable section. We particularly form the development under random effects models.

A further relevant and interesting topic concerns the validity of model assumptions. When these assumptions are violated, estimation and selection results could be biased or incorrect. There are some studies in dealing with misspecified model selection issues (Lv and Liu, 2010). However, little work has been done under the penalized likelihood or penalized composite likelihood framework. In this chapter, we explore the asymptotic results obtained from misspecified models.

The rest of the chapter is organized as follows. Section 4.2 describes the generalized linear mixed models (GLMMs) formulation and notations. We then introduce the formulations of the composite likelihood methods. Section 4.3 presents the penalized composite likelihood and the implementation algorithm. This section also derives the asymptotic results for our penalized composite likelihood approach. Section 4.4 demonstrates the asymptotic results obtained from misspecified models. To evaluate the performance of the proposed methods, we conduct various empirical studies and display the results in Section 4.5. The application of our methods into a real data analysis is illustrated in Section 4.6, and concluding remarks are given in Section 4.7.

4.2 Model Setup

Suppose there are n clusters and J_i subjects within cluster i, $i = 1, 2, \ldots, n$. We assume that each subject is assessed at K specified time points. Let Y_{ijk} denote the response for subject j in cluster i at visit k, $k = 1, 2, \ldots, K$. Take $Y_{ij} = (Y_{ij1}, Y_{ij2}, \ldots, Y_{ijK})^T$, $j = 1, 2, \ldots, J_i$, and $Y_i = (Y_{i1}^T, Y_{i2}^T, \ldots, Y_{iJ_i}^T)^T$, $i = 1, 2, \ldots, n$. Let u_i denote a random effects vector corresponding to cluster i, $i = 1, 2, \ldots, n$. Let $X_{ijk} = (X_{ijk,1}, \ldots, X_{ijk,p})^T$ be the $p \times 1$ fixed effect covariate vector for subject j in cluster i at time k, $X_{ij} = (X_{ij1}^T, X_{ij2}^T, \ldots, X_{ijK}^T)^T$, and $X_i = (X_{i1}^T, X_{i2}^T, \ldots, X_{iJ_i}^T)^T$. Let $Z_{ijk} = (Z_{ijk,1}, \ldots, Z_{ijk,q})^T$ be the $q \times 1$ random effect covariate vector, $Z_{ij} = (Z_{ij1}^T, Z_{ij2}^T, \ldots, Z_{ijk}^T)^T$ and $Z_i = (Z_{i1}^T, Z_{i2}^T, \ldots, Z_{iJ_i}^T)^T$.

4.2.1 Generalized Linear Mixed Models

The usual generalized linear mixed models (GLMMs) consist of two steps of modeling (Laird and Ware, 1982; McCulloch, 1997). In the first step, we assume that conditional on random effects u_i, the Y_{ijk} $(j = 1, \ldots, J_i; k = 1, \ldots, K)$ are independent and have the probability (density) function given by

$$f(y_{ijk}|u_i) = \exp\Big[\{y_{ijk}\tau_{ijk} - b(\tau_{ijk})\}/a(\phi) + c(y_{ijk}; \phi)\Big], \tag{4.1}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are given functions, ϕ is a scale parameter, and τ_{ijk} is the canonical parameter. This leads to $\mathrm{E}(Y_{ijk}|u_i) = b'(\tau_{ijk})$, and $\mathrm{Var}(Y_{ijk}|u_i) = a(\phi)b''(\tau_{ijk})$.

The second step links the conditional mean of Y_{ijk} to the covariates with a regression model

$$h\{\mathrm{E}(Y_{ijk}|u_i)\} = X_{ijk}^T\boldsymbol{\beta} + Z_{ijk}^T u_i, \tag{4.2}$$

where h is a monotone link function, $\boldsymbol{\beta}$ is the vector of $p \times 1$ fixed effect coefficients, and the random effects vector u_i is assumed to follow a certain distribution, such as a multivariate normal distribution. Let $f(u_i; \alpha)$ denote the joint probability density function of u_i, where α is an associated parameter vector.

Different types of random effects models can be obtained by various choices of the Z_{ijk} vector or random effects vector u_i. For instance, (4.2) includes commonly used one-way (Fieuws and Verbeke, 2006), two-way (Sutradhar and Rao, 2003) or three-way (Bellio and Varin, 2005) random effects models:

$$h\{\mathrm{E}(Y_{ijk}|\nu_i)\} = X_{ijk}^T\boldsymbol{\beta} + \nu_i, \tag{4.3}$$

$$h\{\mathrm{E}(Y_{ijk}|\nu_i, \omega_j)\} = X_{ijk}^T\boldsymbol{\beta} + \nu_i + \omega_j, \tag{4.4}$$

or

$$h\{\mathrm{E}(Y_{ijk}|\nu_i, \omega_j, \tau_k)\} = X_{ijk}^T\boldsymbol{\beta} + \nu_i + \omega_j + \tau_k, \tag{4.5}$$

where ν_i, ω_j and τ_k are random effects which respectively facilitate cluster-level, subject-level and time-specific heterogeneity, and are assumed to be independent of each other.

Under the conditional independence assumption that the Y_{ijk} are independent given u_i and covariates, inference can, in principle, be carried out by maximizing the observed likelihood with unobservable random effects integrated out. For example, under model (4.5), the marginal likelihood is given by

$$\prod_{i=1}^{n}\int\left[\iint\Big\{\prod_{j=1}^{J_i}\prod_{k=1}^{K} f(y_{ijk}|\nu_i,\omega_j,\tau_k)\Big\}f(\omega_j)f(\tau_k)\,d\omega_j\,d\tau_k\right]f(\nu_i)d\nu_i. \tag{4.6}$$

Evaluation of this likelihood requires calculation of $n(K+1)+\sum_{i=1}^{n}J_i$ dimensional integrals. Several serious issues would arise here. The number of integrals involved in (4.6) rapidly grows with the number of random effects, creating increasing computational intensity, especially for the case that integrals are intractable. In addition, specifying appropriate distributions for random effects could be difficult, because random effects are not observable. Moreover, the conditional independence assumption for the Y_{ijk} given u_i can be inflexible to handle data with complex association.

To overcome these limitations of GLMMs, we now propose a wider class of models that generalize GLMMs: generalized linear mixed pairwise models (GLMPMs).

4.2.2 Generalized Linear Mixed Pairwise Models

Define $(j,k)<(j',k')$ if $j<j'$ or $j=j'$, $k<k'$. For any $(j,k)<(j',k')$, let $Y_{i;jk;j'k'}=(Y_{ijk},Y_{ij'k'})^T$. Generalized linear mixed pairwise models (GLMPMs) are specified by two steps. In the first step, unlike that GLMMs assume conditional independence among the Y_{ijk} given random effects u_i, GLMPMs assume conditional independence among the $Y_{i;jk;j'k'}$ pairs. To be specific, conditional on random effects, say $\tilde{u}_i$, pairs $Y_{i;jk;j'k'}$ are independent and have the probability (density) function belonging to the bivariate exponential family

$$f(y_{i;jk;j'k'}|\tilde{u}_i)=\exp\left[\tilde{\tau}_{i;jk;j'k'}^{T}y_{i;jk;j'k'}-\tilde{b}(\tilde{\tau}_{i;jk;j'k'})+\tilde{c}(y_{i;jk;j'k'})\right], \tag{4.7}$$

where $\tilde{b}(\cdot)$ and $\tilde{c}(\cdot)$ are known functions, $\tilde{\tau}_{i;jk;j'k'} = (\tilde{\tau}_{ijk}, \tilde{\tau}_{ij'k'})^T$ is a 2×1 vector of canonical parameters. Analogous to the property of GLMMs, it can be shown that

$$\mathrm{E}\Big(Y_{i;jk;j'k'}|\tilde{u}_i\Big) = \frac{\partial}{\partial\tilde{\tau}_{i;jk;j'k'}}\tilde{b}(\tilde{\tau}_{i;jk;j'k'}), \tag{4.8}$$

and

$$\mathrm{Var}\Big(Y_{i;jk;j'k'}|\tilde{u}_i\Big) \quad = \quad \begin{pmatrix} \mathrm{Var}(Y_{ijk}|\tilde{u}_i) & \mathrm{Cov}(Y_{ijk}, Y_{ij'k'}|\tilde{u}_i) \\ \mathrm{Cov}(Y_{ijk}, Y_{ij'k'}|\tilde{u}_i) & \mathrm{Var}(Y_{ij'k'}|\tilde{u}_i) \end{pmatrix}$$

where $\mathrm{Var}(Y_{ijk}|\tilde{u}_i) = \frac{\partial^2}{\partial\tilde{\tau}_{ijk}\partial\tilde{\tau}_{ijk}}\tilde{b}(\tilde{\tau}_{i;jk;j'k'})$, and $\mathrm{Cov}(Y_{ijk}, Y_{ij'k'}|\tilde{u}_i) = \frac{\partial^2}{\partial\tilde{\tau}_{ijk}\partial\tilde{\tau}_{ij'k'}}\tilde{b}(\tilde{\tau}_{i;jk;j'k'})$.

Let $\tilde{\mu}_{i;jk;j'k'} = E(Y_{i;jk;j'k'}|\tilde{u}_i)$ be the conditional mean vector for the pair vector $Y_{i;jk;j'k'}$ given random effects $\tilde{u}_i$. In the second step, we link the conditional mean of $Y_{i;jk;j'k'}$ to the covariates with a bivariate regression model

$$\tilde{h}(\tilde{\mu}_{i;jk;j'k'}) = (X_{ijk}^T\boldsymbol{\beta} + Z_{ijk}^T\tilde{u}_i, X_{ij'k'}^T\boldsymbol{\beta} + Z_{ij'k'}^T\tilde{u}_i), \tag{4.9}$$

where $\tilde{h}$ is a bivariate transformation with a given form.

Model (4.9) accommodates model (4.2) as a special case but requires weaker assumptions. For instance, in model (4.9), if $\tilde{u}_i = \Big\{(\tilde{\nu}_i, \tilde{\omega}_j, \tilde{\tau}_k, \tilde{\omega}_{j'}, \tilde{\tau}_{k'})^T, (j,k) < (j',k')\Big\}$, then setting

$$Z_{ijk}\tilde{u}_i = \tilde{\nu}_i + \tilde{\omega}_j + \tilde{\tau}_k$$

and

$$Z_{ij'k'}\tilde{u}_i = \tilde{\nu}_i + \tilde{\omega}_{j'} + \tilde{\tau}_{k'}$$

leads to model (4.5) if all random effects $\tilde{\nu}_i$, $\tilde{\omega}_j$, $\tilde{\tau}_k$, $\tilde{\omega}_{j'}$, $\tilde{\tau}_{k'}$ are assumed to be independent of each other. This strong independence assumption is, however, not required in forming model (4.9). In other words, in forming (4.5), we require all components in $\tilde{u}_i$ to be mutually independent, but in forming (4.9), we only assume pairwise independence among the $\tilde{u}_i$.

The joint density function of $(Y_{ijk}, Y_{ij'k'})$ is given by

$$f(y_{ijk}, y_{ij'k'}) = \iint f(y_{ijk}|\tilde{u}_i)f(y_{ij'k'}|\tilde{u}_i)f(\tilde{u}_i)d\tilde{u}_i,$$

where $f(\tilde{u}_i)$ is the density function for random effects $\tilde{u}_i$. As a result, the probability density function of $f(y_{ijk})$ is given by $f(y_{ijk}) = \int f(y_{ijk}, y_{ij'k'})\, dy_{ij'k'}$. As an example, with model (4.5), we have the pairwise probability density function

$$\begin{aligned} f(y_{ijk}, y_{ij'k'}) \;\; &= \prod_{(j,k)<(j',k')} \int f(y_{ijk}|\tilde{\nu}_i, \tilde{\omega}_j, \tilde{\tau}_k) f(y_{ij'k'}|\tilde{\nu}_i, \tilde{\omega}_{j'}, \tilde{\tau}_{k'}) f(\tilde{\nu}_i) f(\tilde{\omega}_j) \\ &\quad \cdot f(\tilde{\tau}_k) f(\tilde{\omega}_{j'}) f(\tilde{\tau}_{k'})\, d\tilde{\nu}_i\, d\tilde{\omega}_j\, d\tilde{\tau}_k\, d\tilde{\omega}_{j'}\, d\tilde{\tau}_{k'}. \end{aligned}$$

This formulation considerably simplifies the computation of integrals. To formulate this pairwise likelihood, only 5 dimensional integrals are needed to compute, while the formulation of the full likelihood (4.6) involves $n(K+1) + \sum_{i=1}^{n} J_i$ dimensional integrals.

4.2.3 Pairwise Likelihoods

Now we consider a pairwise modeling strategy instead of fully specifying $f(y_i|x_i, z_i)$. Let $\ell(y_{ijk})$ and $\ell(y_{ijk}, y_{ij'k'})$ be the marginal and pairwise observed log likelihoods for y_{ijk} and $(y_{ijk}, y_{ij'k'})$, given x_i and z_i, respectively. Similar to but not the same as Lindsay et al. (2011), a general form of pairwise log likelihood $\ell_c(y_{ijk}, y_{ij'k'})$ with respect to y_{ijk} and $y_{ij'k'}$ can be written as

$$\ell_c(y_{ijk}, y_{ij'k'}) = B_{jk,j'k'} \ell(y_{ijk}, y_{ij'k'}) - B_{jk} \ell(y_{ijk}) - B_{j'k'} \ell(y_{ij'k'}), \tag{4.10}$$

where $B_{jk,j'k'}$, B_{jk} and $B_{j'k'}$ are scalar weights. We limit our discussion to two specific scenarios. When all $B_{jk,j'k'} = 1$ and $B_{jk} = B_{j'k'} = 0$, (4.10) results in all-pairwise marginal log likelihood (APW). When all $B_{jk,j'k'} = 2$ and $B_{jk} = B_{j'k'} = 1$, (4.10) becomes all-pairwise conditional log likelihood (APC). Thus, estimation of the model parameters can be conducted by optimizing

$$\ell_c(y) = \sum_{i=1}^{n} \ell_c(y_i) = \sum_{i=1}^{n} \sum_{(j,k)<(j',k')} \ell_c(y_{ijk}, y_{ij'k'}). \tag{4.11}$$

4.3 Methodology: Selecting Fixed Effects

In this section, we focus on selecting fixed effects only by treating random effects $\tilde{u}_i$ being adequately specified. Denote $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \boldsymbol{\xi}^T)^T$, where $\boldsymbol{\xi}$ represents all parameters other than $\boldsymbol{\beta}$. To achieve both model selection and parameter estimation in (4.10), we propose to maximize the following penalized pairwise log likelihood function:

$$\ell_{pen1}(y) = \ell_c(y) - n \sum_{s=1}^{p} p_\lambda(|\beta_s|), \tag{4.12}$$

where $p_\lambda(|\beta_s|)$ is the penalty function for the s-th element in $\boldsymbol{\beta}$. Following Fan and Li (2001, 2004), we adopt the SCAD penalty, which has nice properties such as unbiasedness, sparsity and continuity properties. The SCAD penalty is a nonconcave function defined by $p_\lambda(0) = 0$ and for $\beta_s > 0$, its first derivative satisfies

$$p'_\lambda(\beta_s) = \lambda \left\{ I(\beta_s \leq \lambda) + \frac{(a\lambda - \beta_s)_+}{(a-1)\lambda} I(\beta_s > \lambda) \right\} \tag{4.13}$$

where $a > 2$ and $\lambda > 0$.

Following Fan and Li (2001), one may maximize (4.12) by using the Newton-Raphson algorithm, where a second order Taylor's series approximation of $p_\lambda(|\beta_s|)$ is often used. Alternatively, we describe an implementation method that shares the same spirit of the EM algorithm. At the tth iteration for the E-step, let the complete log pairwise likelihood for $(Y_{ijk}, Y_{ij'k'})$ be

$$\ell_{cpl}(y_{ijk}, y_{ij'k'}, \tilde{u}_i; \boldsymbol{\psi}) = \log \left\{ f(y_{ijk}, y_{ij'k'} | \tilde{u}_i; \boldsymbol{\psi}) f(\tilde{u}_i; \boldsymbol{\psi}) \right\},$$

and define

$$\begin{aligned} Q_{cpl}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t-1)}) &= \sum_{i=1}^{n} \sum_{(j,k)<(j',k')} E\{\ell_{cpl}(Y_{ijk}, Y_{ij'k'}, \tilde{u}_i; \boldsymbol{\psi}) | Y_{ijk}, Y_{ij'k'}; \boldsymbol{\psi}^{(t-1)}\} \\ &= \sum_{i=1}^{n} \sum_{(j,k)<(j',k')} \int \ell_{cpl}(y_{ijk}, y_{ij'k'}, \tilde{u}_i; \boldsymbol{\psi}) f(\tilde{u}_i | y_{ijk}, y_{ij'k'}; \boldsymbol{\psi}^{(t-1)}) d\tilde{u}_i. \end{aligned}$$

Then, at the $t-$th iteration, the conditional expectation of the complete log composite likelihood function is given by

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t-1)}) = Q_{cpl}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t-1)}),$$

or

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t-1)}) = 2Q_{cpl}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t-1)}) - \ell_c(Y_{ijk}, \boldsymbol{\psi}) - \ell_c(Y_{ij'k'}, \boldsymbol{\psi}),$$

corresponding to the APW and APC methods respectively.

As a result, at the $t-$th iteration, the penalized Q-function for variable selection is given by

$$Q_\lambda(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t-1)}) = Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t-1)}) - n\sum_{s=1}^{p} p_\lambda(|\beta_s|).$$

In the M-step, we maximize $Q_\lambda(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t-1)})$ with respect to $\boldsymbol{\psi}$ to obtain $\boldsymbol{\psi}^{(t)}$. In this step, we again encounter the non-differentiality of penalty functions. Conventionally, the quadratic approximation approach can be used to approximate the penalty function. The E and M steps are iterated until convergence of $\boldsymbol{\psi}^{(t)}$.

The aforementioned algorithm is implemented with given tuning parameters $(a^{(r)}, \lambda^{(r)})$. In practice, a suitable value of $(a^{(r)}, \lambda^{(r)})$ is not obvious, and one can consider a specified grid of candidates for $(a^{(r)}, \lambda^{(r)})$. For each r, one can use the algorithm above to obtain a solution $\hat{\boldsymbol{\psi}}_r$. The final model selection and estimates $\hat{\boldsymbol{\psi}}$ can be realized based on certain selection criteria. For instance, recent studies (Wang et al., 2007; Bondell et al., 2010; Ma and Li, 2010; Zhang et al., 2010) show that the Bayesian information criterion (BIC) is consistent for model selection given that the true model lies in the class of candidate models. Let $\tilde{H}(\hat{\boldsymbol{\psi}}_r) = -\partial^2\ell_{pen1}(y; \hat{\boldsymbol{\psi}}_r)/\partial\tilde{\boldsymbol{\psi}}_r\partial\tilde{\boldsymbol{\psi}}_r^T$, and $\tilde{J}(\hat{\boldsymbol{\psi}}_r) = \sum_{i=1}^{n}\{\partial\ell_c(y_i; \hat{\boldsymbol{\psi}}_r)/\partial\tilde{\boldsymbol{\psi}}_r\}\{\partial\ell_c(y_i; \hat{\boldsymbol{\psi}}_r)/\partial\tilde{\boldsymbol{\psi}}_r\}^T$, where $\tilde{\boldsymbol{\psi}}_r$ denotes the parameter set in which the 0 elements in $\hat{\boldsymbol{\psi}}_r$ are removed.

Under maximum likelihood inference framework, we can take the number of nonzero parameters in $\hat{\boldsymbol{\psi}}_r$ as the degrees of freedom. However, this strategy may produce biased selection results under our pairwise likelihood framework. Define $\mathrm{df}_{(a^{(r)},\lambda^{(r)})}(\tilde{\boldsymbol{\psi}}_r)$ to be the degrees of freedom given by $\mathrm{tr}\{\tilde{J}(\hat{\boldsymbol{\psi}}_r)\tilde{H}(\hat{\boldsymbol{\psi}}_r)^{-1}\}$. Then we define

$$\mathrm{BIC}_{(a^{(r)},\lambda^{(r)})} = -2\ell_c(y;\hat{\boldsymbol{\psi}}_r) + \log(n) \times \mathrm{df}_{(a^{(r)},\lambda^{(r)})}(\tilde{\boldsymbol{\psi}}_r). \tag{4.14}$$

We then choose the solution $\hat{\boldsymbol{\psi}}$ that minimizes the $\mathrm{BIC}_{(a^{(r)},\lambda^{(r)})}$ criterion.

4.4 Methodology: Selecting Both Fixed and Random Effects

In this section, we discuss the model selection strategy for choosing appropriate random effects as well as fixed effects. For ease of exposition, we set $J_i = J$ for all $i = 1, \ldots, n$. Let D be the $q^* \times q^*$ covariance matrix for $\tilde{u}_i$ $(i = 1, \ldots, n)$, where q^* is the number of random effects variables in $\tilde{u}_i$. Let d_{lm} be the (l, m) element of D.

Cholesky decomposition approach is widely applied in longitudinal data studies to select random effects. Chen and Dunson (2003) use the modified Cholesky decomposition to select random effects in linear mixed models. Bondell et al. (2010) and Ibrahim et al. (2010) combine the decomposition approach with the EM algorithm. However, the Cholesky decomposition strategy may not be proper for longitudinal data arising in clusters. To circumvent this problem, we propose a new decomposition strategy and develop a modified Expectation/Conditional Maximization Either algorithm (ECME) (Liu and Pierce, 1994; Schafer, 1998) for model selection and estimation.

4.4.1 Review of Cholesky Decomposition

The Cholesky decomposition specifies a covariance matrix D as $D = LL^T$, where L is a lower triangular matrix with positive diagonal entries. The modified Cholesky decomposition further assumes the form

$$D = D^*\Gamma\Gamma^T D^*, \tag{4.15}$$

where D^* is a diagonal matrix with $D^* = \text{diag}(d_1^*, d_2^*, \ldots, d_{q^*}^*)$, and Γ is a lower triangular matrix with diagonal elements 1. This relationship immediately implies that once $d_l^* = 0$, then the elements in the lth row or lth column of D would be zero. That is, eliminating the lth random effect can be featured by setting $d_l^* = 0$.

Two issues may arise if the (modified) Cholesky decomposition approach is applied handle longitudinal data arising in clusters. To see this, we consider a simple case involving longitudinal data arising in clusters with $J_i = 2$ for $i = 1, \ldots, n$. Let $f(\tilde{u}_i) = f(\tilde{u}_{i1}, \tilde{u}_{i2})$ be the joint distribution of two random effects $\tilde{u}_{i1}$ and $\tilde{u}_{i2}$. Assume $f(\tilde{u}_i)$ is a bivariate normal density with covariance matrix

$$D = \begin{pmatrix} \sigma_u^2 & a_1\sigma_u^2 \\ a_1\sigma_u^2 & \sigma_u^2 \end{pmatrix}, \tag{4.16}$$

where $0 \leq a_1 < 1$. Note that two random variables $\tilde{u}_{i1}$ and $\tilde{u}_{i2}$ have identical variance, which implies that if we decide to take away one random variable, the other should also be removed.

However, if the modified Cholesky decomposition is applied, we obtain $D = D^*\Gamma\Gamma^T D^*$, with

$$\Gamma = \begin{pmatrix} 1 & \\ a_1/\sqrt{1-a_1^2} & 1 \end{pmatrix}$$

and

$$D^* = \begin{pmatrix} d_1^* & \\ & d_2^* \end{pmatrix},$$

where $d_1^* = \sigma_u$, and $d_2^* = \sigma_u\sqrt{1-a_1^2}$, which are not equal unless $a_1 = 0$. When a_1 is nearly 1, d_2^* is almost equal to 0, and variable selection procedure based on a finite sample may yield $\hat{d}_1^* > 0$ and $\hat{d}_2^* = 0$. Hence, $\tilde{u}_{i2}$ could be removed from the model but $\tilde{u}_{i1}$ is kept. Thus, this model selection returns a contradictory result to the original setting that $\tilde{u}_{i1}$ and $\tilde{u}_{i2}$ are equally important in the model.

To show another drawback related to the Cholesky decomposition under pairwise likelihood framework, we follow the same example and consider two paired observations. The random effects distribution for the two pairs could be $f(\tilde{u}_{i2})$ and $f(\tilde{u}_{i1}, \tilde{u}_{i2})$, which are two normal distributions with variance σ_u^2 and covariance matrix $\begin{pmatrix} \sigma_u^2 & a_1\sigma_u^2 \\ a_1\sigma_u^2 & \sigma_u^2 \end{pmatrix}$, respectively.

If using the modified Cholesky decomposition, the same random effect $\tilde{u}_{i2}$ is represented by σ_u and $\sigma_u\sqrt{1-a_1^2}$ in two diagonal matrices, respectively. Therefore, if $a_1 \neq 0$, the same random effect component $\tilde{u}_{i2}$ would be differently represented in different pairwise likelihood functions, which is obviously problematic.

These examples illustrate that the selection procedure can not meaningfully incorporate the relationship among parameters in covariance matrix D. Special care is often needed to avoid meaningless selection results.

4.4.2 The Algorithm

Covariance Matrix Decomposition

We propose a matrix decomposition for symmetric D based on the fact that

$$\begin{aligned} D &= \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1q^*} \\ d_{21} & d_{22} & \cdots & d_{2q^*} \\ \dots & \cdots & \ddots & \dots \\ d_{q^*1} & d_{q^*2} & \cdots & d_{q^*q^*} \end{pmatrix} \\ &= \begin{pmatrix} d_1^2 & d_1 d_2 r_{12} & \cdots & d_1 d_{q^*} r_{1q^*} \\ d_1 d_2 r_{12} & d_2^2 & \cdots & d_2 d_{q^*} r_{2q^*} \\ \dots & \cdots & \ddots & \dots \\ d_1 d_{q^*} r_{1q^*} & d_2 d_{q^*} r_{2q^*} & \cdots & d_{q^*}^2 \end{pmatrix}, \end{aligned}$$

where $d_l = \sqrt{d_{ll}}$, $(l = 1, \dots, q^*)$ and $r_{lm} = d_{lm}/\sqrt{d_{ll}d_{mm}}$ for $l = 1, \dots, q^*$; $m = 1, \dots, q^*$; $l < m$.

Thus, the decomposition can be written as

$$D = \mathcal{DRD}, \tag{4.17}$$

where $\mathcal{D}$ is a $q^* \times q^*$ diagonal matrix $\mathrm{diag}(d_1, d_2, \ldots, , d_{q^*})$, and $\mathcal{R}$ is a square matrix with

$$\begin{pmatrix} 1 & r_{12} & \cdots & r_{1q^*} \\ r_{12} & 1 & \cdots & r_{2q^*} \\ \ldots & \cdots & \ddots & \ldots \\ r_{1q^*} & r_{2q^*} & \cdots & 1 \end{pmatrix}.$$

The decomposition in (4.17) takes the elements in $\mathcal{D}$ as standard error for each random effect, while $\mathcal{R}$'s elements as correlation coefficients of random effects. According to the description in Section 4.4.1, there could be predetermined identical variance parameters in D for the model of longitudinal data arising clusters. If two random variables are set to have identical variance parameters, say the q_1th and q_2th ($q_1 \neq q_2$) random effects, the decomposition in (4.17) just returns $\tilde{d} = d_{q_1} = d_{q_2}$. If $\tilde{d} = 0$, two random variables are removed simultaneously. Thus, it can be seen that our decomposition circumvents the problems in the Cholesky approach.

Based on the covariance matrix decomposition approach, we introduce the doubly penalized log pairwise likelihood

$$\ell_{pen2}(y; \boldsymbol{\beta}, \mathcal{D}, \mathcal{R}) = \ell_c(y; \boldsymbol{\beta}, \mathcal{D}, \mathcal{R}) - n \sum_{s=1}^{p} p_{\lambda_\beta}(|\beta_s|) - n \sum_{l=1}^{\mathcal{Q}} p_{\lambda_{\tilde{d}}}(|\tilde{d}_l|), \qquad (4.18)$$

where $\ell_c(y; \boldsymbol{\beta}, \mathcal{D}, \mathcal{R})$ is the unpenalized pairwise likelihood functions determined by (4.10), $p_{\lambda_\beta}(|\beta_s|)$ is the penalty function for fixed effects, $\mathcal{Q}$ is the number of distinct variance parameters for random effects, and $p_{\lambda_{\tilde{d}}}(|\tilde{d}_l|)$ is the penalty function for random effects with lth distinct variance parameter. In addition, it is straightforward to obtain the penalized Q-function defined similarly in Section 4.3 as

$$\begin{aligned} &Q_\lambda(\boldsymbol{\beta}, \mathcal{D}, \mathcal{R} | \boldsymbol{\beta}^{(t-1)}, \mathcal{D}^{(t-1)}, \mathcal{R}^{(t-1)}) \\ &= Q(\boldsymbol{\beta}, \mathcal{D}, \mathcal{R} | \boldsymbol{\beta}^{(t-1)}, \mathcal{D}^{(t-1)}, \mathcal{R}^{(t-1)}) - n \sum_{s=1}^{p} p_{\lambda_\beta}(|\beta_s|) - n \sum_{l=1}^{\mathcal{Q}} p_{\lambda_{\tilde{d}}}(|\tilde{d}_l|), \end{aligned}$$

where $Q(\boldsymbol{\beta}, \mathcal{D}, \mathcal{R} | \boldsymbol{\beta}^{(t-1)}, \mathcal{D}^{(t-1)}, \mathcal{R}^{(t-1)})$ is Q-function determined in Section 4.3.

A Modified ECEM Algorithm

We employ a modified Expectation/Conditional Maximization Either algorithm (ECME) (Liu and Pierce, 1994; Schafer, 1998) to maximize the composite likelihood function. Our modified ECME algorithm updates the parameters in composite likelihood via both the Newton-Raphson and the EM approaches in turn. In particular, for the parameters $\boldsymbol{\beta}^{(t-1)}$, $\mathcal{D}^{(t-1)}$, $\mathcal{R}^{(t-1)}$, the algorithm has

1. Fix $(\mathcal{D}^{(t-1)}, \mathcal{R}^{(t-1)})$, and update $\boldsymbol{\beta}^{(t)}$ by maximizing $\ell_{pen2}(Y; \boldsymbol{\beta}^{(t-1)}, \mathcal{D}^{(t-1)}, \mathcal{R}^{(t-1)})$. If $\beta_s^{(t)}$ is very close to 0, then set $\hat{\beta}_s = 0$, and remove its corresponding elements from the iteration.

2. Fix $(\boldsymbol{\beta}^{(t)}, \mathcal{R}^{(t-1)})$, and update $\mathcal{D}^{(t)}$ by maximizing $Q_\lambda(\mathcal{D}|\boldsymbol{\beta}^{(t)}, \mathcal{D}^{(t-1)}, \mathcal{R}^{(t-1)})$. If $\tilde{d}_l^{(t)}$ is very close to 0, then set $\hat{\tilde{d}}_l = 0$, remove corresponding random variables from the model and the related elements in $\mathcal{R}^{(t-1)}$ are also deleted.

3. Fix $(\boldsymbol{\beta}^{(t)}, \mathcal{D}^{(t)})$, and update $\mathcal{R}^{(t)}$ by maximizing $Q_\lambda(\mathcal{R}|\boldsymbol{\beta}^{(t)}, \mathcal{D}^{(t)}, \mathcal{R}^{(t-1)})$.

Iteratively run the updating procedure until convergence, and denote the estimator as $\hat{\boldsymbol{\psi}}$. In practice, the tuning parameters can be selected by the composite BIC strategy determined in Section 4.3.

4.5 Asymptotic Results

We now discuss the asymptotic results for our pairwise variable selection strategy. For ease of exposition, we consider the selection for fixed effects variables only, and the selection for random effects follows analogously with more complex notations involved. Let $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{p0})$ denote the true parameter value of $\boldsymbol{\beta}$, which is written, without loss of generality, as $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0I}^T, \boldsymbol{\beta}_{0II}^T)^T$, where $\boldsymbol{\beta}_{0I} = (\beta_{10}, \ldots, \beta_{p_10})^T$ is the vector consisting of all non-zero values and $\boldsymbol{\beta}_{0II} = (\beta_{p_1+1,0}, \ldots, \beta_{p0})^T = \mathbf{0}^T_{\boldsymbol{\beta}_{0II}}$ includes all zero components of $\boldsymbol{\beta}$.

Correspondingly, write $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_{II}^T)^T$, $\boldsymbol{\psi} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_{II}^T, \xi^T)^T$, and $\boldsymbol{\psi}_0 = (\boldsymbol{\beta}_{0I}^T, \mathbf{0}_{\boldsymbol{\beta}_{0II}}^T, \xi_0^T)$ with ξ_0 being the true value of ξ.

For any square matrix M of the same dimension as $\boldsymbol{\psi}$, let $\tilde{M}$ denote the sub-matrix after removing the (p_1+1)st, $\ldots$, and pth rows and columns from the matrix M. Similarly, for any vector α of the same dimension as $\boldsymbol{\psi}$, we use $\tilde{\alpha}$ to denote the resulting vector after removing the (p_1+1)st, $\ldots$, and pth elements from the vector α. For example, $\tilde{\boldsymbol{\psi}}_0 = (\boldsymbol{\beta}_{0I}^T, \xi_0^T)^T$.

Consistency of the estimator $\hat{\boldsymbol{\psi}}$ is established by the following theorem, and its proof is outline in Appendix B.

Theorem 1: Under regularity conditions in Appendix A, there exists a local maximizer $\hat{\boldsymbol{\psi}}$ of $\ell_{pen1}(Y;\boldsymbol{\psi})$ such that

$$\|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\| = O_p(n^{-1/2}),$$

The sparsity is suggested by the following result, and its proof is outlined in Appendix C.

Theorem 2: Under regularity conditions in Appendix A, with probability tending to 1, for any given $\boldsymbol{\beta}_I$ and ξ satisfying

$$\|\boldsymbol{\beta}_I - \boldsymbol{\beta}_{0I}\| = O_p(n^{-1/2}) \qquad \text{and } \|\xi - \xi_0\| = O_p(n^{-1/2}),$$

we have

$$\ell_{pen1}(Y;\boldsymbol{\beta}_I,\mathbf{0},\xi) = \max_{\|\boldsymbol{\beta}_{II}\| \le Cn^{-1/2}} \ell_{pen1}(Y;\boldsymbol{\beta}_I,\boldsymbol{\beta}_{II},\xi) \qquad \text{for any positive constant } C.$$

Now we come to the oracle property of the estimator $\hat{\boldsymbol{\psi}}$. Let

$$\Sigma = \operatorname{diag}\{p''_{\lambda_n}(|\beta_{01}|), \ldots, p''_{\lambda_n}(|\beta_{0p}|), \mathbf{0}_\xi\},$$

and

$$\mathbf{b} = \Big(\big(p'_{\lambda_n}(|\beta_{01}|)\operatorname{sgn}(\beta_{01}), \ldots, p'_{\lambda_n}(|\beta_{0p}|)\operatorname{sgn}(\beta_{0p})\big)^T, \mathbf{0}_\xi^T\Big)^T,$$

where $\mathbf{0}_\xi$ is a zero vector with the same length as ξ. The asymptotic property is suggested by the following result, and its proof is outlined in Appendix D.

Theorem 3: Under regularity conditions in Appendix A, with probability tending to 1, the root-n consistent local maximizers $\hat{\boldsymbol{\psi}}$ in Theorem 1 must satisfy:

(a). Sparsity: $\hat{\boldsymbol{\beta}}_{II} = \mathbf{0}$.

(b). Asymptotic normality: $\sqrt{n}(\tilde{D}(\tilde{\boldsymbol{\psi}}_0)+\tilde{\Sigma})\{\tilde{\hat{\boldsymbol{\psi}}}-\tilde{\boldsymbol{\psi}}_0+(\tilde{D}(\tilde{\boldsymbol{\psi}}_0)+\tilde{\Sigma})^{-1}\tilde{\mathbf{b}}\} \to_D N(\mathbf{0}, \tilde{M}(\tilde{\boldsymbol{\psi}}_0))$.

where $M(\boldsymbol{\psi}) = E_{Y_i;\boldsymbol{\psi}_0}\left[\left\{\partial\ell_c(Y_i;\boldsymbol{\psi})/\partial\boldsymbol{\psi}\right\}\left\{\partial\ell_c(Y_i;\boldsymbol{\psi})/\partial\boldsymbol{\psi}\right\}^T\right]$, and

$$D(\boldsymbol{\psi}) = E_{Y_i;\boldsymbol{\psi}_0}\left\{-\partial^2\ell_c(Y_i;\boldsymbol{\psi})/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T\right\},$$

and similar definitions are applied to $\tilde{M}(\tilde{\boldsymbol{\psi}}_0)$ and $\tilde{D}(\tilde{\boldsymbol{\psi}}_0)$.

4.6 Numerical Studies

4.6.1 Simulation for Selecting Fixed Effects

Linear Mixed Model

We now conduct a simulation study for the linear mixed model. The data are generated from the model

$$Y_{ijk} = X_{ijk}^T\boldsymbol{\beta} + u_{ij} + \epsilon_{ijk}, \tag{4.19}$$

where the residual $\epsilon_i = (\epsilon_{i11}, \ldots, \epsilon_{ijk}, \ldots, \epsilon_{iJ_iK})^T$ are normally distributed with joint distribution specified in the following examples, $u_i = (u_{i1}, \ldots, u_{ij}, \ldots, u_{iJ_i})^T$ are random effects with a distribution specified in the following examples, and the residual ϵ_i is independent of the random effects u_i. Set $p = 8$, $\sigma_\epsilon^2 = 1$ and $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. Covariates $X_{ijk} = (X_{ijk,1}, X_{ijk,2}, \ldots, X_{ijk,8})^T$ are generated from a multivariate normal distribution with mean zero and covariance matrix $V = [\sigma_{st}^2]$, where $\sigma_{st}^2 = \rho_{st}\sigma^2$. We set $\rho_{st} = \rho^{|s-t|}$, $\rho = 0.5$ and $\sigma^2 = 1$. We particularly consider the following scenarios.

Example 1: $n = 200$, $J_i = J = 1$, for $i = 1, \ldots, n$, and $K = 5$. This corresponds to an ordinary longitudinal setting with 5 visits times. The model is set to be ordinary LMMs with the ϵ_i to be independently distributed with joint distribution $N_5(0, \sigma_\epsilon^2 I_5)$, where I_5 is a 5×5 identity matrix. Random effects u_i become one-dimensional and have a normal distribution $N_1(0, \sigma_u^2)$ with $\sigma_u^2 = 1$.

Example 2: The setup is the same as in Example 1 but we take $n = 500$.

Example 3: $n = 60$, $J_i = J = 3$ and $K = 3$. This corresponds to longitudinal data with 3 subjects in each cluster by following 3 visits times. The model is set to be ordinary LMMs with the ϵ_i to be independently distributed with joint distribution $N_9(0, \sigma_\epsilon^2 I_9)$, where I_9 is a 9×9 identity matrix. For each subject, we set $u_i = (u_{i1}, u_{i2}, u_{i3})$ to be 3-dimensional random effects following a normal distribution $N_3(\mathbf{0}, R)$, where

$$R = \sigma_u^2 \begin{pmatrix} 1 & \rho^* & \rho^* \\ \rho^* & 1 & \rho^* \\ \rho^* & \rho^* & 1 \end{pmatrix},$$

with $\rho^* = 0.5$.

Example 4: The setup is the same as in Example 3 but we take $n = 300$.

We describe a measure that is used to feature the performance of the estimates obtained from different models. Let $\mu = E_{u_i}\big\{E(Y_{ijk}|u_i, X_i, Z_i)\big\} = E_{u_i}\{h^{-1}(X_{ijk}^T \boldsymbol{\beta}_0 + Z_{ijk}^T u_i)\}$, and $\hat{\mu} = E_{u_i}\{h^{-1}(X_{ijk}^T \hat{\boldsymbol{\beta}} + Z_{ijk}^T u_i)\}$, where $h(\cdot)$ is the link function defined in (4.1), $\hat{\boldsymbol{\beta}}$ is an estimator obtained from the proposed method. The expectations are evaluated with respect to the true model. We define

$$\mathrm{MME}(\hat{\mu}) = E_{(X_i, Z_i)}\{\hat{\mu} - \mu\}^2,$$

and use this measure to quantify the marginal model error induced by estimator $\hat{\boldsymbol{\beta}}$, where the expectation is taken with respect to the marginal distribution for (X_i, Z_i).

For each example, we repeat the simulation 500 times and fit each dataset by maximum likelihood (ML), all-pairwise marginal likelihood (APW) and all-pairwise conditional likelihood (APC) approaches. Tuning parameters are selected by fixing $a = 3.7$ but only searching for λ. In Examples 1 and 2, we also explore searching for both a and λ.

Table 4.3 reports the average of zero coefficients. The column labeled "Correct" presents the average of zero coefficients that are correctly estimated, and the column labeled "Incorrect" depicts the average of non-zero coefficients erroneously set to zero. We report the median ratios of MME, denoted by R.MME, for a selected model to that of the un-penalized estimate under the unpenalized model in each of the ML, APW and APC scenarios, respectively. We also report the median of MME, denoted by M.MME, for selected models in each of ML, APW and APC scenarios. Table 4.4 summaries the estimated $(\beta_1, \beta_2, \beta_5)$, their relative biases, empirical standard errors, model-based standard errors, and coverage rates of 95% confidence intervals.

For all six examples, three methods show a good sparsity property. Moreover, compared to the ML method, the APW and the APC approaches produce similar rates of shrinking unimportant coefficients to zero, and higher R.MME. The APC outperforms the APW with higher shrinkage rates and smaller R.MME. It can be seen that the estimates of the β_s have relatively small biases in all cases. The standard error formulas perform well in most cases as they are close to the empirical estimates. It is interesting to note that the APW approach provides slightly larger standard errors than the APC method.

Tables 4.3-4.5 further illustrate the approach with grid searching on both a and λ. No obvious difference from only searching on λ is revealed. Moreover, two tuning parameter selection methods result in a similar model selection and estimation results. Since fixing $a = 3.7$ has a cheaper computation cost, we only use this tuning parameter selection approach in our subsequent studies.

Logistic Mixed Model

We now conduct a simulation study for the logistic mixed model. The data are generated from the model

$$\text{logit}\{P(Y_{ijk} = 1|X_i, Z_i, u_i)\} = X_{ijk}^T\boldsymbol{\beta} + u_{ij},$$

where logit(a) is a logistic link function in a form of $\log\{a/(1-a)\}$, X_{ijk}, $\boldsymbol{\beta}$ and u_{ij}, which are the same as those in linear mixed model simulation. We particularly consider the following two scenarios.

Example 1: $n = 200$, $J_i = J = 1$, for $i = 1, \ldots, n$, and $K = 5$. Other parameter settings follow from Example 1 in linear mixed model.

Example 2: The example is the same as Example 1 except we take $n = 800$.

Example 3: The setup is the same as the one in Example 1, except that we take $n = 200$, $J_i = J = 3$ and $K = 4$, and set $u_i = (u_{i1}, u_{i2}, u_{i3})$ to be 3-dimensional random effects following a normal distribution $N_3(\mathbf{0}, R)$, where

$$R = \sigma_u^2 \begin{pmatrix} 1 & \rho^* & \rho^* \\ \rho^* & 1 & \rho^* \\ \rho^* & \rho^* & 1 \end{pmatrix},$$

with $\rho^* = 0.3$.

Example 4: The setup is the same as Example 3, except we take $n = 400$.

Table 4.6 shows a good sparsity property with estimating results excluding large proportion of the zero coefficients covariates, while all non-zero coefficients corresponded covariates are maintained in the model. Moreover, compared to the ML method, the APW and the APC approaches produce similar rate of shrinking unimportant coefficients to zero, and higher M.MME. The APW outperforms the APC with higher shrinkage rate and smaller M.MME. It can be seen that our estimates of β_s have relatively small biases in all cases. The standard error formulas' performance are slightly lower than the empirical estimates. It is interesting to report that the APC approaches provides slightly larger standard error than the APW method with respect to regression coefficients.

Poisson Mixed Model

We now conduct a simulation study for the Poisson mixed model. The data are generated from the model

$$\log\Big\{E(Y_{ijk}|X_i, Z_i, u_i)\Big\} = X_{ijk}^T\boldsymbol{\beta} + u_{ij}, \tag{4.20}$$

where $\boldsymbol{\beta} = (1.2, 0.6, 0, 0, 0.8, 0, 0, 0)^T$, u_{ij} and X_{ijk} are the same as that of linear mixed model. We consider following scenarios

Example 1: $n = 60$, $J_i = J = 1$ for $i = 1, \ldots, n$, and $K = 5$. Other parameter settings follow from Example 1 in linear mixed model.

Example 2: The setup is the same as the one in Example 1, except we take $n = 500$.

Example 3: $n = 60$, $J_i = J = 3$, $K = 2$, and set $u_i = (u_{i1}, u_{i2}, u_{i3})$ to be 3-dimensional random effect following a normal distribution $N_3(\mathbf{0}, R)$, where R follows the same settings as in the logistic case.

Example 4: The setup is the same as the one in Example 3, except we take $n = 300$.

The results are shown in Table 4.9. All three methods show a good sparsity property. The APC method outperforms the APW approach with higher shrinkage rate and smaller M.MME. It can be seen that the estimates of the β_s have relatively small biases in all cases. The standard error formulas' performance are slightly smaller than empirical estimates. It is interesting to report that the APW approach provides slightly larger standard error than the APC method with respect to regression coefficients.

4.6.2 Simulation for Both Fixed and Random Effects

Linear Mixed Model

We simulate data set consisting of n independent observations according to the model $Y_i = X_i^T\boldsymbol{\beta}+Z_i^T u_i+\epsilon_i$, $i = 1, \ldots, n$, where $\epsilon_i = (\epsilon_{i11}, \ldots, \epsilon_{ijk}, \ldots, \epsilon_{iJ_iK})^T$ are normally distributed with joint distribution specified in the following examples, $u_i = (u_{i1}, \ldots, u_{ij}, \ldots, u_{iJ_i})^T$ are random effects with a distribution specified in the following examples, and the residual ϵ_i is independent of the random effects u_i. Set $\sigma_\epsilon^2 = 1$ and $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. We consider the true model

$$Y_{ijk} = u_{ij,0} + (\beta_1 + u_{ij,1})X_{ijk,1} + (\beta_2 + u_{ij,2})X_{ijk,2} + \beta_5 X_{ijk,5} + \epsilon_{ijk}.$$

Moreover, $u_{ij} = (u_{ij,0}, u_{ij,1}, u_{ij,2})$ for $i = 1, \ldots, n$; $j = 1, \ldots, J_i$ follows multivariate normal random vectors with zero mean and the true covariance matrix

$$D = \begin{pmatrix} 9 & 4.8 & 0.6 \\ 4.8 & 4 & 1 \\ 0.6 & 1 & 1 \end{pmatrix}.$$

The covariates X_{ijk} are generated as in fixed effect selecting case. We set $Z_i = X_i$ plus a random intercept term.

We particularly consider the following cases:

Situation 1: Generate Data from GLMMs

Scenario 1: $n = 100$, $J_i = J = 1$, for $i = 1, \ldots, n$, and $K = 5$. This corresponds to an ordinary longitudinal setting with 5 visits times. The model is set to be ordinary GLMMs with the ϵ_i to be independently distributed with joint distribution $N_5(0, \sigma_\epsilon^2 I_5)$, where I_5 is an 5×5 identity matrix.

Scenario 2: The setup is the same as the one in Scenario 1, but we take $n = 300$.

Scenario 3: The setup is the same as the one in Scenario 1, except that we take $n = 100$, $J_i = J = 3$ and $K = 3$, and set $u_i = (u_{i1}, u_{i2}, u_{i3})^T$ to be random effects with zero mean and covariance matrix D_3, where

$$D_3 = \begin{pmatrix} D & \rho^* D & \rho^* D \\ \rho^* D & D & \rho^* D \\ \rho^* D & \rho^* D & D \end{pmatrix},$$

with $\rho^* = 0.5$.

Scenario 4: The setup is the same as the one in Scenario 3, but $n = 300$.

Situation 2: Generate Data from GLMPMs

Scenario 1: $n = 100$, $J_i = J = 1$, for $i = 1, \ldots, n$, and $K = 5$. The setup is the same as the one in Scenario 1 in GLMMs, but the model is set to be GLMPMs with ϵ_i to have correlated distribution $N_5(0, \sigma_\epsilon^2 A_5)$, with A_5 to have AR(1) structure with correlation coefficient $\rho_e = 0.5$.

Scenario 2: The setup is the same as the one in Scenario 1, but we take $n = 300$.

Scenario 3: The setup is the same as the one in Scenario 1, except that we take $n = 100$, $J_i = J = 3$ and $K = 3$. We set u_i follows Scenario 3 in GLMM, and ϵ_i to have correlated distribution $N_9(0, \sigma_\epsilon^2 A_9)$, with A_9 to have

$$A_9 = \begin{pmatrix} A_e & & \\ & A_e & \\ & & A_e \end{pmatrix},$$

where A_e is 3×3 matrix of AR(1) structure with correlation coefficient $\rho_e = 0.5$.

Scenario 4: The setup is the same the one in Scenario 3, but $n = 300$.

For comparing model selection results, we employ mean squared errors for fixed effects: $MSE_{\boldsymbol{\beta}} = ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2$ and mean squared errors for random effects: $MSE_D = ||\sqrt{\text{diag}(D)} - \sqrt{\text{diag}(\hat{D})}||^2$. We report the median of both quantities, denoted by $M.MSE_{\boldsymbol{\beta}}$ and $M.MSE_D$. Moreover, we report the median ratios of $MSE_{\boldsymbol{\beta}}$ and $M.MSE_D$, denoted by $R.MSE_{\boldsymbol{\beta}}$ and $R.MSE_D$, for a selected model to that of the un-penalized estimate, respectively.

For each scenario, we repeat the simulation 500 times and fit each dataset by maximum likelihood (ML) and all-pairwise marginal likelihood (APW). Tables 4.12 and 4.15 report the average of zero coefficients. The column labeled "Correct1" presents the average of fixed zero coefficients that are correctly estimated. The column labeled "Incorrect1" depicts the average of fixed non-zero coefficients erroneously set to zero. Similarly, columns labeled "Correct2" and "Incorrect2" represent the selection precision average for random effects.

Tables 4.13 and 4.16 summarize the estimated $(\beta_1, \beta_2, \beta_5)$, their relative biases, empirical standard errors, model-based standard errors, and coverage rates of 95% confidence intervals.

For all examples above, two methods show a good sparsity property when sample size increases. Moreover, compared to the ML method, the APW approach produces similar rates of shrinking unimportant fixed and random coefficients to zero under large sample size. It can be seen that the estimates of the β_s have relatively small biases in all cases. The standard error formulas perform well in large sample cases: they are close to the empirical estimates.

Poisson Mixed Model

We now conduct a simulation study for the Poisson mixed model. We consider Y^*_{ijk} to be generated from a Poisson distribution with

$$\log\Big\{E(Y^*_{ijk}|X_i, Z_i, u_i)\Big\} = X^T_{ijk}\boldsymbol{\beta} + Z^T_{ijk}u_{ij}. \qquad (4.21)$$

We also generate another Poisson data Y_{ij}^{**} $(j = 1, \dots, J_i)$ with mean to be 1. Y_{ijk}^{*} and Y_i^{**} are independent. We set $\boldsymbol{\beta} = (1.2, 0.6, 0, 0, 0.8, 0, 0, 0)^T$, while X_{ijk} and Z_{ijk} are the same as that of the linear mixed model.

Situation 1: Generate Data from GLMM

Scenario 1: $n = 250$, $J_i = J = 1$, for $i = 1, \dots, n$,and $K = 9$. We take Y_{ijk}^{*} as the response. u_{ij} follows multivariate normal random vectors with zero mean and the true covariance matrix

$$D = \begin{pmatrix} 0.25 & 0.015 & 0.02 \\ 0.015 & 0.09 & 0.03 \\ 0.02 & 0.03 & 0.04 \end{pmatrix}.$$

Scenario 2: The setup is the same as the one in Scenario 1, but we take $n = 500$.

Scenario 3: The setup is the same as the one in Scenario 1, except that we take $n = 250$, $J_i = J = 3$ and $K = 4$, and set u_i to be random effects with zero mean and covariance matrix D_3, where

$$D_3 = \begin{pmatrix} D & \rho^* D & \rho^* D \\ \rho^* D & D & \rho^* D \\ \rho^* D & \rho^* D & D \end{pmatrix},$$

with $\rho^* = 0.5$.

Scenario 4: The setup is the same as the one in Scenario 3, but $n = 500$.

Situation 2: Generate Data from GLMPMs

Scenario 1: $n = 250$, $J_i = J = 1$, for $i = 1, \dots, n$, and $K = 9$. We take the response $Y_{ijk} = Y_{ijk}^{*} + Y_i^{**}$. Other settings follow Situation 1.

Scenario 2: The setup is the same as the one in Scenario 1, but we take $n = 500$.

Scenario 3: The setup is the same as the one in Scenario 1, except that we take $n = 250$, $J_i = J = 3$ and $K = 4$. We set u_i follows Scenario 3 in GLMMs, and $Y_{ijk} = Y^*_{ijk} + Y^{**}_{ij}$, with Y^{**}_{ij} to be independent for $j = 1, 2, 3$.

Scenario 4: The setup is the same as the one in Scenario 3 but $n = 500$.

Tables 4.18-4.23 show that when data are generated by GLMMs or GLMPMs, our GLMPMs always have good sparsity property, relatively small biases for the estimates of β_s, and good performance for the standard error formulas in most cases. On the other hand, GLMMs perform poor when the data are generated from GLMPMs, where the estimates are significantly biased.

4.6.3 Data Analysis

The National Population Health Survey (NPHS) is a longitudinal study that collects health information and related socio-demographic information by following a group of Canadian household residents. The questions for the NPHS include many aspects of in-depth health information such as health status, use of health services, chronic conditions and activity restrictions. Moreover, social background questions, including age, sex and income level, are contained in the questionnaire. A research interest focuses on modeling the influence of income on population health. The data we analyze here contain observations from 6 cycles, including $n = 1033$ males with age between 50-70 at Cycle 1, and less than 80 at Cycle 6. All the deceased subjects are excluded from the analysis.

Health status (HUI) is measured by the Health Utilities Index Mark after zero-mean normalization. The higher HUI score indicates a better health status. The covariate prone to missingness is household income (INC), which is measured by provincial level of household income with zero-mean normalization. The other covariate, denoted by CYCLE is cycle number after log-transformation, respectively. All observations with incomplete HUI or INC are excluded from the analysis.

Preliminary analysis indicates that random intercept may be sufficient to account for the correlation across cycles, and cubic terms of INC and CYCLE together with their interactions may be relevant in modeling HUI. This motivates us to consider variable selection in the following model

$$Y_{ijk} = X_{ijk}\beta + u_{ij} + \varepsilon_{ijk}, \tag{4.22}$$

where $J_i = 1$ $K = 6$ for all i, Y_{ijk} is the HUI score for subject i measured at Cycle k, X_{ijk} is a 16×1 vector of variables measured at j: Intercept, INC, INC^2, INC^3, CYCLE, CYCLE^2, CYCLE^3, CYCLE, INC × CYCLE, INC^2 × CYCLE, INC^3 × CYCLE, INC × CYCLE^2, INC^2 × CYCLE^2, INC^3 × CYCLE^2, INC × CYCLE^3, INC^2 × CYCLE^3, INC^3 × CYCLE^3. $u_{ij} \sim N(0, \sigma_u^2)$ is the subject specific random effect and $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$ is the independent residual error.

We apply the ML, APW and APC procedures to model (5.12). Table 5.2 displays the model fitting and selection results. The three methods obtain relatively comparable results that exclude all interaction terms, but suggest a cubic influence from cycle time. The ML approach claims that income has only a linear effect on health index, while the APW and APC methods suggest that income also has a cubic influence on HUI as well.

Table 4.1: Analysis results for the NPHS data: entries represent the estimates and standard errors (in brackets)

Variable	ML[†]		APW		APC	
	Full Model	Selected Model	Full Model	Selected Model	Full Model	Selected Model
Intercept	−0.020(0.041)	0.001(0.028)	-0.016(0.043)	0.010(0.030)	0.003(0.042)	0.014(0.028)
INC	0.109(0.064)	0.085(0.014)	0.159(0.072)	0.095(0.022)	0.080(0.071)	0.041(0.014)
INC^2	-0.012(0.027)		-0.023(0.029)	-0.014(0.011)	-0.025(0.027)	
INC^3	-0.003(0.033)		0.009(0.036)	0.040(0.012)	0.015(0.035)	0.026(0.008)
CYCLE	0.349(0.216)	0.073(0.019)	0.226(0.201)	0.081(0.019)	0.295(0.199)	0.079(0.019)
CYCLE^2	-0.284(0.315)		-0.085(0.298)		-0.206(0.295)	
CYCLE^3	0.033(0.116)	-0.040(0.007)	-0.032(0.111)	-0.034(0.007)	0.005(0.110)	-0.041(0.007)
$\text{INC} \times \text{CYCLE}$	-0.258(0.380)		-0.348(0.394)		-0.151(0.379)	
$\text{INC}^2 \times \text{CYCLE}$	-0.130(0.166)		-0.040(0.164)		-0.099(0.161)	
$\text{INC}^3 \times \text{CYCLE}$	0.236(0.203)		0.268(0.218)		0.177(0.210)	
$\text{INC} \times \text{CYCLE}^2$	0.291(0.551)		0.440(0.568)		0.196(0.546)	
$\text{INC}^2 \times \text{CYCLE}^2$	0.165(0.244)		0.039(0.240)		0.139(0.234)	
$\text{INC}^3 \times \text{CYCLE}^2$	-0.353(0.297)		-0.399(0.316)		-0.305(0.305)	
$\text{INC} \times \text{CYCLE}^3$	-0.093(0.202)		-0.149(0.211)		-0.068(0.203)	
$\text{INC}^2 \times \text{CYCLE}^3$	-0.047(0.090)		-0.006(0.088)		-0.040(0.086)	
$\text{INC}^3 \times \text{CYCLE}^3$	0.134(0.109)		0.150(0.117)		0.122(0.113)	

† ML, APW and APC represent maximum likelihood, all-pairwise marginal pairwise likelihood and all-pairwise conditional pairwise likelihood approaches, respectively.

4.7 Model Selection under Misspecified Models

Our previous discussions are all based on the assumption that both the conditional pairwise distribution and the distribution for random effects are correctly specified. When these assumptions are violated, the estimation and the selection results could be biased or incorrect. To be specific, we focus on the estimation and selection bias for fixed effects, and the random effects conclusions can be obtained by the same spirit. With the logistic mixed model with misspecified random effects, Heagerty and Kurland (2001) explore several types of model misspecfication and find that biased results can be yielded. Other studies include Neuhaus et al. (1992, 1994), Verbeke et al. (2001) and Neuhaus and McCulloch (2006). Recently, there are some studies dealing with the misspecified model selection issue. For example, Varin and Vidoni (2005) and Gao and Song (2010) propose pairwise AIC and pairwise BIC for the variable selection with pairwise likelihood, which includes "pseudo" association structures. More generally, Lv and Liu (2010) discuss a semi-Bayesian information criterion (SIC) with a particular decomposition for taking goodness of model fit, model complexity and model misspecification simultaneously. However, little work was done under the penalized likelihood or penalized pairwise likelihood framework.

4.7.1 Misspecified Models

Here we develop theoretical results in the variable selection via penalized pairwise likelihood. We particularly consider the case that the distribution for random effects is misspecified. For ease of notations, we use superscript * to indicate the corresponding quantities under a misspecified model. In particular, let $\ell_c^*(Y;\boldsymbol{\psi}^*)$ be the corresponding version of the log pairwise likelihood function (4.11) when random effects are misspecified, where $\boldsymbol{\psi}^* = (\boldsymbol{\beta}^{*T}, \xi^{*T})^T$, $\boldsymbol{\beta}^*$ represents the $p \times 1$ vector of regression coefficients, and ξ^* represents all the remaining parameters.

In application, we may obtain $\hat{\boldsymbol{\psi}}^*$ via the maximization of penalized pairwise likelihood function

$$\ell^*_{pen1}(y; \boldsymbol{\psi}^*) = \ell^*_c(y; \boldsymbol{\psi}^*) - n \sum_{s=1}^{p} p_\lambda(|\beta^*_s|) \tag{4.23}$$

where $p_\lambda(|\beta^*_s|)$ is taken as the SCAD penalty function for the s-th element in $\boldsymbol{\beta}^*$.

Yi and Reid (2010) demonstrate that the estimator $\hat{\boldsymbol{\psi}}^*$ would converge in probability to a limit $\boldsymbol{\psi}^*_0$. This limit $\boldsymbol{\psi}^*_0 = (\boldsymbol{\beta}^{*T}_0, \xi^{*T}_0)^T$ is, under certain regularity conditions, the solution of

$$E_{(Y;\boldsymbol{\psi}_0)} \left\{ \frac{\partial \ell^*_c(Y; \boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}^*} \right\} = \mathbf{0}, \tag{4.24}$$

where the expectation is taken under the true joint distribution with true parameter value $\boldsymbol{\psi}_0$. According to (4.24), the solution, $\boldsymbol{\psi}^*_0$, is a function of $\boldsymbol{\psi}_0$. The discrepancy amount between $\boldsymbol{\psi}^*_0$ and $\boldsymbol{\psi}_0$ indicates the degree of biased results.

Asymptotic Results for Misspecified Models

Now we examine the asymptotic properties for $\boldsymbol{\psi}^*$. Without loss of generality, we write $\boldsymbol{\beta}^*_0 = (\boldsymbol{\beta}^{*T}_{0I}, \boldsymbol{\beta}^{*T}_{0II})^T$, where $\boldsymbol{\beta}^*_{0I} = (\beta^*_{10}, \ldots, \beta^*_{p^*_1 0})^T$ is the $p^*_1 \times 1$ vector consisting of all non-zero values while $\boldsymbol{\beta}^*_{0II} = (\beta^*_{p^*_1+1,0}, \ldots, \beta^*_{p0})^T = \mathbf{0}^T_{\boldsymbol{\beta}^*_{0II}}$ is the $(p - p^*_1) \times 1$ vector. We comment that p^*_1 could differ from p_1. Thus, we have $\boldsymbol{\psi}^*_0 = (\boldsymbol{\beta}^{*T}_{0I}, \mathbf{0}^{T*}_{\boldsymbol{\beta}_{0II}}, \xi^{*T}_0)$. Correspondingly, write $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^{*T}_I, \boldsymbol{\beta}^{*T}_{II})^T$, $\boldsymbol{\psi}^* = (\boldsymbol{\beta}^{*T}_I, \boldsymbol{\beta}^{*T}_{II}, \xi^{*T})^T$.

Similar to Section 4.5, for any square matrix M of the same dimension as $\boldsymbol{\psi}$, let $\tilde{M}$ denote the sub-matrix after removing the $(p^*_1 + 1)$st, $\ldots$, and pth rows and columns from the matrix M. For any vector α of the same dimension as $\boldsymbol{\psi}^*$, we use $\tilde{\alpha}^*$ to denote the resulting vector after removing the (p^*_1+1)st, $\ldots$, and pth elements from the vector α. For example, $\tilde{\boldsymbol{\psi}}^*_0 = (\boldsymbol{\beta}^{*T}_{0I}, \xi^{*T}_0)^T$.

In Appendices F, G and H, we sketch the proofs of the following results.

Theorem 4: Under the regularity condition outlined in Appendix E, there exists a local maximizer $\hat{\boldsymbol{\psi}}^*$ of $\ell^*_{pen1}(Y; \boldsymbol{\psi}^*)$ such that

$$\|\hat{\boldsymbol{\psi}}^* - \boldsymbol{\psi}^*_0\| = O_p(n^{-1/2}).$$

Theorem 5: Under the regularity condition outlined in Appendix E, with probability tending to 1, for any given $\boldsymbol{\beta}_I^*$ and ξ^* satisfying

$$\|\boldsymbol{\beta}_I^* - \boldsymbol{\beta}_{0I}^*\| = O_p(n^{-1/2}) \qquad \text{and } \|\xi^* - \xi_0^*\| = O_p(n^{-1/2}),$$

we have

$$\ell_{pen1}^*(Y; \boldsymbol{\beta}_I^*, \mathbf{0}, \xi^*) = \max\nolimits_{\|\boldsymbol{\beta}_{II}^*\| \leq Cn^{-1/2}} \ell_{pen1}^*(Y; \boldsymbol{\beta}_I^*, \boldsymbol{\beta}_{II}^*, \xi^*) \qquad \text{for any positive constant } C.$$

Now we define $\Sigma^* = \mathrm{diag}\{p''_{\lambda_n}(|\beta_{10}^*|), \ldots, p''_{\lambda_n}(|\beta_{p0}^*|), \mathbf{0}_{\xi^*}\}$, and

$$\mathbf{b}^* = \Big(\big(p'_{\lambda_n}(|\beta_{10}^*|)\mathrm{sgn}(\beta_{10}^*), \ldots, p'_{\lambda_n}(|\beta_{p0}^*|)\mathrm{sgn}(\beta_{p0}^*) \big)^T, \mathbf{0}_{\xi^*}^T \Big)^T,$$

where $\mathbf{0}_{\xi^*}$ is a zero vector with the same dimension as that of ξ^*.

Theorem 6: Under the regularity condition outlined in Appendix E, with probability tending to 1, the root-n consistent local maximizers $\hat{\boldsymbol{\psi}}^*$ in Theorem 4 must satisfy:

(a). Sparsity: $\hat{\boldsymbol{\beta}}_{II}^* = \mathbf{0}$.

(b). Asymptotic normality: $\sqrt{n}(\tilde{D}^*(\tilde{\boldsymbol{\psi}}_0^*) + \tilde{\Sigma}^*)\{\hat{\tilde{\boldsymbol{\psi}}}^* - \tilde{\boldsymbol{\psi}}_0^* + (\tilde{D}^*(\tilde{\boldsymbol{\psi}}_0^*) + \tilde{\Sigma}^*)^{-1}\tilde{\mathbf{b}}^*\} \to_D N(\mathbf{0}, \tilde{M}^*(\tilde{\boldsymbol{\psi}}_0^*))$,

where $M^*(\boldsymbol{\psi}^*) = E_{Y_i;\boldsymbol{\psi}_0}\left[\Big\{ \partial \ell_c^*(Y_i; \boldsymbol{\psi}^*)/\partial \boldsymbol{\psi}^* \Big\} \Big\{ \partial \ell_c^*(Y_i; \boldsymbol{\psi}^*)/\partial \boldsymbol{\psi}^* \Big\}^T \right]$, and

$$D^*(\boldsymbol{\psi}^*) = E_{Y_i;\boldsymbol{\psi}_0}\left\{ -\partial^2 \ell_c^*(Y_i; \boldsymbol{\psi}^*)/\partial \boldsymbol{\psi}^* \partial \boldsymbol{\psi}^{*T} \right\}.$$

Similar definitions are applied to $\tilde{M}^*(\tilde{\boldsymbol{\psi}}_0^*)$ and $\tilde{D}^*(\tilde{\boldsymbol{\psi}}_0^*)$.

4.7.2 Numerical Studies

Here we conduct a simulation to evaluate the impact of misspecification of random effects. In particular, we consider the case that the true distribution for random effects is skewed-normal but the working distribution is assumed to be normal. Skewed-normal distributions have been studied by many authors such as Azzalini (1985), Azzalini and Valle (1996), Azzalini and Capitanio (1999), Arellano-Valle et al. (2005) and Lin and Lee (2008). Such distributions relax the symmetric assumption and provide flexibility to capture a broad range of non-normal features. A p-dimensional random vector u_i follows a skew-normal distribution $SN_p(\mu, D, \boldsymbol{\alpha})$ with location vector μ, dispersion matrix D (a $p \times p$ positive definite matrix) and skewness vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^T$, if its probability density function is given by

$$f(u_i) = 2\phi_p(u_i; \mu, D)\Phi_1\Big\{\boldsymbol{\alpha}^T \mathfrak{D}^{-1/2}(u_i - \mu)\Big\},$$

where $\mathfrak{D}$ is the diagonal matrix with its components extracting from matrix D's diagonal elements, $\phi_p(u_i; \mu, D)$ is the $n-$dimensional normal density function with mean μ and covariance D for u_i and $\Phi_1(\cdot)$ is the cumulative distribution function for the $N(0, 1)$ distribution.

We generate data from the model

$$Y_{ijk} = X_{ijk}^T \boldsymbol{\beta}_0 + Z_{ijk}^T u_{ij} + \epsilon_{ijk}, \tag{4.25}$$

where the residual $\epsilon_i = (\epsilon_{i11}, \dots, \epsilon_{ijk}, \dots, \epsilon_{iJ_iK})^T$ are independently distributed with marginal distribution $N(0, \sigma_\epsilon^2)$, $u_i = (u_{i1}, \dots, u_{ij}, \dots, u_{iJ_i})^T$ are random effects with a distribution specified in following examples, and the residual ϵ_i is independent of the random effects u_i. Set $\sigma_\epsilon^2 = 1$ and $\boldsymbol{\beta}_0 = (1.2, 0.6, 0, 0, 0.8, 0, 0, 0)^T$. Covariates X_{ijk} are generated the same way as in the correct specified model cases. The matrix of Z_{ijk} is set equal to X_{ijk}. For simplicity, the simulation inference only estimates $\boldsymbol{\beta}$ while we set all other parameters to be known.

The simulation study is conducted under following scenarios.

Scenario 1: $n = 250$, $J_i = J = 1$, and $K = 5$. This corresponds to an ordinary longitudinal setting with 5 visits times. Random effects $u_i = u_{i1}$ follow skewed normal distribution $SN_8(\mathbf{0}, D, \boldsymbol{\alpha})$, where D is a diagonal matrix with element to be 4 and $\boldsymbol{\alpha} = (1, -1.2\sqrt{\frac{\pi}{8-\pi}}, 0, 1, -1.6\sqrt{\frac{\pi}{8-\pi}}, 1, 0, 0)^T$.

Scenario 2: The setup is the same as the one in Scenario 1, but we take $n = 1000$.

Scenario 3: $n = 250$, $J_i = J = 3$, and $K = 3$. This corresponds to longitudinal data arising in clusters with 3 subjects setting with 3 visits times. Random effects $u_i = (u_{i1}^T, u_{i2}^T, u_{i3}^T)^T$ follow skewed normal distribution $SN_{24}(\mathbf{0}, \tilde{D}, \tilde{\boldsymbol{\alpha}})$, where $\tilde{D}$ is a diagonal matrix with element to be 9 and $\tilde{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}^T, \boldsymbol{\alpha}^T, \boldsymbol{\alpha}^T)^T$ with $\boldsymbol{\alpha} = (1, -\sqrt{\frac{18\pi}{90-15\pi}}, 0, 1, -\sqrt{\frac{32\pi}{90-15\pi}}, 1, 0, 0)^T$.

Scenario 4: The setup is the same as the one in Scenario 1, but we take $n = 500$.

When we use a misspecified model (non-skewed normal random effects) to estimate the dataset, the solution of equation (4.24) has $\boldsymbol{\beta}_0^*$, which is different from true value $\boldsymbol{\beta}_0$. Table 4.2 displays different values for $\boldsymbol{\beta}$ in the true and the misspecified models.

Table 4.2: The parameter values for the true model and misspecified model for the simulation study

	$X_{ijk,1}$	$X_{ijk,2}$	$X_{ijk,3}$	$X_{ijk,4}$
True Model	$\beta_1 = 1.2$	$\beta_2 = 0.6$	$\beta_3 = 0$	$\beta_4 = 0$
Misspecified Model (Scenario 1 & 2)	$\beta_1^* = 1.2 + \frac{1}{2}\sqrt{\frac{8-\pi}{\pi}}$	$\beta_2^* = 0$	$\beta_3^* = 0$	$\beta_4^* = \frac{1}{2}\sqrt{\frac{8-\pi}{\pi}}$
Misspecified Model (Scenario 3 & 4)	$\beta_1^* = 1.2 + \frac{\sqrt{2}}{10}\sqrt{\frac{90-15\pi}{\pi}}$	$\beta_2^* = 0$	$\beta_3^* = 0$	$\beta_4^* = \frac{\sqrt{2}}{10}\sqrt{\frac{90-15\pi}{\pi}}$

	$X_{ijk,5}$	$X_{ijk,6}$	$X_{ijk,7}$	$X_{ijk,8}$
True Model	$\beta_5 = 0.8$	$\beta_6 = 0$	$\beta_7 = 0$	$\beta_8 = 0$
Misspecified Model (Scenario 1 & 2)	$\beta_5^* = 0$	$\beta_6^* = \frac{1}{2}\sqrt{\frac{8-\pi}{\pi}}$	$\beta_7^* = 0$	$\beta_8^* = 0$
Misspecified Model (Scenario 3 & 4)	$\beta_5^* = 0$	$\beta_6^* = \frac{\sqrt{2}}{10}\sqrt{\frac{90-15\pi}{\pi}}$	$\beta_7^* = 0$	$\beta_8^* = 0$

For each scenario, we repeat the simulation 500 times and fit each dataset by the maximum likelihood (ML), all-pairwise marginal pairwise likelihood (APW) and all-pairwise conditional pairwise likelihood (APC) approaches. Each method is applied with correct skewed-normal random effects (labeled as "$\surd$") and incorrect normal random effects (labeled as "×").

Table 4.24 reports the model selection precision rate for each variable. The columns labeled "RCS-Nonzero" (rate of correct selection of non-zero coefficients) presents the rate of each non-zero coefficient that is correctly estimated as non-zero, and the column labeled "RCS-Zero" (rate of correct selection of zero coefficients) depicts the rate of each zero coefficients that is correctly set to zero.

For both scenarios, under the correct model, all the three methods show a good sparsity property. They often correctly distinguish the zero and non-zero coefficients. As expected, as the sample size increased, the precision improves. However, when a wrong model is implemented, all the three methods show poor selection results. In particular, the erroneous model always leads our methods to make incorrect selection by setting β_2 and β_5 to zero, but taking β_4 and β_6 to non-zero. Associated standard errors for the misspecified model may not increase as the sample size increases.

Table 4.25 summaries the estimates of β_1, its relative biases, empirical standard errors, model-based standard errors, and coverage rates of 95% confidence intervals. It is observed that the estimates of the β_1 have relatively small biases under the correctly specified model as the sample size increases. The misspecified model, on the other hand, yields remarkably biased estimates regardless of the sample size. It is interesting to note that "RCS-Nonzero" for β_1 is always 100 in the simulation we consider; this is partially due to that both β_1 in the true model and β_1^* in the misspecified model are not zero.

Appendices: Proofs of Theoretical Results

A. Regularity Conditions

In this subsection, we list regularity conditions are needed for the subsequent development.

(C1). For all i, $\ell_c(Y_i; \boldsymbol{\psi})$ is three-times continuously differentiable.

(C2). $\ell_c(Y_i;\boldsymbol{\psi})$, $|\frac{\partial \ell_c(Y_i;\boldsymbol{\psi})}{\partial \psi_j}|^2$, $|\frac{\partial^2 \ell_c(Y_i;\boldsymbol{\psi})}{\partial \psi_j \partial \psi_k}|$, and $|\frac{\partial^3 \ell_c(Y_i;\boldsymbol{\psi})}{\partial \psi_j \partial \psi_k \partial \psi_l}|$ are dominated by some functions $B_i(Y_i, X_i, Z_i)$ for all $j, k, l = 1, \dots, \dim(\boldsymbol{\psi})$, in which ψ_j is the $j-$th element of $\boldsymbol{\psi}$. Moreover, $E_{\boldsymbol{\psi}_0}\{B_i(Y_i, X_i, Z_i)\} < \infty$ for all i.

(C3). $E_{\boldsymbol{\psi}}\Big\{\frac{\partial \ell_c(Y_i;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big\} = \mathbf{0}$,

(C4). Let $M(\boldsymbol{\psi}) = E_{\boldsymbol{\psi}_0}\left[\Big\{\frac{\partial}{\partial \boldsymbol{\psi}}\ell_c(Y_i;\boldsymbol{\psi})\Big\}\Big\{\frac{\partial}{\partial \boldsymbol{\psi}}\ell_c(Y_i;\boldsymbol{\psi})\Big\}^T\right]$, and $D(\boldsymbol{\psi}) = E_{\boldsymbol{\psi}_0}\left\{-\frac{\partial^2 \ell_c(Y_i;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^T}\right\}$. Assume that

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{\partial}{\partial \boldsymbol{\psi}}\ell_c(Y_i;\boldsymbol{\psi})\Big\}\Big\{\frac{\partial}{\partial \boldsymbol{\psi}}\ell_c(Y_i;\boldsymbol{\psi})\Big\}^T = M(\boldsymbol{\psi}) + o_p(1),$$

and

$$-\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{\partial^2 \ell_c(Y_i;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^T}\Big\} = D(\boldsymbol{\psi}) + o_p(1).$$

Similar definitions and requirements are applied to $\tilde{M}(\tilde{\boldsymbol{\psi}})$ and $\tilde{D}(\tilde{\boldsymbol{\psi}})$.

(C5). There exists an open subset ω that contains the true parameter point $\boldsymbol{\psi}_0$ such that for all $\boldsymbol{\psi} \in \omega$, $D(\boldsymbol{\psi})$ and $\tilde{D}(\tilde{\boldsymbol{\psi}})$ are positive definite.

(C6). Let λ_n be the tunning parameter with the dependence on cluster size n explicitly spelled out. Define

$$a_n = \max_{s=1,\dots,p}\{p'_{\lambda_n}(|\beta_{s0}|) : \beta_{s0} \neq 0\},$$
$$b_n = \max_{s=1,\dots,p}\{p''_{\lambda_n}(|\beta_{s0}|) : \beta_{s0} \neq 0\},$$

We assume that

(C6.1). $\lambda_n = o_p(1)$,

(C6.2). $a_n = O_p(n^{-1/2})$,

(C6.3). $b_n = o_p(1)$.

(C7). We assume that

(C7.1). $\mathrm{liminf}_{n\to\infty}\mathrm{liminf}_{\epsilon\to 0^+}p'_{\lambda_n}(\epsilon)/\lambda_n > 0$.

(C7.2). $\lim_{n\to\infty}\sqrt{n}\lambda_n = \infty$.

B. Consistency

Proof: Let $\alpha_n = n^{-1/2} + a_n$. Adapting the arguments by Fan and Li (2001, 2002), we need to show that for any given $\epsilon > 0$, there exists a large constant C_ϵ such that

$$P\left\{\sup_{\|\mathbf{u}\|=C_\epsilon} \ell_{pen1}(Y; \boldsymbol{\psi}_0 + \alpha_n \mathbf{u}) < \ell_{pen1}(Y; \boldsymbol{\psi}_0)\right\} \geq 1 - \epsilon,$$

where $\mathbf{u} = ((u_1, \ldots, u_{p_1}, \ldots, u_p)^T, u_\xi^T)^T$, u_ξ is a vector with the same length as ξ, and $\|x\| = \sqrt{x^T x}$.

Suppose C_ϵ is sufficiently large such that $\|(u_1, \ldots, u_{p_1})\| > 0$. Note that $p_{\lambda_n}(0) = 0$, we consider

$$\begin{aligned}
K_n(\mathbf{u}) &= \ell_{pen1}(Y; \boldsymbol{\psi}_0 + \alpha_n \mathbf{u}) - \ell_{pen1}(Y; \boldsymbol{\psi}_0) \\
&= \ell_c(Y; \boldsymbol{\psi}_0 + \alpha_n \mathbf{u}) - \ell_c(Y; \boldsymbol{\psi}_0) - n \sum_{s=1}^{p} p_{\lambda_n}(|\beta_{s0} + \alpha_n u_s|) + n \sum_{s=1}^{p} p_{\lambda_n}(|\beta_{s0}|) \\
&= \ell_c(Y; \boldsymbol{\psi}_0 + \alpha_n \mathbf{u}) - \ell_c(Y; \boldsymbol{\psi}_0) - n \sum_{s=1}^{p_1} p_{\lambda_n}(|\beta_{s0} + \alpha_n u_s|) - n \sum_{s=p_1+1}^{p} p_{\lambda_n}(|0 + \alpha_n u_s|)) \\
&\quad + n \sum_{s=1}^{p_1} p_{\lambda_n}(|\beta_{s0}|) + n \sum_{s=p_1+1}^{p} p_{\lambda_n}(|0|) \\
&\leq \ell_c(Y; \boldsymbol{\psi}_0 + \alpha_n \mathbf{u}) - \ell_c(Y; \boldsymbol{\psi}_0) - n \sum_{s=1}^{p_1} p_{\lambda_n}(|\beta_{s0} + \alpha_n u_s|) + n \sum_{s=1}^{p_1} p_{\lambda_n}(|\beta_{s0}|), \qquad (4.26)
\end{aligned}$$

because of the fact that $n \sum_{s=p_1+1}^{p} p_{\lambda_n}(|0 + \alpha_n u_s|)) \geq 0$.

By the standard argument on the Taylor expansion and the conditions from (C1) and (C2), we obtain

$$\begin{aligned}
\ell_c(Y; \boldsymbol{\psi}_0 + \alpha_n \mathbf{u}) &= \ell_c(Y; \boldsymbol{\psi}_0) + \alpha_n \left\{\frac{\partial \ell_c(Y; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right\}^T \mathbf{u} + \frac{1}{2} \mathbf{u}^T \left\{\frac{\partial^2 \ell_c(Y; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T}\right\} \mathbf{u} \alpha_n^2 \\
&\quad + \sum_{s=1}^{p} O_p(|\alpha_n u_s|^3) \\
&= \ell_c(Y; \boldsymbol{\psi}_0) + \alpha_n \left\{\frac{\partial \ell_c(Y; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\right\}^T \mathbf{u} + \frac{1}{2} \mathbf{u}^T \left\{\frac{\partial^2 \ell_c(Y; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T}\right\} \mathbf{u} \alpha_n^2 \{1 + o_p(1)\} \qquad (4.27)
\end{aligned}$$

and

$$
\begin{aligned}
& n\sum_{s=1}^{p_1}\Big\{p_{\lambda_n}(|\beta_{s0}+\alpha_n u_s|)\Big\} \\
= \; & n\sum_{s=1}^{p_1}p_{\lambda_n}(|\beta_{s0}|)+n\sum_{s=1}^{p_1}\alpha_n p'_{\lambda_n}(|\beta_{s0}|)\mathrm{sgn}(\beta_{s0})u_s+n\sum_{s=1}^{p_1}\alpha_n^2 p''_{\lambda_n}(|\beta_{s0}|)u_s^2\{1+o(1)\}.
\end{aligned} \tag{4.28}
$$

Substituting (4.27)(4.28) into (4.26), we obtain

$$
\begin{aligned}
K_n(\mathbf{u}) \quad \le \quad & \alpha_n\Big\{\frac{\partial \ell_c(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}}\Big\}^T\mathbf{u}+\frac{1}{2}\mathbf{u}^T\Big\{\frac{\partial^2\ell_c(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}\Big\}\mathbf{u}\alpha_n^2\{1+o_p(1)\} \\
& -\sum_{s=1}^{p_1} n\{\alpha_n p'_{\lambda_n}(|\boldsymbol{\beta}_{0s}|)\mathrm{sgn}(\boldsymbol{\beta}_{0s})u_s+\alpha_n^2 p''_{\lambda_n}(|\boldsymbol{\beta}_{0s}|)u_s^2\{1+o(1)\}\} \\
\overset{denote}{=\!=\!=} \quad & \mathcal{A}+\mathcal{B}-\mathcal{C}.
\end{aligned} \tag{4.29}
$$

Now we individually examine $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. By Hölder's inequality, the $\mathcal{A}$ term on the right-hand side of (4.29) is

$$
\begin{aligned}
\alpha_n\Big\{\frac{\partial \ell_c(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}}\Big\}^T\mathbf{u} \;&=\; n^{1/2}\alpha_n n^{-1/2}\Big\{\frac{\partial \ell_c(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}}\Big\}^T\mathbf{u} \\
&\le\; n^{1/2}\alpha_n\left|n^{-1/2}\Big\{\frac{\partial \ell_c(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}}\Big\}^T\mathbf{u}\right| \\
&\le\; n^{1/2}\alpha_n\left\|n^{-1/2}\frac{\partial \ell_c(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}}\right\|\cdot\left\|\mathbf{u}\right\|.
\end{aligned} \tag{4.30}
$$

By (C1), (C2) and (C3), we obtain that, $n^{-1/2}\frac{\partial \ell_c(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}}=O_p(1)$, $\mathcal{A}$ can be bounded by $n^{1/2}\alpha_n\|\mathbf{u}\|$.

For the $\mathcal{B}$ term, since $\frac{1}{n}\Big\{\frac{\partial^2\ell_c(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}\Big\}=O_p(1)$ by (C1) and (C2), we obtain that $\mathbf{u}^T\Big\{\frac{\partial^2\ell_c(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}\Big\}\mathbf{u}\alpha_n^2$ is bounded by $n\alpha_n^2\|\mathbf{u}\|^2$.

For the $\mathcal{C}$ term, we obtain that, using Hölder's inequality,

$$
\sum_{s=1}^{p_1} n\alpha_n p'_{\lambda_n}(|\beta_{s0}|)\mathrm{sgn}(\beta_{s0})u_s\le n\alpha_n a_n\Big|\sum_{s=1}^{p_1}u_s\Big|\le n\alpha_n a_n\|\mathbf{u}\|\cdot\|\mathbf{1}\|=\sqrt{p_1}n\alpha_n a_n\|\mathbf{u}\|.
$$

Furthermore, by the definition of b_n, we obtain

$$\sum_{s=1}^{p_1} n\alpha_n^2 p''_{\lambda_n}(|\boldsymbol{\beta}_{0s}|)u_s^2\{1+o(1)\} \le n\alpha_n^2 b_n\|\mathbf{u}\|^2\{1+o(1)\}.$$

Note that $n\alpha_n a_n = O_p(n\alpha_n^2)$, and $b_n = o_p(1)$ by (C6.3), therefore, term $\mathcal{C}$ is bounded by $n\alpha_n a_n\|\mathbf{u}\|$.

Since $a_n = O_p(n^{-1/2})$ from (C6.2), all $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are of the order $O_p(n\alpha_n^2)$. If $\|\mathbf{u}\| = C_\epsilon$ is sufficiently large, then $\mathcal{B}$ dominates $\mathcal{A}$ and $\mathcal{C}$. Moreover, by (C4)-(C5), $D(\boldsymbol{\psi}_0)$ is positive definite, then we have

$$P\Big\{\sup_{\|\mathbf{u}\|=C_\epsilon} K_n(\mathbf{u}) < 0\Big\} = P\left\{\sup_{\|\mathbf{u}\|=C_\epsilon} \ell_{pen1}(Y;\boldsymbol{\psi}_0+\alpha_n\mathbf{u}) < \ell_{pen1}(Y;\boldsymbol{\psi}_0)\right\} \ge 1-\epsilon,$$

which indicates at least $1-\epsilon$ that there exists a local maximum in $\{\boldsymbol{\psi}_0+\alpha_n\mathbf{u}\}$. Hence, there exists a local maximizer such that $\|\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0\| = O_p(\alpha_n)$.

C. Sparsity

Proof: By Theorem 1, it suffices to show that with probability tending to 1 as $n\to\infty$, for any given $\boldsymbol{\beta}_I$ satisfying $\|\boldsymbol{\beta}_I-\boldsymbol{\beta}_{0I}\| = O_p(n^{-1/2})$, ξ satisfying $\|\xi-\xi_0\| = O_p(n^{-1/2})$, and for $\epsilon_n = Cn^{-1/2}$, and $s = p_1+1,\ldots,p$, we have

$$\frac{\partial\ell_{pen1}(Y;\boldsymbol{\psi})}{\partial\beta_s} < 0 \quad \text{for } 0<\beta_s<\epsilon_n,$$

and

$$\frac{\partial\ell_{pen1}(Y;\boldsymbol{\psi})}{\partial\beta_s} > 0 \quad \text{for } -\epsilon_n<\beta_s<0.$$

With Taylor Series expansion, we obtain

$$\begin{aligned}\frac{\partial\ell_{pen1}(Y;\boldsymbol{\psi})}{\partial\beta_s} &= \frac{\partial\ell_c(Y;\boldsymbol{\psi})}{\partial\beta_s} - np'_{\lambda_n}(|\beta_s|)\mathrm{sgn}(\beta_s)\\ &= \frac{\partial\ell_c(Y;\boldsymbol{\psi}_0)}{\partial\beta_s} + \Big\{\frac{\partial^2\ell_c(Y;\boldsymbol{\psi}_0)}{\partial\beta_s\partial\boldsymbol{\psi}}\Big\}^T(\boldsymbol{\psi}-\boldsymbol{\psi}_0)\\ &\quad +(\boldsymbol{\psi}-\boldsymbol{\psi}_0)^T\Big\{\frac{\partial^3\ell_{pen1}(Y;\dot{\boldsymbol{\psi}})}{\partial\beta_s\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}\Big\}(\boldsymbol{\psi}-\boldsymbol{\psi}_0) - np'_{\lambda_n}(|\beta_s|)\mathrm{sgn}(\beta_s)\\ &\overset{denote}{=\!=} \mathcal{A}+\mathcal{B}+\mathcal{C} - np'_{\lambda_n}(|\beta_s|)\mathrm{sgn}(\beta_s)\end{aligned}$$

where $\dot{\boldsymbol{\psi}}$ lies "between" $\boldsymbol{\psi}$ and $\boldsymbol{\psi}_0$. By the assumption that $\|\boldsymbol{\psi} - \boldsymbol{\psi}_0\| = O_p(n^{-1/2})$, then it follows that

$$\mathcal{A} = O_p(n^{1/2}), \qquad \mathcal{B} = O_p(n^{1/2}), \qquad \mathcal{C} = O_p(1),$$

and thus

$$(n\lambda_n)^{-1}\mathcal{A} = O_p(n^{-1/2}/\lambda_n), \qquad (n\lambda_n)^{-1}\mathcal{B} = O_p(n^{-1/2}/\lambda_n), \quad \text{and } (n\lambda_n)^{-1}\mathcal{C} = O_p(n^{-1}/\lambda_n).$$

As a result, we obtain

$$\begin{aligned} \frac{\partial \ell_{pen1}(Y;\boldsymbol{\psi})}{\partial \beta_s} &= n\lambda_n\{(n\lambda_n)^{-1}(\mathcal{A}+\mathcal{B}+\mathcal{C}) - \lambda_n^{-1}p'_{\lambda_n}(|\beta_s|)\text{sgn}(\beta_s)\} \\ &= n\lambda_n\{O_p(n^{-1/2}/\lambda_n) - \lambda_n^{-1}p'_{\lambda_n}(|\beta_s|)\text{sgn}(\beta_s)\}. \qquad (4.31) \end{aligned}$$

By the regularity condition (C6), $\text{liminf}_{n\to\infty}\text{liminf}_{\epsilon\to 0^+}p'_{\lambda_n}(\epsilon)/\lambda_n > 0$ and $\lim_{n\to\infty}\sqrt{n}\lambda_n = \infty$, the sign of the derivative in (4.31) is determined by β_s. Thus we have

$$\frac{\partial \ell_{pen1}(Y;\boldsymbol{\psi})}{\partial \beta_s} < 0 \quad \text{for } 0 < \beta_s < \epsilon_n,$$

and

$$\frac{\partial \ell_{pen1}(Y;\boldsymbol{\psi})}{\partial \beta_s} > 0 \quad \text{for } -\epsilon_n < \beta_s < 0.$$

This completes the proof.

D. Asymptotic Distribution

Proof: Part (a) follows from Theorem 1 and Theorem 2. Now we show part (b). By Theorem 1 and Theorem 2, there exists a $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}_I, \mathbf{0}, \hat{\xi})$ that is a root-n consistent local maximizer of $\ell_{pen1}(Y;\boldsymbol{\psi})$, and that satisfies

$$\frac{\partial \ell_{pen1}(Y;\tilde{\boldsymbol{\psi}})}{\partial \tilde{\boldsymbol{\psi}}}\Big|_{\tilde{\boldsymbol{\psi}}=\tilde{\hat{\boldsymbol{\psi}}}} = \mathbf{0}.$$

By Taylor Series expansion, we obtain

$$\frac{\partial \ell_c(Y;\tilde{\boldsymbol{\psi}}_0)}{\partial \tilde{\boldsymbol{\psi}}} + \left\{\frac{\partial^2 \ell_c(Y;\tilde{\boldsymbol{\psi}}_0)}{\partial \tilde{\boldsymbol{\psi}}\partial \tilde{\boldsymbol{\psi}}^T} + o_p(1)\right\}(\tilde{\hat{\boldsymbol{\psi}}} - \tilde{\boldsymbol{\psi}}_0) - n\Big\{\tilde{\mathbf{b}} + \{\tilde{\Sigma} + o_p(1)\}(\tilde{\hat{\boldsymbol{\psi}}} - \tilde{\boldsymbol{\psi}}_0)\Big\} = \mathbf{0}.$$

Thus, we obtain

$$-\frac{1}{\sqrt{n}}\left\{\frac{\partial^2\ell_c(Y;\tilde{\boldsymbol{\psi}}_0)}{\partial\tilde{\boldsymbol{\psi}}\partial\tilde{\boldsymbol{\psi}}^T}+o_p(1)\right\}(\hat{\tilde{\boldsymbol{\psi}}}-\tilde{\boldsymbol{\psi}}_0)+\sqrt{n}\Big[\tilde{\mathbf{b}}+\{\tilde{\Sigma}+o_p(1)\}(\hat{\tilde{\boldsymbol{\psi}}}-\tilde{\boldsymbol{\psi}}_0)\Big]=\frac{1}{\sqrt{n}}\frac{\partial\ell_c(Y;\tilde{\boldsymbol{\psi}})}{\partial\tilde{\boldsymbol{\psi}}}.$$

Applying Slusky's theorem and the Central Limiting Theorem, we obtain

$$\sqrt{n}\{\tilde{D}(\tilde{\boldsymbol{\psi}}_0)(\hat{\tilde{\boldsymbol{\psi}}}-\tilde{\boldsymbol{\psi}}_0)+\tilde{\mathbf{b}}+\tilde{\Sigma}(\hat{\tilde{\boldsymbol{\psi}}}-\tilde{\boldsymbol{\psi}}_0)\}\rightarrow_D N(\mathbf{0},\tilde{M}(\tilde{\boldsymbol{\psi}}_0)),$$

i.e.

$$\sqrt{n}\Big[\{\tilde{D}(\tilde{\boldsymbol{\psi}}_0)+\tilde{\Sigma}\}(\hat{\tilde{\boldsymbol{\psi}}}-\tilde{\boldsymbol{\psi}}_0)+\tilde{\mathbf{b}}\Big]\rightarrow_D N(\mathbf{0},\tilde{M}(\tilde{\boldsymbol{\psi}}_0)).$$

E. Regularity Conditions for Misspecified Model

In this subsection, we list regularity conditions that are needed for the subsequent development.

(C1). For all i, $\ell_c^*(Y_i;\boldsymbol{\psi}^*)$ is three-times continuously differentiable.

(C2). $\ell_c^*(Y_i;\boldsymbol{\psi}^*)$, $|\frac{\partial\ell_c^*(Y_i;\boldsymbol{\psi}^*)}{\partial\psi_j^*}|^2$, $|\frac{\partial^2\ell_c^*(Y_i;\boldsymbol{\psi}^*)}{\partial\psi_j^*\partial\psi_k^*}|$, and $|\frac{\partial^3\ell_c^*(Y_i;\boldsymbol{\psi}^*)}{\partial\psi_j^*\partial\psi_k^*\partial\psi_l^*}|$ are dominated by some functions $B_i(Y_i,X_i,Z_i)$ for all $j,k,l=1,\ldots,\dim(\boldsymbol{\psi}^*)$, in which ψ_j^* is the $j-$th element of $\boldsymbol{\psi}^*$. Moreover, $E_{Y_i;\boldsymbol{\psi}_0}\{B_i(Y_i,X_i,Z_i)\}<\infty$ for all i.

(C3). The solution for $\boldsymbol{\psi}^*$ in the equation $E_{Y_i;\boldsymbol{\psi}_0}\Big\{\frac{\partial\ell_c^*(Y_i;\boldsymbol{\psi}^*)}{\partial\boldsymbol{\psi}^*}\Big\}=\mathbf{0}$ is $\boldsymbol{\psi}_0^*$.

(C4). Let $M^*(\boldsymbol{\psi}^*)=E_{Y_i;\boldsymbol{\psi}_0}\left[\Big\{\frac{\partial}{\partial\boldsymbol{\psi}^*}\ell_c^*(Y_i;\boldsymbol{\psi}^*)\Big\}\Big\{\frac{\partial}{\partial\boldsymbol{\psi}^*}\ell_c^*(Y_i;\boldsymbol{\psi}^*)\Big\}^T\right]$, and

$$D^*(\boldsymbol{\psi}^*)=E_{Y_i;\boldsymbol{\psi}_0}\left\{-\frac{\partial^2\ell_c^*(Y_i;\boldsymbol{\psi}^*)}{\partial\boldsymbol{\psi}^*\partial\boldsymbol{\psi}^{*T}}\right\}.$$

Assume that

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{\partial}{\partial\boldsymbol{\psi}^*}\ell_c^*(Y_i;\boldsymbol{\psi}^*)\Big\}\Big\{\frac{\partial}{\partial\boldsymbol{\psi}^*}\ell_c^*(Y_i;\boldsymbol{\psi}^*)\Big\}^T=M^*(\boldsymbol{\psi}^*)+o_p(1),$$

and

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial^2 \ell_c^*(Y_i;\boldsymbol{\psi}^*)}{\partial\boldsymbol{\psi}^*\partial\boldsymbol{\psi}^{*T}}\right\} = D^*(\boldsymbol{\psi}^*) + o_p(1).$$

Similar definitions and requirements are applied to $\tilde{M}^*(\tilde{\boldsymbol{\psi}}^*)$ and $\tilde{D}^*(\tilde{\boldsymbol{\psi}}^*)$.

(C5). There exists an open subset ω that contains the parameter point $\boldsymbol{\psi}_0^*$ such that for all $\boldsymbol{\psi}^* \in \omega$, $D^*(\boldsymbol{\psi}^*)$ and $\tilde{D}^*(\boldsymbol{\psi}^*)$ are positive definite.

(C6). Let λ_n be the tunning parameter with the dependence on cluster size n explicitly spelled out. Define

$$a_n = \max_{s=1,\ldots,p}\{p'_{\lambda_n}(|\beta_{s0}^*|) : \beta_{s0}^* \neq 0\},$$

$$b_n = \max_{s=1,\ldots,p}\{p''_{\lambda_n}(|\beta_{s0}^*|) : \beta_{s0}^* \neq 0\},$$

We assume that

(C6.1). $\lambda_n = o_p(1)$,

(C6.2). $a_n = O_p(n^{-1/2})$,

(C6.3). $b_n = o_p(1)$.

(C7). We assume that

(C7.1). $\text{liminf}_{n\to\infty}\text{liminf}_{\epsilon\to 0^+} p'_{\lambda_n}(\epsilon)/\lambda_n > 0$.

(C7.2). $\lim_{n\to\infty}\sqrt{n}\lambda_n = \infty$.

F. Consistency under Misspecified Model

Proof: Let $\alpha_n = n^{-1/2} + a_n$. We need to show that for any given $\epsilon > 0$, there exists a large constant C_ϵ such that

$$P\left\{\sup_{\|\mathbf{u}\|=C_\epsilon} \ell_{pen1}^*(Y;\boldsymbol{\psi}_0^* + \alpha_n\mathbf{u}) < \ell_{pen1}^*(Y;\boldsymbol{\psi}_0^*)\right\} \geq 1-\epsilon,$$

where $\mathbf{u} = ((u_1, \ldots, u_{p_1^*}, \ldots, u_p)^T, u_{\xi^*}^T)^T$, u_{ξ^*} is a vector with the same length as ξ^*, and $\|x\| = \sqrt{x^T x}$.

Suppose C_ϵ is sufficiently large such that $\|(u_1, \ldots, u_{p_1^*})\| > 0$. Note that $p_{\lambda_n}(0) = 0$, we consider

$$\begin{aligned}
K_n(\mathbf{u}) &= \ell^*_{pen1}(Y; \boldsymbol{\psi}^*_0 + \alpha_n \mathbf{u}) - \ell^*_{pen1}(Y; \boldsymbol{\psi}^*_0) \\
&= \ell^*_c(Y; \boldsymbol{\psi}^*_0 + \alpha_n \mathbf{u}) - \ell^*_c(Y; \boldsymbol{\psi}^*_0) - n \sum_{s=1}^{p} p_{\lambda_n}(|\beta^*_{s0} + \alpha_n u_s|) + n \sum_{s=1}^{p} p_{\lambda_n}(|\beta^*_{s0}|) \\
&= \ell^*_c(Y; \boldsymbol{\psi}^*_0 + \alpha_n \mathbf{u}) - \ell^*_c(Y; \boldsymbol{\psi}^*_0) - n \sum_{s=1}^{p_1^*} p_{\lambda_n}(|\beta^*_{s0} + \alpha_n u_s|) - n \sum_{s=p_1^*+1}^{p} p_{\lambda_n}(|0 + \alpha_n u_s|)) \\
&\quad + n \sum_{s=1}^{p_1^*} p_{\lambda_n}(|\beta^*_{s0}|) + n \sum_{s=p_1^*+1}^{p} p_{\lambda_n}(|0|) \\
&\leq \ell^*_c(Y; \boldsymbol{\psi}^*_0 + \alpha_n \mathbf{u}) - \ell^*_c(Y; \boldsymbol{\psi}^*_0) - n \sum_{s=1}^{p_1^*} p_{\lambda_n}(|\beta^*_{s0} + \alpha_n u_s|) + n \sum_{s=1}^{p_1^*} p_{\lambda_n}(|\beta^*_{s0}|), \qquad (4.32)
\end{aligned}$$

because of the fact that $n \sum_{s=p_1^*+1}^{p} p_{\lambda_n}(|0 + \alpha_n u_s|)) \geq 0$.

By the standard argument on the Taylor expansion and the conditions from (C1) and (C2), we obtain

$$\begin{aligned}
\ell^*_c(Y; \boldsymbol{\psi}^*_0 + \alpha_n \mathbf{u}) &= \ell^*_c(Y; \boldsymbol{\psi}^*_0) + \alpha_n \left\{ \frac{\partial \ell^*_c(Y; \boldsymbol{\psi}^*_0)}{\partial \boldsymbol{\psi}^*} \right\}^T \mathbf{u} + \frac{1}{2} \mathbf{u}^T \left\{ \frac{\partial^2 \ell^*_c(Y; \boldsymbol{\psi}^*_0)}{\partial \boldsymbol{\psi}^* \partial \boldsymbol{\psi}^{*T}} \right\} \mathbf{u} \alpha_n^2 \\
&\quad + \sum_{s=1}^{p} O_p(|\alpha_n u_s|^3) \\
&= \ell^*_c(Y; \boldsymbol{\psi}^*_0) + \alpha_n \left\{ \frac{\partial \ell^*_c(Y; \boldsymbol{\psi}^*_0)}{\partial \boldsymbol{\psi}^*} \right\}^T \mathbf{u} + \frac{1}{2} \mathbf{u}^T \left\{ \frac{\partial^2 \ell^*_c(Y; \boldsymbol{\psi}^*_0)}{\partial \boldsymbol{\psi}^* \partial \boldsymbol{\psi}^{*T}} \right\} \mathbf{u} \alpha_n^2 \{1 + o_p(1)\} \\
&\qquad (4.33)
\end{aligned}$$

and

$$
\begin{aligned}
& n\sum_{s=1}^{p_1^*}\Big\{p_{\lambda_n}(|\beta_{s0}^*+\alpha_n u_s|)\Big\} \\
= \; & n\sum_{s=1}^{p_1^*}p_{\lambda_n}(|\beta_{s0}^*|)+n\sum_{s=1}^{p_1^*}\alpha_n p'_{\lambda_n}(|\beta_{s0}^*|)\mathrm{sgn}(\beta_{s0}^*)u_s+n\sum_{s=1}^{p_1^*}\alpha_n^2 p''_{\lambda_n}(|\beta_{s0}^*|)u_s^2\{1+o(1)\}.
\end{aligned} \tag{4.34}
$$

Substituting (4.33)(4.34) into (4.32), we obtain

$$
\begin{aligned}
K_n(\mathbf{u}) \quad \leq \quad & \alpha_n\Big\{\frac{\partial\ell_c^*(Y;\boldsymbol{\psi}_0^*)}{\partial\boldsymbol{\psi}^*}\Big\}^T\mathbf{u}+\frac{1}{2}\mathbf{u}^T\Big\{\frac{\partial^2\ell_c^*(Y;\boldsymbol{\psi}_0^*)}{\partial\boldsymbol{\psi}^*\partial\boldsymbol{\psi}^{*T}}\Big\}\mathbf{u}\alpha_n^2\{1+o_p(1)\} \\
& -\sum_{s=1}^{p_1^*}n\{\alpha_n p'_{\lambda_n}(|\beta_{s0}^*|)\mathrm{sgn}(\beta_{s0}^*)u_s+\alpha_n^2 p''_{\lambda_n}(|\beta_{s0}^*|)u_s^2\{1+o(1)\}\} \\
\overset{denote}{=\!=\!=} \quad & \mathcal{A}+\mathcal{B}-\mathcal{C}.
\end{aligned} \tag{4.35}
$$

Now we individually examine $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. By Hölder's inequality, the $\mathcal{A}$ term on the right-hand side of (4.35) is

$$
\begin{aligned}
\alpha_n\Big\{\frac{\partial\ell_c^*(Y;\boldsymbol{\psi}_0^*)}{\partial\boldsymbol{\psi}^*}\Big\}^T\mathbf{u} \quad = \quad & n^{1/2}\alpha_n n^{-1/2}\Big\{\frac{\partial\ell_c^*(Y;\boldsymbol{\psi}_0^*)}{\partial\boldsymbol{\psi}^*}\Big\}^T\mathbf{u} \\
\leq \quad & n^{1/2}\alpha_n\left|n^{-1/2}\Big\{\frac{\partial\ell_c^*(Y;\boldsymbol{\psi}_0^*)}{\partial\boldsymbol{\psi}^*}\Big\}^T\mathbf{u}\right| \\
\leq \quad & n^{1/2}\alpha_n\left\|n^{-1/2}\frac{\partial\ell_c^*(Y;\boldsymbol{\psi}_0^*)}{\partial\boldsymbol{\psi}^*}\right\|\cdot\Big\|\mathbf{u}\Big\|.
\end{aligned} \tag{4.36}
$$

By (C1), (C2) and (C3), we obtain that, $n^{-1/2}\frac{\partial\ell_c^*(Y;\boldsymbol{\psi}_0^*)}{\partial\boldsymbol{\psi}^*}=O_p(1)$, $\mathcal{A}$ can be bounded by $n^{1/2}\alpha_n\|\mathbf{u}\|$.

For the $\mathcal{B}$ term, since $\frac{1}{n}\Big\{\frac{\partial^2\ell_c^*(Y;\boldsymbol{\psi}_0^*)}{\partial\boldsymbol{\psi}^*\partial\boldsymbol{\psi}^{*T}}\Big\}=O_p(1)$ by (C1) and (C2), we obtain that $\mathbf{u}^T\Big\{\frac{\partial^2\ell_c^*(Y;\boldsymbol{\psi}_0^*)}{\partial\boldsymbol{\psi}^*\partial\boldsymbol{\psi}^{*T}}\Big\}\mathbf{u}\alpha_n^2$ is bounded by $n\alpha_n^2\|\mathbf{u}\|^2$.

For the $\mathcal{C}$ term, we obtain that, using Hölder's inequality,

$$
\sum_{s=1}^{p_1^*}n\alpha_n p'_{\lambda_n}(|\beta_{s0}^*|)\mathrm{sgn}(\beta_{s0}^*)u_s\leq n\alpha_n a_n\Big|\sum_{s=1}^{p_1^*}u_s\Big|\leq n\alpha_n a_n\|\mathbf{u}\|\cdot\|\mathbf{1}\|=\sqrt{p_1^*}n\alpha_n a_n\|\mathbf{u}\|.
$$

Furthermore, by the definition of b_n, we obtain

$$\sum_{s=1}^{p_1^*} n\alpha_n^2 p''_{\lambda_n}(|\beta_{s0}^*|)u_s^2\{1+o(1)\} \leq n\alpha_n^2 b_n \|\mathbf{u}\|^2\{1+o(1)\}.$$

Note that $n\alpha_n a_n = O_p(n\alpha_n^2)$, and $b_n = o_p(1)$ by (C6.3). Therefore, term $\mathcal{C}$ is bounded by $n\alpha_n a_n\|\mathbf{u}\|$.

Since $a_n = O_p(n^{-1/2})$ from (C6.2), all $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are of the order $O_p(n\alpha_n^2)$. If $\|\mathbf{u}\| = C_\epsilon$ is sufficiently large, then $\mathcal{B}$ dominates $\mathcal{A}$ and $\mathcal{C}$. Moreover, by (C4)-(C5), $D^*(\boldsymbol{\psi}_0^*)$ is positive definite, then we have

$$P\Big\{\sup_{\|\mathbf{u}\|=C_\epsilon} K_n(\mathbf{u}) < 0\Big\} = P\left\{\sup_{\|\mathbf{u}\|=C_\epsilon} \ell_{pen1}^*(Y;\boldsymbol{\psi}_0^* + \alpha_n\mathbf{u}) < \ell_{pen1}^*(Y;\boldsymbol{\psi}_0^*)\right\} \geq 1-\epsilon,$$

which indicates at least $1-\epsilon$ that there exists a local maximum in $\{\boldsymbol{\psi}_0^* + \alpha_n\mathbf{u}\}$. Hence, there exists a local maximizer such that $\|\hat{\boldsymbol{\psi}}^* - \boldsymbol{\psi}_0^*\| = O_p(\alpha_n)$.

G. Sparsity under Misspecified Model

Proof: By Theorem 4, it suffices to show that with probability tending to 1 as $n \to \infty$, for any given $\boldsymbol{\beta}_I^*$ satisfying $\|\boldsymbol{\beta}_I^* - \boldsymbol{\beta}_{0I}^*\| = O_p(n^{-1/2})$, ξ^* satisfying $\|\xi^* - \xi_0^*\| = O_p(n^{-1/2})$. Then for $\epsilon_n = Cn^{-1/2}$, and $s = p_1^*+1,\ldots,p$, we have to prove

$$\frac{\partial \ell_{pen1}^*(Y;\boldsymbol{\psi}^*)}{\partial \beta_s^*} < 0 \quad \text{for } 0 < \beta_s^* < \epsilon_n,$$

and

$$\frac{\partial \ell_{pen1}^*(Y;\boldsymbol{\psi}^*)}{\partial \beta_s^*} > 0 \quad \text{for } -\epsilon_n < \beta_s^* < 0.$$

With Taylor Series expansion, we obtain

$$\begin{aligned}\frac{\partial \ell^*_{pen1}(Y;\boldsymbol{\psi}^*)}{\partial \beta^*_s} &= \frac{\partial \ell^*_c(Y;\boldsymbol{\psi}^*)}{\partial \beta^*_s} - np'_{\lambda_n}(|\beta^*_s|)\mathrm{sgn}(\beta^*_s) \\ &= \frac{\partial \ell^*_c(Y;\boldsymbol{\psi}^*_0)}{\partial \beta^*_s} + \Big\{\frac{\partial^2 \ell^*_c(Y;\boldsymbol{\psi}^*_0)}{\partial \beta^*_s \partial \boldsymbol{\psi}^*}\Big\}^T(\boldsymbol{\psi}^* - \boldsymbol{\psi}^*_0) \\ &\quad +(\boldsymbol{\psi}^* - \boldsymbol{\psi}^*_0)^T\Big\{\frac{\partial^3 \ell^*_{pen1}(Y;\dot{\boldsymbol{\psi}}^*)}{\partial \beta^*_s \partial \boldsymbol{\psi}^* \partial \boldsymbol{\psi}^{*T}}\Big\}(\boldsymbol{\psi}^* - \boldsymbol{\psi}^*_0) - np'_{\lambda_n}(|\beta^*_s|)\mathrm{sgn}(\beta^*_s) \\ &\overset{denote}{=\!=} \mathcal{A} + \mathcal{B} + \mathcal{C} - np'_{\lambda_n}(|\beta^*_s|)\mathrm{sgn}(\beta^*_s)\end{aligned}$$

where $\dot{\boldsymbol{\psi}}^*$ lies “between” $\boldsymbol{\psi}^*$ and $\boldsymbol{\psi}^*_0$. By the assumption that $\|\boldsymbol{\psi}^* - \boldsymbol{\psi}^*_0\| = O_p(n^{-1/2})$, then it follows that

$$\mathcal{A} = O_p(n^{1/2}), \qquad \mathcal{B} = O_p(n^{1/2}), \qquad \mathcal{C} = O_p(1),$$

and thus

$$(n\lambda_n)^{-1}\mathcal{A} = O_p(n^{-1/2}/\lambda_n), \qquad (n\lambda_n)^{-1}\mathcal{B} = O_p(n^{-1/2}/\lambda_n), \quad \text{and } (n\lambda_n)^{-1}\mathcal{C} = O_p(n^{-1}/\lambda_n).$$

As a result, we obtain

$$\begin{aligned}\frac{\partial \ell^*_{pen1}(Y;\boldsymbol{\psi}^*)}{\partial \beta^*_s} &= n\lambda_n\{(n\lambda_n)^{-1}(\mathcal{A}+\mathcal{B}+\mathcal{C}) - \lambda_n^{-1}p'_{\lambda_n}(|\beta^*_s|)\mathrm{sgn}(\beta^*_s)\} \\ &= n\lambda_n\{O_p(n^{-1/2}/\lambda_n) - \lambda_n^{-1}p'_{\lambda_n}(|\beta^*_s|)\mathrm{sgn}(\beta^*_s)\}. \qquad (4.37)\end{aligned}$$

By the regularity condition (C6), $\mathrm{liminf}_{n\to\infty}\mathrm{liminf}_{\epsilon\to 0+}p'_{\lambda_n}(\epsilon)/\lambda_n > 0$ and $\lim_{n\to\infty}\sqrt{n}\lambda_n = \infty$, the sign of the derivative in (4.37) is determined by β^*_s. Thus we have

$$\frac{\partial \ell^*_{pen1}(Y;\boldsymbol{\psi}^*)}{\partial \beta^*_s} < 0 \quad \text{for } 0 < \beta^*_s < \epsilon_n,$$

and

$$\frac{\partial \ell^*_{pen1}(Y;\boldsymbol{\psi}^*)}{\partial \beta^*_s} > 0 \quad \text{for } -\epsilon_n < \beta^*_s < 0.$$

This completes the proof.

H. Asymptotic Distribution under Misspecified Model

Proof: Part (a) follows from Theorem 4 and Theorem 5. Now we show part (b). By Theorem 4 and Theorem 5, there exists a $\hat{\boldsymbol{\psi}}^* = (\hat{\boldsymbol{\beta}}_I^*, \mathbf{0}, \hat{\xi}^*)$ that is a root-n consistent local maximizer of $\ell^*_{pen1}(Y; \boldsymbol{\psi}^*)$, and that satisfies

$$\frac{\partial \ell^*_{pen1}(Y; \tilde{\boldsymbol{\psi}}^*)}{\partial \tilde{\boldsymbol{\psi}}^*}\Big|_{\tilde{\boldsymbol{\psi}}^* = \tilde{\hat{\boldsymbol{\psi}}}^*} = \mathbf{0}.$$

By Taylor Series expansion, we obtain

$$\frac{\partial \ell_c^*(Y; \tilde{\boldsymbol{\psi}}_0^*)}{\partial \tilde{\boldsymbol{\psi}}^*} + \left\{ \frac{\partial^2 \ell_c^*(Y; \tilde{\boldsymbol{\psi}}_0^*)}{\partial \tilde{\boldsymbol{\psi}}^* \partial \tilde{\boldsymbol{\psi}}^{*T}} + o_p(1) \right\} (\tilde{\hat{\boldsymbol{\psi}}}^* - \tilde{\boldsymbol{\psi}}_0^*) - n \Big\{ \tilde{\mathbf{b}}^* + \{\tilde{\Sigma}^* + o_p(1)\}(\tilde{\hat{\boldsymbol{\psi}}}^* - \tilde{\boldsymbol{\psi}}_0^*) \Big\} = \mathbf{0}.$$

Thus, we obtain

$$\frac{1}{\sqrt{n}} \left\{ \frac{\partial^2 \ell_c^*(Y; \tilde{\boldsymbol{\psi}}_0^*)}{\partial \tilde{\boldsymbol{\psi}}^* \partial \tilde{\boldsymbol{\psi}}^{*T}} + o_p(1) \right\} (\tilde{\hat{\boldsymbol{\psi}}}^* - \tilde{\boldsymbol{\psi}}_0^*) - \sqrt{n} \Big[\tilde{\mathbf{b}}^* + \{\tilde{\Sigma}^* + o_p(1)\}(\tilde{\hat{\boldsymbol{\psi}}}^* - \tilde{\boldsymbol{\psi}}_0^*) \Big] = -\frac{1}{\sqrt{n}} \frac{\partial \ell_c^*(Y; \tilde{\boldsymbol{\psi}}^*)}{\partial \boldsymbol{\psi}^*}.$$

Applying Slusky's theorem and the Central Limiting Theorem, we obtain

$$\sqrt{n}\{\tilde{D}^*(\tilde{\boldsymbol{\psi}}_0^*)(\tilde{\hat{\boldsymbol{\psi}}}^* - \tilde{\boldsymbol{\psi}}_0^*) + \tilde{\mathbf{b}}^* + \tilde{\Sigma}^*(\tilde{\hat{\boldsymbol{\psi}}}^* - \tilde{\boldsymbol{\psi}}_0^*)\} \rightarrow_D N(\mathbf{0}, \tilde{M}^*(\tilde{\boldsymbol{\psi}}_0^*)),$$

i.e.

$$\sqrt{n}\Big[\{\tilde{D}^*(\tilde{\boldsymbol{\psi}}_0^*) + \tilde{\Sigma}^*\}(\tilde{\hat{\boldsymbol{\psi}}}^* - \tilde{\boldsymbol{\psi}}_0^*) + \tilde{\mathbf{b}}^* \Big] \rightarrow_D N(\mathbf{0}, \tilde{M}^*(\tilde{\boldsymbol{\psi}}_0^*)).$$

Appendices: Simulation Results

Simulation for Selecting Fixed Effects

Linear Mixed Model

Table 4.3: Simulation results for the fixed effects selection under linear mixed model: model selection

				Avg. No. of 0 Coefficients	
	Method	R.MME(%)[‡]	1000×M.MME	Correct[*]	Incorrect[**]
Example 1 $n = 60, J = 1, K = 5$	$\text{ML}^{\dagger}_{(a,\lambda)}$	47.770	13.423	4.738	0
	ML_{λ}	48.230	13.553	4.735	0
	$\text{APW}_{(a,\lambda)}$	54.311	16.968	4.614	0
	APW_{λ}	55.270	17.301	4.598	0
	$\text{APC}_{(a,\lambda)}$	52.870	14.907	4.637	0
	APC_{λ}	53.646	15.225	4.629	0
Example 2 $n = 500, J = 1, K = 5$	$\text{ML}_{(a,\lambda)}$	34.297	1.153	5	0
	ML_{λ}	34.699	1.152	5	0
	$\text{APW}_{(a,\lambda)}$	33.859	1.250	5	0
	APW_{λ}	34.543	1.250	5	0
	$\text{APC}_{(a,\lambda)}$	34.340	1.145	5	0
	APC_{λ}	34.679	1.152	5	0
Example 3 $n = 60, J_i = 3, K = 3$	ML_{λ}	45.515	7.186	4.806	0
	APW_{λ}	48.875	10.019	4.682	0
	APC_{λ}	46.875	8.613	4.747	0
Example 4 $n = 300, J_i = 3, K = 3$	ML_{λ}	33.806	1.103	5	0
	APW_{λ}	34.303	1.442	4.999	0
	APC_{λ}	33.620	1.241	5	0

[†] ML, APW, APC represent maximum likelihood, all-pairwise marginal pairwise likelihood, all-pairwise conditional pairwise likelihood, respectively. $(a, \lambda), \lambda$ denote the tuning parameter selection by both a, λ and only searching λ with fixing $a = 3.7$, respectively.

[‡] R.MME represents the median of ratios of MME of a selected model to that of the un-penalized estimate under the full model in ML, APW, APC methods, respectively. M.MME denotes the median of MME for selected models in ML, APW and APC scenarios.

[*] "Correct" presents the average restricted to the true zero coefficients. 0 represents no true zero coefficient is shrink, while 5 implies all true zero coefficients are restricted into zero.

[**] "Incorrect" depicts the average of significant coefficients erroneously set to zero. 0 represents no significant coefficient is shrink, while 3 implies all significant coefficients are erroneously set to zero.

Table 4.4: Simulation results for the fixed effects selection under linear mixed model: estimation of selected regression coefficients $\boldsymbol{\beta}$

		β_1				β_2				β_5			
	Method	Bias(%)*	ESE‡	ASE$^♯$	CP(%)$^♭$	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)
Example 1 $n = 60, J = 1, K = 5$	ML$^†_{(a,\lambda)}$	-0.268	0.076	0.072	94.2	-0.805	0.079	0.072	93.7	-0.758	0.068	0.064	92.8
	ML$_\lambda$	-0.251	0.076	0.072	94.4	-0.790	0.078	0.073	93.5	-0.740	0.068	0.064	93.0
	APW$_{(a,\lambda)}$	-0.333	0.080	0.075	93.0	-0.937	0.082	0.075	91.9	-0.904	0.074	0.066	90.0
	APW$_\lambda$	-0.322	0.080	0.075	93.3	-0.959	0.082	0.075	92.0	-0.928	0.074	0.066	90.3
	APC$_{(a,\lambda)}$	-0.275	0.076	0.072	93.8	-0.874	0.079	0.072	92.3	-0.839	0.070	0.063	90.8
	APC$_\lambda$	-0.259	0.076	0.072	93.7	-0.882	0.080	0.072	92.0	-0.855	0.070	0.063	90.9
Example 2 $n = 500, J = 1, K = 5$	ML$_{(a,\lambda)}$	-0.044	0.025	0.025	95.5	-0.021	0.025	0.025	95.7	-0.085	0.023	0.022	94.2
	ML$_\lambda$	-0.021	0.025	0.025	95.8	0.017	0.025	0.025	95.7	-0.039	0.023	0.022	94.0
	APW$_{(a,\lambda)}$	-0.038	0.026	0.026	95.6	-0.038	0.026	0.027	95.5	-0.088	0.023	0.023	94.6
	APW$_\lambda$	-0.009	0.026	0.026	95.5	0.010	0.026	0.027	95.5	-0.031	0.023	0.023	94.8
	APC$_{(a,\lambda)}$	-0.044	0.025	0.025	95.5	-0.021	0.025	0.025	95.5	-0.085	0.023	0.022	93.8
	APC$_\lambda$	-0.018	0.025	0.025	95.4	0.021	0.025	0.025	95.3	-0.034	0.023	0.022	93.9
Example 3 $n = 60, J_i = 3, K = 3$	ML$_\lambda$	-0.216	0.055	0.056	95.4	0.152	0.057	0.057	95.0	-0.266	0.053	0.050	92.6
	APW$_\lambda$	-0.106	0.063	0.062	93.7	-0.349	0.067	0.063	92.6	-0.394	0.059	0.055	92.1
	APC$_\lambda$	-0.089	0.059	0.058	93.5	-0.237	0.063	0.059	93.0	-0.274	0.055	0.052	92.9
Example 4 $n = 300, J_i = 3, K = 3$	ML$_\lambda$	-0.015	0.025	0.025	96.0	-0.008	0.025	0.025	96.4	0.013	0.023	0.022	94.4
	APW$_\lambda$	0.017	0.027	0.028	95.4	-0.157	0.028	0.028	95.6	-0.055	0.025	0.025	95.5
	APC$_\lambda$	0.009	0.025	0.027	95.6	-0.139	0.026	0.027	96.1	-0.055	0.023	0.023	95.3

† ML, APW, APC represent maximum likelihood, all-pairwise marginal pairwise likelihood and all-pairwise conditional pairwise likelihood approaches, respectively. $(a, \lambda), \lambda$ denote the tuning parameter selection by both a, λ and only λ with fixing $a = 3.7$, respectively.

∗ Relative bias defined by $(\hat{\beta} - \beta_{true})/\beta_{true} \times 100$.

‡ ASE is the average standard error for 1000 simulations, which is defined as $1000^{-1} \sum_{i=1}^{1000} \sqrt{\widehat{Var}(\hat{\beta}^i)}$, where $\sqrt{\widehat{Var}(\hat{\beta}^i)}$ is the standard error estimates in ith simulation result.

♯ ESE is the empirical standard error for 1000 simulations, which is defined as $(1000-1)^{-1} \sum_{i=1}^{1000} (\hat{\beta}^i - \bar{\hat{\beta}})^2$, where $\hat{\beta}^i$ is the ith simulation result, and $\bar{\hat{\beta}} = 1000^{-1} \sum_{i=1}^{1000} \hat{\beta}^i$.

♭ CP(%) is 95% coverage rate.

Table 4.5: Simulation results for the fixed effects selection under linear mixed model: estimation of random effects parameters

		σ_ϵ^2				σ_u^2				ρ			
	Method	Bias(%)*	ESE‡	ASE$^\sharp$	CP(%)$^\flat$	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)
Example 1 $n = 60, J = 1, K = 5$	$\text{ML}^\dagger_{(a,\lambda)}$	-2.187	0.092	0.089	91.5	-0.435	0.218	0.219	91.9	-	-	-	-
	ML_λ	-2.186	0.092	0.089	91.5	-0.436	0.218	0.219	91.9	-	-	-	-
	$\text{APW}_{(a,\lambda)}$	-1.933	0.092	0.088	90.3	-1.576	0.217	0.207	89.2	-	-	-	-
	APW_λ	-1.942	0.092	0.088	90.3	-0.436	0.217	0.207	89.2	-	-	-	-
	$\text{APC}_{(a,\lambda)}$	-2.313	0.092	0.088	89.5	-0.412	0.219	0.211	89.8	-	-	-	-
	APC_λ	-2.323	0.092	0.088	89.5	-0.411	0.219	0.211	89.8	-	-	-	-
Example 2 $n = 500, J = 1, K = 5$	$\text{ML}_{(a,\lambda)}$	-0.209	0.031	0.032	95.7	0.399	0.075	0.076	95.8	-	-	-	-
	ML_λ	-0.209	0.031	0.032	95.7	0.398	0.075	0.076	95.8	-	-	-	-
	$\text{APW}_{(a,\lambda)}$	-0.180	0.031	0.031	95.5	0.315	0.075	0.076	95.9	-	-	-	-
	APW_λ	-0.180	0.031	0.032	95.5	0.315	0.075	0.076	95.9	-	-	-	-
	$\text{APC}_{(a,\lambda)}$	-0.209	0.031	0.031	95.5	0.398	0.075	0.076	96.0	-	-	-	-
	APC_λ	-0.209	0.031	0.031	95.5	0.398	0.075	0.076	96.0	-	-	-	-
Example 3 $n = 60, J_i = 3, K = 3$	ML_λ	-1.291	0.075	0.074	93.2	-0.311	0.161	0.161	94.0	-0.812	0.090	0.102	96.2
	APW_λ	-0.703	0.074	0.073	93.4	-0.915	0.160	0.154	92.7	-0.929	0.095	0.099	94.2
	APC_λ	-0.898	0.074	0.073	93.1	-0.497	0.160	0.155	92.7	-0.750	0.095	0.099	94.4
Example 4 $n = 300, J_i = 3, K = 3$	ML_λ	0.121	0.033	0.033	94.8	-0.580	0.071	0.072	94.2	0.376	0.044	0.045	96.0
	APW_λ	0.010	0.034	0.033	94.9	-0.015	0.073	0.071	93.4	0.127	0.045	0.045	94.5
	APC_λ	-0.022	0.033	0.033	94.8	0.039	0.073	0.071	93.5	0.157	0.045	0.045	94.5

† ML, APW, APC represent maximum likelihood, all-pairwise marginal pairwise likelihood and all-pairwise conditional pairwise likelihood approaches, respectively. $(a, \lambda), \lambda$ denote the tuning parameter selection by both a, λ and only λ with fixing $a = 3.7$, respectively.

∗ Relative bias defined by $(\hat{\beta} - \beta_{true})/\beta_{true} \times 100$.

‡ ASE is the average standard error for 1000 simulations, which is defined as $1000^{-1} \sum_{i=1}^{1000} \sqrt{\widehat{Var}(\hat{\beta}^i)}$, where $\sqrt{\widehat{Var}(\hat{\beta}^i)}$ is the standard error estimates in ith simulation result.

♯ ESE is the empirical standard error for 1000 times simulation, which is defined by $(1000-1)^{-1} \sum_{i=1}^{1000} (\hat{\beta}^i - \bar{\hat{\beta}})^2$, where $\hat{\beta}^i$ is the ith simulation result, and $\bar{\hat{\beta}} = 1000^{-1} \sum_{i=1}^{1000} \hat{\beta}^i$.

♭ CP(%) is 95% coverage rate.

Logistic Mixed Model

Table 4.6: Simulation results for the fixed effects selection under logistic mixed model: model selection

				Avg. No. of 0 Coefficients	
	Method	R.MME(%)	1000×M.MME	Correct	Incorrect
	$\mathrm{ML}_\lambda^\dagger$	40.816	0.486	4.907	0
Example 1	APW_λ	43.364	0.507	4.813	0
$n = 200, J_i = J = 1, K = 5$	APC_λ	49.784	0.592	4.703	0
	ML_λ	36.357	0.097	4.967	0
Example 2	APW_λ	36.572	0.097	4.961	0
$n = 800, J_i = J = 1, K = 5$	APC_λ	37.150	0.098	4.948	0
	ML_λ	44.060	0.247	4.886	0
Example 3	APW_λ	45.738	0.398	4.895	0
$n = 200, J_i = J = 3, K = 4$	APC_λ	48.837	0.270	4.891	0
	ML_λ	41.212	0.068	4.940	0
Example 4	APW_λ	44.589	0.132	4.942	0
$n = 400, J_i = J = 3, K = 4$	APC_λ	44.982	0.127	4.970	0

† ML, APW, APC represent maximum likelihood, all-pairwise marginal pairwise likelihood and all-pairwise conditional pairwise likelihood approaches, respectively. λ denotes the tuning parameter selection by only searching λ with fixing $a = 3.7$.

Table 4.7: Simulation results for the fixed effects selection under logistic mixed model: estimation of regression coefficients $\boldsymbol{\beta}$

		β_1				β_2				β_5			
	Method	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)
	$\mathrm{ML}_\lambda^\dagger$	2.014	0.294	0.290	96.1	2.213	0.198	0.191	94.5	1.864	0.226	0.210	94.1
Example 1	APW_λ	1.132	0.309	0.282	92.9	1.116	0.206	0.187	91.6	0.849	0.240	0.204	90.5
$n = 200, J_i = J = 1, K = 5$	APC_λ	0.060	0.318	0.272	90.9	-0.191	0.215	0.181	87.9	-0.497	0.250	0.197	87.5
	ML_λ	0.582	0.142	0.141	95.1	0.706	0.092	0.093	96.0	0.599	0.103	0.102	94.5
Example 2	APW_λ	0.452	0.144	0.142	94.7	0.590	0.093	0.094	94.9	0.474	0.106	0.103	94.3
$n = 800, J_i = J = 1, K = 5$	APC_λ	0.434	0.143	0.141	94.3	0.567	0.093	0.093	95.4	0.428	0.106	0.102	94.1
	ML_λ	1.788	0.237	0.220	95.4	2.281	0.157	0.141	94.8	1.949	0.169	0.158	94.8
Example 3	APW_λ	0.998	0.292	0.241	89.7	1.007	0.164	0.152	91.9	0.725	0.206	0.171	88.7
$n = 200, J_i = J = 3, K = 4$	APC_λ	-1.045	0.239	0.203	90.2	-0.983	0.170	0.135	88.4	-0.696	0.185	0.148	86.1
	ML_λ	0.916	0.135	0.124	93.4	1.125	0.085	0.080	94.8	1.122	0.094	0.089	93.4
Example 4	APW_λ	1.103	0.175	0.172	95.2	1.139	0.111	0.108	94.9	1.129	0.131	0.122	92.5
$n = 400, J_i = J = 3, K = 4$	APC_λ	0.128	0.169	0.151	93.0	0.317	0.110	0.099	94.0	0.365	0.123	0.109	90.0

† ML, APW, APC represent maximum likelihood, all-pairwise marginal pairwise likelihood and all-pairwise conditional pairwise likelihood approaches, respectively. λ denotes the tuning parameter selection by only searching λ with fixing $a = 3.7$.

Table 4.8: Simulation results for the fixed effects selection under logistic mixed model: estimation of random effects parameters

		σ_u				ρ			
	Method	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)
	$\text{ML}_\lambda^\dagger$	2.431	0.253	0.255	98.1	-	-	-	-
Example 1	APW_λ	-1.079	0.276	0.260	96.0	-	-	-	-
$n = 200, J_i = J = 1, K = 5$	APC_λ	-3.158	0.278	0.253	96.1	-	-	-	-
	ML_λ	0.406	0.124	0.123	95.8	-	-	-	-
Example 2	APW_λ	-0.210	0.128	0.126	95.2	-	-	-	-
$n = 800, J_i = J = 1, K = 5$	APC_λ	-0.081	0.125	0.123	94.9	-	-	-	-
	ML_λ	2.429	0.252	0.223	94.6	0.121	0.091	0.087	94.0
Example 3	APW_λ	0.988	0.301	0.245	89.2	2.994	0.098	0.094	93.3
$n = 200, J_i = J = 3, K = 4$	APC_λ	-2.122	0.252	0.198	88.4	7.595	0.100	0.093	95.4
	ML_λ	1.330	0.144	0.125	93.2	-0.668	0.055	0.050	90.4
Example 4	APW_λ	1.270	0.179	0.174	93.5	0.596	0.067	0.064	94.9
$n = 400, J_i = J = 3, K = 4$	APC_λ	-0.004	0.166	0.147	90.0	4.438	0.065	0.064	95.0

† ML, APW, APC represent maximum likelihood, all-pairwise marginal pairwise likelihood and all-pairwise conditional pairwise likelihood approaches, respectively. λ denotes the tuning parameter selection by only searching λ with fixing $a = 3.7$.

Poisson Mixed Model

Table 4.9: Simulation results for the fixed effects selection under Poisson mixed model: model selection

				Avg. No. of 0 Coefficients	
	Method	R.MME(%)	M.MME	Correct	Incorrect
	$\text{ML}_\lambda^\dagger$	75.460	47.159	4.633	0
Example 1	APW_λ	82.396	65.310	4.302	0
$n = 60, J_i = J = 1, K = 5$	APC_λ	82.222	53.579	4.328	0
	ML_λ	75.448	4.532	5	0
Example 2	APW_λ	77.834	7.108	4.998	0
$n = 500, J_i = J = 1, K = 5$	APC_λ	74.370	5.033	5	0
	ML_λ	86.560	75.026	4.484	0
Example 3	APW_λ	109.717	173.273	4.055	0
$n = 60, J_i = J = 3, K = 2$	APC_λ	96.514	103.141	4.200	0
	ML_λ	75.118	11.706	4.930	0
Example 4	APW_λ	71.482	17.412	4.866	0
$n = 300, J_i = J = 3, K = 2$	APC_λ	71.078	15.439	4.892	0

† ML, APW, APC represent maximum likelihood, all-pairwise marginal pairwise likelihood and all-pairwise conditional pairwise likelihood approaches, respectively. λ denotes the tuning parameter selection by only searching λ with fixing $a = 3.7$.

Table 4.10: Simulation results for the fixed effects selection under Poisson mixed model: estimation of regression coefficients $\boldsymbol{\beta}$

		β_1				β_2				β_5			
	Method	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)
	$\mathrm{ML}_\lambda^\dagger$	-0.797	0.046	0.043	93.1	-1.097	0.046	0.040	91.7	-1.699	0.043	0.037	90.6
Example 1	APW_λ	-0.905	0.055	0.050	92.0	-1.447	0.053	0.045	90.7	-2.122	0.052	0.043	89.1
$n = 60, J_i = J = 1, K = 5$	APC_λ	-0.699	0.048	0.043	90.5	-0.951	0.048	0.040	89.9	-1.625	0.046	0.037	87.3
	ML_λ	-0.006	0.014	0.014	95.2	-0.044	0.013	0.013	95.1	-0.050	0.012	0.012	95.1
Example 2	APW_λ	0.005	0.017	0.017	95.3	-0.123	0.015	0.015	95.5	-0.040	0.014	0.014	95.0
$n = 500, J_i = J = 1, K = 5$	APC_λ	-0.014	0.015	0.015	94.5	-0.129	0.014	0.014	94.8	-0.075	0.013	0.013	94.3
	ML_λ	-1.097	0.059	0.054	92.8	-2.427	0.058	0.052	92.0	-2.532	0.055	0.047	89.0
Example 3	APW_λ	-1.976	0.072	0.065	89.8	-3.273	0.070	0.062	90.5	-5.192	0.070	0.058	84.2
$n = 60, J_i = J = 3, K = 2$	APC_λ	-1.647	0.066	0.058	89.4	-2.729	0.064	0.055	89.6	-3.576	0.061	0.051	87.3
	ML_λ	-0.059	0.024	0.024	94.6	-0.157	0.023	0.023	95.4	-0.079	0.020	0.021	96.0
Example 4	APW_λ	-0.176	0.028	0.028	93.3	-0.260	0.029	0.027	92.7	-0.494	0.027	0.024	92.0
$n = 300, J_i = J = 3, K = 2$	APC_λ	-0.196	0.027	0.026	93.6	-0.243	0.026	0.025	93.4	-0.427	0.025	0.022	92.8

† ML, APW, APC represent maximum likelihood, all-pairwise marginal pairwise likelihood and all-pairwise conditional pairwise likelihood approaches, respectively. λ denotes the tuning parameter selection by only searching λ with fixing $a = 3.7$.

Table 4.11: Simulation results for the fixed effects selection under Poisson mixed model: estimation of random effects parameters

		σ_u				ρ			
	Method	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)
	$\text{ML}_\lambda^\dagger$	-0.845	0.107	0.106	93.7	-	-	-	-
Example 1	APW_λ	-2.010	0.111	0.103	90.7	-	-	-	-
$n = 60, J_i = J = 1, K = 5$	APC_λ	-0.524	0.115	0.111	93.2	-	-	-	-
	ML_λ	-0.118	0.038	0.037	94.6	-	-	-	-
Example 2	APW_λ	-0.243	0.040	0.038	93.4	-	-	-	-
$n = 500, J_i = J = 1, K = 5$	APC_λ	-0.213	0.040	0.039	94.0	-	-	-	-
	ML_λ	-1.452	0.076	0.075	91.4	-3.263	0.119	0.113	93.0
Example 3	APW_λ	-1.630	0.078	0.073	90.9	-4.481	0.123	0.112	91.1
$n = 60, J_i = J = 3, K = 2$	APC_λ	-1.419	0.077	0.073	90.9	-2.104	0.126	0.114	90.6
	ML_λ	-0.902	0.034	0.033	93.6	-2.915	0.049	0.051	95.8
Example 4	APW_λ	-0.382	0.036	0.034	93.5	0.258	0.055	0.054	94.4
$n = 300, J_i = J = 3, K = 2$	APC_λ	-0.352	0.036	0.034	93.0	-0.615	0.053	0.054	95.4

† ML, APW and APC represent maximum likelihood, all-pairwise marginal pairwise likelihood and all-pairwise conditional pairwise likelihood approaches, respectively. λ denotes the tuning parameter selection by only searching λ with fixing $a = 3.7$.

Simulation for Both Fixed and Random Effects

Linear Mixed Model

Situation 1: Generate Data from GLMMs

Table 4.12: Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMMs: model selection in the linear mixed model

						Fixed Coefficients		Random Coefficients	
		$M.MSE_{\boldsymbol{\beta}}^{\ddagger}$	$R.MSE_{\boldsymbol{\beta}}$	$M.MSE_D$	$R.MSE_D$	Correct1**	Incorrect1	Correct2	Incorrect2
Scenario 1: $n = 100, J_i = J = 1, K = 5$									
GLMM	ML$_\lambda$	0.086	0.531	0.366	0.517	4.652	0	4.150	0
	APW$_\lambda$	0.127	0.653	0.104	0.171	3.812	0	5.538	0
GLMPM	ML$_\lambda$	0.075	0.518	0.395	0.554	4.588	0	4.020	0
	APW$_\lambda$	0.126	0.649	0.105	0.175	3.548	0	5.516	0
Scenario 2: $n = 300, J_i = J = 1, K = 5$									
GLMM	ML$_\lambda$	0.018	0.371	0.033	0.061	4.968	0	5.802	0
	APW$_\lambda$	0.026	0.438	0.025	0.044	4.764	0	5.976	0
GLMPM	ML$_\lambda$	0.020	0.400	0.038	0.070	4.960	0	5.806	0
	APW$_\lambda$	0.026	0.439	0.024	0.044	4.718	0	5.950	0
Scenario 3: $n = 100, J_i = J = 3, K = 3$									
GLMM	ML$_\lambda$	0.049	0.492	0.033	0.059	4.744	0	5.912	0
	APW$_\lambda$	0.112	0.746	0.009	0.016	3.522	0	5.264	0
GLMPM	ML$_\lambda$	0.049	0.495	0.035	0.064	4.736	0	5.798	0
	APW$_\lambda$	0.110	0.745	0.011	0.018	3.454	0	4.994	0
Scenario 4: $n = 300, J_i = J = 3, K = 3$									
GLMM	ML$_\lambda$	0.012	0.390	0.009	0.017	4.970	0	6	0
	APW$_\lambda$	0.021	0.472	0.003	0.005	4.670	0	6	0
GLMPM	ML$_\lambda$	0.012	0.384	0.009	0.017	4.974	0	5.998	0
	APW$_\lambda$	0.021	0.478	0.003	0.005	4.666	0	5.594	0

† ML and APW represent maximum likelihood and all-pairwise marginal pairwise likelihood, respectively. λ denotes the tuning parameter selection by searching λ with fixing $a = 3.7$, respectively.

‡ $MSE_{\boldsymbol{\beta}} = ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2$, $MSE_D = ||\sqrt{\text{diag}(D)} - \sqrt{\text{diag}(\hat{D})}||^2$. $M.MSE_{\boldsymbol{\beta}}$ and $M.MSE_D$ are the median of both quantities. $R.MSE_{\boldsymbol{\beta}}$ and $R.MSE_D$ are the median ratios of $MSE_{\boldsymbol{\beta}}$ and MSE_D, for a selected model to that of the un-penalized estimate, respectively.

* "Correct1" presents the average restricted to the true fixed effects zero coefficients. 0 represents no true fixed effects zero coefficient is shrink, while 5 implies that all true fixed effects zero coefficients are restricted into zero. "Incorrect1" depicts that the average of significant fixed effects coefficients erroneously set to zero. 0 represents that no significant fixed effects coefficient is shrink, while 3 implies that all significant fixed effects coefficients are erroneously set to zero.

** "Correct2" presents the average restricted to the true random effects zero coefficients. 0 represents that no true random effects zero coefficient is shrink, while 6 implies that all true random effects zero coefficients are restricted into zero. "Incorrect2" depicts the average of significant random effects coefficients that are erroneously set to zero. 0 represents that no significant random effects coefficient is shrink, while 3 implies that all significant random effects coefficients are erroneously set to zero.

Table 4.13: Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMMs: estimation of selected regression coefficients $\boldsymbol{\beta}$ in the linear mixed model

		β_1				β_2				β_5			
		Bias(%)*	ESE‡	ASE$^\sharp$	CP(%)$^\flat$	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)
Scenario 1: $n = 100, J_i = J = 1, K = 5$													
GLMM	ML$_\lambda$	-2.145	0.219	0.181	88.0	-4.420	0.195	0.159	85.8	-1.480	0.134	0.114	88.3
	APW$_\lambda$	-1.084	0.192	0.208	95.8	-1.748	0.187	0.186	92.4	-1.771	0.144	0.137	91.0
GLMPM	ML$_\lambda$	-2.269	0.207	0.180	88.1	-3.432	0.193	0.158	87.1	-1.646	0.132	0.113	88.9
	APW$_\lambda$	-1.220	0.191	0.215	95.4	-1.821	0.187	0.206	92.4	-1.748	0.144	0.139	91.4
Scenario 2: $n = 300, J_i = J = 1, K = 5$													
GLMM	ML$_\lambda$	0.031	0.119	0.106	91.8	-0.139	0.102	0.095	94.4	-0.581	0.064	0.064	94.8
	APW$_\lambda$	-0.075	0.113	0.111	94.2	-0.732	0.111	0.101	91.9	-1.106	0.081	0.071	91.7
GLMPM	ML$_\lambda$	-0.205	0.119	0.107	92.8	-0.624	0.107	0.095	92.4	-0.083	0.071	0.064	93.4
	APW$_\lambda$	-0.284	0.112	0.112	94.8	-0.937	0.108	0.100	93.4	-0.508	0.082	0.071	89.8
Scenario 3: $n = 100, J_i = J = 3, K = 3$													
GLMM	ML$_\lambda$	-1.289	0.165	0.146	90.4	-2.805	0.152	0.129	88.4	-0.905	0.098	0.088	92.6
	APW$_\lambda$	-0.490	0.175	0.160	90.2	-1.232	0.167	0.151	91.4	-0.523	0.124	0.115	92.8
GLMPM	ML$_\lambda$	-1.298	0.165	0.146	90.0	-2.803	0.151	0.129	89.0	-0.951	0.099	0.088	91.6
	APW$_\lambda$	0.076	0.156	0.160	95.2	-1.308	0.158	0.151	93.4	-0.736	0.129	0.116	93.8
Scenario 4: $n = 300, J_i = J = 3, K = 3$													
GLMM	ML$_\lambda$	-0.511	0.093	0.086	91.6	-0.554	0.086	0.076	92.0	-0.395	0.052	0.051	94.4
	APW$_\lambda$	-0.228	0.094	0.093	94.8	-0.389	0.092	0.087	92.6	-0.430	0.066	0.063	93.8
GLMPM	ML$_\lambda$	-0.481	0.093	0.085	91.4	-0.505	0.085	0.076	92.2	-0.378	0.052	0.051	94.6
	APW$_\lambda$	-0.180	0.095	0.093	94.6	-0.406	0.090	0.087	93.6	-0.366	0.067	0.063	93.4

Table 4.14: Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMM: estimation of random effects parameters in the linear mixed model

		d_1		d_2		d_3		r_{12}		r_{13}		r_{23}		σ^2		ρ^*		ρ_e	
		Bias†	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE
Scenario 1: $n = 100, J_i = J = 1, K = 5$																			
GLMM	ML$_\lambda$	-0.057	0.240	-0.038	0.194	0.010	0.163	-0.015	0.067	0.000	0.163	-0.038	0.153	-0.352	0.364	-	-	-	-
	APW$_\lambda$	-0.023	0.250	-0.005	0.199	-0.002	0.115	0.000	0.055	-0.009	0.164	-0.024	0.133	-0.225	0.320	-	-	-	-
GLMPM	ML$_\lambda$	-0.070	0.248	-0.037	0.203	-0.001	0.159	-0.009	0.063	-0.010	0.168	-0.049	0.168	-0.353	0.397	-	-	-0.002	0.103
	APW$_\lambda$	-0.023	0.249	-0.011	0.199	-0.003	0.119	-0.002	0.055	-0.011	0.165	-0.024	0.136	-0.245	0.341	-	-	-0.032	0.107
Scenario 2: $n = 300, J_i = J = 1, K = 5$																			
GLMM	ML$_\lambda$	-0.019	0.140	-0.007	0.116	0.013	0.087	-0.004	0.033	-0.005	0.095	-0.015	0.084	-0.047	0.194	-	-	-	-
	APW$_\lambda$	-0.000	0.137	0.008	0.109	0.007	0.057	-0.006	0.029	-0.008	0.089	0.002	0.067	-0.103	0.184	-	-	-	-
GLMPM	ML$_\lambda$	-0.017	0.139	-0.010	0.121	0.009	0.089	-0.007	0.033	-0.005	0.097	-0.014	0.091	-0.047	0.204	-	-	-0.000	0.054
	APW$_\lambda$	0.005	0.136	0.007	0.109	0.004	0.059	-0.005	0.029	-0.010	0.090	0.001	0.069	-0.123	0.196	-	-	-0.024	0.062
Scenario 3: $n = 100, J_i = J = 3, K = 3$																			
GLMM	ML$_\lambda$	0.001	0.158	0.009	0.130	-0.002	0.073	-0.006	0.032	-0.004	0.080	-0.001	0.058	-0.132	0.233	-0.006	0.051	-	-
	APW$_\lambda$	-0.018	0.097	-0.015	0.058	-0.012	0.026	0.002	0.011	0.002	0.024	-0.001	0.014	-0.091	0.184	-0.001	0.016	-	-
GLMPM	ML$_\lambda$	0.013	0.161	0.010	0.128	-0.006	0.074	-0.003	0.031	-0.004	0.077	-0.001	0.053	-0.169	0.263	-0.006	0.050	-0.025	0.094
	APW$_\lambda$	-0.025	0.093	0.059	-0.015	0.026	-0.010	0.001	0.011	-0.000	0.025	-0.001	0.014	-0.086	0.185	-0.002	0.014	-0.001	0.119
Scenario 4: $n = 300, J_i = J = 3, K = 3$																			
GLMM	ML$_\lambda$	0.014	0.095	0.011	0.071	0.001	0.037	-0.004	0.016	-0.005	0.046	0.004	0.034	-0.089	0.143	-0.004	0.029	-	-
	APW$_\lambda$	0.003	0.065	-0.001	0.039	-0.004	0.017	0.001	0.007	0.000	0.016	-0.001	0.009	-0.035	0.128	0.001	0.009	-	-
GLMPM	ML$_\lambda$	0.029	0.098	0.013	0.068	-0.003	0.035	-0.001	0.015	-0.005	0.040	0.003	0.028	-0.122	0.162	-0.003	0.027	-0.022	0.051
	APW$_\lambda$	-0.004	0.058	-0.003	0.036	-0.004	0.016	0.001	0.006	0.000	0.015	-0.001	0.009	-0.037	0.118	0.000	0.009	-	-

† Bias is defined by $\hat{\beta} - \beta_{true}$.

Situation 2: Generate Data from GLMPMs

Table 4.15: Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMPMs: model selection in the linear mixed model

						Fixed Coefficients		Random Coefficients	
		$M.MSE_{\boldsymbol{\beta}}^{\ddagger}$	$R.MSE_{\boldsymbol{\beta}}$	$M.MSE_D$	$R.MSE_D$	Correct1**	Incorrect1	Correct2	Incorrect2
Scenario 1: $n = 100, J_i = J = 1, K = 5$									
GLMM	ML$_\lambda$	0.068	0.563	0.173	0.323	4.686	0	5.576	0
	APW$_\lambda$	0.092	0.611	0.108	0.219	3.988	0	5.998	0
GLMPM	ML$_\lambda$	0.059	0.538	0.289	0.441	4.668	0	4.700	0
	APW$_\lambda$	0.092	0.625	0.124	0.201	3.810	0	5.366	0
Scenario 2: $n = 300, J_i = J = 1, K = 5$									
GLMM	ML$_\lambda$	0.017	0.433	0.062	0.145	4.984	0	6	0
	APW$_\lambda$	0.024	0.471	0.056	0.129	4.828	0	6	0
GLMPM	ML$_\lambda$	0.016	0.455	0.032	0.064	4.968	0	5.914	0
	APW$_\lambda$	0.023	0.472	0.027	0.048	4.788	0	5.770	0
Scenario 3: $n = 100, J_i = J = 3, K = 3$									
GLMM	ML$_\lambda$	0.040	0.534	0.132	0.264	4.806	0	6	0
	APW$_\lambda$	0.097	0.729	0.071	0.144	3.604	0	6	0
GLMPM	ML$_\lambda$	0.038	0.501	0.035	0.063	4.834	0	5.904	0
	APW$_\lambda$	0.093	0.735	0.010	0.018	3.722	0	4.916	0
Scenario 4: $n = 300, J_i = J = 3, K = 3$									
GLMM	ML$_\lambda$	0.010	0.429	0.117	0.236	4.988	0	6	0
	APW$_\lambda$	0.021	0.497	0.051	0.106	4.730	0	6	0
GLMPM	ML$_\lambda$	0.010	0.422	0.009	0.016	4.994	0	5.996	0
	APW$_\lambda$	0.019	0.504	0.003	0.006	4.766	0	5.422	0

† ML and APW represent maximum likelihood and all-pairwise marginal pairwise likelihood, respectively. λ denotes the tuning parameter selection by searching λ with fixing $a = 3.7$, respectively.

‡ $MSE_{\boldsymbol{\beta}} = ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2$, $MSE_D = ||\sqrt{\text{diag}(D)} - \sqrt{\text{diag}(\hat{D})}||^2$. $M.MSE_{\boldsymbol{\beta}}$ and $M.MSE_D$ are the median of both quantities. $R.MSE_{\boldsymbol{\beta}}$ and $R.MSE_D$ are the median ratios of $MSE_{\boldsymbol{\beta}}$ and MSE_D, for a selected model to that of the un-penalized estimate, respectively.

∗ "Correct1" presents the average restricted to the true fixed effects zero coefficients. 0 represents no true fixed effects zero coefficient is shrink, while 5 implies that all true fixed effects zero coefficients are restricted into zero. "Incorrect1" depicts that the average of significant fixed effects coefficients erroneously set to zero. 0 represents that no significant fixed effects coefficient is shrink, while 3 implies that all significant fixed effects coefficients are erroneously set to zero.

∗∗ "Correct2" presents the average restricted to the true random effects zero coefficients. 0 represents that no true random effects zero coefficient is shrink, while 6 implies that all true random effects zero coefficients are restricted into zero. "Incorrect2" depicts the average of significant random effects coefficients that are erroneously set to zero. 0 represents that no significant random effects coefficient is shrink, while 3 implies that all significant random effects coefficients are erroneously set to zero.

Table 4.16: Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMPMs: estimation of selected regression coefficients $\boldsymbol{\beta}$ in the linear mixed model

		β_1				β_2				β_5			
		Bias(%)*	ESE‡	ASE$^\sharp$	CP(%)$^\flat$	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)
Scenario 1: $n = 100, J_i = J = 1, K = 5$													
GLMM	ML$_\lambda$	-2.758	0.222	0.175	84.2	-4.180	0.188	0.146	84.4	-1.465	0.112	0.095	89.2
	APW$_\lambda$	-0.792	0.185	0.184	94.2	-1.621	0.167	0.159	93.6	-1.211	0.123	0.110	91.0
GLMPM	ML$_\lambda$	-1.938	0.203	0.172	88.3	-3.266	0.174	0.141	86.7	-1.001	0.098	0.091	92.4
	APW$_\lambda$	-0.755	0.183	0.191	94.4	-1.526	0.165	0.187	94.4	-1.154	0.120	0.118	91.4
Scenario 2: $n = 300, J_i = J = 1, K = 5$													
GLMM	ML$_\lambda$	-0.065	0.113	0.102	92.8	-0.188	0.096	0.086	93.8	0.109	0.056	0.055	94.2
	APW$_\lambda$	-0.182	0.115	0.109	93.2	-0.465	0.101	0.093	91.6	-0.475	0.066	0.062	93.6
GLMPM	ML$_\lambda$	-0.189	0.118	0.100	90.6	-0.074	0.097	0.083	91.4	-0.065	0.052	0.051	95.0
	APW$_\lambda$	-0.174	0.115	0.113	93.2	-0.391	0.099	0.101	91.6	-0.421	0.064	0.063	94.0
Scenario 3: $n = 100, J_i = J = 3, K = 3$													
GLMM	ML$_\lambda$	-1.200	0.161	0.136	88.8	-1.992	0.141	0.112	86.4	-0.453	0.081	0.074	92.8
	APW$_\lambda$	0.491	0.160	0.159	94.1	-1.150	0.156	0.145	93.5	-1.090	0.127	0.107	90.7
GLMPM	ML$_\lambda$	-0.951	0.159	0.137	90.0	-1.569	0.135	0.116	90.8	-0.415	0.079	0.072	91.6
	APW$_\lambda$	0.331	0.156	0.158	95.8	-1.109	0.152	0.143	93.2	-0.905	0.117	0.103	91.8
Scenario 4: $n = 300, J_i = J = 3, K = 3$													
GLMM	ML$_\lambda$	-0.465	0.088	0.080	93.4	-0.354	0.075	0.066	90.6	-0.024	0.044	0.043	94.0
	APW$_\lambda$	-0.300	0.097	0.093	93.0	-0.130	0.090	0.084	92.2	-0.289	0.065	0.059	92.8
GLMPM	ML$_\lambda$	-0.432	0.084	0.080	94.2	-0.243	0.072	0.068	92.8	-0.011	0.043	0.042	93.6
	APW$_\lambda$	-0.300	0.097	0.092	92.7	-0.114	0.088	0.082	92.7	-0.326	0.062	0.057	93.3

Table 4.17: Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMPMs: estimation of random effects parameters in the linear mixed model

		d_1		d_2		d_3		r_{12}		r_{13}		r_{23}		σ^2		ρ^*		ρ_e	
		Bias†	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE
Scenario 1: $n = 100, J_i = J = 1, K = 5$																			
GLMM	ML$_\lambda$	0.176	0.251	-0.011	0.194	0.009	0.147	-0.054	0.066	-0.025	0.141	-0.010	0.156	-1.360	0.264	-	-	-	-
	APW$_\lambda$	0.178	0.255	-0.031	0.193	-0.040	0.116	-0.030	0.060	-0.033	0.159	0.047	0.119	-1.240	0.275	-	-	-	-
GLMPM	ML$_\lambda$	-0.112	0.302	-0.030	0.186	-0.005	0.145	0.005	0.067	0.008	0.164	-0.013	0.145	0.155	0.954	-	-	0.053	0.117
	APW$_\lambda$	0.004	0.255	-0.019	0.194	-0.005	0.121	-0.003	0.054	-0.018	0.165	-0.018	0.127	-0.355	0.441	-	-	-0.048	0.120
Scenario 2: $n = 300, J_i = J = 1, K = 5$																			
GLMM	ML$_\lambda$	0.198	0.141	-0.001	0.109	-0.009	0.081	-0.045	0.035	-0.020	0.082	0.025	0.085	-1.221	0.151	-	-	-	-
	APW$_\lambda$	0.199	0.152	-0.021	0.106	-0.039	0.067	-0.033	0.034	-0.031	0.086	0.065	0.063	-1.181	0.161	-	-	-	-
GLMPM	ML$_\lambda$	-0.011	0.174	-0.009	0.105	-0.002	0.084	-0.005	0.038	-0.003	0.088	-0.005	0.082	-0.031	0.492	-	-	-0.008	0.064
	APW$_\lambda$	0.018	0.152	-0.001	0.105	-0.004	0.064	-0.002	0.029	-0.012	0.087	0.001	0.062	-0.184	0.255	-	-	-0.030	0.064
Scenario 3: $n = 100, J_i = J = 3, K = 3$																			
GLMM	ML$_\lambda$	0.321	0.166	-0.026	0.129	-0.076	0.086	-0.042	0.042	-0.057	0.093	0.064	0.062	-1.601	0.181	-0.051	0.055	-	-
	APW$_\lambda$	0.171	0.186	-0.024	0.140	-0.078	0.068	-0.000	0.035	-0.026	0.076	0.012	0.041	-0.937	0.249	-0.021	0.049	-	-
GLMPM	ML$_\lambda$	0.045	0.171	0.005	0.119	-0.011	0.081	-0.002	0.031	-0.007	0.077	-0.002	0.052	-0.243	0.302	-0.007	0.048	-0.029	0.068
	APW$_\lambda$	-0.025	0.099	-0.015	0.063	-0.010	0.029	0.001	0.010	0.001	0.024	-0.000	0.014	-0.083	0.211	-0.002	0.016	0.004	0.090
Scenario 4: $n = 300, J_i = J = 3, K = 3$																			
GLMM	ML$_\lambda$	0.320	0.096	-0.021	0.076	-0.077	0.050	-0.043	0.024	-0.059	0.052	0.070	0.035	-1.575	0.102	-0.048	0.033	-	-
	APW$_\lambda$	0.192	0.097	-0.019	0.090	-0.074	0.034	0.000	0.097	-0.025	0.090	0.018	0.034	-0.921	0.065	-0.017	0.032	-	-
GLMPM	ML$_\lambda$	0.034	0.095	0.011	0.065	-0.008	0.042	0.001	0.015	-0.004	0.038	0.002	0.026	-0.155	0.172	-0.002	0.026	-0.021	0.036
	APW$_\lambda$	-0.008	0.097	-0.005	0.088	-0.005	0.031	0.002	0.097	0.001	0.088	-0.000	0.031	-0.044	0.062	-0.000	0.029	0.005	0.004

† Bias is defined by $\hat{\beta} - \beta_{true}$.

Poisson Mixed Model

Situation 1: Generate Data from GLMM

Table 4.18: Simulation results for doubly selecting fixed and random effects under the case that data are generated by GLMMs: model selection in the Poisson mixed model

		100×		100×		Fixed Coefficients		Random Coefficients	
		$M.MSE_{\boldsymbol{\beta}}$[‡]	$R.MSE_{\boldsymbol{\beta}}$	$M.MSE_D$	$R.MSE_D$	Correct1[**]	Incorrect1	Correct2	Incorrect2
Scenario 1: $n = 250, J_i = J = 1, K = 9$									
GLMM	ML$_\lambda$	0.060	0.306	0.126	0.086	5	0	5.942	0
	APW$_\lambda$	0.101	0.357	0.044	0.029	4.988	0	5.610	0
GLMPM	ML$_\lambda$	0.060	0.300	0.122	0.081	5	0	5.982	0
	APW$_\lambda$	0.101	0.357	0.044	0.029	4.988	0	5.754	0
Scenario 2: $n = 500, J_i = J = 1, K = 9$									
GLMM	ML$_\lambda$	0.033	0.325	0.116	0.079	5	0	5.994	0
	APW$_\lambda$	0.058	0.364	0.014	0.009	5	0	5.870	0
GLMPM	ML$_\lambda$	0.033	0.326	0.114	0.077	5	0	6	0
	APW$_\lambda$	0.058	0.364	0.014	0.009	5	0	5.910	0
Scenario 3: $n = 250, J_i = J = 3, K = 4$									
GLMM	ML$_\lambda$	0.054	0.303	0.223	0.005	5	0	5.762	0
	APW$_\lambda$	0.087	0.345	0.257	0.057	4.990	0	4.650	0
GLMPM	ML$_\lambda$	0.055	0.302	0.255	0.006	5	0	5.834	0
	APW$_\lambda$	0.091	0.332	2.525	0.056	5	0	5.124	0
Scenario 4: $n = 500, J_i = J = 3, K = 4$									
GLMM	ML$_\lambda$	0.028	0.316	0.009	0.002	5	0	5.962	0
	APW$_\lambda$	0.054	0.387	0.251	0.056	5	0	5.030	0
GLMPM	ML$_\lambda$	0.027	0.302	0.009	0.002	5	0	5.982	0
	APW$_\lambda$	0.057	0.409	0.049	0.011	5	0	5.038	0

[†] ML and APW represent maximum likelihood and all-pairwise marginal pairwise likelihood, respectively. λ denotes the tuning parameter selection by searching λ with fixing $a = 3.7$, respectively.

[‡] $MSE_{\boldsymbol{\beta}} = ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2$, $MSE_D = ||\sqrt{\text{diag}(D)} - \sqrt{\text{diag}(\hat{D})}||^2$. $M.MSE_{\boldsymbol{\beta}}$ and $M.MSE_D$ are the median of both quantities. $R.MSE_{\boldsymbol{\beta}}$ and $R.MSE_D$ are the median ratios of $MSE_{\boldsymbol{\beta}}$ and MSE_D, for a selected model to that of the un-penalized estimate, respectively.

[*] "Correct1" presents the average restricted to the true fixed effects zero coefficients. 0 represents no true fixed effects zero coefficient is shrink, while 5 implies that all true fixed effects zero coefficients are restricted into zero. "Incorrect1" depicts that the average of significant fixed effects coefficients erroneously set to zero. 0 represents that no significant fixed effects coefficient is shrink, while 3 implies that all significant fixed effects coefficients are erroneously set to zero.

[**] "Correct2" presents the average restricted to the true random effects zero coefficients. 0 represents that no true random effects zero coefficient is shrink, while 6 implies that all true random effects zero coefficients are restricted into zero. "Incorrect2" depicts the average of significant random effects coefficients that are erroneously set to zero. 0 represents that no significant random effects coefficient is shrink, while 3 implies that all significant random effects coefficients are erroneously set to zero.

Table 4.19: Simulation results for doubly selection fixed and random effects under the case that data are generated by GLMMs: estimation of selected regression coefficients $\boldsymbol{\beta}$ in the Poisson mixed model

		β_1				β_2				β_5			
		Bias(%)*	ESE‡	ASE$^\sharp$	CP(%)$^\flat$	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)
Scenario 1: $n = 250, J_i = J = 1, K = 9$													
GLMM	ML$_\lambda$	-0.032	0.017	0.017	94.4	-0.106	0.017	0.017	95.2	-0.095	0.014	0.013	93.6
	APW$_\lambda$	0.831	0.021	0.021	94.4	1.032	0.020	0.021	94.2	0.545	0.017	0.016	92.6
GLMPM	ML$_\lambda$	-0.044	0.017	0.017	94.2	-0.132	0.017	0.017	95.6	-0.084	0.014	0.013	93.6
	APW$_\lambda$	0.832	0.021	0.021	94.4	-1.033	0.020	0.021	94.2	0.536	0.017	0.016	92.6
Scenario 2: $n = 500, J_i = J = 1, K = 9$													
GLMM	ML$_\lambda$	-0.068	0.013	0.012	93.2	-0.255	0.013	0.012	94.4	-0.133	0.009	0.009	94.0
	APW$_\lambda$	0.807	0.016	0.015	90.2	1.011	0.014	0.015	93.0	0.585	0.012	0.012	92.6
GLMPM	ML$_\lambda$	-0.080	0.013	0.012	93.4	-0.277	0.013	0.012	94.2	-0.126	0.009	0.009	94.0
	APW$_\lambda$	0.807	0.016	0.015	90.2	1.011	0.014	0.015	93.0	0.586	0.012	0.012	92.6
Scenario 3: $n = 250, J_i = J = 3, K = 4$													
GLMM	ML$_\lambda$	-0.117	0.016	0.016	95.2	-0.110	0.016	0.016	94.2	-0.065	0.013	0.013	95.2
	APW$_\lambda$	0.843	0.019	0.019	91.6	-1.057	0.020	0.019	92.2	0.723	0.015	0.015	93.4
GLMPM	ML$_\lambda$	-0.114	0.016	0.016	95.2	-0.106	0.016	0.016	94.2	-0.063	0.013	0.013	95.2
	APW$_\lambda$	0.814	0.020	0.019	91.6	1.128	0.020	0.019	93.8	0.744	0.015	0.015	94.4
Scenario 4: $n = 500, J_i = J = 3, K = 4$													
GLMM	ML$_\lambda$	-0.010	0.012	0.012	94.6	-0.067	0.012	0.011	93.6	-0.083	0.009	0.009	94.6
	APW$_\lambda$	0.978	0.014	0.014	87.0	1.146	0.014	0.013	94.2	0.682	0.011	0.011	92.2
GLMPM	ML$_\lambda$	-0.009	0.012	0.012	94.2	0.069	0.012	0.011	95.0	-0.053	0.009	0.009	95.8
	APW$_\lambda$	0.947	0.014	0.014	86.9	1.338	0.013	0.013	91.5	0.682	0.011	0.011	92.3

Table 4.20: Simulation results for doubly selection fixed and random effects under the case that data are generated by GLMMs: estimation of random effects parameters in the Poisson mixed model

		d_1		d_2		d_3		r_{12}		r_{13}		r_{23}		ρ^*		ρ_e	
		Bias†	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE
Scenario 1: $n = 250, J_i = J = 1, K = 9$																	
GLMM	ML$_\lambda$	-0.034	0.023	-0.005	0.005	-0.005	0.005	0.014	0.089	0.008	0.085	-0.027	0.049	-	-	-	-
	APW$_\lambda$	-0.008	0.019	-0.001	0.001	-0.001	0.002	-0.029	0.039	-0.027	0.039	-0.010	0.014	-	-	-	-
GLMPM	ML$_\lambda$	-0.035	0.022	-0.004	0.004	-0.005	0.004	0.019	0.079	0.011	0.075	-0.022	0.038	-	-	0.000	0.001
	APW$_\lambda$	-0.008	0.019	-0.001	0.001	-0.001	0.002	-0.029	0.039	-0.027	0.039	-0.010	0.014	-	-	0.000	0.001
Scenario 2: $n = 500, J_i = J = 1, K = 9$																	
GLMM	ML$_\lambda$	-0.032	0.017	-0.005	0.004	-0.005	0.004	0.011	0.069	0.005	0.061	-0.026	0.039	-	-	-	-
	APW$_\lambda$	-0.006	0.014	-0.001	0.001	-0.001	0.001	-0.026	0.028	-0.025	0.028	-0.009	0.010	-	-	-	-
GLMPM	ML$_\lambda$	-0.033	0.016	-0.004	0.003	-0.005	0.003	0.015	0.061	0.006	0.055	-0.022	0.031	-	-	0.000	0.001
	APW$_\lambda$	-0.006	0.014	-0.001	0.001	-0.001	0.001	-0.026	0.028	-0.025	0.028	-0.009	0.010	-	-	0.001	0.001
Scenario 3: $n = 250, J_i = J = 3, K = 4$																	
GLMM	ML$_\lambda$	-0.007	0.015	-0.001	0.001	-0.001	0.001	-0.008	0.028	-0.007	0.027	-0.004	0.013	-0.000	0.016	-	-
	APW$_\lambda$	0.003	0.016	-0.000	0.001	0.000	0.001	-0.002	0.019	-0.000	0.020	-0.001	0.007	0.002	0.005	-	-
GLMPM	ML$_\lambda$	-0.007	0.017	-0.001	0.002	-0.001	0.002	-0.007	0.031	-0.006	0.031	-0.005	0.015	0.000	0.018	0.000	0.001
	APW$_\lambda$	0.001	0.011	-0.000	0.001	0.000	0.001	-0.002	0.014	-0.001	0.014	-0.001	0.005	0.001	0.004	0.000	0.002
Scenario 4: $n = 500, J_i = J = 3, K = 4$																	
GLMM	ML$_\lambda$	-0.006	0.012	-0.001	0.001	-0.001	0.001	-0.009	0.020	-0.007	0.020	-0.005	0.009	0.002	0.011	-	-
	APW$_\lambda$	0.004	0.012	0.000	0.001	0.000	0.001	-0.001	0.014	0.001	0.014	-0.000	0.005	0.003	0.004	-	-
GLMPM	ML$_\lambda$	-0.006	0.013	-0.001	0.001	-0.001	0.001	-0.006	0.023	-0.006	0.022	-0.005	0.010	0.002	0.014	0.000	0.001
	APW$_\lambda$	0.002	0.009	-0.000	0.000	0.000	0.000	-0.001	0.011	-0.000	0.011	-0.000	0.004	0.002	0.003	0.000	0.002

† Bias is defined by $\hat{\beta} - \beta_{true}$.

Situation 2: Generate Data from GLMPMs

Table 4.21: Simulation results for doubly selection fixed and random effects under the case that data are generated by GLMPMs: model selection in the Poisson mixed model

		100×		100×		Fixed Coefficients		Random Coefficients	
		$M.MSE_{\boldsymbol{\beta}}^{\ddagger}$	$R.MSE_{\boldsymbol{\beta}}$	$M.MSE_D$	$R.MSE_D$	Correct1**	Incorrect1	Correct2	Incorrect2
Scenario 1: $n = 250, J_i = J = 1, K = 9$									
GLMM	ML$_\lambda$	**3.451**	0.999	**32.883**	0.696	4.378	0	**1.110**	0
	APW$_\lambda$	**4.187**	0.892	**54.010**	0.929	4.996	0	**2.106**	0
GLMPM	ML$_\lambda$	0.066	0.303	0.155	0.103	5	0	5.968	0
	APW$_\lambda$	0.113	0.356	0.045	0.029	4.988	0	5.596	0
Scenario 2: $n = 500, J_i = J = 1, K = 9$									
GLMM	ML$_\lambda$	**2.571**	0.864	**31.856**	0.668	4.688	0	**1.488**	0
	APW$_\lambda$	**4.139**	0.896	**54.839**	0.927	5	0	**1.938**	0
GLMPM	ML$_\lambda$	0.032	0.304	0.140	0.095	5	0	5.998	0
	APW$_\lambda$	0.066	0.376	0.014	0.009	4.996	0	5.862	0
Scenario 3: $n = 250, J_i = J = 3, K = 4$									
GLMM	ML$_\lambda$	**25.047**	1.036	**31.264**	0.336	4.994	0	**0.352**	0
	APW$_\lambda$	**17.815**	0.922	**51.366**	0.309	4.971	0	**0.367**	0
GLMPM	ML$_\lambda$	0.056	0.310	0.027	0.006	5	0	5.884	0
	APW$_\lambda$	0.129	0.396	0.252	0.056	4.968	0	5.162	0
Scenario 4: $n = 500, J_i = J = 3, K = 4$									
GLMM	ML$_\lambda$	**21.617**	1.042	**31.129**	0.336	5	0	**0.470**	0
	APW$_\lambda$	**17.715**	0.932	**52.018**	0.309	5	0	**0.096**	0
GLMPM	ML$_\lambda$	0.028	0.290	0.007	0.002	5	0	5.972	0
	APW$_\lambda$	0.073	0.390	0.250	0.056	5	0	5.324	0

† ML and APW represent maximum likelihood and all-pairwise marginal pairwise likelihood, respectively. λ denotes the tuning parameter selection by searching λ with fixing $a = 3.7$, respectively.

‡ $MSE_{\boldsymbol{\beta}} = ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2$, $MSE_D = ||\text{diag}(D) - \text{diag}(\hat{D})||^2$. $M.MSE_{\boldsymbol{\beta}}$ and $M.MSE_D$ are the median of both quantities. $R.MSE_{\boldsymbol{\beta}}$ and $R.MSE_D$ are the median ratios of $MSE_{\boldsymbol{\beta}}$ and MSE_D, for a selected model to that of the un-penalized estimate, respectively.

∗ “Correct1” presents the average restricted to the true fixed effects zero coefficients. 0 represents no true fixed effects zero coefficient is shrink, while 5 implies that all true fixed effects zero coefficients are restricted into zero. “Incorrect1” depicts that the average of significant fixed effects coefficients erroneously set to zero. 0 represents that no significant fixed effects coefficient is shrink, while 3 implies that all significant fixed effects coefficients are erroneously set to zero.

∗∗ “Correct2” presents the average restricted to the true random effects zero coefficients. 0 represents that no true random effects zero coefficient is shrink, while 6 implies that all true random effects zero coefficients are restricted into zero. “Incorrect2” depicts the average of significant random effects coefficients that are erroneously set to zero. 0 represents that no significant random effects coefficient is shrink, while 3 implies that all significant random effects coefficients are erroneously set to zero.

Table 4.22: Simulation results for doubly selection fixed and random effects under the case that data are generated by GLMPMs: estimation of selected regression coefficients $\boldsymbol{\beta}$ in the Poisson mixed model

		β_1				β_2				β_5			
		Bias(%)*	ESE‡	ASE$^\sharp$	CP(%)$^\flat$	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)
Scenario 1: $n = 250, J_i = J = 1, K = 9$													
GLMM	ML$_\lambda$	**-10.925**	0.035	0.018	**0**	**-4.495**	0.033	0.017	**51.9**	**-13.401**	0.034	0.015	**1.0**
	APW$_\lambda$	**-4.561**	0.024	0.021	**27.8**	**0.959**	0.021	0.019	**89.4**	**-24.372**	0.022	0.0.019	**0**
GLMPM	ML$_\lambda$	-0.009	0.018	0.018	93.8	-0.150	0.018	0.017	93.2	-0.068	0.014	0.013	95.0
	APW$_\lambda$	1.005	0.021	0.022	91.0	1.361	0.022	0.021	91.8	0.657	0.017	0.017	93.4
Scenario 2: $n = 500, J_i = J = 1, K = 9$													
GLMM	ML$_\lambda$	**-10.464**	0.027	0.013	**0**	**-3.934**	0.025	0.011	**45.8**	**-12.232**	0.026	0.010	**0.2**
	APW$_\lambda$	**-4.485**	0.016	0.015	**5.4**	0.972	0.014	0.013	90.4	**-24.452**	0.016	0.013	**0**
GLMPM	ML$_\lambda$	-0.004	0.013	0.012	95.1	-0.272	0.013	0.012	92.2	0.054	0.009	0.009	95.0
	APW$_\lambda$	1.028	0.015	0.015	88.2	1.087	0.015	0.015	92.4	0.792	0.012	0.012	91.8
Scenario 3: $n = 250, J_i = J = 3, K = 4$													
GLMM	ML$_\lambda$	**-32.013**	0.090	0.019	**0**	**-31.587**	0.035	0.017	**0**	**-30.180**	0.020	0.015	**0**
	APW$_\lambda$	**-25.724**	0.027	0.022	**0**	**-12.793**	0.022	0.019	**0**	**-30.172**	0.021	0.017	**0**
GLMPM	ML$_\lambda$	-0.112	0.017	0.0170	94.6	-0.095	0.017	0.017	93.6	-0.227	0.013	0.013	94.8
	APW$_\lambda$	1.180	0.021	0.020	89.4	1.740	0.021	0.020	90.4	0.937	0.017	0.016	90.0
Scenario 4: $n = 500, J_i = J = 3, K = 4$													
GLMM	ML$_\lambda$	**-29.656**	0.086	0.014	**0**	**-30.103**	0.028	0.012	**0**	**-29.507**	0.018	0.010	**0**
	APW$_\lambda$	**-25.609**	0.019	0.016	**0**	**-62.998**	0.015	0.013	**0**	**-53.435**	0.015	0.012	**0**
GLMPM	ML$_\lambda$	-0.067	0.012	0.012	94.6	-0.151	0.012	0.012	94.2	-0.144	0.009	0.009	94.8
	APW$_\lambda$	1.190	0.015	0.015	85.6	1.625	0.014	0.014	92.2	0.797	0.012	0.012	90.6

Table 4.23: Simulation results for doubly selection fixed and random effects under the case that data are generated by GLMPMs: estimation of random effects parameters in the Poisson mixed model

		d_1		d_2		d_3		r_{12}		r_{13}		r_{23}		ρ^*		ρ_e	
		Bias†	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE	Bias	ESE
Scenario 1: $n = 250, J_i = J = 1, K = 9$																	
GLMM	ML$_\lambda$	0.480	0.030	0.214	0.026	0.100	0.023	-0.903	0.036	-0.992	0.048	0.443	0.015	-	-	-	-
	APW$_\lambda$	0.616	0.029	0.347	0.020	0.178	0.012	-0.967	0.013	-1.061	0.013	0.472	0.003	-	-	-	-
GLMPM	ML$_\lambda$	-0.036	0.023	-0.004	0.004	-0.004	0.004	0.026	0.081	0.017	0.077	-0.018	0.039	-	-	0.003	0.065
	APW$_\lambda$	-0.004	0.022	-0.001	0.001	-0.001	0.002	-0.031	0.043	-0.028	0.042	-0.011	0.015	-	-	-0.012	0.069
Scenario 2: $n = 500, J_i = J = 1, K = 9$																	
GLMM	ML$_\lambda$	0.471	0.023	0.215	0.019	0.098	0.016	-0.895	0.028	-0.986	0.038	0.444	0.010	-	-	-	-
	APW$_\lambda$	0.619	0.022	0.349	0.016	0.180	0.010	-0.967	0.009	-1.061	0.009	0.472	0.002	-	-	-	-
GLMPM	ML$_\lambda$	-0.037	0.017	-0.004	0.003	-0.004	0.003	0.031	0.059	0.019	0.060	-0.017	0.029	-	-	0.003	0.043
	APW$_\lambda$	-0.003	0.015	-0.001	0.001	-0.001	0.001	-0.028	0.029	-0.025	0.029	-0.011	0.010	-	-	-0.012	0.045
Scenario 3: $n = 250, J_i = J = 3, K = 4$																	
GLMM	ML$_\lambda$	0.516	0.026	0.137	0.040	0.044	0.018	-0.161	0.339	-0.272	0.196	0.244	0.047	0.271	0.044	-	-
	APW$_\lambda$	0.681	0.033	0.179	0.005	0.059	0.002	-0.462	0.025	-0.495	0.023	-0.284	0.006	0.267	0.002	-	-
GLMPM	ML$_\lambda$	-0.006	0.019	-0.001	0.002	-0.001	0.002	-0.010	0.030	-0.009	0.031	-0.006	0.015	0.002	0.019	0.001	0.038
	APW$_\lambda$	0.001	0.010	-0.000	0.001	-0.000	0.001	-0.003	0.014	-0.002	0.014	-0.001	0.005	0.001	0.003	-0.036	0.043
Scenario 4: $n = 500, J_i = J = 3, K = 4$																	
GLMM	ML$_\lambda$	0.511	0.022	0.150	0.036	0.046	0.014	-0.273	0.311	-0.332	0.157	0.250	0.036	0.257	0.042	-	-
	APW$_\lambda$	0.686	0.023	0.178	0.006	0.058	0.003	-0.460	0.024	-0.495	0.022	0.283	0.006	0.267	0.002	-	-
GLMPM	ML$_\lambda$	-0.003	0.012	-0.001	0.001	-0.001	0.001	-0.008	0.023	-0.007	0.021	-0.005	0.010	0.002	0.014	0.000	0.027
	APW$_\lambda$	0.003	0.008	-0.000	0.001	-0.000	0.001	-0.002	0.011	-0.001	0.011	-0.001	0.004	0.001	0.003	-0.032	0.035

Model Selection under Misspecified Models

Table 4.24: Simulation results for the fixed effects selection under misspecified linear mixed model: model selection precision rate (%) for each variable

		RCS-Nonzero			RCS-Zero				
	Method	β_1	β_2	β_5	β_3	β_4	β_6	β_7	β_8
	ML($\surd$)[†]	100[*]	93.2	99.0	96.0[**]	96.8	96.2	98.4	98.0
	ML($\times$)	100	2.6	3.6	96.0	7.0	6.0	97.4	97.8
Example 1	APW($\surd$)	100	98.4	99.2	78.6	80.6	81.0	85.8	86.0
$n = 250, J_i = J = 1, K = 5$	APW($\times$)	100	20.8	24.4	77.0	2.8	1.8	80.8	83.6
	APC($\surd$)	100	98.6	99.4	81.6	83.0	84.2	87.6	88.0
	APC($\times$)	100	18.4	19.8	81.6	2.6	2.0	83.6	85.2
	ML($\surd$)	100	100	100	98.6	97.8	99.4	99.6	99.6
	ML($\times$)	100	1.2	0.6	98.8	0	0	99.2	100
Example 2	APW($\surd$)	100	100	100	95.4	94.0	97.2	98.4	97.6
$n = 1000, J_i = J = 1, K = 5$	APW($\times$)	100	5.6	10.0	95.0	0	0	96.4	97.4
	APC($\surd$)	100	100	100	97.4	95.8	98.4	99.0	98.2
	APC($\times$)	100	2.6	4.6	97.8	0	0	98.4	98.0
	ML($\surd$)	100	85.4	98.8	94.8	96.4	97.0	98.2	98.4
	ML($\times$)	100	6.4	7.2	92.8	20.6	18.4	95.2	96.4
Example 3	APW($\surd$)	100	97.6	100	69.2	74.2	77.0	81.2	77.8
$n = 250, J_i = J = 3, K = 3$	APW($\times$)	100	26.6	28.8	71.4	5.8	4.8	79.0	77.4
	APC($\surd$)	100	97.8	100	70.8	75.2	78.0	82.2	79.0
	APC($\times$)	100	25.6	27.4	71.6	5.2	4.6	79.2	78.6
	ML($\surd$)	100	97.8	99.8	98.0	97.8	98.4	99.2	98.6
	ML($\times$)	100	1.8	3.6	97.2	6.8	7.6	98.2	97.8
Example 4	APW($\surd$)	100	99.6	100	80.6	79.8	84.0	89.2	90.4
$n = 500, J_i = J = 3, K = 3$	APW($\times$)	100	19.4	26.0	80.6	1.8	1.6	85.6	88.0
	APC($\surd$)	100	99.8	100	84.0	79.2	86.4	90.6	88.6
	APC($\times$)	100	20.6	25.8	81.8	1.4	1.0	86.8	88.6

[†] ML, APW, APC represent maximum likelihood, all-pairwise marginal pairwise likelihood, all-pairwise conditional pairwise likelihood, respectively. ($\surd$) and ($\times$) denote the model with correct skewed-normal random effects and the model with incorrect normal random effects, respectively.

[*] The values under "RCS-Nonzero" column presents the rate of corresponding non-zero coefficient that is correctly estimated as non-zero.

[**] The values under "RCS-Zero" depicts the rate of related zero coefficient that is correctly set to zero.

Table 4.25: Simulation results for the fixed effects selection under misspecified linear mixed model: regression coefficient estimation for β_1

		β_1			
	Method	Bias(%)	ESE	ASE	CP(%)
Example 1 $n = 250, J_i = J = 1, K = 5$	ML(√)[†]	-1.362	0.171	0.160	93.8
	ML(×)	50.331	0.174	0.158	3.4
	APW(√)	-9.910	0.197	0.148	80.2
	APW(×)	38.843	0.191	0.153	22.2
	APC(√)	-9.014	0.187	0.154	82.2
	APC(×)	40.816	0.183	0.145	15.2
Example 2 $n = 1000, J_i = J = 1, K = 5$	ML(√)	-0.258	0.081	0.080	95.0
	ML(×)	51.557	0.082	0.080	0
	APW(√)	-2.426	0.095	0.093	91.2
	APW(×)	49.022	0.096	0.087	0
	APC(√)	-1.935	0.089	0.088	93.2
	APC(×)	50.045	0.090	0.086	0
Example 3 $n = 250, J_i = J = 3, K = 3$	ML(√)	1.315	0.188	0.179	93.2
	ML(×)	42.819	0.190	0.173	15.6
	APW(√)	-9.899	0.205	0.145	80.6
	APW(×)	28.163	0.202	0.162	52.2
	APC(√)	-9.634	0.199	0.150	81.4
	APC(×)	28.729	0.196	0.158	50.0
Example 4 $n = 500, J_i = J = 3, K = 3$	ML(√)	0.556	0.123	0.129	96.4
	ML(×)	43.653	0.119	0.124	1.2
	APW(√)	-6.133	0.139	0.130	88.0
	APW(×)	33.048	0.134	0.121	14.2
	APC(√)	-5.737	0.142	0.146	86.4
	APC(×)	33.331	0.135	0.127	12.4

[†] ML, APW, APC represent maximum likelihood, all-pairwise marginal pairwise likelihood, all-pairwise conditional pairwise likelihood, respectively. (√) and (×) denote the model with correct skewed-normal random effects and the model with incorrect normal random effects, respectively.

Chapter 5

Variable Selection via Composite Likelihood for Incomplete Longitudinal Data Arising in Clusters

5.1 Introduction

Longitudinal data arising in clusters are typically collected by following up subjects in clusters over a period of time. Incomplete data and variable selection issues are important for such data. Incompleteness of data presents a challenge in standard analysis methods, because analysis with missingness ignored may lead to biased results. On the other hand, irrelevantly incorporating a large number of covariates to the model may result in the difficulty of computation, interpretation and prediction, thus parsimonious models are typically desirable. Many existing methods focus on handling either the missing data or the variable selection, but not both (e.g. Wu and Carroll, 1988; Diggle and Kenward, 1994; Little, 1995; Akaike, 1973; Tibshirani, 1997; Fan and Li, 2001). Ni et al. (2010) propose a double-penalized likelihood approach to deal with the model selection for incomplete response data with missing at random (MAR)(Little and Rubin, 2002), but the method is not applicable if missingness occurs in both the response and the covariates under missing not at random (MNAR) scenarios.

Another particular issue for longitudinal data arising in clusters may be attributed to substantially increased modeling complexity and computational difficulty. With clusters present in longitudinal studies, the likelihood function may become cumbersome. Fieuws and Verbeke (2006) argue that for longitudinal data arising in clusters under random effects models, computation will become difficult as the dimension of the random effects vector increases.

It is desirable to develop methods that can accommodate missingness, variable selection and complex modeling issues. In this chapter, we propose a unified penalized missingness modified composite likelihood framework (Lindsay, 1988; Arnold and Strauss, 1991; Cox and Reid, 2004; Lindsay et al., 2011) to handle various features. In particular, our method can accommodate data missing not at random (MNAR) for both the response and the covariates. Moreover, it is flexible to handle the situation when the response and the covariates are missing not simultaneously. For the missing not at random (MNAR) case, our inference requires only some "structural" assumptions for the missing data process. Under the assumptions, we do not need to specify a specific model form for the missing data process, which circumvents the misspecification and non-identifiability problems (Fitzmaurice et al., 1996). We further add penalized terms in the likelihood functions to facilitate the variable selection, while the composite likelihood formulations involve simpler model form and cheapness in computation.

The remainder of this chapter is organized as follows. In Section 5.2, we introduce notations and models. In Section 5.3, we provide details on the inference strategy. A study of the NPHS data will be illustrated in Section 5.4. Numerical studies concerning asymptotic bias will be given in Section 5.5. Concluding remarks are given in Section 5.6.

5.2 Model Formulation

5.2.1 Generalized Linear Mixed Models

Suppose that there are n clusters and J_i subjects within cluster i, $i = 1, 2, \ldots, n$. Further suppose that there are K visits planed. Let Y_{ijk} denote the response for subject j in cluster i at the visit k, $k = 1, 2, \ldots, K$. Take $Y_{ij} = (Y_{ij1}, Y_{ij2}, \ldots, Y_{ijK})^T$, $j = 1, 2, \ldots, J_i$. and $Y_i = (Y_{i1}^T, Y_{i2}^T, \ldots, Y_{iJ_i}^T)^T$, $i = 1, 2, \ldots, n$. Let $X_{ijk} = (X_{ijk,1}, \ldots, X_{ijk,p})^T$ be the $p \times 1$ fixed effect covariate vector for subject j in cluster i at time k, $X_{ij} = (X_{ij1}^T, X_{ij2}^T, \ldots, X_{ijK}^T)^T$, and $X_i = (X_{i1}^T, X_{i2}^T, \ldots, X_{iJ_i}^T)^T$. Let $Z_{ijk} = (Z_{ijk,1}, \ldots, Z_{ijk,q})^T$ be the $q \times 1$ random effect covariates vector. Z_{ij} and Z_i are defined by following the similar pattern as X_{ij} and X_i. Let u_i denote a random effects vector corresponding to cluster i, $i = 1, 2, \ldots, n$.

Conditional on random effects u_i and covariate vectors, Y_{ijk} follows the distribution given by

$$f(Y_{ijk}|X_i, Z_i, u_i) = \exp\Big[\{Y_{ijk}\tau_{ijk} - b(\tau_{ijk})\}/a(\phi) + c(Y_{ijk}; \phi)\Big], \tag{5.1}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are some specific functions, ϕ is a scale parameter, τ_{ijk} is a canonical parameter, $\mathrm{E}(Y_{ijk}|X_i, Z_i, u_i) = b'(\tau_{ijk})$ and $\mathrm{Var}(Y_{ijk}|X_i, Z_i, u_i) = a(\phi)b''(\tau_{ijk})$. We further consider a regression model

$$h\{\mathrm{E}(Y_{ijk}|X_i, Z_i, u_i)\} = X_{ijk}^T\boldsymbol{\beta} + Z_{ijk}^T u_i,$$

where h is a link function and $\boldsymbol{\beta}$ is a $p \times 1$ vector for fixed effect coefficients. Note that when $J_i = 1$ for all $i = 1, \ldots, n$, the aforementioned model becomes ordinary generalized linear mixed models (GLMMs) (Laird and Ware, 1982). Under the conditional independence assumption given random effects u_i, we have

$$f(Y_i|X_i, Z_i, u_i) = \prod_{j=1}^{J_i}\prod_{k=1}^{K} f(Y_{ijk}|X_i, Z_i, u_i),$$

and thus the statistical inference can be applied by the likelihood of Y_i with integrating out the unobservable random effects given by

$$f(Y_i|X_i, Z_i) = \int f(Y_i|X_i, Z_i, u_i)f(u_i)\, du_i \tag{5.2}$$

where $f(u_i)$ is the distribution of random effects u_i.

5.2.2 Missing Data Models

In longitudinal studies, individuals in clusters may not be completely observed at all occasions. Missingness can occur in both the response and the covariates measurements. Here we start with the case that all subjects in all clusters are observed at visit 1, but they can be missing at any other visit. Extensions to accommodating more general cases are discussed in Section 5.2.4.

Let $R_{ijk} = 1$ if the observation for cluster i, subject j at occasion k are complete (both the response and the covariates are fully observed) and $R_{ijk} = 0$ otherwise. Then we take $R_i = (R_{i11}, R_{i12}, \ldots, R_{iJ_iK})^T$. Write $Y_i = (Y_i^{obs}, Y_i^{mis})$, $X_i = (X_i^{obs}, X_i^{mis})$, and $Z_i = (Z_i^{obs}, Z_i^{mis})$, to distinguish the observed and unobserved components of Y_i, X_i and Z_i, respectively. The full likelihood for (Y_i, X_i, Z_i, R_i) in the ith cluster can be written as

$$f(Y_i, X_i, Z_i, R_i) = f(R_i|Y_i, X_i, Z_i; \boldsymbol{\phi}) f(X_i, Z_i|\boldsymbol{\upsilon}) f(Y_i|X_i, Z_i; \boldsymbol{\psi}),$$

where parameters $\boldsymbol{\phi}$, $\boldsymbol{\upsilon}$ and $\boldsymbol{\psi}$ are assumed to be functionally independent.

If the missing data mechanism is missing not at random (MNAR), we have

$$f(R_i|Y_i, X_i, Z_i; \boldsymbol{\phi}) = f(R_i|Y_i^{obs}, Y_i^{mis}, X_i^{obs}, X_i^{mis}, Z_i^{obs}, Z_i^{mis}; \boldsymbol{\phi}),$$

where the missing data probability depends on the unobserved components of Y_i, X_i and Z_i.

Therefore, the statistical inference can use the observed data full likelihood function

$$f(R_i, Y_i^{obs}, X_i^{obs}, Z_i^{obs}) = \iiint f(R_i|Y_i, X_i, Z_i; \boldsymbol{\phi}) f(X_i, Z_i|\boldsymbol{\upsilon}) f(Y_i|X_i, Z_i; \boldsymbol{\psi})\, dY_i^{mis}\, dX_i^{mis}\, dZ_i^{mis}, \tag{5.3}$$

where the integrals are taken for all unobserved responses and covariates. The observed data likelihood function in (5.3) requires fully specification of response process $f(Y_i|X_i, Z_i; \boldsymbol{\psi})$, covariates process $f(X_i, Z_i|\boldsymbol{\upsilon})$ and missing data process $f(R_i|Y_i, X_i, Z_i; \boldsymbol{\phi})$. Moreover, the full likelihood estimation for interested parameter $\boldsymbol{\psi}$ involves a large set of nuisance parameters $\boldsymbol{\phi}$ and $\boldsymbol{\upsilon}$. To circumvent the difficulties in the full likelihood, we propose a composite likelihood strategy.

5.2.3 Composite Likelihood

In the spirit of the conditional likelihood discussed in Fitzmaurice et al. (2005), we assume the missing mechanism satisfies

$$P(R_{ijk} = 1|Y_{ijk}, Y_{ij'1}, X_i, Z_i) = P(R_{ijk} = 1|Y_{ijk}, X_i, Z_i), \tag{5.4}$$

for all $k \neq 1$, $j = 1, \ldots, J_i$ and $j' = 1, \ldots, J_i$.

Under the assumption in (5.4), we can prove that the conditional likelihood form for $Y_{ij'1}$ ($j' = 1, \ldots, J_i$) given observed Y_{ijk} ($j = 1, \ldots, J_i; k \neq 1$) has

$$f(Y_{ij'1} \mid Y_{ijk}, X_i, Z_i, R_{ijk} = 1; \boldsymbol{\psi}, \boldsymbol{\phi}) = f(Y_{ij'1} \mid Y_{ijk}, X_i, Z_i; \boldsymbol{\psi}).$$

This implies that in cluster i, the conditional distribution of $Y_{ij'1}$ given Y_{ijk} in a complete observation for subject j at occasion k, equals to the conditional distribution of $f(Y_{ij'1} \mid Y_{ijk}, X_i, Z_i; \boldsymbol{\psi})$. Therefore, it can be shown that the log likelihood obtained from the complete observation $I(R_{ijk} = 1) \log f(Y_{ij'1} \mid Y_{ijk}, X_i, Z_i; \boldsymbol{\psi})$ leads to unbiased estimating equations. The proof is sketched in Appendix 1.

However, the assumption in (5.4) may not secure such equalities for the marginal form of Y_{ijk} to have

$$f(Y_{ijk} \mid X_i, Z_i, R_{ijk} = 1; \boldsymbol{\psi}, \boldsymbol{\phi}) = f(Y_{ijk} \mid X_i, Z_i; \boldsymbol{\psi}),$$

and the conditional form for Y_{ijk} given $Y_{ij'1}$ to have

$$f(Y_{ijk} \mid Y_{ij'1}, X_i, Z_i, R_{ijk} = 1; \boldsymbol{\psi}, \boldsymbol{\phi}) = f(Y_{ijk} \mid Y_{ij'1}, X_i, Z_i; \boldsymbol{\psi}).$$

Thus, the log likelihood functions obtained from the complete observation

$$I(R_{ijk}=1)\log f(Y_{ijk}\mid X_i, Z_i;\boldsymbol{\psi}),$$

and

$$I(R_{ijk}=1)\log f(Y_{ijk}\mid Y_{ij'1}, X_i, Z_i;\boldsymbol{\psi})$$

may lead to biased estimation equations.

Therefore, the inference can take the log composite likelihood for subject i as

$$\begin{aligned} &\ell(Y_i)\\ =\;& \log\left\{\prod_{j<j'}\left\{f(Y_{ij1}|Y_{ij'1}, X_i, Z_i)f(Y_{ij'1}|Y_{ij1}, X_i, Z_i)\right\}\times\prod_{\substack{j=1,\ldots,J_i\\ j'=1,\ldots,J_i\\ k\neq 1}} f(Y_{ij'1}|Y_{ijk}, X_i, Z_i)^{I(R_{ijk}=1)}\right\}. \end{aligned} \tag{5.5}$$

According to (5.5), we need a composite likelihood modeling strategy by implementing pairwise conditional log likelihood forms. The key difference between the composite likelihood and full likelihood methods is, instead of working on the *full* distribution structure, the composite likelihood approach only centers on *partial* structures of the probability distributions. In particular, the log likelihood (5.5) only requires the specification of conditional distribution form as Y_{ijk} given $Y_{ij'k'}$, X_i, and Z_i, which can be obtained from (5.1) by

$$\begin{aligned} f(Y_{ijk}|Y_{ij'k'}, X_i, Z_i) &= \frac{f(Y_{ijk}, Y_{ij'k'}, X_i, Z_i)}{f(Y_{ij'k'}, X_i, Z_i)}\\ &= \frac{\int f(Y_{ijk}|X_i, Z_i, u_i)f(Y_{ij'k'}|X_i, Z_i, u_i)f(u_i)\,du_i}{\int f(Y_{ij'k'}|X_i, Z_i, u_i)f(u_i)\,du_i}. \end{aligned}$$

Note that comparing with the full likelihood (5.3), our log composite likelihood function (5.5) does not involve the specification of the covariates process $f(X_i, Z_i|\boldsymbol{\upsilon})$ and the missing data process $f(R_i|Y_i, X_i, Z_i;\boldsymbol{\phi})$. Moreover, the integrals for unobserved response Y^{mis} and covariates X^{mis}, Z^{mis} are not included in (5.5). Thus, the composite likelihood shows modeling tractability and computational cheapness.

5.2.4 Extensions

Previous discussions assume that all subjects in clusters are observed at visit 1, but they can be missing at any other visits. In applications, this requirement may be too restrictive. Moreover, assumption (5.4) is quite strong since it does not allow the missingness depends on any of the response in the first occasion.

In fact, the missingness modified composite likelihood approach can be applied as long as for each cluster i, there exists some j, k which are free of missingness. Let $\mathcal{S}_i$ be the subset of $(i11, i12, \dots, iJ_iK)$ that includes the missingness-free occasions for cluster i, while $\mathcal{R}_i$ is the complementary of $\mathcal{S}_i$ to display the missingness-prone occasions for cluster i. Then assume

$$P(R_{ijk} = 1|Y_{ijk}, Y_{ij'k'}, X_i, Z_i) = P(R_{ijk} = 1|Y_{ijk}, X_i, Z_i), \tag{5.6}$$

for some $ijk \in \mathcal{R}_i$ and $ij'k' \in \mathcal{S}_i$. Let $(ijk, ij'k') \in \mathcal{A}$ if they meet the assumption (5.6). We obtain a general composite likelihood for cluster i as

$$\begin{aligned} \ell(Y_i) &= \log\Bigg[\prod_{\substack{ijk\in\mathcal{S}_i \\ ij'k'\in\mathcal{S}_i}} \left\{ f(Y_{ijk}|Y_{ij'k'}, X_i, Z_i) f(Y_{ij'k'}|Y_{ijk}, X_i, Z_i) \right\} \\ &\quad \times \prod_{(ijk,ij'k')\in\mathcal{A}} f(Y_{ij'k'}|Y_{ijk}, X_i, Z_i)^{I(R_{ijk}=1)} \Bigg]. \end{aligned} \tag{5.7}$$

Therefore, the composite likelihood function has

$$\ell(Y) = \sum_{i=1}^{n} \ell(Y_i). \tag{5.8}$$

5.3 Selecting Fixed Effects Using the Composite Likelihood

In this section, we focus on selecting fixed effect. Denote $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \xi^T)^T$, where ξ represents all parameters other than $\boldsymbol{\beta}$. To achieve both the model selection and the parameter estimation, we propose to maximize the following penalized composite log likelihood function:

$$\ell_{pen}(Y;\boldsymbol{\psi}) = \ell(Y;\boldsymbol{\psi}) - n\sum_{s=1}^{p} p_\lambda(|\beta_s|), \tag{5.9}$$

where $\ell(Y;\boldsymbol{\psi})$ is defined in (5.8), $p_\lambda(|\beta_s|)$ is the penalty function for the s-th element in $\boldsymbol{\beta}$. Following Fan and Li (2001, 2004), we adopt the SCAD penalty. The SCAD penalty is a nonconcave function defined by $p_\lambda(0) = 0$ and for $\beta_s > 0$, its first derivative satisfies

$$p'_\lambda(\beta_s) = \lambda\left\{I(\beta_s \leq \lambda) + \frac{(a\lambda - \beta_s)_+}{(a-1)\lambda} I(\beta_s > \lambda)\right\}$$

for some $a > 2$ and $\lambda > 0$. In practice, 2-dimensional grid searching for optimal tuning parameter (a, λ) can be computational expensive. Based on the calculation of Bayesian risk, Fan and Li (2001) suggests setting $a = 3.7$, and only searching for λ.

Given known values of tuning parameter $a = 3.7$ and $\lambda^{(r)}$, the estimate of $\boldsymbol{\psi}$, denoted by $\hat{\boldsymbol{\psi}}_r$, is the maximizer of the penalized composite likelihood. That is

$$\hat{\boldsymbol{\psi}}_r = \text{argmax}_{\boldsymbol{\psi}} \ell_{pen}(Y;\boldsymbol{\psi}).$$

The maximization can be implemented using the Newton-Raphson algorithm. However, the SCAD penalty function is singular at the origin, and does not have continuous second order derivatives. We can apply the local quadratic approximation approach proposed by Fan and Li (2001) to circumvent this problem with a modified Newton-Raphson algorithm for the tth iteration:

$$\boldsymbol{\psi}_r^{(t+1)} = \boldsymbol{\psi}_r^{(t)} - \left\{\frac{\partial^2 \ell(Y;\boldsymbol{\psi}^{(t)})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T} - nW^{(t)}\right\}^{-1}\left\{\frac{\partial \ell(Y;\boldsymbol{\psi}^{(t)})}{\partial\boldsymbol{\psi}} - nU^{(t)}\right\}, \qquad t = 0, 1, \ldots \tag{5.10}$$

where $W^{(t)} = \text{diag}\{p'_{\lambda_r}(|\beta_1^{(t)}|)/|\beta_1^{(t)}|, \ldots, p'_{\lambda_r}(|\beta_p^{(t)}|)/|\beta_p^{(t)}|, \mathbf{0}_\xi\}$, $U^{(t)} = W^{(t)} \cdot (\boldsymbol{\beta}^{(t)T}, \mathbf{0}_\xi^T)^T$, and $\mathbf{0}_\xi$ is the 0-vector with equal length as ξ. If $\beta_s^{(t+1)}$ is very close to 0, then set $\hat{\beta}_s = 0$, and remove its corresponding elements in (5.10) from the iteration. The estimates $\hat{\boldsymbol{\psi}}_r$ is obtained when all parameters converge to a stable set.

The aforementioned maximization algorithm is conducted based on a fixed tuning parameters $(a^{(r)}, \lambda^{(r)})$. In practice, $(a^{(r)}, \lambda^{(r)})$ is chosen on a grid and the solution $\hat{\boldsymbol{\psi}}_r$ is obtained for each r. The final model selection and estimates $\hat{\boldsymbol{\psi}}$ can be realized based on certain selection critera. For instance, recent studies (Wang et al., 2007; Bondell et al., 2010; Ma and Li, 2010; Zhang et al., 2010) show that the Bayesian information criterion (BIC) is consistent for model selection given the true model lies in the class of candidate models. The BIC criterion has the form

$$\text{BIC}_{(a^{(r)},\lambda^{(r)})} = -2\ell(Y;\hat{\boldsymbol{\psi}}_r) + \log(n) \times \text{df}_{(a^{(r)},\lambda^{(r)})}(\tilde{\boldsymbol{\psi}}_r), \tag{5.11}$$

where $\tilde{\boldsymbol{\psi}}_r$ denotes the parameter set in which the 0 elements in $\hat{\boldsymbol{\psi}}_r$ are removed, $\text{df}_{(a^{(r)},\lambda^{(r)})}(\tilde{\boldsymbol{\psi}}_r)$ is the effective number of degrees of freedom given by $\text{tr}(\tilde{J}(\hat{\boldsymbol{\psi}}_r)\tilde{H}(\hat{\boldsymbol{\psi}}_r)^{-1})$, (Varin and Vidoni, 2005; Gao and Song, 2010) where $\tilde{H}(\hat{\boldsymbol{\psi}}_r) = -\frac{\partial^2 \ell_{pen}(Y;\hat{\boldsymbol{\psi}}_r)}{\partial\tilde{\boldsymbol{\psi}}_r \partial\tilde{\boldsymbol{\psi}}_r^T}$, and

$$\tilde{J}(\hat{\boldsymbol{\psi}}_r) = \sum_{i=1}^{n} \frac{\partial \ell(Y_i;\hat{\boldsymbol{\psi}}_r)}{\partial\tilde{\boldsymbol{\psi}}_r} \{\frac{\partial \ell(Y_i;\hat{\boldsymbol{\psi}}_r)}{\partial\tilde{\boldsymbol{\psi}}_r}\}^T.$$

Under some mild regulation conditions (see Appendix 2), the asymptotic properties for our method can be established. The proof is sketched in Appendix 3.

5.4 Application

The National Population Health Survey (NPHS) is a longitudinal study that collects health information and related socio-demographic information by following a group of Canadian household residents. The questions for the NPHS include many aspects of in-depth health information such as health status, use of health services, chronic conditions and activity restrictions. Moreover, social background questions, including age, sex and income level, are contained in the questionnaire. A research interest focuses on modeling the influence of income on population health. The data we analyze here contain 6 cycles' observations (from Cycle 1 to Cycle 6), including n = 1033 males with age between 50-70 at Cycle 1, and less than 80 at Cycle 6. All the deceased subjects are excluded from the analysis.

Health status (HUI) is measured by the Health Utilities Index Mark after zero-mean normalization. The higher HUI score indicates a better health status. The covariate prone to missingness is household income (INC), which is measured by provincial level of household income with zero-mean normalization. The other covariate, denoted by CYCLE is cycle number after log-transformation, respectively.

In the data set, the first two occasions are complete for all subjects. However, only 43.2% of the individuals have complete observations for both the response and the covariate in the following 4 cycles. The missingness proportions in the response in the following 4 cycles are 11.9%, 16.8%, 22.3%, and 25.6%, respectively, while the missingness proportions in the covariate are 17.1%, 24.0%, 29.0% and 33.4%, respectively. Various types of missingness patterns are present. A sample of summarized proportions is displayed in Table 5.1.

Table 5.1: Missing data proportions for HUI and INC variables in the NPHS data (%)

Percentage	HUI						INC					
	1	2	3	4	5	6	1	2	3	4	5	6
43.2%	√	√	√	√	√	√	√	√	√	√	√	√
...												
2%	√	√	√	√	√	√	√	√	×	√	√	√
1%	√	√	√	√	×	√	√	√	√	√	×	√
1%	√	√	√	√	×	×	√	√	√	√	×	×

√ Observed; × Missing

Orpana et al. (2009) indicate that random intercept is sufficient to account for the correlation across cycles. Moreover, both cubic terms of INC and CYCLE with interactions are of interest in the modeling of HUI. This motivates us to consider variable selection in the following model

$$Y_{ijk} = X_{ijk}\beta + u_{ij} + \varepsilon_{ijk}, \tag{5.12}$$

where $J_i = 1$ for all i, Y_{ijk} is the HUI score for subject i measured at Cycle k, X_{ijk} is a 16×1 vector of variables measured at j: Intercept, INC, INC^2, INC^3, CYCLE, CYCLE^2, CYCLE^3, CYCLE, INC × CYCLE, INC^2 × CYCLE, INC^3 × CYCLE, INC × CYCLE^2, INC^2 × CYCLE^2, INC^3 × CYCLE^2, INC × CYCLE^3, INC^2 × CYCLE^3, INC^3 × CYCLE^3. $u_{ij} \sim N(0, \sigma_u^2)$ is the subject specific random effect and $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$ is the independent residual error.

We apply our composite likelihood procedure (CL) to model (5.12). As a comparison, we employ a naive approach that is often used by analysts to handle data with missing observations. That is, we apply the naive maximum likelihood method to the complete data only, and denote this method by NML. In the NML approach, all incomplete observations are ignored and only complete data are used for estimation, where the likelihood formula for the ith cluster can be written as

$$L_i^{NML} = \int \left[\prod_{j=1}^{J_i} \prod_{k=1}^{K} \{ f(Y_{ijk} | X_i, Z_i, u_i) \}^{R_{ijk}} \right] f(u_i)\, du_i.$$

Table 5.2 displays the model fitting and selection results. Two methods obtain relatively comparable results that income has only a linear effect on health index. They also suggest a cubic temporal effect. The NML approach excludes all of the interaction terms, while CL maintains some higher order interaction terms.

Table 5.2: Analysis results for the NPHS data: entries represent the estimates and standard errors (in brackets)

Variable	NML[†]		CL	
	Full Model	Selected Model	Full Model	Selected Model
Intercept	−0.020(0.041)	0.001(0.028)	-0.006(0.041)	0.016(0.031)
INC	0.109(0.064)	0.085(0.014)	0.092(0.068)	0.080(0.019)
INC^2	-0.012(0.027)		-0.020(0.028)	
INC^3	-0.003(0.033)		0.002(0.034)	
CYCLE	0.349(0.216)	0.073(0.019)	0.563(0.232)	0.080(0.019)
CYCLE^2	-0.284(0.315)		-0.669(0.380)	
CYCLE^3	0.033(0.116)	-0.039(0.007)	0.164(0.147)	-0.044(0.007)
INC × CYCLE	-0.258(0.380)		-0.084(0.468)	
INC^2 × CYCLE	-0.130(0.166)		-0.229(0.201)	0.017(0.008)
INC^3 × CYCLE	0.236(0.203)		0.398(0.262)	
INC × CYCLE^2	0.291(0.551)		-0.067(0.778)	
INC^2 × CYCLE^2	0.165(0.244)		0.422(0.339)	
INC^3 × CYCLE^2	-0.353(0.297)		-0.717(0.441)	-0.013(0.006)
INC × CYCLE^3	-0.092(0.202)		0.039(0.307)	
INC^2 × CYCLE^3	-0.047(0.090)		-0.149(0.132)	
INC^3 × CYCLE^3	0.134(0.109)		0.269(0.174)	-0.022(0.005)

[†] NML and CL represent naive maximum likelihood to complete data and our composite likelihood, respectively.

5.5 Simulation Studies of the Proposed Methods

5.5.1 Measure of Marginal Model Error

In this section, we implement the proposed method to various models, including linear mixed models and Poisson mixed models. First, we describe a measure that is used to feature the performance of the estimates obtained from different models.

Let $\mu(\cdot) = E_{u_i}\big\{E(Y_{ijk}|X_i, Z_i, u_i)\big\} = E_{u_i}\{h^{-1}(X_{ijk}^T\boldsymbol{\beta}_0+Z_{ijk}^T u_i)\}$, and $\hat{\mu}(\cdot) = E_{u_i}\{h^{-1}(X_{ijk}^T\hat{\boldsymbol{\beta}}+$

$Z_{ijk}^T u_i)\}$, where $h(\cdot)$ is the link function defined in (5.1), $\hat{\boldsymbol{\beta}}$ is an estimator obtained from the proposed method. The expectations are evaluated with respect to the true model. We define

$$\mathrm{MME}(\hat{\mu}(\cdot)) = E_{X_i,Z_i}\{\hat{\mu}(\cdot) - \mu(\cdot)\}^2,$$

and use this measure to quantify the marginal model error induced by estimator $\hat{\boldsymbol{\beta}}$. It can be seen that MME is a generalized model error measure (Fan and Li, 2001, 2002, 2004) that takes the random effects into considerations. Other available model error measure can be found from Bondell et al. (2010).

5.5.2 Linear Mixed Model

We now conduct a simulation study for the linear mixed model. The data are generated from the model

$$Y_{ijk} = X_{ijk}\boldsymbol{\beta} + u_{ij} + \epsilon_{ijk}, \tag{5.13}$$

where the ϵ_{ijk} are independently distributed with $N(0, \sigma_\epsilon^2)$, and independent of the u_{ij}. $u_i = (u_{i1}, \ldots, u_{iJ_i})^T$ are random effects with a given distributions. Set $\sigma_\epsilon^2 = 4$ and $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. Covariates $X_{ijk} = (X_{ijk,1}, X_{ijk,2}, \ldots, X_{ijk,8})^T$ are generated from a multivariate normal distribution with zero mean and covariance matrix $V = [\sigma_{st}^2]$, where $\sigma_{st}^2 = \rho_{st}\sigma^2$. We set $\rho_{st} = \rho^{|s-t|}$, $\rho = 0.5$ and $\sigma^2 = 1$.

We particularly consider the following scenarios.

Example 1: $n = 200$, $J = 1$, and $K = 10$. This corresponds to an ordinary longitudinal setting with 10 visits times. Random effects u_i are one-dimensional having a normal distribution $N_1(0, \sigma_u^2)$ with $\sigma_u^2 = 1$. For each subject, we set the first 2 occasions to be always observed while the rest 8 occasions to be subjected to missingness. In every missingness-prone observation, the probability of observing complete covariates expit$\{\gamma_0 + \gamma_1 Y_{ij} + \gamma_2 X_{ijk,1}\}$, and the probability of observing response $P(R^y_{ijk}=1)$ are set to be

$$\text{expit}\{\gamma_0+\gamma_1 Y_{ij}+\gamma_3\sum_{p=1}^{8} X_{ijk,p}+\gamma_4 R^x_{ijk}\},$$

where $\gamma_0=2.5$, $\gamma_1=-1.5$, $\gamma_2=-0.5$, $\gamma_3=-0.1$ and $\gamma_4=0.4$.

Example 2: The setup follows from Example 1 but $n=800$.

Example 3: The setup follows from Example 1, except that we take $J_i=3$ and $K=5$. Take the first occasion to be always observed for each subject and set $u_i=(u_{i1},u_{i2},u_{i3})$ to be 3-dimensional random effects with $N_3(\mathbf{0},R)$, where

$$R=\sigma_u^2\begin{pmatrix}1&\rho^*&\rho^*\\ \rho^*&1&\rho^*\\ \rho^*&\rho^*&1\end{pmatrix},$$

with $\rho^*=0.5$.

Example 4: The setup follows from Example 3 but $n=800$.

5.5.3 Poisson Mixed Model

We now conduct a simulation study for the Poisson mixed model. The data are generated from the model

$$\log\{E(Y_{ijk}|X_i,Z_i,u_i)\}=X_{ijk}\boldsymbol{\beta}+u_{ij}. \tag{5.14}$$

where $\boldsymbol{\beta}=(1.2,0.6,0,0,0.8,0,0,0)^T$, u_{ij} and X_{ijk} are the same as that of linear mixed model.

We consider following scenario:

Example 1: $n=120$, $J=1$ and $K=10$. Other parameter settings follows from Example 1 in the linear mixed model. For each subject, we set the first 2 occasions to be always observed while the rest 8 occasions to be subjected to missingness. In every observation, the probability of observing complete covariates X_{ijk} is expit$\{\gamma_0 + \gamma_1 Y_{ij} + \gamma_2 X_{ijk,1}\}$, and the probability of observing complete response Y_{ijk} is

$$\text{expit}\{\gamma_0 + \gamma_1 Y_{ij} + \gamma_3 \sum_{p=1}^{8} X_{ijk,p} + \gamma_4 R^x_{ijk}\},$$

where $\gamma_0 = -1$, $\gamma_1 = 2$, $\gamma_2 = -0.5$, $\gamma_3 = -0.1$ and $\gamma_4 = 0.25$.

Example 2: The setup follows from Example 1, but $n = 500$.

Example 3: $n = 120$, $J_i = 3$, and $K = 5$. and set $u_i = (u_{i1}, u_{i2}, u_{i3})$ to be 3-dimensional random effect with $N_3(\mathbf{0}, R)$, where

$$R = \sigma_u^2 \begin{pmatrix} 1 & \rho^* & \rho^* \\ \rho^* & 1 & \rho^* \\ \rho^* & \rho^* & 1 \end{pmatrix},$$

with $\rho^* = 0.3$.

Example 4: The setup follows from Example 3, but $n = 300$.

We assess the performance of the proposed composite likelihood (CL) approach, in contrast to the naive maximum likelihood based on complete data (NML). All simulation results are included in Appendix 4. Tables 5.3 and 5.4 report the average of zero coefficients. The column labeled “Correct” presents the average of zero coefficients that are correctly estimated, and the column labeled “Incorrect” depicts the average of non-zero coefficients that are erroneously set to zero. We report the median ratios of MME, denoted by R.MME, for a selected model to that of the un-penalized estimate under the unpenalized model scenarios, respectively. We also report the median of MME, denoted by M.MME. Tables 5.5 and 5.6 summarize the estimated $(\beta_1, \beta_2, \beta_5)$, their relative biases, empirical, model-based standard errors and 95% coverage rate.

For all examples, both methods show a good sparsity property. The results show that our CL approach yields small biases and satisfactory coverage probabilities for both the mean and the association parameters. ASE and ESE agree reasonably well for the method, suggesting the consistency of variance estimates. The NML method, on the other hand, yields remarkably biased estimates and low coverage rate.

Appendix

1. Consistency

The proof involves two steps. For the first step, we prove

$$f(Y_{ij'1} \mid Y_{ijk}, X_i, Z_i, R_{ijk} = 1; \boldsymbol{\psi}, \boldsymbol{\phi}) = f(Y_{ij'1} \mid Y_{ijk}, X_i, Z_i; \boldsymbol{\psi}), \tag{5.15}$$

for all $k \neq 1$, $j = 1, \ldots, J_i$ and $j' = 1, \ldots, J_i$.

Since we have

$$\begin{aligned}
& f(Y_{ij'1} \mid Y_{ijk}, X_i, Z_i, R_{ijk} = 1; \boldsymbol{\psi}, \boldsymbol{\phi}) \\
= \; & \frac{f(Y_{ijk}, Y_{ij'1}, R_{ijk} = 1 \mid X_i, Z_i; \boldsymbol{\psi}, \boldsymbol{\phi})}{f(Y_{ijk}, R_{ijk} = 1 \mid X_i, Z_i; \boldsymbol{\psi}, \boldsymbol{\phi})} \\
= \; & \frac{f(Y_{ijk}, Y_{ij'1} \mid X_i, Z_i; \boldsymbol{\psi}) f(R_{ijk} = 1 \mid Y_{ijk}, Y_{ij'1}, X_i, Z_i; \boldsymbol{\phi})}{f(Y_{ijk} \mid X_i, Z_i; \boldsymbol{\psi}) f(R_{ijk} = 1 \mid Y_{ijk}, X_i, Z_i; \boldsymbol{\phi})} \\
& \text{(By the assumption in (5.4))} \\
= \; & \frac{f(Y_{ijk}, Y_{ij'1} \mid X_i, Z_{ij}; \boldsymbol{\psi})}{f(Y_{ijk} \mid X_i, Z_i; \boldsymbol{\psi}},
\end{aligned}$$

which implies the conclusion.

Then we prove that the estimating equations obtained in (5.5) are unbiased estimating equations. We then obtain

$$
\begin{aligned}
& E_{Y_i,R_i|X_i,Z_i;\boldsymbol{\psi},\boldsymbol{\phi}}\left\{\frac{\partial \ell_i(Y_i|X_i,Z_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right\} \\
= \;& E_{Y_i,R_i|X_i,Z_i;\boldsymbol{\psi},\boldsymbol{\phi}}\Big[\sum_{j<j'}\Big\{\frac{\partial \log f(Y_{ij1}|Y_{ij'1},X_i,Z_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}+\frac{\partial \log f(Y_{ij'1}|Y_{ij1},X_i,Z_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big\} \\
& +\sum_{\substack{j=1,\ldots,J_i\\ j'=1,\ldots,J_i\\ k\neq 1}} I(R_{ijk}=1)\cdot\frac{\partial \log f(Y_{ij'1}\mid Y_{ijk},X_i,Z_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big] \\
= \;& \sum_{j<j'} E_{Y_i,R_i|X_i,Z_i;\boldsymbol{\psi},\boldsymbol{\phi}}\Big\{\frac{\partial \log f(Y_{ij1}|Y_{ij'1},X_i,Z_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}+\frac{\partial \log f(Y_{ij'1}|Y_{ij1},X_i,Z_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big\} \\
& +\sum_{\substack{j=1,\ldots,J_i\\ j'=1,\ldots,J_i\\ k\neq 1}} E_{R_{ijk}|X_i,Z_i;\boldsymbol{\psi},\boldsymbol{\phi}}\Bigg[I(R_{ijk}=1)E_{Y_i|X_i,Z_i,R_{ijk};\boldsymbol{\psi},\boldsymbol{\phi}}\Big\{\frac{\partial \log f(Y_{ij'1}\mid Y_{ijk},X_i,Z_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big\}\Bigg] \\
& \text{(By the settings that observations in visit 1 are always complete)} \\
= \;& 0+\sum_{\substack{j=1,\ldots,J_i\\ j'=1,\ldots,J_i\\ k\neq 1}} E_{R_{ijk}|X_i,Z_i;\boldsymbol{\psi},\boldsymbol{\phi}}\Big[I(R_{ijk}=1)E_{Y_{ij'1}|Y_{ijk},X_i,Z_i,R_{ijk};\boldsymbol{\psi},\boldsymbol{\phi}}\Big\{\frac{\partial \log f(Y_{ij'1}\mid Y_{ijk},X_i,Z_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big\}\Big] \\
& \text{(By the conclusion from (5.15))} \\
= \;& \sum_{\substack{j=1,\ldots,J_i\\ j'=1,\ldots,J_i\\ k\neq 1}} E_{R_{ijk}|X_i,Z_i;\boldsymbol{\psi},\boldsymbol{\phi}}\Big[I(R_{ijk}=1)E_{Y_{ij'1}|Y_{ijk},X_i,Z_i;\boldsymbol{\psi}}\Big\{\frac{\partial \log f(Y_{ij'1}\mid Y_{ijk},X_i,Z_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\Big\}\Big] \\
= \;& 0.
\end{aligned}
$$

2. Regularity Conditions

Now we establish the asymptotic distribution of the resulting estimator. Let $\boldsymbol{\beta}_0=(\beta_{10},\ldots,\beta_{p0})$ be the true parameter value of $\boldsymbol{\beta}$, and we write, without loss of generality, $\boldsymbol{\beta}_0=(\boldsymbol{\beta}_{0I}^T,\boldsymbol{\beta}_{0II}^T)^T$, where $\boldsymbol{\beta}_{0I}=(\beta_{10},\ldots,\beta_{p_10})^T$ is the $p_1\times 1$ vector consisting of all non-zero values while $\boldsymbol{\beta}_{0II}=(\beta_{p_1+1,0},\ldots,\beta_{p0})^T=\mathbf{0}_{\boldsymbol{\beta}_{0II}}^T$ is the $(p-p_1)\times 1$ vector. Thus, we have $\boldsymbol{\psi}_0=(\boldsymbol{\beta}_{0I}^T,\mathbf{0}_{\boldsymbol{\beta}_{0II}}^T,\xi_0^T)$ with ξ_0 being the true value of ξ. Correspondingly, write $\boldsymbol{\beta}=(\boldsymbol{\beta}_I^T,\boldsymbol{\beta}_{II}^T)^T$, $\boldsymbol{\psi}=(\boldsymbol{\beta}_I^T,\boldsymbol{\beta}_{II}^T,\xi^T)^T$.

For any square matrix M of the same dimension as $\boldsymbol{\psi}$, let $\tilde{M}$ denote the sub-matrix after removing the $(p_1+1,\dots,p)$ rows and columns from the matrix M. Similarly, for any vector α of the same dimension as $\boldsymbol{\psi}$, we use $\tilde{\alpha}$ to denote the resulting vector after removing the $(p_1+1,\dots,p)$ elements from the vector α. For example, $\tilde{\boldsymbol{\psi}}_0 = (\boldsymbol{\beta}_{0I}^T, \xi_0^T)^T$.

The following conditions are needed to establish the asymptotic properties of $\hat{\boldsymbol{\psi}}$.

(C1). For all i, $\ell(Y_i;\boldsymbol{\psi})$ is three-times continuously differentiable.

(C2). $\ell(Y_i;\boldsymbol{\psi})$, $|\frac{\partial \ell(Y_i;\boldsymbol{\psi})}{\partial \psi_j}|^2$, $|\frac{\partial^2 \ell(Y_i;\boldsymbol{\psi})}{\partial \psi_j \partial \psi_k}|$, and $|\frac{\partial^3 \ell(Y_i;\boldsymbol{\psi})}{\partial \psi_j \partial \psi_k \partial \psi_l}|$ are dominated by some functions $B_i(Y_i, X_i, Z_i)$ for all $j,k,l = 1,\dots,\dim(\boldsymbol{\psi})$, in which ψ_j is the $j-$th element of $\boldsymbol{\psi}$. Moreover, $E_{\boldsymbol{\psi}_0}\{B_i(Y_i,X_i,Z_i)\} < \infty$ for all i.

(C3). $E_{\boldsymbol{\psi}}\Big\{\frac{\partial \ell(Y_i;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big\} = \mathbf{0}$,

(C4). Let $M(\boldsymbol{\psi}) = E_{\boldsymbol{\psi}_0}\left[\Big\{\frac{\partial}{\partial \boldsymbol{\psi}}\ell(Y_i;\boldsymbol{\psi})\Big\}\Big\{\frac{\partial}{\partial \boldsymbol{\psi}}\ell(Y_i;\boldsymbol{\psi})\Big\}^T\right]$, and $D(\boldsymbol{\psi}) = E_{\boldsymbol{\psi}_0}\left\{-\frac{\partial^2 \ell(Y_i;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^T}\right\}$. Assume that

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{\partial}{\partial \boldsymbol{\psi}}\ell(Y_i;\boldsymbol{\psi})\Big\}\Big\{\frac{\partial}{\partial \boldsymbol{\psi}}\ell(Y_i;\boldsymbol{\psi})\Big\}^T = M(\boldsymbol{\psi}) + o_p(1),$$

and

$$-\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{\partial^2 \ell(Y_i;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^T}\Big\} = D(\boldsymbol{\psi}) + o_p(1).$$

Similar definitions and requirements are applied to $\tilde{M}(\tilde{\boldsymbol{\psi}})$ and $\tilde{D}(\tilde{\boldsymbol{\psi}})$.

(C5). There exists an open subset ω that contains the true parameter point $\boldsymbol{\psi}_0$ such that for all $\boldsymbol{\psi} \in \omega$, $D(\boldsymbol{\psi})$ and $\tilde{D}(\tilde{\boldsymbol{\psi}})$ are positive definite.

(C6). Let λ_n be the tunning parameter with the dependence on cluster size n explicitly spelled out. Define

$$a_n = \max_{s=1,\dots,p}\{p'_{\lambda_n}(|\beta_{s0}|) : \beta_{s0} \neq 0\},$$
$$b_n = \max_{s=1,\dots,p}\{p''_{\lambda_n}(|\beta_{s0}|) : \beta_{s0} \neq 0\},$$

We assume that

(C6.1). $\lambda_n = o_p(1)$,

(C6.2). $a_n = O_p(n^{-1/2})$,

(C6.3). $b_n = o_p(1)$.

(C7). We assume that

(C7.1). $\text{liminf}_{n\to\infty}\text{liminf}_{\epsilon\to 0^+} p'_{\lambda_n}(\epsilon)/\lambda_n > 0$.

(C7.2). $\text{lim}_{n\to\infty}\sqrt{n}\lambda_n = \infty$.

3. Asymptotic Results

Consistency

Theorem 1: There exists a local maximizer $\hat{\boldsymbol{\psi}}$ of $\ell_{pen}(Y;\boldsymbol{\psi})$ such that

$$\|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\| = O_p(n^{-1/2} + a_n).$$

Proof: Let $\alpha_n = n^{-1/2} + a_n$. Adapting the arguments by Fan and Li (2001, 2002), we need to show that for any given $\epsilon > 0$, there exists a large constant C_ϵ such that

$$P\left\{\text{sup}_{\|\mathbf{u}\|=C_\epsilon}\ell_{pen}(Y;\boldsymbol{\psi}_0 + \alpha_n\mathbf{u}) < \ell_{pen}(Y;\boldsymbol{\psi}_0)\right\} \geq 1 - \epsilon,$$

where $\mathbf{u} = ((u_1,\ldots,u_{p_1},\ldots,u_p)^T, u_\xi^T)^T$, u_ξ is a vector with the same length as ξ, and $\|x\| = \sqrt{x^T x}$.

Suppose C_ϵ is sufficiently large such that $\|(u_1,\ldots,u_{p_1})\| > 0$. Note that $p_{\lambda_n}(0) = 0$, we consider

$$
\begin{aligned}
K_n(\mathbf{u}) &= \ell_{pen}(Y;\boldsymbol{\psi}_0+\alpha_n\mathbf{u})-\ell_{pen}(Y;\boldsymbol{\psi}_0)\\
&= \ell_c(Y;\boldsymbol{\psi}_0+\alpha_n\mathbf{u})-\ell_c(Y;\boldsymbol{\psi}_0)-n\sum_{s=1}^{p}p_{\lambda_n}(|\beta_{s0}+\alpha_n u_s|)+n\sum_{s=1}^{p}p_{\lambda_n}(|\beta_{s0}|)\\
&= \ell_c(Y;\boldsymbol{\psi}_0+\alpha_n\mathbf{u})-\ell_c(Y;\boldsymbol{\psi}_0)-n\sum_{s=1}^{p_1}p_{\lambda_n}(|\beta_{s0}+\alpha_n u_s|)-n\sum_{s=p_1+1}^{p}p_{\lambda_n}(|0+\alpha_n u_s|))\\
&\quad +n\sum_{s=1}^{p_1}p_{\lambda_n}(|\beta_{s0}|)+n\sum_{s=p_1+1}^{p}p_{\lambda_n}(|0|)\\
&\le \ell_c(Y;\boldsymbol{\psi}_0+\alpha_n\mathbf{u})-\ell_c(Y;\boldsymbol{\psi}_0)-n\sum_{s=1}^{p_1}p_{\lambda_n}(|\beta_{s0}+\alpha_n u_s|)+n\sum_{s=1}^{p_1}p_{\lambda_n}(|\beta_{s0}|), \qquad (5.16)
\end{aligned}
$$

because of the fact that $n\sum_{s=p_1+1}^{p}p_{\lambda_n}(|0+\alpha_n u_s|))\ge 0$.

By the standard argument on the Taylor expansion and the conditions from (C1) and (C2), we obtain

$$
\begin{aligned}
\ell_c(Y;\boldsymbol{\psi}_0+\alpha_n\mathbf{u}) &= \ell_c(Y;\boldsymbol{\psi}_0)+\alpha_n\left\{\frac{\partial\ell(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}}\right\}^T\mathbf{u}+\frac{1}{2}\mathbf{u}^T\Big\{\frac{\partial^2\ell(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}\Big\}\mathbf{u}\alpha_n^2\\
&\quad +\sum_{s=1}^{p}O_p(|\alpha_n u_s|^3)\\
&= \ell_c(Y;\boldsymbol{\psi}_0)+\alpha_n\left\{\frac{\partial\ell(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}}\right\}^T\mathbf{u}+\frac{1}{2}\mathbf{u}^T\Big\{\frac{\partial^2\ell(Y;\boldsymbol{\psi}_0)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}\Big\}\mathbf{u}\alpha_n^2\{1+o_p(1)\} \qquad (5.17)
\end{aligned}
$$

and

$$
\begin{aligned}
& n\sum_{s=1}^{p_1}\Big\{p_{\lambda_n}(|\beta_{s0}+\alpha_n u_s|)\Big\}\\
= & n\sum_{s=1}^{p_1}p_{\lambda_n}(|\beta_{s0}|)+n\sum_{s=1}^{p_1}\alpha_n p'_{\lambda_n}(|\beta_{s0}|)\mathrm{sgn}(\beta_{s0})u_s+n\sum_{s=1}^{p_1}\alpha_n^2 p''_{\lambda_n}(|\beta_{s0}|)u_s^2\{1+o(1)\}. \qquad (5.18)
\end{aligned}
$$

Substituting (5.17) and (5.18) into (6.1), we obtain

$$\begin{aligned} K_n(\mathbf{u}) \quad &\leq \quad \alpha_n \Big\{ \frac{\partial \ell(Y;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} \Big\}^T \mathbf{u} + \frac{1}{2} \mathbf{u}^T \Big\{ \frac{\partial^2 \ell(Y;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} \Big\} \mathbf{u} \alpha_n^2 \{1 + o_p(1)\} \\ & \quad\quad -\sum_{s=1}^{p_1} n\{\alpha_n p'_{\lambda_n}(|\boldsymbol{\beta}_{0s}|)\text{sgn}(\boldsymbol{\beta}_{0s}) u_s + \alpha_n^2 p''_{\lambda_n}(|\boldsymbol{\beta}_{0s}|) u_s^2 \{1 + o(1)\}\} \\ &\overset{denote}{=\!=\!=} \quad \mathcal{A} + \mathcal{B} - \mathcal{C}. \end{aligned} \tag{5.19}$$

Now we individually examine $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. By Hölder's inequality, the $\mathcal{A}$ term on the right-hand side of (5.19) is

$$\begin{aligned} \alpha_n \Big\{ \frac{\partial \ell(Y;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} \Big\}^T \mathbf{u} \quad &= \quad n^{1/2} \alpha_n n^{-1/2} \Big\{ \frac{\partial \ell(Y;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} \Big\}^T \mathbf{u} \\ &\leq \quad n^{1/2} \alpha_n \left| n^{-1/2} \Big\{ \frac{\partial \ell(Y;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} \Big\}^T \mathbf{u} \right| \\ &\leq \quad n^{1/2} \alpha_n \left\| n^{-1/2} \frac{\partial \ell(Y;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} \right\| \cdot \left\| \mathbf{u} \right\|. \end{aligned} \tag{5.20}$$

By (C1), (C2) and (C3), we obtain that, $n^{-1/2} \frac{\partial \ell(Y;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}} = O_p(1)$, $\mathcal{A}$ can be bounded by $n^{1/2}\alpha_n \|\mathbf{u}\|$.

For the $\mathcal{B}$ term, since $\frac{1}{n}\Big\{ \frac{\partial^2 \ell(Y;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} \Big\} = O_p(1)$ by (C1) and (C2), we obtain that $\mathbf{u}^T \Big\{ \frac{\partial^2 \ell(Y;\boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} \Big\} \mathbf{u} \alpha_n^2$ is bounded by $n\alpha_n^2 \|\mathbf{u}\|^2$.

For the $\mathcal{C}$ term, we obtain that, using Hölder's inequality,

$$\sum_{s=1}^{p_1} n\alpha_n p'_{\lambda_n}(|\beta_{s0}|)\text{sgn}(\beta_{s0}) u_s \leq n\alpha_n a_n \Big| \sum_{s=1}^{p_1} u_s \Big| \leq n\alpha_n a_n \|\mathbf{u}\| \cdot \|\mathbf{1}\| = \sqrt{p_1} n \alpha_n a_n \|\mathbf{u}\|.$$

Furthermore, by the definition of b_n, we obtain

$$\sum_{s=1}^{p_1} n\alpha_n^2 p''_{\lambda_n}(|\boldsymbol{\beta}_{0s}|) u_s^2 \{1 + o(1)\} \leq n\alpha_n^2 b_n \|\mathbf{u}\|^2 \{1 + o(1)\}.$$

Note that $n\alpha_n a_n = O_p(n\alpha_n^2)$, and $b_n = o_p(1)$ by (C6.3). Therefore, term $\mathcal{C}$ is bounded by $n\alpha_n a_n \|\mathbf{u}\|$.

Since $a_n = O_p(n^{-1/2})$ from (C6.2), all $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are of the order $O_p(n\alpha_n^2)$. If $\|\mathbf{u}\| = C_\epsilon$ is sufficiently large, then $\mathcal{B}$ dominates $\mathcal{A}$ and $\mathcal{C}$. Moreover, by (C4)-(C5), $D(\boldsymbol{\psi}_0)$ is positive definite, then we have

$$P\Big\{\sup_{\|\mathbf{u}\|=C_\epsilon} K_n(\mathbf{u}) < 0\Big\} = P\left\{\sup_{\|\mathbf{u}\|=C_\epsilon} \ell_{pen}(Y;\boldsymbol{\psi}_0 + \alpha_n\mathbf{u}) < \ell_{pen}(Y;\boldsymbol{\psi}_0)\right\} \geq 1-\epsilon,$$

which indicates at least $1-\epsilon$ that there exists a local maximum in $\{\boldsymbol{\psi}_0 + \alpha_n\mathbf{u}\}$. Hence, there exists a local maximizer such that $\|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\| = O_p(\alpha_n)$.

Sparsity

Lemma 1: With probability tending to 1, for any given $\boldsymbol{\beta}_I$ and ξ satisfying

$$\|\boldsymbol{\beta}_I - \boldsymbol{\beta}_{0I}\| = O_p(n^{-1/2}), \qquad \text{and } \|\xi - \xi_0\| = O_p(n^{-1/2}),$$

we have

$$\ell_{pen}(Y;\boldsymbol{\beta}_I, \mathbf{0}, \xi) = \max_{\|\boldsymbol{\beta}_{II}\| \leq Cn^{-1/2}} \ell_{pen}(Y;\boldsymbol{\beta}_I, \boldsymbol{\beta}_{II}, \xi) \qquad \text{for any constant } C.$$

Proof: By Theorem 1, it suffices to show that with probability tending to 1 as $n \to \infty$, for any given $\boldsymbol{\beta}_I$ satisfying $\|\boldsymbol{\beta}_I - \boldsymbol{\beta}_{0I}\| = O_p(n^{-1/2})$, ξ satisfying $\|\xi - \xi_0\| = O_p(n^{-1/2})$, and for $\epsilon_n = Cn^{-1/2}$, and $s = p_1+1, \ldots, p$, we have

$$\frac{\partial \ell_{pen}(Y;\boldsymbol{\psi})}{\partial \beta_s} < 0 \quad \text{for } 0 < \beta_s < \epsilon_n,$$

and

$$\frac{\partial \ell_{pen}(Y;\boldsymbol{\psi})}{\partial \beta_s} > 0 \quad \text{for } -\epsilon_n < \beta_s < 0.$$

With Taylor Series expansion, we obtain

$$\begin{aligned}
\frac{\partial \ell_{pen}(Y;\boldsymbol{\psi})}{\partial \beta_s} &= \frac{\partial \ell(Y;\boldsymbol{\psi})}{\partial \beta_s} - np'_{\lambda_n}(|\beta_s|)\text{sgn}(\beta_s) \\
&= \frac{\partial \ell(Y;\boldsymbol{\psi}_0)}{\partial \beta_s} + \left\{\frac{\partial^2 \ell(Y;\boldsymbol{\psi}_0)}{\partial \beta_s \partial \boldsymbol{\psi}}\right\}^T (\boldsymbol{\psi} - \boldsymbol{\psi}_0) \\
&\quad + (\boldsymbol{\psi} - \boldsymbol{\psi}_0)^T \left\{\frac{\partial^3 \ell_{pen}(Y;\boldsymbol{\psi}^*)}{\partial \beta_s \partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T}\right\} (\boldsymbol{\psi} - \boldsymbol{\psi}_0) - np'_{\lambda_n}(|\beta_s|)\text{sgn}(\beta_s) \\
&\overset{denote}{=\!=} \mathcal{A} + \mathcal{B} + \mathcal{C} - np'_{\lambda_n}(|\beta_s|)\text{sgn}(\beta_s)
\end{aligned}$$

where $\boldsymbol{\psi}^*$ lies "between" $\boldsymbol{\psi}$ and $\boldsymbol{\psi}_0$. By the assumption that $\|\boldsymbol{\psi} - \boldsymbol{\psi}_0\| = O_p(n^{-1/2})$, then it follows that

$$\mathcal{A} = O_p(n^{1/2}), \qquad \mathcal{B} = O_p(n^{1/2}), \qquad \mathcal{C} = O_p(1),$$

and thus

$$(n\lambda_n)^{-1}\mathcal{A} = O_p(n^{-1/2}/\lambda_n), \qquad (n\lambda_n)^{-1}\mathcal{B} = O_p(n^{-1/2}/\lambda_n), \quad \text{and } (n\lambda_n)^{-1}\mathcal{C} = O_p(n^{-1}/\lambda_n).$$

As a result, we obtain

$$\begin{aligned} \frac{\partial \ell_{pen}(Y;\boldsymbol{\psi})}{\partial \beta_s} &= n\lambda_n\{(n\lambda_n)^{-1}(\mathcal{A}+\mathcal{B}+\mathcal{C}) - \lambda_n^{-1}p'_{\lambda_n}(|\beta_s|)\text{sgn}(\beta_s)\} \\ &= n\lambda_n\{O_p(n^{-1/2}/\lambda_n) - \lambda_n^{-1}p'_{\lambda_n}(|\beta_s|)\text{sgn}(\beta_s)\}. \qquad (5.21) \end{aligned}$$

By the regularity condition (C6), $\text{liminf}_{n\to\infty}\text{liminf}_{\epsilon\to 0^+}p'_{\lambda_n}(\epsilon)/\lambda_n > 0$ and $\lim_{n\to\infty}\sqrt{n}\lambda_n = \infty$, the sign of the derivative in (5.21) is determined by β_s. Thus we have

$$\frac{\partial \ell_{pen}(Y;\boldsymbol{\psi})}{\partial \beta_s} < 0 \quad \text{for } 0 < \beta_s < \epsilon_n,$$

and

$$\frac{\partial \ell_{pen}(Y;\boldsymbol{\psi})}{\partial \beta_s} > 0 \quad \text{for } -\epsilon_n < \beta_s < 0.$$

This completes the proof.

Asymptotic Distribution

Now we come to the proof of oracle property. Denote

$$\Sigma = \text{diag}\{p''_{\lambda_n}(|\beta_{01}|), \ldots, p''_{\lambda_n}(|\beta_{0p}|), \mathbf{0}_\xi\},$$

and

$$\mathbf{b} = \Big(\big(p'_{\lambda_n}(|\beta_{01}|)\text{sgn}(\beta_{01}), \ldots, p'_{\lambda_n}(|\beta_{0p}|)\text{sgn}(\beta_{0p})\big)^T, \mathbf{0}_\xi^T\Big)^T.$$

Theorem 2: With probability tending to 1, the root-n consistent local maximizers $\hat{\boldsymbol{\psi}}$ in Theorem 1 must satisfy:

(a). Sparsity: $\hat{\boldsymbol{\beta}}_{II} = \mathbf{0}$.

(b). Asymptotic normality: $\sqrt{n}(\tilde{D}(\tilde{\boldsymbol{\psi}}_0)+\tilde{\Sigma})\{\tilde{\hat{\boldsymbol{\psi}}}-\tilde{\boldsymbol{\psi}}_0+(\tilde{D}(\tilde{\boldsymbol{\psi}}_0)+\tilde{\Sigma})^{-1}\tilde{\mathbf{b}}\} \rightarrow_D N(\mathbf{0}, \tilde{M}(\tilde{\boldsymbol{\psi}}_0))$.

Proof: Part (a) follows from Lemma 1. Now we show part (b). By Theorem 1, there exists a $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}_I, \mathbf{0}, \hat{\xi})$ that is a root-n consistent local maximizer of $\ell_{pen}(Y; \boldsymbol{\psi})$, and that satisfies

$$\frac{\partial \ell_{pen}(Y; \tilde{\boldsymbol{\psi}})}{\partial \tilde{\boldsymbol{\psi}}}\Big|_{\tilde{\boldsymbol{\psi}}=\tilde{\hat{\boldsymbol{\psi}}}} = \mathbf{0}.$$

By Taylor Series expansion, we obtain

$$\frac{\partial \ell(Y; \tilde{\boldsymbol{\psi}}_0)}{\partial \tilde{\boldsymbol{\psi}}} + \left\{ \frac{\partial^2 \ell(Y; \tilde{\boldsymbol{\psi}}_0)}{\partial \tilde{\boldsymbol{\psi}} \partial \tilde{\boldsymbol{\psi}}^T} + o_p(1) \right\} (\tilde{\hat{\boldsymbol{\psi}}} - \tilde{\boldsymbol{\psi}}_0) - n\Big\{ \tilde{\mathbf{b}} + \{\tilde{\Sigma} + o_p(1)\}(\tilde{\hat{\boldsymbol{\psi}}} - \tilde{\boldsymbol{\psi}}_0) \Big\} = \mathbf{0}.$$

Thus, we obtain

$$\frac{1}{\sqrt{n}} \left\{ \frac{\partial^2 \ell(Y; \tilde{\boldsymbol{\psi}}_0)}{\partial \tilde{\boldsymbol{\psi}} \partial \tilde{\boldsymbol{\psi}}^T} + o_p(1) \right\} (\tilde{\hat{\boldsymbol{\psi}}} - \tilde{\boldsymbol{\psi}}_0) - \sqrt{n}\Big[\tilde{\mathbf{b}} + \{\tilde{\Sigma} + o_p(1)\}(\tilde{\hat{\boldsymbol{\psi}}} - \tilde{\boldsymbol{\psi}}_0) \Big] = -\frac{1}{\sqrt{n}} \frac{\partial \ell(Y; \tilde{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}}.$$

Applying Slusky's theorem and the Central Limiting Theorem, we obtain

$$\sqrt{n}\{\tilde{D}(\tilde{\boldsymbol{\psi}}_0)(\tilde{\hat{\boldsymbol{\psi}}} - \tilde{\boldsymbol{\psi}}_0) + \tilde{\mathbf{b}} + \tilde{\Sigma}(\tilde{\hat{\boldsymbol{\psi}}} - \tilde{\boldsymbol{\psi}}_0)\} \rightarrow_D N(\mathbf{0}, \tilde{M}(\tilde{\boldsymbol{\psi}}_0)),$$

i.e.

$$\sqrt{n}\Big[\{\tilde{D}(\tilde{\boldsymbol{\psi}}_0) + \tilde{\Sigma}\}(\tilde{\hat{\boldsymbol{\psi}}} - \tilde{\boldsymbol{\psi}}_0) + \tilde{\mathbf{b}} \Big] \rightarrow_D N(\mathbf{0}, \tilde{M}(\tilde{\boldsymbol{\psi}}_0)).$$

4. Simulation Results

Variable Selection

Table 5.3: Simulation results for the incomplete data via the linear mixed model: model selection

	Method	R.MME(%)[‡]	M.MME	Avg. No. of 0 Coefficients	
				Correct[*]	Incorrect[**]
Example 1	$\text{NML}_\lambda^\dagger$	81.310	0.050	4.889	0
$n = 200, J = 1, K = 10$	CL_λ	54.329	0.046	4.478	0
Example 2	NML_λ	93.009	0.042	4.993	0
$n = 800, J = 1, K = 10$	CL_λ	39.108	0.008	4.923	0
Example 3	NML_λ	84.830	0.047	4.908	0
$n = 200, J_i = 3, K = 5$	CL_λ	51.222	0.031	4.540	0
Example 4	NML_λ	96.086	0.043	4.990	0
$n = 800, J_i = 3, K = 5$	CL_λ	39.133	0.005	4.920	0

[†] NML and CL represent naive maximum likelihood to complete data and the proposed composite likelihood, respectively. λ denotes the tuning parameter selection by only λ with fixing $a = 3.7$.

[‡] C.MME represents the median of ratios of MME of a selected model to NML and CL, respectively. A.MME denotes the median of ratios of MME of a selected model to that of the un-penalized full model with CL estimate.

[*] "Correct" presents the average restricted to the true zero coefficients. 0 represents that no true zero coefficient is shrink, while 5 implies that all true zero coefficients are restricted into zero.

[**] "Incorrect" depicts the average of significant coefficients that are erroneously set to 0. 0 represents that no significant coefficient is shrink, while 3 implies that all significant coefficients are erroneously set to zero.

Table 5.4: Simulation results for the incomplete data via the Poisson mixed model: model selection

				Avg. No. of 0 Coefficients	
	Method	R.MME(%)	M.MME	Correct	Incorrect
Example 1	$\text{NML}_\lambda^\dagger$	98.842	91.409	4.984	0
$n = 120, J = 1, K = 10$	CL_λ	78.626	27.118	4.690	0
Example 2	NML_λ	97.637	77.657	5	0
$n = 500, J = 1, K = 10$	CL_λ	74.057	5.787	4.999	0
Example 3	NML_λ	98.701	105.598	4.990	0
$n = 120, J_i = 3, K = 5$	CL_λ	84.812	32.023	4.724	0
Example 4	NML_λ	98.081	101.778	5	0
$n = 300, J_i = 3, K = 5$	CL_λ	73.716	9.272	4.964	0

† NML and CL represent naive maximum likelihood to complete data and the proposed composite likelihood, respectively. λ denotes the tuning parameter selection by only λ with fixing $a = 3.7$.

Parameter Estimation

Table 5.5: Simulation results for the incomplete data via linear mixed model: model estimation on regression coefficients

		β_1				β_2				β_5			
	Method	Bias(%)*	ESE‡	ASE$^\sharp$	CP(%)$^\flat$	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)
Example 1	NML$_\lambda^\dagger$	-4.384	0.071	0.067	50.2	-4.605	0.070	0.066	80.0	-4.321	0.063	0.058	67.6
$n = 200, J = 1, K = 10$	CL$_\lambda$	-0.814	0.128	0.121	92.9	-2.324	0.132	0.120	90.6	-2.687	0.127	0.108	88.2
Example 2	NML$_\lambda$	-4.357	0.035	0.034	3.8	-4.444	0.033	0.033	46.5	-4.292	0.029	0.029	14.0
$n = 800, J = 1, K = 10$	CL$_\lambda$	0.026	0.060	0.062	95.5	-0.513	0.064	0.062	93.5	-0.350	0.056	0.054	94.3
Example 3	NML$_\lambda$	-4.479	0.056	0.056	34.4	-4.286	0.058	0.055	78.0	-4.348	0.052	0.048	54.2
$n = 200, J_i = 3, K = 5$	CL$_\lambda$	-0.740	0.102	0.102	94.4	-1.374	0.114	0.102	92.0	-2.102	0.107	0.090	88.2
Example 4	NML$_\lambda$	-4.518	0.028	0.028	0	-4.347	0.026	0.027	33.3	-4.463	0.024	0.024	4.6
$n = 800, J_i = 3, K = 5$	CL$_\lambda$	0.017	0.050	0.050	95.2	-0.090	0.052	0.052	95.2	0.016	0.047	0.045	93.8

† NML and CL represent naive maximum likelihood to complete data and the proposed composite likelihood, respectively. λ denotes the tuning parameter selection by only λ with fixing $a = 3.7$.

$*$ Relative bias defined by $(\hat{\beta} - \beta_{true})/\beta_{true} \times 100$.

$**$ ASE is the average standard error for 1000 simulations, which is defined as $1000^{-1}\sum_{i=1}^{1000}\sqrt{\widehat{Var}(\hat{\beta}^i)}$, where $\sqrt{\widehat{Var}(\hat{\beta}^i)}$ is the standard error estimates in ith simulation result.

$***$ ESE is the empirical standard error for 1000 simulations, which is defined as $(1000-1)^{-1}\sum_{i=1}^{1000}(\hat{\beta}^i - \bar{\hat{\beta}})^2$, where $\hat{\beta}^i$ is the ith simulation result, and $\bar{\hat{\beta}} = 1000^{-1}\sum_{i=1}^{1000}\hat{\beta}^i$.

Table 5.6: Simulation results for the incomplete data via Poisson mixed model: model estimation on regression coefficients

		β_1				β_2				β_5			
	Method	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)	Bias(%)	ESE	ASE	CP(%)
Example 1	NML$_\lambda^\dagger$	-3.045	0.021	0.018	45.9	-3.185	0.017	0.017	78.4	-3.174	0.016	0.015	61.3
$n = 120, J = 1, K = 10$	CL$_\lambda$	-0.352	0.035	0.032	93.0	-0.587	0.034	0.029	89.4	-0.859	0.032	0.027	89.3
Example 2	NML$_\lambda$	-2.813	0.009	0.008	2.7	-2.877	0.009	0.008	41.2	-2.837	0.008	0.007	11.1
$n = 500, J = 1, K = 10$	CL$_\lambda$	-0.073	0.017	0.016	91.7	-0.022	0.015	0.015	93.7	-0.129	0.014	0.013	94.2
Example 3	NML$_\lambda$	-3.345	0.019	0.017	34.0	-3.382	0.017	0.016	75.8	-3.468	0.015	0.015	50.2
$n = 120, J_i = 3, K = 5$	CL$_\lambda$	-0.371	0.035	0.034	94.4	-0.908	0.034	0.031	93.0	-1.124	0.033	0.028	90.2
Example 4	NML$_\lambda$	-3.201	0.012	0.011	6.67	-3.355	0.011	0.010	49.4	-3.281	0.010	0.009	17.7
$n = 300, J_i = 3, K = 5$	CL$_\lambda$	-0.121	0.021	0.021	95.8	-0.310	0.021	0.020	93.6	-0.192	0.020	0.018	91.6

† NML and CL represent naive maximum likelihood to complete data and the proposed composite likelihood, respectively. λ denotes the tuning parameter selection by only λ with fixing $a = 3.7$.

Chapter 6

Discussion and Future Research

6.1 Composite Likelihood Analysis for Incomplete Longitudinal Data

In Chapter 2 and Chapter 3, we develop two estimation approaches using the pairwise likelihood to handle longitudinal data with missing values in both the response and the covariate variables. The analysis of the NPHS data using the proposed methods demonstrates their utility of real applications. Simulation studies show reliable and satisfactory performance of our methods. The PL method is appealing for its higher efficiency, while the TS approach is easier to implement. Our empirical studies show, as expected, that relative to the maximum likelihood method, both the PL and the TS approaches may incur efficiency loss, especially when repeated measurements are strongly correlated. However, this limitation is compensated by the robustness of our methods as against the full likelihood method. The proposed methods would still lead to consistent estimates even when third order association structures for the response process are mis-modeled, whereas the likelihood method would break down if the full distribution of data is misspecified.

The proposed methods can be extended to accommodate circumstances with multiple covariates being subject to missingness. In particular, let $X_{ij} = (X_{ij1}, \ldots, X_{ijp})'$ with $p \geq 2$, and $H_{ijr}^x = \{X_{ij1}, \ldots, X_{ij,r-1}\}$ with $r = 2, \ldots, p$. Noticing the factorization

$$\begin{aligned} P(X_{ij} = \mathbf{x}_{ij}, X_{ik} = \mathbf{x}_{ik}|Z_i) &= P(X_{ij1} = x_{ij1}, X_{ik1} = x_{ik1}|Z_i) \\ &\quad \cdot \prod_{r=2}^{p} P(X_{ijr} = x_{ijr}, X_{ikr} = x_{ikr}|H_{ijr}^x, H_{ikr}^x, Z_i) \end{aligned}$$

where $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$, we only need to model a sequence of conditional bivariate distributions

$$\big\{P(X_{ijr} = x_{ijr}, X_{ikr} = x_{ikr}|H_{ijr}^x, H_{ikr}^x, Z_i), \qquad r = 2, \ldots, p\big\} \tag{6.1}$$

in order to determine the distribution of $P(X_{ij} = \mathbf{x}_{ij}, X_{ik} = \mathbf{x}_{ik}|Z_i)$ for $j < k$. Analogous to the formulation in Section 3.2.2, we can postulate the bivariate distributions for (6.1). A similar strategy applies to modeling the missing data processes.

As opposed to the full likelihood, $\binom{m}{2}$ terms are involved in the pairwise likelihood formulation. Although the number of these terms grows quadratically in m, the computation of pairwise likelihoods are often much cheaper than that for the full likelihood. In general, the computational cost to produce the full likelihood is heavily dependent on the dimension m. It can grow exponentially in m, and this may occur, for instance, when calculation of the distribution of the marginal subset requires integration over a set of m unobserved random variables. In this case, the pairwise likelihood method has a clear computational gain over the full likelihood approach. For more discussion on computational expense associated with a composite likelihood formulation, see Lindsay et al. (2011) and Bellio and Varin (2005).

Finally, we comment that our discussion is focused on bivariate normal or probit models for the responses. The proposed methods can be modified to handle other types of data. For example, if the data is continuous and non-normal, bivariate skew normal distributions (Azzalini and Valle, 1996) may be employed. With longitudinal ordinal data, one may employ the model discussed by Qu et al. (1995) using the bivariate probit model and adopt the development here for data analysis.

6.2 Variable Selection via Composite Likelihood for Analysis of Longitudinal Data Arising in Clusters

In chapter 4 we develop composite likelihood framework to handle longitudinal data arising in clusters with variable selection. The asymptotic properties of our methods are proved and simulation studies show their satisfactory performance in both the model selection and the estimation. Comparing with maximum likelihood approach, our methods are less efficient, but they outperform the full likelihood method in robustness and convenience in the model specification.

Moreover, we also study the variable selection for both fixed and random effects. Although Cholesky decomposition strategy is widely used in selecting random effects (Bondell et al., 2010; Ibrahim et al., 2010), our study shows that they may not be proper for longitudinal data arising in clusters. In addition, the Cholesky decomposition may lead to inappropriate results for the composite likelihood. Thus, to circumvent this problem, we propose a standard error-correlation coefficient decomposition strategy. Furthermore, a modified ECME algorithm (Liu and Pierce, 1994) is employed for the model selection and the estimation.

Furthermore, this chapter shows that the model is misspecified, the parameter estimation and the variable selection results could be biased or incorrect. Based on the framework proposed by Yi and Reid (2010), we prove that, under certain regularity conditions, the misspecified model may asymptotically lead to biased results. The simulation studies in this chapter demonstrate that if we misspecify the random effect distributions in the statistical inference, biased selection and estimation outcomes may occur.

6.3 Variable Selection via Composite Likelihood for Incomplete Longitudinal Data arising in Clusters

In chapter 5 we develop estimation approach using the missingness modified composite likelihood to handle incomplete longitudinal data arising in clusters with variable selection. Simulation studies show reliable and satisfactory performance of our methods. It provides valid variable selection and parameter estimation results, while naive estimation approach may result in biased estimation outcomes.

Moreover, our method outperforms other approaches because it does not require the specification and estimation of missing data process, which is often employed in the inference under missing not at random (MNAR) scenario. This simplification results in the augmentation for the estimation procedure. Firstly, the estimators can avoid the bias from the misspecification of the missing data processes described in Chapter 2 and Chapter 3. Secondly, the estimation procedure does not include a large set of nuisance parameters to postulate the missing data process. Thirdly, our missingness modified composite likelihood functions does not involve integrals which can be intractable for the computation.

However, our missingness modified composite likelihood is not assumption free for all missing mechanisms. To be specific, it assumes

$$P(R_{ijk} = 1|Y_{ijk}, Y_{ij'k'}, X_i, Z_i) = P(R_{ijk} = 1|Y_{ijk}, X_i, Z_i), \tag{6.2}$$

for some ijk in the missingness-prone set and $ij'k'$ in the missingness-free set. This assumption can not be directly tested from the dataset. To evaluate the validity of the missing data assumption, Qu et al. (2011) propose an assessment approach for weighted generalized estimating equations. However, this method can not be directly used in the composite likelihood framework with model selection, and further study in this area is needed.

Another typical drawback of our method is that it only uses the observations with complete response and covariates, while all other incomplete records are not included.

This leads to a significant efficiency loss, especially for the data with high missing rate. Therefore, studies for improving efficiency under missing data scenarios are required.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *International Symposium on Information Theory*, 2:267–281.

Albert, P. S. and Follmann, D. A. (2003). A random effects transition model for longitudinal binary data with informative missingness. *Statistica Neerlandica*, 57:100–111.

Almasi, G. S. and Gottlieb, A. (1989). *Highly Parallel Computing*. Benjamin-Cummings Publishing Co., Inc: Redwood City.

Arellano-Valle, R. B., Bolfarine, H., and Lachos, V. H. (2005). Skew-normal linear mixed models. *Journal of Data Science*, 3:415–438.

Arnold, B. C. and Strauss, D. (1991). Pseudolikelihood estimation: Some examples. *Sankhy: The Indian Journal of Statistics, Series B*, 53:233–243.

Ashford, J. R. and Sowden, R. R. (1970). Multi-variate probit analysis. *Biometrics*, 26:535–546.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12:171–178.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61:579–602.

Azzalini, A. and Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika*, 83:715–726.

Bellio, R. and Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, 5:217–227.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24:179–195.

Besag, J. (1977). Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, 64:616–618.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 1:1–9.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society, Series B*, 61:265–285.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B (Methodological)*, 22:302–306.

Cameron, R., Brown, K. S., Best, J. A., Pelkman, C. L., Madill, C. L., Manske, S. R., and Payne, M. E. (1999). Effectiveness of a social influences smoking prevention program as a function of provider type, training method, and social risk. *American Journal of Public Health*, 89:1827–1831.

Carpenter, J. R. and Kenward, M. G. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A*, 169:571–584.

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

Chatelain, F., Lambert-Lacroix, S., and Tourneret, J.-Y. (2008). Pairwise likelihood estimation for multivariate mixed poisson models generated by gamma intensities. *Statistics and Computing*, 19:283–301.

Chen, B., Yi, G. Y., and Cook, R. J. (2010). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association*, 105:336–353.

Chen, Q., Ibrahim, J. G., Chen, M., and Senchaudhuri, P. (2008). Theory and inference for regression models with missing responses and covariates. *Journal of Multivariate Analysis*, 99:1302–1331.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59:762–769.

Choi, N. H., Li, W., and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105:354–364.

Cook, R. J., Zeng, L., and Yi, G. Y. (2004). Marginal analysis of incomplete longitudinal binary data: A cautionary note on locf imputation. *Biometrics*, 60:820–828.

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

Curriero, F. C. and Lele, S. (1999). A composite likelihood approach to semivariogram estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, 4:9–28.

Daniels, M. J. and Hogan, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman and Hall: London.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38.

Diehr, P. and Patrick, D. L. (2003). Trajectories of health for older adults over time: Accounting fully for death. *Annals of Internal Medicine*, 139:416–420.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, 43:49–93.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data, 2nd edition.* Oxford University Press: Oxford.

Dufouil, C., Brayne, C., and Clayton, D. (2004). Analysis of longitudinal studies with death and drop-out: a case study. *Statistics in Medicine*, 23:2215–2226.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32:407–451.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30:74–99.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, 99:710–723.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *Proceedings of the Madrid International Congress of Mathematicians.*

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

Feeny, D., Furlong, W., Torrance, G. W., Goldsmith, C. H., Zhu, Z., DePauw, S., Denton, M., and Boyle, M. (2002). Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Medical Care*, 40:113–128.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62:424–431.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied Longitudinal Analysis*. Wiley-Interscience: New York.

Fitzmaurice, G. M., Laird, N. M., and Zahner, G. E. P. (1996). Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, 91:99–108.

Fitzmaurice, G. M., Lipsitz, S. R., Molenberghs, G., and Ibrahim, J. G. (2005). A protective estimator for longitudinal binary data subject to non-ignorable non-monotone missingness. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 168:723–735.

Fitzmaurice, G. M., Molenberghs, G., and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:691–704.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135.

Fu, W. J. (2003). Penalized estimating equations. *Biometrics*, 59:126–132.

Gao, X. and Song, P. X. (2010). Composite likelihood bayesian information criteria for model selection in high dimensional data. *Journal of American Statistical Association*, 105:1531–1540.

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

Geys, H., Molenberghs, G., and Lipsitz, S. (1998). A note on the comparison of pseudo-likelihood and generalized estimating equations for marginally specified odds ratio models with exchangeable association structure. *Journal of Statistical Computation and Simulation*, 62:45–71.

Geys, H., Molenberghs, G., and Ryan, L. M. (1997). Pseudo-likelihood inference for clustered binary data. *Communications in Statistics - Theory and Methods*, 26:2743–2767.

Glonek, G. F. V. (1999). On identifiability in models for incomplete binary data. *Statistics and Probability Letters*, 41:191–197.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association*, 88:984–993.

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

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, 63:277–284.

Godambe, V. P. (1991). *Estimating Functions*. Oxford University Press: New York.

Godambe, V. P. and Thompson, M. E. (1984). Robust estimation through estimating equations. *Biometrika*, 71:115–125.

Goldstein, H. (2002). *Multilevel Statistical Models*. Wiley: New York.

Hanfelt, J. J. (2004). Composite conditional likelihood for sparse clustered data. *Journal of the Royal Statistical Society, Series B*, 66:259–273.

Harel, O., Hofer, S. M., Hoffman, L., and Pedersen, N. L. (2007). Population inference with mortality and attrition in longitudinal studies on aging: A two-stage multiple imputation method. *Experimental Aging Research*, 33:187–203.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–338.

Hawkins, D. L. (1989). Using u statistics to derive the asymptotic distribution of fisher's z statistic. *The American Statistician*, 43:235–237.

He, W. and Yi, G. Y. (2011). A pairwise likelihood method for correlated binary data with/without missing observations under generalized partially linear single-index models. *Statistica Sinica*, 21:207–229.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88:973–985.

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

Henmi, M. and Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*, 91:929–941.

Ibrahim, J. G., Chen, M., and Lipsitz, S. R. (1999). Missing covariates in generalized linear models when the missing data mechanism is non- ignorable. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61:173–190.

Ibrahim, J. G., Chen, M., and Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, 88:551–564.

Ibrahim, J. G., Chen, M., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100:332–346.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. *The Canadian Journal of Statistics*, 30:55–78.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2010). Fixed and random effects selection in mixed effects models. *Biometrics*, 1:1–9.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall: London.

Joe, H. and Lee, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis*, 100:670–685.

Johnson, B. A., Lin, D. Y., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, 103:672–680.

Kim, J.-O. and Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods Research*, 6:215–240.

Kinney, S. K. and Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, 63:690–698.

Kong, E. and Xia, Y. (2007). Variable selection for the single-index model. *Biometrika*, 94:217–229.

Konishi, S., Ando, T., and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, 91:27–43.

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

Kuk, A. Y. C. (2007). A hybrid pairwise likelihood method. *Biometrika*, 94:939–952.

Kurland, B. F. and Heagerty, P. J. (2005). Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics*, 6:241–258.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963–974.

Landrum, M. B. and Becker, M. P. (2001). A multiple imputation strategy for incomplete longitudinal data. *Statistics in Medicine*, 20:15–30.

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

Lin, T. I. and Lee, J. C. (2008). Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data. *Statistics in Medicine*, 27:1490–1507.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80:221–239.

Lindsay, B. G., Yi, G. Y., and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21:71–105.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90:1112–1121.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data, Second Edition.* Wiley-Interscience: New York.

Liu, H., Weiss, R. E., Jennrich, R. I., and Wenger, N. S. (1999). Press model selection in repeated measures data. *Computational Statistics and Data Analysis*, 30:169–184.

Liu, Q. and Pierce, D. A. (1994). A note on gauss-hermite quadrature. *Biometrika*, 81:624–629.

Longford, N. T. (1994). Logistic regression with random coefficients. *Computational Statistics and Data Analysis*, 17:1–15.

Lv, J. and Liu, J. S. (2010). Model selection principles in misspecified models. Technical report, University of Southern California and Harvard University.

Ma, Y. and Li, R. (2010). Variable selection in measurement error models. *Bernoulli*, 16:274–300.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall: London.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer: New York.

Naik, P. A. and Tsai, C.-L. (2001). Single-index model selections. *Biometrika*, 88:821–832.

Neuhaus, J. M., Hauck, W. W., and Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*, 79:755–762.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1994). Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *The Canadian Journal of Statistics*, 22:139–148.

Neuhaus, J. M. and McCulloch, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 5:859–872.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*. North-Holland.

Ni, X., Zhang, D., and Zhang, H. H. (2010). Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics*, 66:79–88.

Ochi, Y. and Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika*, 71:531–543.

Orpana, H. M., Ross, N., Feeny, D., McFarland, B., Bernier, J., and Kaplan, M. (2009). The natural history of health-related quality of life: A 10-year cohort study. Technical report, Statistics Canada.

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:659–677.

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

Parzen, M., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., and Troxel, A. B. (2006). Pseudo-likelihood methods for longitudinal binary data with non-ignorable missing responses and covariates. *Statistics in Medicine*, 25:2784–2796.

Parzen, M., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Troxel, A. B., and Molenberghs, G. (2007). Pseudo-likelihood methods for the analysis of longitudinal binary data subject to nonignorable non-monotone missingness. *Journal of Data Science*, 5:1–21.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

Pauler, D. K. (1998). The schwarz criterion and related methods for normal linear models. *Biometrika*, 85:13–27.

Payment, P., Richardson, L., Siemiatycki, J., Dewar, R., Edwards, M., and Franco, E. . (1991). A randomized trial to evaluate the risk of gas- trointestinal disease due to consumption of drinking water meeting cur- rent microbiological standards. *American Journal of Public Health*, 81:703–708.

Pepe, M. S. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation*, 23:939–951.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4:12–35.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed Effects Models in S and S-Plus*. Springer: New York.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometric*, 44:1033–1048.

Press, S. J. and Scott, A. J. (1976). Missing variables in bayesian regression. *Journal of the American Statistical Association*, 71:366–369.

Qu, A., Yi, G. Y., Song, P. X.-K., and Wang, P. (2011). Assessing the validity of weighted generalized estimating equations. *Biometrika*, 98:215–224.

Qu, Y., Piedmonte, M. R., and Medendorp, S. V. (1995). Latent variable models for clustered ordinal data. *Biometrics*, 51:268–275.

Renard, D., Molenberghs, G., and Geys, H. (2002). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, 44:649–667.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90:106–121.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93:1321–1339.

Roy, J. and Lin, X. (2002). Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: Changes in methadone treatment practices. *Journal of the American Statistical Association*, 97:40–52.

Roy, S. and Banerjee, T. (2009). Analysis of misclassified correlated binary data using a multivariate probit model when covariates are subject to measurement error. *Biometrical Journal*, 51:420–432.

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

Schafer, J. L. (1998). Some improved procedures for linear mixed models. Technical report, Department of Statistics, The Pennsylvania State University.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33:545–571.

Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, 11:437–457.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Shardell, M. and Miller, R. R. (2008). Weighted estimating equations for longitudinal studies with death and non-monotone missing time-dependent covariates and outcomes. *Statistics in Medicine*, 27:1008–1025.

Shen, X. and Ye, J. (2002). Adaptive model selection. *Journal of the American Statistical Association*, 97:210–221.

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97:1141–1153.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 40:961–971.

Stubbendick, A. L. and Ibrahim, J. G. (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, 59:1140–1150.

Stubbendick, A. L. and Ibrahim, J. G. (2006). Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica*, 16:1143–1167.

Sutradhar, B. C. and Rao, R. P. (2003). On quasi-likelihood inference in generalized linear mixed models with two components of dispersion. *Canadian Journal of Statistics*, 31:415–435.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics*, 3:245–265.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16:385–395.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society. Series B (Methodological)*, 73:273–282.

Tong, X., He, X., Sun, L., and Sun, J. (2009). Variable selection for panel count data via non-concave penalized estimating function. *Scandinavian Journal of Statistics*, 36:620–635.

Troxel, A. B., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Sinha, D., and Molenberghs, G. (2010). A weighted combination of pseudo-likelihood estimators for longitudinal binary data subject to non-ignorable non-monotone missingness. *Statistics in Medicine*, 29:1511–1521.

Troxel, A. B., Lipsitz, S. R., and Harrington, D. P. (1998). Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. *Biometrika*, 85:661–672.

Vaida, F. and Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika*, 92:351–370.

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

Varin, C., Host, G., and Skare, O. (2005). Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics and Data Analysis*, 49:1173 –1191.

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

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

Verbeke, G., Spiessens, B., and Lesaffre, E. (2001). Conditional linear mixed models. *The American Statistician*, 55:25–34.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94:553–568.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84:1065–1073.

Weiss, R. E., Wang, Y., and Ibrahim, J. G. (1997). Predictive model selection for repeated measures random effects models using bayes factors. *Biometrics*, 53:592–602.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44:175–188.

Yafune, A., Funatogawa, T., , and Ishiguro, M. (2005). Extended information criterion (eic) approach for linear mixed effects models under restricted maximum likelihood (reml) estimation. *Statistics in Medicine*, 24:3417–3429.

Yan, X. and Su, X. (2009). *Linear Regression Analysis: Theory and Computing*. World Scientific: New York.

Yi, G. Y. and Cook, R. J. (2002). Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association*, 97:1071–1080.

Yi, G. Y., Liu, W., and Wu, L. (2011a). Simultaneous inference and bias analysis for longitudinal data with covariate measurement error and missing responses. *Biometrics*, 67:67–75.

Yi, G. Y., Ma, Y., and Carroll, R. J. (2012). A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*, 99:151–165.

Yi, G. Y. and Reid, N. (2010). A note on misspecified estimating functions. *Statistica Sinica*, 20:1749–1769.

Yi, G. Y., Zeng, L., and Cook, R. J. (2011b). A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. *Canadian Journal of Statistics*, 39:34–51.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44:1049–1060.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38:894–942.

Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105:312–323.

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

Zhao, Y. and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *The Canadian Journal of Statistics*, 33:335–356.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:301–320.