©Copyright 2014
Rui Zhang

Marginalizable mixed effect models for clustered binary, categorical and survival data

Rui Zhang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

KC Gary Chan, Chair

Patrick J. Heagerty

Jon A. Wellner

Program Authorized to Offer Degree:
Department of Biostatistics, University of Washington

University of Washington

Abstract

Marginalizable mixed effect models for clustered binary, categorical and survival data

Rui Zhang

Chair of the Supervisory Committee:
Associate Professor KC Gary Chan
Department of Biostatistics

In this thesis, I propose new models for clustered data, design estimators of covariate effects, implement model inference algorithms, and show asymptotic properties of my estimators, including consistency and asymptotic normality. I also present Monte Carlo simulations and real dataset analyses for demonstration.

In this thesis three inherently related marginalizable mixed effects models for datasets with clustered binary, ordinal and right-censored, i.e. survival outcomes are proposed. These models can evaluate population-level covariate effects, and model a diverse variety of cluster correlation structures. The marginalizable property of the models is obtained via a pair of conjugate distributions.

Chapter I contains literature review of various models designed for clustered datasets, focusing on binary, ordinal and survival data. Chapter II discusses and compares different inference methods for models in Part I.

In Chapter III, I first give the motivation of the new marginalizable mixed effects model. Then I introduce the multivariate exponential random variables, which serve as random effects in the new models. I formally present the new model formulation for binary data, the inference procedure followed by a brief discussion, as well as relevant asymptotic theorems of my estimators.

Chapter IV and Chapter V discuss clustered ordinal data and survival data, and are organized similarly.

In Chapter VI I present some Monte Carlo simulation results along with real dataset applications: the Madras longitudinal schizophrenia study, the British Social Attitudes Panel Survey have

clustered binary outcomes; the Arthritis data and the Television, the School and Family Smoking Prevention and Cessation Project have clustered ordinal outcomes; the Rats and Lung datasets for have clustered survival data.

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

$|| \beta ||$: $\beta \in \mathbb{R}^p$, the Euclidean norm of β.

$1\{A\}$: an indicator function equals to one if A is true and zero otherwise.

AR(1): auto-correlation structure with degree one.

E$\cdot$: the operation of taking expectation of a random variable.

EM ALGORITHM: expectation maximization algorithm.

γ: unknown variance parameter of a Gamma distribution.

Γ: correlation matrix of multivariate normal random variables.

GLM: generalized linear models.

h: inverse link function.

$i.i.d.$: independently and identically distributed.

MLE: maximum likelihood estimation.

MC: Monte Carlo.

MSE: mean square error.

NPMCLE: non-parametric maximum composite likelihood estimation.

NPMLE: non-parametric maximum likelihood estimation.

OR: odds ratio.

$R(\rho)$: correlation matrix of frailties.

$(\boldsymbol{Y}, \boldsymbol{Z})$: a set of correlated observations from the same cluster/institute, where $\boldsymbol{Y}$ is a set of binary or ordered categorical outcomes and $\boldsymbol{Z}$ is corresponding covariates.

(Y, Z): one observation where Y is a binary or ordered categorical outcome and Z is corresponding covariates.

τ: study time.

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to University of Washington, where she has had the opportunity to work on interesting problems in biostatistics, to KC Gary Chan, her advisor, who has proposed this set of interesting problems; to Patrick J. Heagerty, who has shared datasets, to Jon A. Wellner, who has supervised her with theoretical proof; to Norman Breslow, who has introduced many interesting points and alternative ways to think about her study.

DEDICATION

to my family

Chapter 1

LITERATURE REVIEW: MODELS

1.1 Clustered Data and Model Overview

In this thesis, I am interested in clustered data; i,e, observations can be grouped into independent clusters. Clustered data arise from many applications. For example, due to a limited study time many clinical trials are performed in multiple institutions to enroll enough patients. Even though study protocols are strictly structured, types of supportive care and other factors can vary across institutions and these institution effects are generally latent. Thus observations from the same institution are correlated and more sophisticated models are needed in place of models assuming independent observations. In this thesis new models on clustered data are proposed and I assume the number of observations in each cluster is bounded by a finite number. I propose estimators of association levels between covariates and outcome, and study asymptotic behaviour of the estimators as the number of independent clusters goes to infinity.

Existing models in this field can be grouped into two big categories. The first category contains marginal models, evaluating population-averaged covariate effects. The second category includes conditional models, in which covariate effects are only interpretable conditioning on latent cluster effects. Review of models evaluating associations in binary or ordinal data forms the first part of Chapter I, in which the i^{th} cluster observations are denoted by $\boldsymbol{O}_i := (\boldsymbol{Y}_i,\ \boldsymbol{Z}_i)$; $O_{ij} = (Y_{ij},\ Z_{ij})$ is the j^{th} observation or component of $\boldsymbol{O}_i$, $j = 1, \ldots, n_i$; n_i is the total number of observations in this cluster. Ordinal outcome Y comes from $G + 1$ categories; G is a positive integer. The second part of Chapter I contains reviews on survival models for right truncated data, evaluating associations between the failure time T and covariates Z. In the i^{th} cluster, right-truncated failure times are $Y_{ij} := \min\{T_{ij}, C_{ij}\}$ where C is the censoring time and failure indicators are

$$\Delta_{ij} = \begin{cases} 1 & T_{ij} \leq C_{ij} \\ 0 & T_{ij} > C_{ij} \end{cases},$$

Each observation is denoted by $O_{ij} = (Y_{ij}, \Delta_{ij}, Z_{ij})$; observations from the i^{th} cluster are denoted

by $\boldsymbol{O}_i = \{(Y_{ij}, \Delta_{ij}, Z_{ij}); j = 1, \ldots, n_i\} = \{O_{ij}; j = 1, \ldots, n_i\}$. For clarity, bold notations stand for clustered observations in this thesis.

1.2 Models for Independent Data

Before delving into clustered data literatures, the following is a brief overview of models for independent data $O_i := (Y_i, Z_i)$, $i = 1, \ldots, m$. These models will be generalized to accommodate clustered data in this thesis.

One of the most popular models for binary data is the logistic model and the logit link is

$$Z_i^T \beta = \log \left(\frac{\mathrm{pr}(Y_i = 1 \mid Z_i)}{1 - \mathrm{pr}(Y_i = 1 \mid Z_i)} \right) .$$

Another choice is the complementary-log-log link:

$$Z_i^T \beta = \log \left\{ -\log \left(1 - \mathrm{pr}(Y_i = 1 \mid Z_i) \right) \right\} .$$

When $Z_i^T \beta$ is not high, probabilities $\mathrm{pr}(Y_i = 1 \mid Z_i)$ from the two models are similar; as $Z_i^T \beta$ goes to infinity, the probability from the second model approaches one slower than its counterpart.

For categorical data, μ_{ig} stands for $\mathrm{pr}(Y_i = g \mid Z_i)$ and the baseline-category logit model has this link:

$$\alpha_g + Z_i^T \beta_g := \log \left(\frac{\mu_{ig}}{\mu_{i,G+1}} \right) , \quad g = 1, \ldots, G ,$$

where category $G + 1$ is used as the baseline category. In contrary to binary models, there is no intercept in the covariate vector Z_i due to an identifiability problem. Suppose there is a common intercept α^*, we can set new category specific intercepts as $\alpha'_g := \alpha_g + \alpha^*$ and the new common intercept becomes zero. It can model both nominal and ordinal data and the choice of a baseline category is arbitrary. For ordinal data, the adjacent-categories logit model sets the odds of adjacent categories as:

$$\alpha_g + Z_i^T \beta := \log \left(\frac{\mu_{ig}}{\mu_{i,g+1}} \right) , \quad g = 1, \ldots, G .$$

β remains constant across different pairs of categories. Also for ordinal data, the cumulative logit model is

$$\begin{aligned} \alpha_g + Z_i^T \beta := \quad & \log \left(\frac{\mu_{i1} + \ldots + \mu_{ig}}{\mu_{i,g+1} + \ldots + \mu_{i,G+1}} \right) , \quad g = 1, \ldots, G , \\ \text{constrained by} \quad & -\infty < \alpha_1 \leq \ldots \leq \alpha_G < \infty . \end{aligned}$$

Inference based on the above models is straightforward using the quasi-likelihood by Wedderburn (1974). In this thesis, I will generalize this inference procedure into my models.

For independent survival data, Cox (1972) proposed the proportional hazards model; this model and its generalizations are the most popular models in this field; the new model in this thesis is also a generalization. In the proportional hazards mode, the hazard rate has two parts: a baseline hazard rate function λ_0, i.e. the hazard rate at zero-value of covariates; and the regression parameters β, describing how the hazard varies in response to covariates. Hazard rates from different sub-populations are assumed to be proportional over time: given a covariate Z_i, the hazard rate $\lambda(t \mid Z_i)$ and the cumulative hazard $\Lambda(t \mid Z_i)$ at time t are:

$$\lambda(t \mid Z_i) = \lambda_0(t) e^{Z_i^T \beta}, \quad \Lambda(t \mid Z_i) = \int_0^t \lambda(s \mid Z_i) ds := \Lambda_0(t) e^{Z_i^T \beta} .$$

For model inference, Cox (1972, 1975) proposed the elegant partial likelihood approach; its estimator achieves the optimal estimation efficiency and a straightforward estimator on the asymptotic covariance of the estimator was also proposed. A generalization of partial likelihood approach is applied in this thesis.

Partially due to limitations of the time-invariant proportional hazards assumption, Bennett (1983) generalized the cumulative logit model from ordinal data into survival data. In this model, given a covariate Z_i, the log survival odds at time t is

$$-\text{logit}\left\{S(t \mid Z_i)\right\} = G(t) + Z_i^T \beta, \quad \text{where } \text{logit}(x) := \log\left(\frac{x}{1-x}\right) .$$

where $G(t)$ is the baseline log failure odds function. Hazard rate ratios between sub-populations converge to one as time passes. Bennett (1983) suggested using this model for some effective cure with which the mortality rate in a diseased group is speculated to converge to control's over time. To estimate parameters under interest, Murphy et al. (1997) proposed the non-parametric maximum likelihood estimation (NPMLE): the baseline log failure odds function is estimated non-parametrically as a non-decreasing, right-continuous, step function; the estimates haven been shown to be consistent and achieve the optimal estimation efficiency. In this thesis I use something similar to NPMLE for model inference. Ad the work in Murphy et al. (1997) is quite helpful.

1.3 Marginal Models for Clustered Data

1.3.1 Continuous Outcome

The multivariate normal distribution is the most common choice for modeling clustered continuous outcomes. And most models for clustered dis-continuous outcomes, including the new models in this thesis, can be viewed as generalizations of this method.

1.3.2 Binary or Ordinal Outcome

In marginal models, people are generally interested in covariate effects from the marginal mean model $\mathrm{E}(Y_{ij} \mid Z_{ij}) = h(Z_{ij}^T\beta)$, where h is the inverse link function. And the question is how to introduce or account for underlying correlations in the clustered data. The most straightforward way is to work on the full or partially-specified joint model of outcomes. Liang and Zeger (1986) worked on partially-specified joint models which only assume the marginal mean model. They proposed the Generalized Estimating Equation (GEE) for model inference:

$$\sum_{i=1}^{m} D_i^T V_i^{-1}(\boldsymbol{Y}_i - h(\boldsymbol{Z}_i^T\beta)) = 0\ . \tag{1.1}$$

$D_i = \partial h(\boldsymbol{Z}_i^T\beta)/\partial\beta$ and $V_i = A_i^{1/2} R A_i^{1/2}/\phi$, where A_i is a diagonal matrix with elements $\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i)$, ϕ is the parameter of dispersion and R is a working correlation matrix. Their method is quite general and can adopt to many different types of outcomes. However, for binary or ordinal outcomes, the cluster-common working correlation matrix R is not proper since correlations between two outcomes are related to their marginal means. Besides, the Fréchet bound on binary data correlations is not taken into account; see Chaganty and Joe (2006) for more details. Consequently, in these cases, estimates of the working matrix R are usually not interpreted since estimated correlations may exceed the Fréchet bound. When the estimates of nuisance parameters in R exceed the Fréchet bound, estimates do not correspond to any plausible joint models and thus are meaningless. Yet GEE is one of the most popular inference methods due to its robustness and in this thesis I replace the working weighting matrix by the real covariance matrix retrieved from my parametric joint models, achieving high estimation efficiency while maintaining inference robustness.

People also worked on fully specified joint model. Considering the simplest clustered ordinal data: the paired ordinal data (Y_{i1}, Y_{i2}), $i = 1, \ldots, m$, Dale (1986) proposed to model the marginal

probabilities $\mathrm{pr}(Y_{ij} \le g \mid \boldsymbol{Z}_i)$, $j = 1, 2$, $g = 1, \ldots, G$; for correlation modeling, Dale proposed the global cross-ratios (GCR):

$$\psi^{(i)}_{g_1 g_2 \boldsymbol{Z}} = \frac{\mathrm{pr}(Y_{i1} \le g_1, Y_{i2} \le g_2 \mid \boldsymbol{Z}_i)\mathrm{pr}(Y_{i1} > g_1, Y_{i2} > g_2 \mid \boldsymbol{Z}_i)}{\mathrm{pr}(Y_{i1} > g_1, Y_{i2} \le g_2 \mid \boldsymbol{Z}_i)\mathrm{pr}(Y_{i1} \le g_1, Y_{i2} > g_2 \mid \boldsymbol{Z}_i)}, \quad g_1, g_2 = 1, \ldots, G\ .$$

These three groups of quantities, two marginal means plus the GCR, together specify the distribution of paired ordinal data. Molenberghs and Lesaffre (1994) applied Dale's model to three real datasets and demonstrated its flexibility. More generally, Bishop et al. (1975) gave the saturated log-linear model for clustered binary random variables, which incorporates all possible joint distributions, as discussed by Prentice and Zhao (1991), Liang et al. (1992), Heagerty and Zeger (1996), Heagerty and Zeger (1998), Fitzmaurice and Laird (1993), Prentice and Zhao (1991), etc. Given a binary random vector $\boldsymbol{Y} = \boldsymbol{y}$, the log-linear model is

$$\mathrm{pr}(\boldsymbol{y} = (y_1, \ldots, y_n)) = \exp\{u_0 + \sum_{j=1}^{n} u_j y_j + \sum_{j<k} u_{jk} y_j y_k + \ldots + u_{12\ldots n} y_1 y_2 \ldots y_n\}\ , \qquad (1.2)$$

where u_0 is a normalizing constant. Other parameters are interpreted via conditional odds and odds ratios (OR):

$$\begin{aligned}
u_j &= \log \mathrm{odds}\,(y_j = 1 \mid y_k = 0, k \ne j) := \log \frac{\mathrm{pr}(y_j = 1 \mid y_k = 0, k \ne j)}{\mathrm{pr}(y_j = 0 \mid y_k = 0, k \ne j)}, \quad j = 1, \ldots, n; \\
u_{jk} &= \log \mathrm{OR}(y_j, y_k \mid y_l = 0, l \ne j, k), \quad j < k = 2, \ldots, n; \\
u_{123} &= \log \mathrm{OR}(y_1, y_2 \mid y_3 = 1, y_l = 0, l > 3) - \log \mathrm{OR}(y_1, y_2 \mid y_3 = 0, y_l = 0, l > 3)\ .
\end{aligned}$$

Interpretation and inference on $\boldsymbol{u}$ need a balanced design, as discussed by Fitzmaurice and Laird (1993); interpretation of $\boldsymbol{u}$ changes as the number of observations from a cluster varies, making study replications quite hard. Parameters are interpreted conditional on cluster, rather than marginally. Zhao and Prentice (1990) directly modeled clustered binary data by their marginal means and covariances, showing there is an one-to-one relationship between marginal means, covariances and $(u_j,\ u_{jk},\ j < k,\ j, k = 1 \ldots, n)$. They simply fixed the other canonical parameters involving higher order products, and derived the popular quadratic exponential distribution:

$$l := \log\left(\mathrm{pr}(\boldsymbol{y} = (y_1, \ldots, y_n))\right) = \boldsymbol{y}^T \boldsymbol{\theta} + \boldsymbol{v}^T \boldsymbol{\lambda} + c(\boldsymbol{y}) - \log(\Delta)\ ,$$

where $\boldsymbol{v} = (y_1 y_2, \ldots, y_{n-1} y_n)$. Here they rewrote $\boldsymbol{\theta} = (u_1, \ldots, u_n)^T$, $\boldsymbol{\lambda} = (u_{12}, \ldots, u_{n-1,n})^T$ and $\Lambda = u_0$. Yet extra caution is needed to model marginal pairwise products due to the Fréchet

bound. Enlightened by Dale's binarization idea, Heagerty and Zeger (1996) modeled marginal means and pairwise odds ratios of binarized ordinal variables. Their model has three advantages: marginal pairwise odds ratios are unbounded; a balanced design is not a must; properly chosen regression parameters are orthogonal to the other nuisance canonical parameters $(u_{123}, \ldots, u_{12\ldots n})$, resulting in elevated estimation efficiency. Similar to GEE, these models only require a partial specification of the underlying joint distribution: they only need the pairwise joint distribution to be specified, and thus introduce some degree of robustness. In contrary, proposed models in this thesis are parametric, but on the other hand, I develop robust inference methods that are similar to the above inference methods.

The above models mostly are semi-parametric, leaving the correlation structures unknown or partially specified and sometimes working versions are applied. People have also worked with other parametric joint models using copula idea. Starting from the bi-variate case, the burn injury data from Fan and Gijbels (1996) motivated the copula model in Song et al. (2009). In this dataset, every independent patient has two dependent outcomes: the area of burns and the death indicator. The authors jointly modeled continuous and discrete outcomes by the Gaussian copula:

$$F(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i; \beta, \phi, \Gamma) = \Psi_2 \left\{ \Psi^{-1}[h_1(\boldsymbol{Z}_i^T \beta_1)/\phi], \Psi^{-1}[h_2(\boldsymbol{Z}_i^T \beta_2)/\phi], \Gamma \right\} \ ,$$

where ϕ is a deviance parameter; Ψ and Ψ_2 are normal cumulative distribution functions, univariate and bivariate. Γ is the correlation matrix of Ψ_2, which is quite flexible. In this copula, latent normal variables are transformed into continuous variables. Chaganty and Joe (2004) considered the following latent variable modeling for dependent binary data:

$$Y_{ij} = 1\{Y_{ij}^* < Z_{ij}^T \beta\}; \quad \text{thus } \mathrm{E}(Y_{ij} \mid Z_{ij}) = \Psi(Z_{ij}^T \beta) \ .$$

This model corresponds to the probit marginal link. And it is convenient to impose any legitimate correlation structures onto multivariate normal variables. Other marginal link functions were also discussed. Logistic distributed latent random variables give the popular logistic marginal link:

$$Y_{ij} := 1\{Y_{ij}^* < Z_{ij}^T \beta\}, \quad \boldsymbol{Y}_i^* = \log \frac{F(\boldsymbol{t}_i)}{1 - F(\boldsymbol{t}_i)}; \quad \boldsymbol{t}_i \text{ can be any multivariate random variable.}$$

O'Brien and Dunson (2004) took $\boldsymbol{t}$ to be t-distributed, allowing flexible correlation structures. These parametric models all use the copula idea: they all transform latent continuous outcomes into observed discrete outcomes and cover all marginal parametric model. In this thesis I give a model

under the logit link for clustered binary outcomes and marginally it corresponds to the joint logistic distribution model.

The correlation modeling of the above model is based on latent continuous outcomes, but interpretations are not always straightforward. My new models can be viewed as latent variable models in which marginally outcomes correspond to a multivariate logistic distribution and conditionally they correspond to the multivariate Gumbel distribution. Correlations are originally introduced onto the conditional multivariate Gumbel distribution, leaving the interpretations of the outcomes in complicated forms.

1.3.3 Longitudinal Binary or Ordinal Outcome

Markov models are useful for longitudinal outcomes. These models assume a subject's current outcome depends on its previous outcome(s), see Bishop et al. (1975), Diggle et al. (1994), Azzalini (1994) and Albert (2000) for more details. In this thesis, I am not discussing such situations and this is a future direction of research.

1.3.4 Survival Data

Here I restrict my discussions on models for clustered data without time-varying covariates.

Lin (1994) proposed a marginal model by only specifying the marginal failure of each single observation under the proportional hazards model. This model ignores underlying correlations.

Li and Lin (2006) considered a marginal proportional hazards model for spatial data, introducing correlations by the copula transformation of the observation-specific cumulative hazard functions. They modeled each hazard rate function as

$$\Lambda_i(t) = \Lambda_0(t)\exp\left(Z_i^T\beta\right) .$$

Correlations were introduced by the probit-type transformation of cumulative hazards:

$$T_i^* = \Phi^{-1}\left\{1 - \exp\left(-\Lambda_i(t)\right)\right\} \sim \mathcal{N}(0,1) .$$

Flexible correlation modeling of T_i^* can be imposed.

In this thesis I introduce underlying correlations into a marginal model by correlated frailties.

1.3.5 Discussions

Most marginal mean models of clustered data make the assumption that

$$\mathrm{E}[Y_{ij} \mid \boldsymbol{Z}_i] = \mathrm{E}[Y_{ij} \mid Z_{ij}] \, .$$

These models include assuming a non-independent working correlation matrix R in (1.1), and the model proposed by Zhao and Prentice. However, this assumption may be violated in longitudinal studies: a patient's health status in the first month will affect his or her outcomes in the following months. The violation causes inconsistent estimates. For more details, please refer to Pepe and Anderson (1994). Markov models are useful in this case. And my models also takes the above assumption. One ad-hoc solution is to put in covariate from previous correlated observations into the mean model.

1.4 Conditional Models for Clustered Data

1.4.1 Mixed Effects Models

In this thesis I propose several new mixed effects models, and mixed effects models are the most well-known conditional models, in which latent cluster effects that introduce correlations are incorporated into the mean model as random effects. Fully parametric mixed effects models are specified given: 1) outcome distribution conditional on covariates and random effects and 2) the distribution of random effects. Here I give several examples of the conditional mean models, following notations from Heagerty (1999) and Heagerty and Zeger (2000a):

1. shared random intercept b:

$$\mathrm{E}[Y_{ij} \mid Z_{ij}, b_i] = h(Z_{ij}^T \beta + b_i) \, ;$$

2. individual random intercept $\boldsymbol{b}$:

$$\mathrm{E}[Y_{ij} \mid Z_{ij}, b_{ij}] = h(Z_{ij}^T \beta + b_{ij}), \quad \boldsymbol{b}_i = (b_{ij})_{j=1,\ldots,n_i} \, ;$$

3. shared random slope b:

$$\mathrm{E}[Y_{ij} \mid Z_{ij}, X_{ij}, b_i] = h(Z_{ij}^T \beta + X_{ij}^T b_i) \, , \quad X_{ij} \text{ is another set of known covariates} \, ;$$

4. individual random slope $\boldsymbol{b}$:

$$\mathrm{E}[Y_{ij} \mid Z_{ij}, X_{ij}, b_{ij}] = h(Z_{ij}^T\beta + X_{ij}^T b_{ij}) \, .$$

Without loss of generality, all random effects are assumed to be centered, i.e., mean zero. In general people also assume that conditional on random effects, outcomes are independent of each other. Predictions of underlying cluster effects in the form of posterior expectations of random effects given the data can be made. Besides, these models can evaluate correlations among three or more observations.

More generally, people wrote the mixed effects model by putting observations as $\boldsymbol{Y} := (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_m^T)^T$, $\boldsymbol{Z} := (\boldsymbol{Z}_1^T, \ldots, \boldsymbol{Z}_m^T)^T$ and $\boldsymbol{X} := (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_m^T)^T$. The conditional model becomes

$$\mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{b}] = h(\boldsymbol{Z}^T\beta + \boldsymbol{X}^T\boldsymbol{b}); \quad \boldsymbol{Z}, \boldsymbol{X} \text{ are known design matrices.}$$

With only random intercepts, $\boldsymbol{X}$ is a matrix with ones and zeroes. Suppose there are two clusters and each has three observations, in the case of shared random intercepts, $\boldsymbol{X}$ and $\boldsymbol{b}$ are

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \, , \quad \boldsymbol{b} = (b_1, b_2)^T \, ;$$

in the case of individual random intercepts, they are

$$\boldsymbol{X} = I_6, \quad \boldsymbol{b} = (b_{11}, b_{12}, b_{13}, b_{21}, b_{22}, b_{23})^T \, .$$

I_6 is the 6×6 identity matrix. Breslow and Clayton (1993) discussed semi-parametric mixed effects models, and they only specified the outcome conditional distribution up to the second order of moments.

1.4.2 Linear Mixed Effects Model

The identity link is quite popular for continuous outcomes:

$$\boldsymbol{Y} = \boldsymbol{Z}^T\beta + \boldsymbol{X}^T\boldsymbol{b} + \boldsymbol{\epsilon} \, , \tag{1.3}$$

where $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are centered multivariate normal random variables with covariance matrices D and W.

This model can approximate generalized linear links by Taylor Expansion, transforming original outcomes into working continuous outcomes $\boldsymbol{Y}^*$. Denoting the inverse link function by h, Breslow and Clayton (1993) used this transformation:

$$\boldsymbol{Y}^* := h^{-1}(\boldsymbol{Y}) \approx \boldsymbol{Z}^T\beta + \boldsymbol{X}^T\boldsymbol{b} + (\boldsymbol{Y} - \boldsymbol{\mu^b})\{h^{-1}\}'(\boldsymbol{\mu^b}) \,, \tag{1.4}$$

where $\boldsymbol{\mu^b} = h(\boldsymbol{Z}^T\beta + \boldsymbol{X}^T\boldsymbol{b})$ and $\{h^{-1}\}'$ is the derivative of the link function w.r.t. β. β is both conditionally and marginally intepretable, representing the change in the transformed or approximate outcome expectation with regards to one unit change in the corresponding covariate. Due to the adhoc approximations, this thesis does not discuss the linear link for non-linear outcomes.

1.4.3 Generalized Linear Mixed Effects Models

In the shared random effects mixed effects models, β are interpreted by comparing two subjects from the same cluster. In the individual random effects mixed effects models, regression parameters are interpreted conditional on observations. And these issues motivates this thesis. Since the conditional link function is no longer the identity link, the interpretation of the conditional regression parameter β is not readily generalizable into its marginal counterpart.

1.4.4 Frailty Models

In this thesis one of the new mixed effects model is a frailty model and here I give a brief overview of its history. Clayton (1978a) introduced the shared relative risk, motivated from a clustered survival dataset of father-son pairs. Clayton and Cuzick (1985) incorporated covariates into this model. Such models are equivalent to incorporating some cluster-common random factors multiplicatively into the hazard rates. Vaupel et al. (1979) named these random factors frailties. To be specific, in frailty models, given a cluster-common frailty w_i and a covariate z_{ij}, the conditional hazard rate is assumed to be $w_i\lambda_0(t)\exp\left(z_{ij}^T\beta\right)$; β are conditional log hazard rate ratios. As mixed effects models, observations are assumed to be independent, conditioning on corresponding frailties and covariates. These models are named frailty models. They are mixed effects models with a conditional complementary-log-log link; logarithms of frailties are random intercepts. I also view these models as imposing cluster-specific but proportional baseline hazard rate functions. For bivariate cases, shared frailty models are flexible: Oakes (1982) showed many bivariate survival distributions

uniquely correspond to some shared frailty models. However, with larger clusters, shared frailty models can only impose an exchangeable correlation structure among observations. This is not flexible for more complicated designed familial studies or other applications such as spatial epidemiology. For example, in the adoption study from Nielsen et al. (1992), each cluster contains the adopted child, the adopting parents and the biological parents. There are two kinds of correlations: the first one involves shared environmental factors among adopted parents and the child; the second one represents shared genetic factors among biological parents and the child. Vaida and Xu (2000) proposed an individual frailty model with log-normal frailties. Petersen (1998), Parner (2001) and Parner (1998) discussed the correlated frailty model, in which multivariate Gamma frailties replace the shared frailty. The j^{th} observation in the i^{th} cluster has the hazard rate function:

$$\lambda_{ij}(t \mid z_{ij}, w_{ij}) = w_{ij}\lambda_0(t)\exp\left(z_{ij}^T\beta\right) ,$$

where W_{ij} is a summation of properly scaled independent Gamma random variables. For the above adoption study, Parner (2001) proposed the individual frailties as

$$\begin{aligned} \text{Adoptive mother (AM):} \quad & W_{\mathrm{AM}} = W_1 + W_{01} \ ; \\ \text{Adoptive child (AC):} \quad & W_{\mathrm{AC}} = W_1 + W_2 + W_{02} \ ; \\ \text{Biological mother (BM):} \quad & W_{\mathrm{BM}} = W_2 + W_{03} \ . \end{aligned}$$

He assumed the independent frailty components follow Gamma distributions: $W_k \sim \mathrm{Gamma}(\gamma_k, \eta)$ and $W_{0k} \sim \mathrm{Gamma}(\gamma_{0k}, \eta)$. To ensure model identifiability, he also restricted that $\gamma_1 + \gamma_{01} = \gamma_1 + \gamma_2 + \gamma_{02} = \gamma_2 + \gamma_{03} = \eta$. This method of multivariate Gamma variable generation is relatively straightforward for decomposing cluster effects into different physical sources: W_1 presents the environmental effects and W_2 stands for the shared genetic factors. Yet this simple adding operation has some limitations. Following the same idea, the frailty of the adoptive child's biological father can be decomposed into

$$\text{Biological father (BF):} \quad W_{\mathrm{BF}} = W_3 + W_{04} \ .$$

However, the parameter of W_3 is restricted by the parameter γ_{02} since

$$\text{Adoptive child (AC):} \quad W_{\mathrm{AC}} = W_1 + W_2 + W_{02} = W_1 + W_2 + W'_{02} + W_3 \ .$$

Thus, replicability of this correlated frailty model is limited. It is very hard to generalize this model into spatial data. In this thesis, I propose a new way to generate multivariate Gamma random variables such that cluster size needs not to be specified/fixed before hand.

Researchers have generalized some survival models into clustered ordinal data, since ordinal data arise from dividing study times into several intervals and transforming continuous survival times into corresponding time interval indexes. Heagerty and Zeger (2000b), and Ten Have (1996) gave good examples. And this thesis also connects survival times to ordinal data and binary data.

1.5 Marginalizable Models

For general scientific questions, Neuhaus et al. (1991) and Heagerty (1999) argued that marginal parameters are more directly interpretable and are preferred to answer public health questions. Back to the clinical trial example, suppose the primary interest is to evaluate treatment efficacy in the targeted population, and the secondary interest is making predictions, or evaluating correlation levels among two or more observations. Marginal models can answer the first question while the secondary questions can be solved by mixed effects models. In general these two classes of models are making different assumptions on the population: a parametric mixed effects model naturally corresponds to a marginal model,

$$f(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{Z}_i = \boldsymbol{z}_i) = \int_{b_{in_i}} \cdots \int_{b_{i1}} \left\{ \prod_{j=1}^{n_i} f(y_{ij} \mid z_{ij}, b_{ij}) \right\} f(b_{i1}, \ldots, b_{in_i} \mid \boldsymbol{Z}_i = \boldsymbol{z}_i) db_{i1} \cdots db_{in_i} . \tag{1.5}$$

But such marginal models are rarely interpretable. Consider a shared frailty model where Gamma frailties have mean one and variance 0.5, the marginal survival probability is

$$S(t \mid Z_i = z_i) = \left(\frac{1}{1 + \Lambda_0(t) e^{z_i^T \beta - log2}} \right)^2 .$$

β has no direct interpretations marginally.

Since marginalizable models have the advantages of both models, they have wider applicability, motivating this thesis.

Several marginalizable models have been discussed. Lang and Agresti (1994) considered modeling the joint distribution plus the marginal distribution of ordinal data simultaneously. The authors defined a quantity called profile, which is the joint outcome of a cluster. For example, each cluster

contains n observations and the j^{th} observation comes from I_j categories; then there are $\sum_{j=1}^{n} I_j$ profiles. They considered a contingency table, in which the columns correspond to different profiles and the rows corresponds to different covariate levels. Without any model constraints, the cell frequencies from the data are the MLE of cell probabilities $\boldsymbol{\mu}$. The authors viewed a proposed model as giving constraints in maximizing the observed data likelihood. For example, when the outcomes are binary, each cluster has two outcomes, and the only covariate is the intercept, the following equation

$$C\log(A\boldsymbol{\pi}) = \boldsymbol{Z}^T\beta = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \log \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \alpha$$

corresponds to the marginal logit model. And it imposes a constraint on maximizing the observed log likelihood in the sense that with a matrix U whose columns are orthogonal complement to the space spanned by the columns of $\boldsymbol{Z}$, the following equation is true

$$U^T C\log(A\boldsymbol{\pi}) = \mathbf{0}\ .$$

Another constraints is that the summation of $\boldsymbol{\pi}$ is one with the same covariate level. The authors then proposed maximizing data log likelhood via Lagrange multipliers. The applicability of this methodology largely depends on the size of contingency table and is computationally intensive when the covariates are continuous.

To accommodate continuous covariates, in this thesis I propose marginalizable mixed effects models and in the following I discuss existing mixed effects models that yield interpretable marginal covariate effects. To be clear, I denote the conditional covariate effects by β^C and its marginal counterpart by β^M. Heagerty (1999), and Heagerty and Zeger (2000a) proposed marginalizable conditional models by first modeling the marginal model and the random effect distribution and then solving for the conditional model via deconvolution:

$$h(Z_{ij}^T\beta^M) = \int h\left(\Delta(Z_{ij}) + b_{ij}\right) dF(b_{ij})\ .$$

However, only certain pairs of marginal model and random effect distribution yield analytic results, such as the probit marginal link plus Gaussian random effects, the Poisson conditional link and Gamma random effects Tsutakawa (1988).

On the other hand, marginal and conditional distributions determine the random effect distribution. With a shared random intercept, Wang and Louis (2004) set the marginal and conditional links identical:

$$\mathrm{E}(Y_{ij} \mid Z_{ij}) = h(Z_{ij}^T \beta^M), \quad \mathrm{E}(Y_{ij} \mid Z_{ij}, b_i) = h(Z_{ij}^T \beta_i^C + b_i) \; .$$

They solved for the random effect distribution density function f from (1.5) and named it the bridge distribution:

$$f(b) = \frac{1}{2\pi} \int e^{-ib\xi} \frac{\mathcal{F}h'(\xi/\phi)}{\mathcal{F}h'(\xi)} d\xi \; ,$$

where $\mathcal{F}$ denotes the Fourier transformation and h' is the derivative of the inverse link function h. To account for heterogeneity among clusters, cluster-specific dispersion parameters ϕ_i were introduced. For the i^{th} cluster, $\beta^M = \phi_i \beta_i^C$ and $\phi_i := \left(1 + \sigma_i^2/\sigma_h^2\right)^{-1/2}$ and σ_i^2 is the variance of the random intercept b_i; σ_h^2 is variance of the distribution corresponding to h. For example, logistic link invokes a standard logistic distribution and $\sigma_h^2 = \pi^2/3$. ϕ_i characterizes the effect of cluster-level heterogeneity: the smaller the ϕ_i, the bigger the σ_i^2 and β_i^C. Suppose the population-average covariate effect β^M is a fixed quantity so with a more dispersed random intercept, a stronger cluster-specific covariate effect β_i^C is needed to remove the extra dispersion.

Parzen et al. (2011) discussed the bridge distribution in the case of individual random intercepts with logit link. The conditional mean model and the cumulative distribution function of random effect b are

$$\mathrm{E}(Y_{ij} \mid Z_{ij}, b_{ij}) = \mathrm{logit}(b_{ij} + \phi^{-1} Z_{ij}^T \beta^C) \; , \quad F(b) = 1 - \frac{1}{\pi\phi} \left(\frac{\pi}{2} - \arctan\left(\frac{e^{\phi b} + \cos(\phi\pi)}{\sin(\phi\pi)} \right) \right) \; .$$

The novel bridge distribution has no physical interpretations and since model inference is based on MLE, the bridge distribution computing has to be coded manually.

In frailty models, the exponentiated version of the above bridge distribution is the stable distribution, first introduced by Hougaard (1986). The joint distribution of clustered failure times is:

$$F_i(t_{i1}, \ldots, t_{in_i}) = \exp\left\{ - \left(\sum_{j=1}^{n_i} \Lambda_{ij}(t_{ij}) \right)^{\theta} \right\} \; , \quad \text{i.e. } \beta^M = \theta \beta^C \; .$$

The marginal model is still the proportional hazards model, with scaled marginal covariate effects. Liu et al. (2011) discussed a study comparing the ratio of observed to expected deaths, known as

the standardized mortality ratio, between a group of U.S. kidney transplant centers and the national average. Since smaller centers tend to have more heterogeneous survival observations, a single frailty distribution is not sufficient. They modeled shared frailties by a covariate-dependent stable distribution.

A stable distribution on random effects can be justified when people believe there are many latent cluster effects under some long tail distributions, and these effects act together to produce an additive final effect which imposes a multiplicative effect on the hazard rate function. However, distributions without any finite moments are not widely used in bio-medical applications. Multivariate stable distribution is not fully studied yet and thus it cannot model correlations flexibly.

New marginalizable models discussed in this thesis impose highly flexible correlation structures onto clustered binary, ordinal and survival data. People can estimate the marginal cumulative log odds, and evaluate correlation levels from this set of models. To estimate parameters under interest in binary or ordinal data, I take the advantage of the marginalizable property and propose a robust inference procedure, which has small estimation efficiency loss compared to maximum likelihood estimation (MLE).

Chapter 2

LITERATURE REVIEW: MODEL INFERENCE

2.1 Likelihood-Based Methods

2.1.1 Direct Maximization of the Log-likelihood

Linear mixed effects models generally use an identity link. People usually assumed $\boldsymbol{b}_i \sim \mathcal{N}(0, D)$, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, W)$ and $\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i, \boldsymbol{X}_i) = \boldsymbol{X}_i^T D \boldsymbol{X}_i + W := V_i$, where $\mathcal{N}$ stands for the multivariate normal distribution. Harville (1977) found the MLE as

$$\hat{\beta}_{BLUE} = (\boldsymbol{Z}_i^T V_i^{-1} \boldsymbol{Z}_i)^{-1} \boldsymbol{Z}_i^T V_i^{-1} \boldsymbol{Y}_i \, . \tag{2.1}$$

Since the weighting matrix is the inverse covariance matrix of $\boldsymbol{Y}_i$, $\hat{\beta}_{BLUE}$ has the optimal estimation efficiency. The best linear unbiased prediction (BLUP) of $\boldsymbol{b}_i$ is

$$\hat{\boldsymbol{b}}_{BLUP} = \mathrm{E}(\boldsymbol{b}_i \mid \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{X}_i) = \boldsymbol{X}_i^T D V_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{Z}_i^T \beta) \, . \tag{2.2}$$

Properties of these estimates were well studied.

Wolfinger and O'connell (1993) transformed non-continuous outcome $\boldsymbol{Y}$ into a continuous working outcome $\boldsymbol{Y}^*$ by:

$$\boldsymbol{Y}^* := h^{-1}(\boldsymbol{Y}) \approx \boldsymbol{Z}^T \beta + \boldsymbol{X}^T \boldsymbol{b} + (\boldsymbol{Y} - \boldsymbol{\mu}^{\boldsymbol{b}}) \{h^{-1}\}'(\boldsymbol{\mu}^{\boldsymbol{b}}) \, , \quad \text{where } \boldsymbol{\mu}^{\boldsymbol{b}} = h(\boldsymbol{Z}^T \beta + \boldsymbol{X}^T \boldsymbol{b}) \, .$$

The inference algorithm iterates between the transformation procedure and the estimation procedure in (2.1) and (2.2) until convergence.

Latent variable models are parametric and MLE inference is readily applicable for some cases. For example, Chaganty and Joe (2004) proposed the model

$$Y_{ij} = 1\{Y_{ij}^* < Z_{ij}^T \beta\}; \quad \text{leading to } \mathrm{E}(Y_{ij} \mid Z_{ij}) = \Psi(Z_{ij}^T \beta) \, ,$$

where $\boldsymbol{Y}_i^*$ is a multivariate normal random vector with a standard variance and a correlation matrix R. The joint likelihood of a cluster with n observations is in the form of Ψ_n, the cumulative distribution function of a n-dimensional normal random vector. R software has a package **mvtnorm** computing Ψ_n up to $n = 100$.

MLE inference is also used in simple joint model, such as the paired ordinal data model in Dale (1986) and Markov models in Azzalini (1994) and Heagerty (2002). However, MLE has heavy computation burdens in the case of many mixed effects models due to: 1) non-normal distribution and thus there is no package for computing the joint likelihood, 2) correlated random effects leading to high dimensional integration. This is the exact case for this thesis: I propose models with a marginal logit link, invoking a logistic distribution, corresponding to the marginal logit link, and at the same time there are individual random effects.

2.1.2 Numeric Maximization of the Log-likelihood

Mixed effects models are generally parametric and MLE is applicable, yielding asymptotically consistent and efficient estimates. The likelihood of the i^{th} cluster is:

$$f(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{Z}_i = \boldsymbol{z}_i) = \int_{b_{in_i}} \cdots \int_{b_{i1}} \left\{ \prod_{j=1}^{n_i} f(y_{ij} \mid z_{ij}, b_{ij}) \right\} f(b_{i1}, \ldots, b_{in_i} \mid \boldsymbol{Z}_i = \boldsymbol{z}_i) db_{i1} \cdots db_{in_i} . \tag{2.3}$$

This integral has a closed form with conjugate pairs of distribution, such as Gaussian random effects and the identity link, Gamma random effects and the Poisson conditional link (Cox and Reid, 1987), log-Gamma random effects with the complementary-log-log conditional link in Ten Have (1996). But in general, there is no closed-form expression, resulting in complicated computations. Some researchers approximated (2.3) by Monte Carlo methods. For example, with Gaussian random effects, Gauss-Hermite quadrature from Anderson and Aitkin (1985) is useful. For an integral in the form of

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx ,$$

Gauss-Hermite Quadrature selects a set of n values $(x_1, \ldots, x_n)$, called nodes, which are roots of the Hermite polynomial $H_n(x)$, and then approximates this integrals by a weighted summation $\sum_{i=1}^{n} a_i f(x_i)$: $a_i = 2^{n-1} n! \sqrt{n} / \{n^2 [H_{n-1}(x_i)]^2\}$. Hartzel et al. (2001) found its calculation increases exponentially with the integral dimension; thus likelihoods of mixed effects models with individual random effects are harder to compute. This method is also computationally intensive and the approximation works best when random effects are normally distributed. Since in my case the random effects follow the the highly skewed Gumbel distribution, approximation will not work best and thus I do not consider this method.

2.1.3 Expectation-Maximization Algorithm

Dempster et al. (1977) invented the expectation-maximization (EM) algorithm, an iterative method finding the MLE from some model that depends on latent variables. Here I denote the observed data by $Y \sim Q_\theta$ on $(\mathcal{Y}, \mathcal{B})$, and the full data by $X \sim P_\theta$ on $(\mathcal{X}, \mathcal{A})$; I also assume both probability measures Q_θ and P_θ have respective densities q_θ and p_θ with regards to a dominating measure ν. Usually researchers are interested in finding

$$\hat{\theta}^Y = \underset{\theta}{\operatorname{argmax}} \log q_\theta(Y) \ .$$

Yet it is quite difficult to compute while the following is much more computationally friendly:

$$\hat{\theta}^X = \underset{\theta}{\operatorname{argmax}} \log p_\theta(X) \ .$$

The EM algorithm is designed for such scenario, alternating between an expectation (E) step and a maximization (M) step until convergence. In the $(r+1)^{st}$ iteration,

1. in the E-step,
   one computes, for $\theta \in \Theta$, $\mathrm{E}\left(\log p_\theta(X) \mid Y; \theta^{(r)}\right)$; $\theta^{(r)}$ is the estimate from the r^{th} iteration;

2. in the M-step,
   one maximizes $\mathrm{E}\left(\log p_\theta(X) \mid Y; \theta^{(r)}\right)$ in θ.

This algorithm is in wide use due to its convenient application and the property of at least converging to some local maximum. When Q_θ is some unimodal probability distribution, this algorithm guarantees finding the MLE. In cases of multimodal probability distributions, researchers usually try several starting values and compare the corresponding log-likelihood values.

This algorithm can be applied to mixed effects models, treating random effects as missing data. Given a conjugate pair of the conditional link and the random effect distribution, the E-step is relatively easy to compute. For example, in a shared frailty model with Gamma distributed frailties W's (at this moment, I assumed the Gamma distribution is known), the full log-likelihood contribution with frailties is

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} \delta_{ij}(\log\lambda_0(t_{ij}) + z_{ij}^T\beta) + \delta_{ij}w_i - \Lambda_0(t_{ij})\exp(z_{ij}^T\beta)w_i \ .$$

Only the term $\Lambda_0(t_{ij})e^{z_{ij}^T\beta}w_i$ involves both w_i and parameters $(\beta, \Lambda_0(\cdot))$ under interest, thus in E-step only $\mathrm{E}(w_i \mid \mathrm{Data})$ needs to be found out. Given the data, frailty posterior distribution is still Gamma with different parameters. In the M-step, partial likelihood inference is applicable, offset by *imputed* frailties from the E-step. This is the inference method for the frailty model in this thesis. For binary and ordinal models, I choose to use a more robust inference method.

2.1.4 Composite Likelihood Estimation

Lindsay (1988) proposed the composite likelihood, treating the product of a collection of component likelihoods as the true likelihood, where a component likelihood is some conditional or marginal likelihood.

Given a set of correlated observations $(\boldsymbol{Y}, \boldsymbol{Z}) = ((Y_1, Z_1), \ldots, (Y_n, Z_n))$, the joint likelihood is quite complicated while pairwise joint likelihoods are easier to find. When a component likelihood is set up as a pairwise likelihood, the composite likelihood is

$$\prod_{j<k} L(Y_j, Z_j, Y_k, Z_k)^{a_{jk}} , \tag{2.4}$$

where a_{jk}'s are positive weights assigned by researchers and L stands for the likelihood. In this case the composite likelihood corresponds to the likelihood of a new dataset with independent observations: $\boldsymbol{O}^* := ((Y_j, Z_j, Y_k, Z_k), 1 \le j < k \le n)$. When the component likelihood is correctly specified, MLE from the composite likelihood are consistent. However, since generally there is no corresponding real joint distributions, asymptotic behavior of estimates under model mis-specification cannot be conveniently justified.

The composite likelihood based on marginal likelihoods is particularly useful to avoid the high-dimensional integral in (2.3) caused by individual random effects. And I use the composite likelihood idea in all model inferences to avoid the complicated joint distribution calculation.

As for the new frailty model inference under EM algorithm, I generalize it to adopt the composite likelihood, based on the work in Gao and Song (2011). The M-step is quite straightforward but the E-step is tricky. Denote the missing data by $\boldsymbol{W}$, consider the new dataset outlined in the above, the

posterior expectation of the pairwise log-likelihood is

$$\begin{aligned} & \mathrm{E}\left\{l(Y_j, Z_j, W_j, Y_k, Z_k, W_k) \mid ((Y_1, Z_1), \ldots, (Y_n, Z_n))\right\} \\ = \quad & \mathrm{E}\left\{l(Y_j, Z_j, W_j, Y_k, Z_k, W_k) \mid (Y_j, Z_j, Y_k, Z_k)\right\} \ . \end{aligned}$$

2.1.5 Profile Likelihood

Profile likelihood is usually discussed when people are interested in deriving the MLE of β from a likelihood $L(\boldsymbol{Y}, \boldsymbol{Z}; \lambda, \beta)$, where λ is some unknown nuisance parameter. The straightforward derivation of MLE is to jointly maximize the likelihood in all parameter. However, for some cases the computation is tedious and the profile likelihood serves as an alternative. In the profile likelihood inference procedure, one can first get an MLE of λ with a fixed β, denoted by $\hat{\lambda}(\beta)$. Plugging this estimate back into the original likelihood gives a profile likelihood as a function of parameter β: $L_p(\boldsymbol{Y}, \boldsymbol{Z}; \beta)$. The profile MLE β_{profile} is found from $\arg\max_{\beta} L_p(\boldsymbol{Y}, \boldsymbol{Z}; \beta) := \arg\max_{\beta} L(\boldsymbol{Y}, \boldsymbol{Z}; \hat{\lambda}(\beta), \beta)$.

Patefield (1977) discussed $\hat{\beta}_{\text{profile}}$ asymptotic behavior and profile likelihood ratio test in parametric models. Patefield proved that profile likelihood can be used as a real likelihood. For example, curvature of the profile likelihood function can estimate the asymptotic covariance of $\hat{\beta}_{\text{profile}}$.

For semi-parametric models, Murphy and van der Vaart (2000) showed a profile likelihood behaves similarly as a real likelihood, and it gives the efficient score function and the efficient Fisher information for $\hat{\beta}_{\text{profile}}$. For demonstration, they discussed the proportional hazards model for independent survival data. The likelihood contribution, derived by removing terms involving only censoring times and covariates from the likelihood, is written as $L = \prod_{i=1}^{m} \left(\lambda_0(y_i) e^{z_i^T \beta}\right)^{\delta_i} \exp\left(-\Lambda_0(y_i) e^{z_i^T \beta}\right)$. Fixing β, the non-parametric MLE (NPMLE) of $\Lambda_0(\cdot)$ is a function of β:

$$\hat{\Lambda}_0^{\text{NPMLE}}(t; \beta) = \int_0^t \frac{d\frac{1}{m}\sum_{i=1}^{m} 1\{y_i \delta_i = s\}}{\frac{1}{m}\sum_{i=1}^{m} 1\{y_i \geq s\} e^{z_i^T \beta}} \ , \tag{2.5}$$

which is indeed the Breslow estimator. Plugging this estimator back into the likelihood contribution,

$\hat{\beta}_{\text{profile}}$ is the solution of:

$$\sum_{i=1}^{m} \int_0^{\tau} \left(z_i - \frac{\sum_{j=1}^{m} z_j 1\{y_j \geq u\} e^{z_j^T \beta}}{\sum_{j=1}^{m} 1\{y_j \geq u\} e^{z_j^T \beta}} \right) d1\{y_i \delta_i = u\} = 0 \, .$$

τ is the study time. $\hat{\beta}_{\text{profile}}$ is consistent and asymptotically normal with a covariance matrix derived from the profile Fisher information matrix.

In the profile likelihood, nuisance parameter is treated as a function of the parameter under interest and the data. And models from $L_p(\boldsymbol{Y}, \boldsymbol{Z}; \beta)$ form a subset of the original model set. Profile likelihood has a drawback that sometimes:

$$\mathrm{E}\left(\frac{\partial l(\beta,\lambda)}{\partial \beta}\right) \neq \mathrm{E}\left(\frac{\partial l_p(\beta)}{\partial \beta}\right) = \mathrm{E}\left(\frac{\partial^2 l(\lambda,\beta)}{\partial \beta \partial \lambda}[\hat{\lambda}(\beta) - \lambda]\right) + \mathrm{E}\left(\frac{1}{2}\frac{\partial^3 l(\lambda,\beta)}{\partial \beta \partial \lambda^2}[\hat{\lambda}(\beta) - \lambda]^2\right) + \cdots \, .$$

When the above inequality is true, the profile likelihood estimates is different from MLE, which is obtained by jointly maximizing the likelihood over all the parameters. So when researchers worked with profile likelihood to derive MLE, they needed the extra step to show these two are equivalent. For example, people have shown in the proportional hazards model, $\hat{\beta}_{\text{profile}} = \hat{\beta}_{\text{NPMLE}}$. Due to this extra step, I do not use profile likelihood inference in thesis.

2.1.6 Partial Likelihood

A partial likelihood is a component of the full likelihood. It is derived by writing a likelihood as a product of terms with and without parameters of interest; terms with parameters under interest form the partial likelihood. Cox (1975) proposed it for the proportional hazards model inference via iterative conditioning, assuming: *given covariates, failure and censoring events are independent*.

In the proportional hazards model, regression parameter β is under interest, and $\Lambda_0(t)$ is nuisance. Assuming no ties, order the Q failure times as $t_{(1)}, \ldots, t_{(Q)}$, where (q) is the anti-rank of failure times: $(q) = i$ such that $\delta_i t_i = t_{(q)}$. m_q denotes the number of subjects censored in time interval $[t_{(q)}, t_{(q+1)})$, listing the ordered censoring times as $c_{(q,1)} < \ldots < c_{(q,m_i)}$ and (i,j) is the anti-rank of censor time points $c_{(i,j)}$. I denoted $R(t_j) = \{i : y_i \geq t_j\}$, i.e. the subjects at risk at time t_j.

I denote events $V_i = \{c_{(i-1,j)}, (i-1,j), t_{(i)}, j = 1, \ldots, m_{i-1}\}$, $W_i = (i)$; $i = 1, \ldots, k$. The likelihood contribution is

$$L \propto \text{pr}(W_1 \mid V_1; \beta, \Lambda_0) \prod_{q=2}^{Q} \text{pr}(W_q \mid V_1, W_1, \ldots, V_q; \beta, \Lambda_0) \times \text{pr}(V_1) \prod_{q=2}^{Q} \text{pr}(V_q \mid V_1, W_1, \ldots, V_{q-1}) \,.$$

Only the first two terms have parameters under interest. Suppose $(i) = j$ and I wrote

$$\begin{aligned}
& \text{pr}(W_q \mid V_1, W_1, \ldots, V_q; \beta, \Lambda_0) \\
= & \lim_{\delta \to \infty} \frac{\text{pr}(t_j \in [t_{(q)}, t_{(q)} + \delta) \mid z_j, c_j \geq t, t_j \geq t) \prod_{l \in R(t_{(q)}) \setminus j} [1 - \text{pr}(t_l \in [t_{(q)}, t_{(q)} + \delta) \mid z_l, c_l \geq t, t_l \geq t)]}{\sum_{k \in R(t_{(q)})} \text{pr}(t_k \in [t_{(q)}, t_{(q)} + \delta) \mid z_k, c_k \geq t, t_k \geq t) \prod_{l \in R(t_{(q)}) \setminus k} [1 - \text{pr}(t_l \in [t_{(q)}, t_{(q)} + \delta) \mid z_l, c_l \geq t, t_l \geq t)]} \\
& \lim_{\delta \to \infty} \text{pr}(t_j \in [t_{(q)}, t_{(q)} + \delta) \mid z_j, c_j \geq t, t_j \geq t; \beta, \Lambda_0) = \lim_{\delta \to \infty} \frac{\text{pr}(t_j \in [t_{(q)}, t_{(q)} + \delta), c_j \geq t_j \mid z_j; \beta, \Lambda_0)}{\text{pr}(c_j \geq t, t_j \geq t \mid z_j; \beta, \Lambda_0)} \\
= & \lim_{\delta \to \infty} \frac{\text{pr}(t_j \in [t_{(q)}, t_{(q)} + \delta) \mid z_j; \beta, \Lambda_0) \text{pr}(c_j \geq t_j \mid z_j; \beta, \Lambda_0)}{\text{pr}(t_j \geq t \mid z_j; \beta, \Lambda_0) \text{pr}(c_j \geq t \mid z_j; \beta, \Lambda_0)} = \lambda_0(t_{(q)}) e^{z_j^T \beta} \,.
\end{aligned}$$

This gives the partial likelihood:

$$\text{pr}(W_q \mid V_1, W_1, \ldots, V_i; \beta, \Lambda_0) = \frac{e^{z_j^T \beta}}{\sum_{k \in R(t_{(q)})} e^{z_k^T \beta}} \,; \quad L_{\text{partial}} = \prod_{i=1}^{m} \left(\frac{e^{z_i^T \beta}}{\sum_{k \in R(t_i)} e^{z_k^T \beta}} \right)^{\delta_i} \,.$$

Maximization of L_{partial} gives an estimate of β and $\partial L_{\text{partial}} / \partial \beta = 0$ coincides with the profile score equation of β. In contrary to profile likelihood, partial likelihood is more convenient since its estimates are the MLE. I choose partial likelihood for my new frailty model inference in this thesis.

2.1.7 Summary

Even though numerical approximation methods evaluate the likelihood in (2.3) with inevitable MC errors, these methods still yield MLE; thus the their consistency, asymptotic normality and variances, likelihood-based tests, such as the score test and the likelihood ratio test, have been well studied. And estimates have optimal estimation efficiency among all asymptotically linear estimates. Composite likelihood estimates need extra work: their inference involves solving the respective composite score equation. Since all my new models involve composite likelihood inference, I apply to the Z-estimator theorems from van der Vaart (1995) to show asymptotic normality of my estimators.

2.2 Approximate Likelihood Inference

2.2.1 Penalized Likelihood Estimation

Here I discuss methods approximating the integral in (2.3). Unlike the previous numerical approximations, these methods maximize some simpler formulas that approximate the integral and avoid intractable integrals from exact calculations.

For a dataset $(\boldsymbol{Y}, \boldsymbol{Z}) := (Y_i, Z_i)_{i=1,\ldots,m}$: Breslow and Clayton (1993), Ripatti and Palmgren (2000) assumed random intercepts or slopes following a (multivariate) normal distribution and applied the Laplace approximation. Breslow and Clayton (1993) worked with the model:

$$\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{b}) = h(\boldsymbol{Z}^T\beta + \boldsymbol{X}^T\boldsymbol{b}) =: \boldsymbol{\mu}^{\boldsymbol{b}}; \quad \boldsymbol{b} \sim N(0, D(\theta)) : \theta \text{ unknown.}$$

Assuming a mean-variance relationship by specifying a function v of the mean, for each observation i, they modeled the variance by

$$\mathrm{var}(Y_i \mid \boldsymbol{b}) = \phi a_i v(\mu_i^{\boldsymbol{b}}); \quad a_i \text{ is a known constant.}$$

Observed data quasi-likelihood (quasi-likelihood is in the next section) is

$$e^{ql(\beta,\theta)} \propto |D|^{-1/2} \int \exp\left(-\frac{1}{2\phi}\sum_{i=1}^{m} d(y_i; \mu_i^{\boldsymbol{b}}) - \frac{1}{2}\boldsymbol{b}^T D^{-1}\boldsymbol{b}\right) d\boldsymbol{b}; \quad e^{ql(\beta,\theta)} = c|D|^{-1/2} \int e^{-\kappa(\boldsymbol{b})} d\boldsymbol{b}\,,$$

where

$$d(y_i; \mu_i^{\boldsymbol{b}}) = -2\int_{y_i}^{\mu_i^{\boldsymbol{b}}} \frac{y_i - u}{a_i v(u)} du\,.$$

To use the Laplace approximation, the authors found some $\tilde{\boldsymbol{b}}$ minimizing $\kappa(\boldsymbol{b})$, i.e. $\tilde{\boldsymbol{b}}$ is the solution of

$$\kappa'(\boldsymbol{b}) = -\sum_{i=1}^{m} \frac{(y_i - \mu_i^{\boldsymbol{b}})\boldsymbol{z}_i}{\phi a_i v(\mu_i^{\boldsymbol{b}})\{h^{-1}\}'(\mu_i^{\boldsymbol{b}})} + D^{-1}\boldsymbol{b} = 0\,,$$

and evaluated $ql(\beta, \theta)$ at $\tilde{\boldsymbol{b}}$, ignoring the multiplicative constant c and terms with expectation 0:

$$ql(\beta,\theta) \approx -\frac{1}{2} log|I + \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}D| - \frac{1}{2\phi}\sum_{i=1}^{m} d(y_i; \mu_i^{\tilde{\boldsymbol{b}}}) - \frac{1}{2}\tilde{\boldsymbol{b}}^T D^{-1}\tilde{\boldsymbol{b}}\,, \tag{2.6}$$

where W is a diagonal matrix with elements $\left(\phi a_i v(\mu_i^{\tilde{\boldsymbol{b}}})[\{h^{-1}\}'(\mu_i^{\tilde{\boldsymbol{b}}})]^2\right)^{-1}$. Assuming the first term varies quite slowly as a function of the mean, (2.6) is the summation between the quasi-likelihood

and a penalty term: $\tilde{\boldsymbol{b}}^T D^{-1}\tilde{\boldsymbol{b}}/2$. This penalty term helps avoid over-fitting problem. In summary, predictions of $\boldsymbol{b}$ and inference of β are derived from minimizing (2.6), and β inference profiling on $\boldsymbol{b}$ is the GLM estimates, offset by $\boldsymbol{b}$ predictions.

Ripatti and Palmgren (2000) generalized this method into survival data with log-normal frailties. The log-likelihood contribution approximates

$$\begin{aligned} l(\lambda_0(t),\beta,\theta) \quad \approx \quad & -\frac{1}{2}\log|D(\theta)| - \frac{1}{2}\log|\sum_{i=1}^{m}\Lambda_0(t)\exp\left(Z_i^T\beta + X_i^T\tilde{\boldsymbol{b}}\right)Z_iZ_i^T - D(\theta)^{-1}| \\ & + \sum_{i=1}^{m}\Delta_i[\log(\lambda_0(t) + Z_i^T\beta + X_i^T\tilde{\boldsymbol{b}})] - \Lambda_0(t)\exp\left(Z_i^T\beta + X_i^T\tilde{\boldsymbol{b}}\right) - \frac{1}{2}\tilde{\boldsymbol{b}}^T D(\theta)^{-1}\tilde{\boldsymbol{b}}\,. \end{aligned}$$

Summation of the last three terms gives the log-likelihood contribution for a Cox model offset by known frailties, and a penalty term. To estimate β, assuming the changing rate of the first two terms are negligible compared to the last three, the authors maximized the last three terms, which is equivalent to maximizing the partial likelihood with a penalty term:

$$\sum_{i=1}^{n}\Delta_i\left((Z_i^T\beta + X_i^T\boldsymbol{b}) - \log\sum_{j\in R(t_i)}\exp\left(Z_i^T\beta + X_i^T\boldsymbol{b}\right)\right) - \frac{1}{2}\boldsymbol{b}^T D(\theta)^{-1}\boldsymbol{b}\,.$$

However, all the relevant methods in this sections requires the frailty density function to be in an explicit form so that the approximation step can be started. In my new models, frailties are multivariate Gamma distributed and there is no closed form so I have to give up exploring this method.

2.2.2 Restricted Maximum Likelihood Estimation (REML)

Consider m i.i.d. p-variate normal random vectors with unknown mean and variance, and one wanted to estimate the variance parameter. Its MLE is biased due to non-orthogonality between mean and variance parameters. Even though the MLE of variance parameters is consistent, researchers invented REML for removal or shrinking this bias in finite samples.

In linear mixed effects models, REML decomposes the outcome into two parts; mean and variance estimations are carried out separately using different outcome parts. The projection of outcome $\boldsymbol{Y}$ onto the column space of $\boldsymbol{Z}$ is used to estimate mean regression parameters, and the residual part is applied to the variance component estimation, as discussed by McCulloch et al. (2008), Patterson and Thompson (1971). In this case, REML is equivalent to the profile likelihood correction

proposed by Cox and Reid (1987). Breslow and Clayton (1993) extended REML into generalized mixed effects models by the approximation in (1.4). To be specific, they replaced deviance of the quasi-likelihood by the sum of squared Pearson Residuals $\sum_{i=1}^{m}(y_i - \mu_i^{\boldsymbol{b}})^2/a_i v(\mu_i^{\boldsymbol{b}})$ and used the working outcome $\boldsymbol{Y}^*$ derived from (1.4):

$$ql(\hat{\beta}(\theta),\theta) \approx -\frac{1}{2}log|V| - \frac{1}{2}(\boldsymbol{Y}^* - \boldsymbol{Z}^T\hat{\beta})^T V^{-1}(\boldsymbol{Y}^* - \boldsymbol{Z}^T\hat{\beta}) \ , \quad V = W + \boldsymbol{X}^T D(\theta)\boldsymbol{X} \ ,$$

Using REML, they estimated variance component θ by maximizing:

$$ql(\hat{\beta}(\theta),\theta) \approx -\frac{1}{2}log|V| - \frac{1}{2}log|\boldsymbol{Z}^T V^{-1}\boldsymbol{Z}| - \frac{1}{2}(\boldsymbol{Y}^* - \boldsymbol{Z}^T\hat{\beta})^T V^{-1}(\boldsymbol{Y}^* - \boldsymbol{Z}^T\hat{\beta}) \ .$$

θ is estimated by solving its score equation.

With non-identity link, REML estimates of variance terms are still biased, but since non-orthogonality is taken into account, REML estimates should be less biased than the MLE. In my new models, I standardize the variation of frailties to be one for model marginalizability; thus I do not need to consider REML.

2.2.3 Summary

This section discusses approximate likelihood inference methods which do not maximize any real likelihood; consequently no theorems are directly applicable and asymptotic properties should be studied case-by-case.

These methods can be computed much faster than the numerical methods from the previous section but tend to give more biased estimates. Hartzel et al. (2001) suggested using these methods to provide starting values for numerical approximations.

2.3 Estimating Equation: A Simplified Version of Score Functions

Most likelihood-based inference methods are not robust to model mis-specification. Researchers also focused on estimating equation inference over the years, which gives consistent estimates of parameters as long as part of the model is correct. Estimating equations discussed in this section are derived from some approximate score equation, causing small estimation efficiency loss compared to MLE. For binary and ordinal inferences, I adopt to the estimating equations, which are also

approximations of the real score equations. But later I show these methods estimation efficiency is just a little bit lower than MLE.

2.3.1 Quasi-Likelihood Estimating Equation for Independent Data

For independent data $(Y_i, Z_i)_{i=1,\ldots,m}$ with a mean model $\mathrm{E}[Y_i \mid Z_i] = \mu_i := h(Z_i^T\beta)$, Wedderburn (1974) proposed the quasi-likelihood:

$$\sum_{i=1}^{m} K_i(y_i) := \sum_{i=1}^{m} \int_{y_i}^{\mu_i} \frac{y_i - \mu_i'}{\phi v(\mu_i')} d\mu_i' \,,$$

assuming $\mathrm{var}(y_i) = \phi v(\mu_i)$. For model inference, similar to MLE, Wedderburn (1974) proposed the quasi-score equation for β:

$$\sum_{i=1}^{m} \frac{\partial K_i(y_i)}{\partial \beta} = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T \frac{\partial K_i(y_i)}{\partial \mu_i} = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T \frac{y_i - \mu_i}{\phi v(\mu_i)} = 0 \,.$$

For outcomes following some exponential family distribution, this estimating equation corresponds to the real score equation. And compared to MLE, quasi-likelihood inference has two degrees from freedom. First, only the mean model needs to be specified; i.e., the first moment of outcomes. Second, the mean-to-variance relationship only needs to be specified up to a constant, according to McCulloch et al. (2008).

Interestingly, the above estimating equation can also be derived from minimizing the following divergence to find $\hat{\beta}$:

$$(\boldsymbol{Y} - \boldsymbol{\mu}(\beta))^T V^{-1}(\beta)(\boldsymbol{Y} - \boldsymbol{\mu}(\beta))/\phi, \quad \text{where } \boldsymbol{\mu}(\beta) = \mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{Z}), \quad V(\beta) = \mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{Z}) \,.$$

Differentiating this divergence with regards to β for minimization and pretending V^{-1} not a function of β give

$$-2D^T(\beta)V^{-1}(\beta)(\boldsymbol{Y} - \boldsymbol{\mu}(\beta))/\phi = 0, \quad \text{where } D(\beta) = \frac{\partial \boldsymbol{\mu}(\beta)}{\partial \beta} \,.$$

2.3.2 Estimating Equation Inference on Marginal Means for Clustered Binary Data

For correlated data, solely interested in marginal means $\boldsymbol{\mu} := \mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{Z}] := h(\boldsymbol{Z}^T\beta)$, Liang and Zeger (1986) extended the quasi-score equation into GEE in (1.1). Correlation structures are incorporated to increase estimation efficiency.

2.3.3 Estimating Equations on Marginal Means and Correlations for Clustered Binary Data

Researchers discussed model inference on marginal means as well as correlations, covariances, or odds ratios from the quadratic exponential model by Zhao and Prentice (1990). For the i^{th} cluster, the quadratic exponential log-likelihood of the clustered binary outcome $\boldsymbol{y}_i$ is

$$l_i := \log\left[\mathrm{pr}(\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i}))\right] = \boldsymbol{y}_i^T\boldsymbol{\theta}_i + \boldsymbol{v}_i^T\boldsymbol{\lambda}_i + c(\boldsymbol{y}_i) - \log(\Delta_i)\,, \quad \boldsymbol{v}_i = (y_{i1}y_{i2}, y_{i1}y_{i3}, \ldots, y_{i,n_i-1}y_{i,n_i})\,, \tag{2.7}$$

where $c(\boldsymbol{y}_i)$ is assumed to contain no parameter under interest. They denoted the centered pairwise products as $\boldsymbol{s}_i = (s_{i12}, \ldots, s_{i,n_i-1,n_i})$ where $s_{ijk} = (y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})$: $j < k$. Their marginal mean models are $\mathrm{E}(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i) := \boldsymbol{\mu}_i = h(\boldsymbol{Z}_i^T\beta)$ and $\mathrm{E}(\boldsymbol{s}_i \mid \boldsymbol{Z}_i) := \boldsymbol{\sigma}_i = h_1(\boldsymbol{Z}_i; \beta, \alpha)$. The authors derived a set of score equations for parameters (β, α) from (2.7):

$$\sum_{i=1}^{m} \begin{pmatrix} \partial h(\boldsymbol{Z}_i^T\beta)/\partial\beta & 0 \\ \partial h_1(\boldsymbol{Z}_i; \beta, \alpha)/\partial\beta & \partial h_1(\boldsymbol{Z}_i; \beta, \alpha)/\partial\alpha \end{pmatrix}^T \begin{pmatrix} \mathrm{cov}(\boldsymbol{y}_i) & \mathrm{cov}(\boldsymbol{y}_i, \boldsymbol{s}_i) \\ \mathrm{cov}(\boldsymbol{s}_i, \boldsymbol{y}_i) & \mathrm{cov}(\boldsymbol{s}_i) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y}_i - \boldsymbol{\mu}_i \\ \boldsymbol{s}_i - \boldsymbol{\sigma}_i \end{pmatrix} = \mathbf{0}\,. \tag{2.8}$$

These estimating equations need the third and fourth orders of outcome means, which are left intractable in $c(\boldsymbol{y}_i)$. Similar to GEE, the authors used working formula for $\mathrm{cov}(\boldsymbol{s}_i, \boldsymbol{y}_i)$ and $\mathrm{cov}(\boldsymbol{s}_i)$. Likewise, Gray and Ron (2000) considered modeling the marginal means and marginal correlations.

It gives relatively simple inference methods for regression parameters based on the saturated log-linear model in (1.2), and the estimating equations are close enough to the real score equations, resulting in relatively high estimation efficiency. But this method is not impeccable. First, not many correlation structures can be introduced conveniently by modeling marginal correlations or covariances. Second, simultaneously solving for estimating equations of (β, α) in (2.8) not only adds computation burden, but also reduces robustness. When only the marginal mean model is correctly specified, estimates are not consistent. Third, the above methods do not take the Fréchet bound into account. Fourth, even if the true covariance matrix of outcomes and centered pairwise products are plugged in, estimating equations in GEE and (2.8) are not the true score equation of (1.2), since $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ are not orthogonal to the other parameters.

The first problem, i.e. relatively rigid correlation structures, arises from only partially specifying a model. Full specification of higher-order canonical parameters gives flexibility; however, to specify these parameters, a balanced design is needed and the parameter number increases as the

cluster size increases; i.e. study replication is not convenient.

Regarding the second problem, Prentice (1988) set $\partial h_1(\boldsymbol{Z}_i;\beta,\alpha)/\partial\beta = 0$ and the covariances between outcome means and pairwise products as zero, named GEE1 by Heagerty and Zeger (1996). This simplification is equivalent to solving for β and α estimating equations iteratively and is robust: when only the marginal mean model is correct, β is consistently estimated. Zhao and Prentice (1990) pointed out GEE1 has small estimation efficiency loss but saves a lot of computation time.

Concerned with the third problem, Fitzmaurice and Laird (1993) estimated the marginal means and the conditional log odds ratios λ_{jk}, $j < k$, $j,k = 1\ldots,n$, which are unbounded. However, their model is restricted to studies with a balanced design.

For the last two problems, Heagerty and Zeger (1996) modeled marginal means and marginal pairwise odds ratios: $(\boldsymbol{\mu},\boldsymbol{\nu})$, and proposed a set of estimating equations in the same form as (2.8). Odds ratios $\boldsymbol{\nu}$ are unbounded and the authors pointed out $(\boldsymbol{\mu},\boldsymbol{\nu})$ are orthogonal to the other nuisance parameters; thus their estimating equations in theory are the real score equations and achieve the optimal estimation efficiency. Yet still the working weighting matrix was used since higher-order canonical parameters are un-specified. For robustness, the authors considered GEE1 a better choice.

To model pairwise odds ratios, Carey et al. (1993) developed the alternating logistic regression. Marginal mean regression parameter β inference is the same as GEE 1. To estimate the marginal odds ratio α, they used this relationship:

$$\operatorname{logit}\operatorname{pr}(Y_{ij}=1 \mid Y_{ik}=y_{ik},Z_{ij},Z_{ik}) = \alpha y_{ik} + \log\left(\frac{\mu_{ij}-\mu_{ijk}}{1-\mu_{ij}-\mu_{ik}-\mu_{ijk}}\right), \quad j\neq k\,,$$

where $\mu_{ij} = \operatorname{pr}(Y_{ij}=1 \mid Z_{ij})$ and $\mu_{ijk} = \operatorname{pr}(Z_{ij}=1, Z_{ik}=1 \mid Z_{ij},Z_{ik})$. The authors adopted the GLM regression and the above rightmost term is an offset. This method and the work in Heagerty and Zeger (1996) are quite similar. The working covariance matrix of pairwise odds ratio here was set up as independent, with diagonal elements being proportional to the variance of pairwise odds ratios.

Kuk (2007) generalized Carey's method into a hybrid inference: the marginal mean regression parameter inference remains the same while for correlation parameter inference he maximized a composite likelihood, which is the product of pairwise likelihoods.

These two methods both iterate between solving for marginal mean regression parameters and correlation parameters until convergence.

2.3.4 Estimating Equations for Ordinal Data

For ordinal data, most researchers adopted the binarization idea in Dale (1986) and generalized binary estimating equations for inference. Heagerty and Zeger (1996) and Heagerty and Zeger (1998) give good examples.

2.3.5 Summary

The above methods end up in solving one or several estimating equations, generating Z-estimators. van der Vaart (1995) proposed respective theorems regarding asymptotic normality. To prove the consistency of estimates, in this thesis I show the estimating equations form a P_0-Glivenko-Cantelli function class in the parameter space and sample space, where P_0 is the underlying model generating the data.

2.4 Bayesian Framework: the Gibbs Sampler

This thesis does not discuss anything Bayesian, i.e., every parameter is viewed as a constant. I write the following Bayesian part for potential future research topics.

2.4.1 Overview of the Gibbs Sampler

Gibbs sampler is a Monte Carlo method. Suppose the complicated joint distribution $f(U, V, W)$ is under interested, and the conditionals: $f(U \mid V, W)$, $f(V \mid U, W)$, $f(W \mid U, V)$ are in simpler forms. In Gibbs sampler, one started with $(U^{(0)}, V^{(0)}, W^{(0)})$ and updated them with regards to the respective conditionals: sampling $U^{(i+1)}$ from $f(U \mid V^{(i)}, W^{(i)})$, sample $V^{(i+1)}$ from $f(V \mid U^{(i+1)}, W^{(i)})$, and sample $W^{(i+1)}$ from $f(W \mid U^{(i+1)}, V^{(i+1)})$. Geman and Geman (1993) showed the joint distribution of $(U^{(B)}, V^{(B)}, W^{(B)})$ converges in an exponential rate to the joint distribution of (U, V, W) as $B \to \infty$, giving an empirical estimate of the joint distribution $(U^{(k)}, V^{(k)}, W^{(k)})_{k=B+1}^{k=B+M}$ in which B and M are sufficiently large. Rounds 1 through B are burn-in rounds and discarded for concern of unstable draws.

With a minor modification, lower dimensional marginal distributions are estimable via

$$\hat{f}(U) = \frac{1}{M} \sum_{k=B+1}^{B+M} f(U \mid V^{(k)}, W^{(k)}) \, . \tag{2.9}$$

Higher order marginals such as $f(U, V)$ can be derived from:

$$\hat{f}(U, V) = \frac{1}{M} \sum_{k=B+1}^{B+M} f(V \mid U^{(k)}) \; ;$$

$(U^{(k)}, V^{(k)})_{k=B+1}^{k=B+M}$ gives an alternative solution.

2.4.2 Example: A Mixed Effects Model

Zeger and Karim (1991) cast a mixed effects model into the Bayesian framework by using the Gibbs sampler to approximate (2.3). They specified a conditional mean model as $h(Z_{ij}^T \beta + X_{ij}^T b_i)$ and the corresponding conditional distribution. They also assumed Gaussian random effects with covariance matrix D, and under a Bayesian framework, prior densities $f(\beta)$ and $f(D)$ were also assumed. Their objective is to derive posterior distribution $f(\beta, D \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m)$ from the data:

$$f(\beta, D \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m) = \frac{\prod_{i=1}^{m} \int f(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i, b_i, \beta) f(b_i \mid D) f(\beta) f(D) db_i}{\int \prod_{i=1}^{m} \int f(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i, b_i, \beta) g(b_i \mid D) f(\beta) f(D) db_i d\beta dD} \tag{2.10}$$

whose nominators and denominators are both hard to evaluate. The authors adopted the Gibbs sampler.

Chapter 3

MARGINALIZABLE MIXED EFFECTS MODEL FOR BINARY DATA

3.1 Overview

In this chapter, I first give the motivation of the new marginalizable mixed effects model. Then I introduce the multivariate exponential random variables, which serve as random effects in the new models. I present the new model formulation for binary data, the inference procedure followed by a brief discussion followed by a brief discussion, as well as relevant asymptotic theorems of my estimators.

3.2 Motivated by Frailty Models

A close examination of frailty models motivated this set of new marginalizable mixed effects models. In a frailty model, the j^{th} observation from the i^{th} cluster, conditioning on the latent frailty $W_{ij} = w_{ij}$, has the hazard rate function $w_{ij}\lambda_0(t)\exp\left(z_{ij}^T\beta\right)$, where $\lambda_0(\cdot)$ is some unspecified positive baseline conditional hazard rate function. Frailty models are mixed effects models where logarithms of frailties are random intercepts and has the complementary-log-log conditional link.

Usually frailties are assumed to follow a Gamma distribution with the density function $f_\gamma(W) = \gamma^\gamma W^{\gamma-1}e^{-\gamma W}/\Gamma(\gamma)$, γ unknown. See Clayton (1978b), Oakes (1982), Hougaard (1984), Vaida and Xu (2000) and Klein (1992) for relevant discussions. The conditional survival probability at a time point t is

$$S(t \mid Z_{ij} = z_{ij}, W_{ij} = w_{ij}) = \exp\left(-w_{ij}\Lambda_0(t)e^{z_{ij}^T\beta}\right), \quad \text{where } \Lambda_0(t) := \int_0^t \lambda_0(s)ds. \tag{3.1}$$

Integrating over W_{ij} gives the marginal survival probability at time t

$$S(t \mid Z_{ij} = z_{ij}) = \int_0^\infty \exp\left(-w_{ij}\Lambda_0(t)e^{z_{ij}^T\beta}\right)\frac{\gamma^\gamma}{\Gamma(\gamma)}w_{ij}^{\gamma-1}e^{-\gamma w_{ij}}dw_{ij} = \left(\frac{1}{1+\Lambda_0(t)e^{z_{ij}^T\beta-\log\gamma}}\right)^\gamma.$$

Fixing $\gamma = 1$, frailties are standard exponential random variables and the marginal survival proba-

bility simplifies into

$$S(t \mid Z_{ij} = z_{ij}) = \frac{1}{1 + \Lambda_0(t) e^{z_{ij}^T \beta}} , \quad \text{i.e. } \frac{1 - S(t \mid Z_{ij} = z_{ij})}{S(t \mid Z_{ij} = z_{ij})} = \Lambda_0(t) e^{z_{ij}^T \beta} .$$

β is also the marginal log failure odds ratio.

3.3 Correlated Random Effects

To flexibly model correlations within a cluster, for all new marginalizable mixed effects models, I include observation-specific latent cluster effects in the form of random intercepts. Exponentiated random intercepts, i.e. individual frailties, follow a multivariate exponential distribution with variance one. However, Furman and Landsman (2005) found there is generally no analytic form of the multivariate Gamma random variable density function, making model inference very hard. In the following I introduce a subclass of multivariate Gamma random variables that have an analytic form of Laplace transformation, which, as shown later, is critical to generate the likelihood of my models.

3.3.1 Correlated Random Effect Generation

Generation of a multivariate exponential random vector from a set of independently and identically distributed (i.i.d.) multivariate normal random vectors was discussed by Krishnamoorthy and Parthasarathy (1951), and Henderson and Shimakura (2003). I denote $V_1, V_2 \in \mathbb{R}^p$ to be two i.i.d. mean zero p-variate normal random vectors, each is written as $V_j = (V_{j1}, \ldots, V_{jp})$, $j = 1, 2$, with a $p \times p$ covariance matrix Γ having one's on the diagonal line. Setting $W_d = (V_{1d}^2 + V_{2d}^2)/2$, $d = 1, \ldots, p$, then every W_d is a standard exponential random variable. The correlation matrix R of the random vector $(W_1, \ldots, W_p)$ is an element-wise square of Γ, according to Henderson and Shimakura (2003). This multivariate exponential distribution can accommodate highly flexible positive correlation structures.

3.3.2 Flexible Correlation Structure

I parametrize R by ρ, which can be a vector-valued parameter and rewrite it as $R(\rho)$. In cases of the exchangeable correlation structure and the auto-regressive with order one correlation structure, i.e.

AR(1), the respective correlation matrices are

$$R(\rho) = \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \vdots & \ddots & \dots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{pmatrix}, \quad R(\rho) = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{p-1} \\ \rho & 1 & \rho & \dots & \rho^{p-2} \\ \vdots & \vdots & \ddots & \dots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \dots & 1 \end{pmatrix}.$$

More complicated correlation structures are also compatible. One example is modeling clustered data with multiple levels and I explicitly discuss it in the next section. Another application evolves datasets containing sophisticated correlation structures. In the adoption study from Nielsen et al. (1992), each cluster contains the adopted child, the adopting parents and the biological parents. There are two kinds of correlations: the first one involves shared environmental factors among adopted parents and the child; the second one is caused by the shared genetic factors among biological parents and the child. The correlation matrix $R(\rho)$ between the child, the adopting parents and the biological parents (not living together) is

$$R(\rho) = \begin{pmatrix} 1 & \rho_1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & 0 & 0 \\ \rho_1 & \rho_1 & 1 & 0 & 0 \\ \rho_2 & 0 & 0 & 1 & 0 \\ \rho_3 & 0 & 0 & 0 & 1 \end{pmatrix},$$

where ρ_1 is the correlation parameter of shared environmental factors, ρ_2 for shared genetic materials between the biological mother and the child, and ρ_3 for shared genetic materials between the biological father and the child.

3.3.3 Laplace Transformation of Multivariate Exponential Random Variables

Henderson and Shimakura (2003) found the Laplace transformation of the multivariate exponential random variable $\boldsymbol{W}$ as

$$\mathscr{L}_{\boldsymbol{W}}(\boldsymbol{u}) := \mathrm{E}_{\boldsymbol{W}}\left(e^{-\boldsymbol{u}^T\boldsymbol{W}}\right) = |I + \Gamma \mathrm{diag}(\boldsymbol{u})|^{-1}.$$

This quantity is critical for inference since the new models use the complementary-log-log conditional link and the data likelihood is in the form of the Laplace transformation.

3.4 Model Formulation

3.4.1 Generalization from the Frailty Model

In the absence of censoring and if the survival probability at a certain time point t^* is under interest, clustered survival outcomes are transformed into clustered binary outcomes: $Y_{ij} = 1\{T_{ij} > t^*\}$, where T_{ij} a survival outcome. The conditional model in (3.1) transforms into

$$\text{pr}(Y_{ij} = 1 \mid Z_{ij} = z_{ij}, w_{ij}) = \exp\left(-w_{ij}e^{-z_{ij}^T\beta}\right) , \tag{3.2}$$

where W_{ij} is a standard exponential random variable and $\boldsymbol{W}_i = (W_{ij})_{j=1,\ldots,n_i}$ is a standard multivariate exponential random variable. Assuming given the vector of frailties $\boldsymbol{W}_i$, the covariates are non-informative and independent of frailties, and the outcomes in cluster i are independent Bernoulli random variables:

$$\text{pr}(\boldsymbol{Y}_i = 1 \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, \boldsymbol{W}_i = \boldsymbol{w}_i) = \prod_{j=1}^{n_i}\text{pr}(Y_{ij} = 1 \mid Z_{ij} = z_{ij}, w_{ij}) = \prod_{j=1}^{n_i}\exp\left(-w_{ij}e^{-z_{ij}^T\beta}\right) . \tag{3.3}$$

In this formulation an intercept is included into the covariate vector Z_{ij}, corresponding to $\log\{\Lambda_0(t^*)\}$ in (3.1).

The marginal survival probability at time t^* is

$$\text{pr}(Y_{ij} = 1 \mid Z_{ij} = z_{ij}) = \frac{e^{z_{ij}^T\beta}}{1 + e^{z_{ij}^T\beta}} . \tag{3.4}$$

Therefore the marginal model is the logistic model with the same β coefficients as in the working conditional model (3.2). I described the conditional model as a working model, because in the following section I proposed a robust estimator of β from the marginal model (3.4), which stays consistent even when the working conditional model in (3.2) is mis-specified.

3.4.2 Model Interpretation

Marginally, parameter β represents the log odds ratio of survival at time t^* with regards to one unit change in the corresponding covariate.

I rewrite the conditional model (3.2) as

$$\text{pr}(Y_{ij} = 1 \mid Z_{ij} = z_{ij}, W_{ij} = w_{ij}) = \exp\left(-w_{ij}e^{-z_{ij}^T\beta}\right) = \exp\left(-e^{b_{ij}-z_{ij}^T\beta}\right) ,$$

where b_{ij} is a random intercept following the reverse standard extreme value (*a.k.a.* Gumbel) distribution. I denote a latent outcome $y^*_{ij} := -z^T_{ij}\beta + b_{ij} + \epsilon_{ij}$, where ϵ_{ij} is a Gumbel distributed random variable. I dichotomize the latent outcome Y^*_{ij} by $Y_{ij} := 1\{Y^*_{ij} \leq 0\}$, then

$$\text{pr}(Y_{ij} = 1 \mid Z_{ij} = z_{ij}, b_{ij}) = \text{pr}(Y^*_{ij} \leq 0) = \exp\left(-e^{-z^T_{ij}\beta + b_{ij}}\right) .$$

Therefore, conditionally, β is interpreted as the covariate effects on the mean of a latent outcome which follows an extreme value distribution.

3.4.3 Generalization into Three-level Clustered Data

For simplicity, my earlier discussions focus on two-level clustered data. Since the proposed model allows for flexible modeling of correlations among observations, it can be readily extended to datasets with higher levels of clustering. Here I discuss a three-level clustered dataset where the first level consists of multiple independent institutions, inside every institution there are multiple clusters representing the second level, and multiple individuals observed in each cluster form the third level. For example, in the Television, School and Family Smoking Prevention and Cessation Project (TVSFP), the first level represents schools participating in this project and from each school, multiple classes are sampled, serving as the second level, and students drawn from multiple classes form the third level. Data from the i^{th} school are denoted by $(\boldsymbol{Y}_i, \boldsymbol{Z}_i) = \text{vec}(Y_{ijk}, Z_{ijk}) : j = 1, \ldots, n_i$ indexes classes from the i^{th} school and $k = 1, \ldots, n_{ij}$ counts its students.

I assume a similar working conditional model:

$$\text{pr}(Y_{ijk} = 1 \mid Z_{ijk} = z_{ijk}, w_{ijk}) = \exp\left(-w_{ijk}e^{-z^T_{ijk}\beta}\right), \quad w_{ijk} \sim \text{Exp}(1) .$$

It is easy to show that the marginalization property of the working model still holds in this case:

$$\text{pr}(Y_{ijk} = 1 \mid Z_{ijk} = z_{ijk}) = \frac{1}{1 + e^{-z^T_{ijk}\beta}} .$$

One way to model correlations among W_{ijk}'s is to assume the classes are exchangeably correlated, and students from every class are also exchangeably correlated; i.e.

$$\text{cor}(W_{ijk}, W_{ij'k'}) = \rho_2 , \quad j \neq j' \tag{3.5}$$

$$\text{cor}(W_{ijk}, W_{ij'k'}) = \rho_2 + \rho_3 , \quad j = j' , \quad k \neq k' . \tag{3.6}$$

3.5 Model Inference

3.5.1 Estimating Equation for β with An Optimal Weighting Matrix.

I denote the whole set of parameters by $\theta := (\beta, \rho)$ and denote h as the inverse of the logit link function:

$$h(z_{ij}^T\beta) := \text{pr}(Y_{ij} = 1 \mid Z_{ij} = z_{ij}) = \exp(z_{ij}^T\beta)/\left\{1 + \exp(z_{ij}^T\beta)\right\} \ .$$

For marginal parameter β inference, I fix the correlation parameter ρ as a constant and derive a set of robust estimating equations from the quadratic exponential distribution of $(\boldsymbol{Y}, \boldsymbol{Z})$, originally proposed by Zhao and Prentice (1990). Making the same simplifications as in Heagerty and Zeger (1996), the corresponding quadratic exponential score function of β based on the i^{th} cluster is:

$$D(\boldsymbol{z}_i;\beta)^T V^{-1}(\boldsymbol{z}_i;\theta) S(\boldsymbol{z}_i, \boldsymbol{y}_i;\beta) \ .$$

where $D(\boldsymbol{z}_i;\beta) = \partial h(\boldsymbol{z}_i^T\beta)/\partial\beta$, $S(\boldsymbol{z}_i, \boldsymbol{y}_i;\beta) = \boldsymbol{y}_i - h(\boldsymbol{z}_i^T\beta)$ and $V(\boldsymbol{z}_i;\theta)$ is the $n_i \times n_i$ covariance matrix of the outcome $\boldsymbol{Y}_i$. To be more specific, the j^{th} diagonal entry of $V(\boldsymbol{z}_i;\theta)$ is

$$V_{jj}(\boldsymbol{z}_i;\theta) = \frac{e^{z_{ij}^T\beta}}{\left(1 + e^{z_{ij}^T\beta}\right)^2} \ .$$

Its j^{th} row and k^{th} column entry is

$$V_{jk}(\boldsymbol{z}_i;\theta) = \left[\frac{1}{(1-\rho_{jk})e^{-(z_{ij}+z_{ik})^T\beta} + e^{-z_{ij}^T\beta} + e^{-z_{ik}^T\beta} + 1} - \frac{1}{1+e^{-z_{ij}^T\beta}}\frac{1}{1+e^{-z_{ik}^T\beta}}\right] \ , \quad j \neq k \, ,$$

where ρ_{jk} is the correlation between W_{ij} and W_{ik}. In the case of an exchangeable correlation structure, ρ_{jk} identically equals to a scalar ρ. In the case of an AR(1) correlation structure, ρ_{jk} is a function of a scalar parameter ρ. In more general correlation structures, such as the un-structured correlation structure, ρ_{jk} are functions of some vector-valued parameter ρ.

Given a dataset of m independent clusters, and fixing the correlation parameter ρ as a constant, I solve for β from the equation

$$\sum_{i=1}^{m} D(\boldsymbol{z}_i;\beta)^T V^{-1}(\boldsymbol{z}_i;\theta) S(\boldsymbol{z}_i, \boldsymbol{y}_i;\beta) = 0 \ . \tag{3.7}$$

3.5.2 *Estimating ρ via the Composite Likelihood.*

I choose to maximize a composite log-likelihood function over ρ with a fixed β. I define the composite log-likelihood for each cluster as the summation of all pairwise log-likelihoods. I denote

$$\begin{aligned} p_{ij} &= \text{pr}(Y_{ij}=1 \mid Z_{ij}=z_{ij}) = h(z_{ij}^T\beta)\,, \\ p_{ijk} &= \text{pr}(Y_{ij}=Y_{ik}=1 \mid Z_{ij}=z_{ij}, Z_{ik}=z_{ik}) = \left[(1-\rho_{jk})e^{-(z_{ij}+z_{ik})^T\beta} + e^{-z_{ij}^T\beta} + e^{-z_{ik}^T\beta} + 1\right]^{-1}\,. \end{aligned}$$

For a dataset containing m independent clusters, the empirical composite log-likelihood is

$$\begin{aligned} & \sum_{i=1}^{m}\sum_{j<k} l_{jk}(\boldsymbol{z}_i, \boldsymbol{y}_i; \theta) \\ = & \sum_{i=1}^{m}\sum_{j<k} \Big\{ y_{ij}y_{ik}\text{log}p_{ijk} + (1-y_{ij})y_{ik}\text{log}(p_{ik}-p_{ijk}) \\ & \qquad + y_{ij}(1-y_{ik})\text{log}(p_{ij}-p_{ijk}) + (1-y_{ij})(1-y_{ik})\text{log}(1-p_{ij}-p_{ik}+p_{ijk}) \Big\} \end{aligned} \tag{3.8}$$

To estimate ρ, I solve for ρ from the composite score equation, fixing β as a constant,

$$\sum_{i=1}^{m}\sum_{j<k} \frac{\partial l_{jk}}{\partial \rho}(\boldsymbol{z}_i, \boldsymbol{y}_i; \theta) = 0\,. \tag{3.9}$$

3.5.3 *Overall Inference Procedure*

To estimate β and ρ jointly, Kuk (2007) suggested alternating between solving (3.7) with a fixed plug-in ρ from (3.9), and solving (3.9) with a fixed plug-in β from (3.7), until convergence, obtaining final estimates. I suggest starting with solving for (4.11) fixing $\rho = 0$, i.e. GLM inference, and then estimating ρ from (4.12).

I denote the final estimate as $(\hat{\beta}_m, \hat{\rho}_m)$ where m indicates the estimate is based on a dataset containing m independent clusters. This method is a generalization of alternating logistic regression proposed by Carey et al. (1993).

3.5.4 *Generalization to Three-Level of Clustering*

This robust inference procedure can be generalized into this case. I denote $N_i = \sum_{j=1}^{n_i} n_{ij}$ as the total number of observations from cluster i. For notational simplicity, I concatenate level-two observations in the cluster and denote $(\boldsymbol{Y}_i, \boldsymbol{Z}_i) = \{\text{vec}(Y_{is}, Z_{is}) : s = 1, \ldots, N_i\}$; i.e., I merge the

double index jk into a single index s. I consider two distinct observations s_1, s_2 from sub-clusters j_1, j_2 in the i^{th} cluster, and I write their indexes as $\sum_{j=1}^{j_l-1} n_{ij} < s_l \leq \sum_{j=1}^{j_l} n_{ij}$, $l = 1, 2$.

Entries of the covariance matrix $V(\boldsymbol{Z}_i, \beta, \rho)$ are given by:

$$V_{s_1 s_1}(\boldsymbol{Z}_i = \boldsymbol{z}_i; \beta, \rho) = \frac{e^{-z_{is_1}^T \beta}}{\left(1 + e^{-z_{is_1}^T \beta}\right)^2} ,$$

$$\begin{aligned} & V_{s_1 s_2}(\boldsymbol{Z}_i = \boldsymbol{z}_i; \beta, \rho) \\ = \quad & \frac{1}{\{1 - \text{cor}(w_{is_1}, w_{is_2})\} e^{-(z_{is_1} + z_{is_2})^T \beta} + e^{-z_{is_1}^T \beta} + e^{-z_{is_2}^T \beta} + 1} - \frac{1}{1 + e^{-z_{is_1}^T \beta}} \frac{1}{1 + e^{-z_{is_2}^T \beta}} . \end{aligned}$$

Following the exchangeable correlation formulation in (3.5) and (3.6), I write

$$\begin{aligned} & V_{s_1 s_2}(\boldsymbol{Z}_i = \boldsymbol{z}_i; \beta, \rho) \\ = \quad & \begin{cases} \left\{(1 - \rho_2 - \rho_3) e^{-(z_{is_1} + z_{is_2})^T \beta} + e^{-z_{is_1}^T \beta} + e^{-z_{is_2}^T \beta} + 1\right\}^{-1} - \left\{\left(e^{-z_{is_1}^T \beta} + 1\right)\left(e^{-z_{is_2}^T \beta} + 1\right)\right\}^{-1} , & j_1 = j_2 \\ \left\{(1 - \rho_2) e^{-(z_{is_1} + z_{is_2})^T \beta} + e^{-z_{is_1}^T \beta} + e^{-z_{is_2}^T \beta} + 1\right\}^{-1} - \left\{\left(e^{-z_{is_1}^T \beta} + 1\right)\left(e^{-z_{is_2}^T \beta} + 1\right)\right\}^{-1} , & j_1 \neq j_2 . \end{cases} \end{aligned}$$

Similar to (3.8), I write

$$\begin{aligned} & \sum_{i=1}^{m} \sum_{s_1 < s_2} l_{s_1 s_2}(\boldsymbol{Z}_i, \boldsymbol{Y}_i, \theta) \\ = \quad & \sum_{i=1}^{m} \sum_{s_1 < s_2} \{y_{is_1} y_{is_2} \log p_{is_1 s_2} + (1 - y_{is_1}) y_{is_2} \log(p_{is_2} - p_{is_1 s_2}) \\ & + y_{is_1}(1 - y_{is_2}) \log(p_{is_1} - p_{is_1 s_2}) + (1 - y_{is_1})(1 - y_{is_2}) \log(1 - p_{is_1} - p_{is_2} + p_{is_1 s_2})\} , \end{aligned}$$

where $p_{is_1} = \left(1 + e^{-z_{is_1}^T \beta}\right)^{-1}$ and

$$p_{is_1 s_2} = \begin{cases} \left\{(1 - \rho_2 - \rho_3) e^{-(z_{is_1} + z_{is_2})^T \beta} + e^{-z_{is_1}^T \beta} + e^{-z_{is_2}^T \beta} + 1\right\}^{-1} , & j_1 = j_2 , \\ \left\{(1 - \rho_2) e^{-(z_{is_1} + z_{is_2})^T \beta} + e^{-z_{is_1}^T \beta} + e^{-z_{is_2}^T \beta} + 1\right\}^{-1} , & j_1 \neq j_2 . \end{cases}$$

Similar to the case of two-level clustering, estimates are obtained by solving

$$\begin{cases} \sum_{i=1}^{m} D(\boldsymbol{z}_i; \beta) V^{-1}(\boldsymbol{z}_i; \beta, \rho) S(\boldsymbol{z}_i, \boldsymbol{y}_i; \beta) = 0 , \\ \sum_{i=1}^{m} \sum_{s_1 < s_2} \frac{\partial l_{s_1 s_2}}{\partial \rho}(\boldsymbol{z}_i, \boldsymbol{y}_i; \beta, \rho) = 0 . \end{cases}$$

Other correlation structures can also be used. For example, suppose the level-two observations are exchangeably correlated units and the level-three observations are auto-regressively correlated with

order one, then I can model

$$\begin{aligned}\mathrm{cor}(W_{is_1}, W_{is_2}) &= \rho_2\,, \quad j_1 \neq j_2\,,\\ \mathrm{cor}(W_{is_1}, W_{is_2}) &= \rho_2 + \rho_3^{|s_1-s_2|}\,, \quad j_1 = j_2\,.\end{aligned}$$

Entries in the inverse weighting matrix for estimating β can be written as

$$\begin{aligned}&V_{s_1s_2}(\boldsymbol{Z}_i = \boldsymbol{z}_i; \beta, \rho)\\ =\quad & \frac{1}{\{1-\mathrm{cor}(W_{is_1}, W_{is_2})\}\, e^{-(z_{is_1}+z_{is_2})^T\beta} + e^{-z_{is_1}^T\beta} + e^{-z_{is_2}^T\beta} + 1} - \frac{1}{1+e^{-z_{is_1}^T\beta}}\frac{1}{1+e^{-z_{is_2}^T\beta}}\,.\end{aligned}$$

I write

$$\begin{aligned}&V_{s_1s_2}(\boldsymbol{Z}_i = \boldsymbol{z}_i; \beta, \rho)\\ =\quad & \begin{cases} \left\{(1-\rho_2-\rho_3^{|s_1-s_2|})e^{-(z_{is_1}+z_{is_2})^T\beta} + e^{-z_{is_1}^T\beta} + e^{-z_{is_2}^T\beta} + 1\right\}^{-1} - \left\{\left(e^{-z_{is_1}^T\beta}+1\right)\left(e^{-z_{is_2}^T\beta}+1\right)\right\}^{-1}, & j_1 = j_2\,,\\ \left\{(1-\rho_2)e^{-(z_{is_1}+z_{is_2})^T\beta} + e^{-z_{is_1}^T\beta} + e^{-z_{is_2}^T\beta} + 1\right\}^{-1} - \left\{\left(e^{-z_{is_1}^T\beta}+1\right)\left(e^{-z_{is_2}^T\beta}+1\right)\right\}^{-1}, & j_1 \neq j_2\,; \end{cases}\end{aligned}$$

and

$$p_{is_1s_2} = \begin{cases} \left\{(1-\rho_2-\rho_3^{|s_1-s_2|})e^{-(z_{is_1}+z_{is_2})^T\beta} + e^{-z_{is_1}^T\beta} + e^{-z_{is_2}^T\beta} + 1\right\}^{-1}, & j_1 = j_2\,,\\ \left\{(1-\rho_2)e^{-(z_{is_1}+z_{is_2})^T\beta} + e^{-z_{is_1}^T\beta} + e^{-z_{is_2}^T\beta} + 1\right\}^{-1}, & j_1 \neq j_2\,. \end{cases}$$

3.5.5 Discussion of Inference Methods and Further Remarks

The estimating equation in (3.7) is a simplified version of the quadratic exponential score equation. This simplification is commonly used in a lot of marginal model inference methods due to limited model assumptions, but in this thesis I adopt to it mainly for robustness: under mis-specification of the conditional mean model or the random effect distribution, the estimating equation (3.7) guarantees consistency of the marginal parameter estimator, while the inverse weighting matrix V is still a genuine covariance matrix, but corresponds to a mis-specified model. Under correct model specification, (3.7) is close to the score equation of the quadratic exponential distribution, so compared to MLE, there is small estimation efficiency loss.

With a parametric model, it is natural to consider the maximum likelihood estimation (MLE) for model inference, as discussed by Conaway (1990) and Coull et al. (2006). However, MLE has two major drawbacks. First, obtaining consistent MLE requires a correct specification of the

whole parametric model, even when the marginal parameters are of primary interest. Besides, the likelihood function involves up to 2^n-1 terms for a cluster with n observations. It may be practically unfeasible to compute MLE even for a moderate cluster size, since the computation burden grows exponentially with cluster size.

There are also alternatives for inference on ρ; an example is the second-order GEE proposed by Zhao and Prentice (1990). However, its computational burden is in the order of $O(n^6)$. 4.3 gives a more thorough discussion of the differences and similarities between GEE1 and Kuk's methods.

The above two issues motivate me to use the composite likelihood inference for correlation parameter inference: inference by estimating equations (3.7) and (3.9) reduces the computing order to n^2 for the cluster.

In numeric studies, estimation efficiencies and computing times were compared between my inference method and MLE, under correct model assumptions as well as a mis-specified model.

3.6 Large Sample Properties

I provide several theoretical results concerning the asymptotic behavior of $(\hat{\beta}_m, \hat{\rho}_m)$.

Theorem 3.6.1. *Suppose conditions C1 $\sim$ C6 stated in Appendix A are satisfied, then when $m \to \infty$,*

(a) the solution $\hat{\theta}_m = (\hat{\beta}_m, \hat{\rho}_m)$ of equations in (3.7) *and* (3.9) *is consistent for $\theta_0 := (\beta_0, \rho_0)$;*

(b) $\sqrt{m}\left\{(\hat{\beta}_m - \beta_0)^T, (\hat{\rho}_m - \rho_0)^T\right\}^T$ converges weakly to a normal distribution of mean zero and with covariance matrix V given by

$$V = \{\mathrm{E}(B)\}^{-1}\{\mathrm{E}(C)\}\left\{\mathrm{E}(B)^T\right\}^{-1},$$

where

$$B = \begin{pmatrix} D(\boldsymbol{Z};\beta_0)^T V^{-1}(\boldsymbol{Z};\theta_0) D(\boldsymbol{Z};\beta_0) & 0 \\ -\sum_{j<k} \frac{\partial^2 l_{jk}}{\partial\beta\partial\rho}(\boldsymbol{Z},\boldsymbol{Y};\theta)\mid_{\theta_0} & -\sum_{j<k} \frac{\partial^2 l_{jk}}{\partial\rho^2}(\boldsymbol{Z},\boldsymbol{Y};\theta)\mid_{\theta_0} \end{pmatrix},$$

$$C = \begin{pmatrix} D(\boldsymbol{Z};\beta_0)^T V^{-1}(\boldsymbol{Z};\theta_0) S(\boldsymbol{Z},\boldsymbol{Y};\beta_0) \\ \sum_{j<k} \frac{\partial l_{jk}}{\partial\rho}(\boldsymbol{Z},\boldsymbol{Y};\theta)\mid_{\theta_0} \end{pmatrix}^{\otimes 2}.$$

The proof is in Appendix A .

The next theorem is for a mis-specified parametric model but a correct marginal mean model.

Theorem 3.6.2. *Suppose only the marginal mean model* (3.4) *is true, and all the other conditions in Theorem 1 are satisfied. When* $m \to \infty$*,*
(a) the solution $\hat{\theta}_m = (\hat{\beta}_m, \hat{\rho}_m)$ *of equations* (3.7) *and* (3.9) *is consistent for* (β_0, ρ_1)*, and* ρ_1 *is derived from*

$$\rho_1 = \arg\min_{\rho} KL_{\text{composite}}(L, L^*) := \arg\min_{\rho} P_0 \log \left(\frac{\prod_{j<k} L(Y_j, Y_k, Z_j, Z_k; \beta, \rho')}{\prod_{j<k} L^*(Y_j, Y_k, Z_j, Z_k; \beta, \rho)} \right) ,$$

where L *denotes the likelihood of the true pairwise joint model,* L^* *for the mis-specified one, and* ρ' *is some other parameters in the true model.*
(b) $\sqrt{m} \left\{ (\hat{\beta}_m - \beta_0)^T, (\hat{\rho}_m - \rho_1)^T \right\}^T$ *converges weakly to a normal distribution of mean zero and a covariance matrix* W *given by*

$$W = \{\mathrm{E}(B_1)\}^{-1} \{\mathrm{E}(C_1)\} \left\{\mathrm{E}(B_1)^T\right\}^{-1} ,$$

where

$$B_1 = \begin{pmatrix} D(\boldsymbol{Z}; \beta_0)^T V^{-1}(\boldsymbol{Z}; \beta_0, \rho_1) D(\boldsymbol{Z}; \beta_0) & 0 \\ -\sum_{j<k} \frac{\partial^2 l^*_{jk}}{\partial \beta \partial \rho}(\boldsymbol{Z}, \boldsymbol{Y}; \theta) \mid_{(\beta_0, \rho_1)} & -\sum_{j<k} \frac{\partial^2 l^*_{jk}}{\partial \rho^2}(\boldsymbol{Z}, \boldsymbol{Y}; \theta) \mid_{(\beta_0, \rho_1)} \end{pmatrix} ,$$

$$C_1 = \begin{pmatrix} D(\boldsymbol{Z}; \beta_0)^T V^{-1}(\boldsymbol{Z}; \beta_0, \rho_1) S(\boldsymbol{Z}, \boldsymbol{Y}; \beta_0) \\ \sum_{j<k} \frac{\partial l^*_{jk}}{\partial \rho}(\boldsymbol{Z}, \boldsymbol{Y}; \theta) \mid_{(\beta_0, \rho_1)} \end{pmatrix}^{\otimes 2} .$$

As suggested in Theorem 3.5.1, when the pairwise conditional model is correct, the asymptotic covariance of $\sqrt{m}(\hat{\beta}_m - \beta_0)$ can be estimated by

$$\hat{V}^{\beta}_m := m \left(\sum_{i=1}^{m} D(\boldsymbol{Z}_i; \hat{\beta}_m)^T V^{-1}(\boldsymbol{Z}_i; \hat{\theta}_m) D(\boldsymbol{Z}_i; \hat{\beta}_m) \right)^{-1} . \tag{3.10}$$

Allowing for a potentially mis-specified parametric joint model, a robust estimate of the asymptotic

covariance of $\sqrt{m}(\hat{\beta}_m - \beta_0)$ is given in a sandwich form:

$$\begin{aligned} \hat{V}_m^{\text{robust}} &:= m A_m^{-1} B_m A_m^{-1} \; ; \\ A_m &:= \sum_{i=1}^{m} D(\boldsymbol{z}_i; \hat{\beta}_m)^T V^{-1}(\boldsymbol{z}_i; \hat{\theta}_m) D(\boldsymbol{z}_i; \hat{\beta}_m) \; , \\ B_m &:= \sum_{i=1}^{m} \left[D(\boldsymbol{z}_i; \hat{\beta}_m)^T V^{-1}(\boldsymbol{z}_i; \hat{\theta}_m) S(\boldsymbol{z}_i, \boldsymbol{y}_i; \hat{\beta}_m) \right]^{\otimes 2} \; . \end{aligned} \tag{3.11}$$

Chapter 4

MARGINALIZABLE MIXED EFFECTS MODEL FOR ORDINAL DATA

4.1 Overview

I present the formulation of the new model on clustered ordinal data, which is a generalization of the previous model. Inference procedure is also described and discussed briefly. In the end I present asymptotic theorems.

4.2 Model Formulation

4.2.1 Generalization from the Frailty Model

In the absence of censoring and suppose survival probabilities at several pre-specified time points $0 < t_1 < \ldots < t_G < t_{G+1} = +\infty$ are under interest, clustered survival data are transformed into clustered ordinal data: $Y_{ij} := \min\{g : T_{ij} \leq t_g, g = 1, \ldots, G+1\}$. $\log\{\Lambda_0(t_g)\}$ is the category-specific intercept, and I denote it by α_g's, $g = 1, \ldots, G$. Conditional probability of surviving up to time point t_g is written as

$$S(t_g \mid Z_{ij} = z_{ij}, W_{ij} = w_{ij}) = \text{pr}(T_{ij} > t_g \mid z_{ij}, w_{ij}) = \text{pr}(Y_{ij} > g \mid z_{ij}, w_{ij}) = \exp\left(-w_{ij} e^{z_{ij}^T \beta + \alpha_g}\right) .$$

That is,

$$\text{pr}(Y_{ij} \leq g \mid Z_{ij} = z_{ij}, W_{ij} = w_{ij}) = 1 - \exp\left(-w_{ij} e^{z_{ij}^T \beta + \alpha_g}\right) \quad \text{for } g = 1, \ldots G \,, \tag{4.1}$$

corresponding to the complementary log-log link on cumulative probabilities. I assume that given the vector of frailties $\boldsymbol{W}_i$, the outcomes in cluster i are independent multinomial random variables with marginal cumulative probability odds as

$$\frac{\text{pr}(Y_{ij} \leq g \mid Z_{ij} = z_{ij})}{1 - \text{pr}(Y_{ij} \leq g \mid Z_{ij} = z_{ij})} = \exp(z_{ij}^T \beta + \alpha_g) \,, \tag{4.2}$$

with the same regression parameters as in the working conditional model: $\eta := (\alpha_1, \ldots, \alpha_G, \beta)$. Marginally β represents the log cumulative odds ratio and α are the log cumulative category-specific

baseline odds. Similar to the binary case, I describe the conditional model in (4.3) as a working model.

4.2.2 Conditional Model

In the following I work with a new conditional which is a small modification of (4.1):

$$\text{pr}(Y_{ij} \leq g \mid Z_{ij} = z_{ij}, W_{ij} = w_{ij}) = \exp\left(-w_{ij}e^{-z_{ij}^T\beta - \alpha_g}\right) \quad \text{for } g = 1, \ldots G\,, \tag{4.3}$$

And it also yields the same marginal model in (4.2).

4.2.3 Model Constraints and Assumption.

So as to have a sensible model, I put this constraint:

$$-\infty < \alpha_1 \leq \ldots \leq \alpha_G < \infty\,.$$

And both marginal and conditional models in (4.3) and (4.2) make the assumption that covariate effects on the binary variable $1\{Y_{ij} \leq g\}$ remain constant with different values of g.

4.2.4 Model Interpretation

Marginally, for each $g = 1, \ldots, G$, β represents log odds ratio of $\text{pr}(Y_{ij} \leq g \mid Z_{ij})$ w.r.t. one unit change in some corresponding covariate; α_g stands for its baseline log odds.

At the same time, $(\beta, \alpha_1, \ldots, \alpha_G)$ can be interpreted conditionally by introducing a set of latent outcomes. Let $Y_{ij}^g = -\alpha_g - z_{ij}^T\beta + b_{ij} + \epsilon_{ij}$, $g = 1, \ldots, G$, and suppose ϵ_{ij} follows the Gumbel distribution. I dichotomize the latent outcome by $1\{Y_{ij} \leq g\} = 1\{Y_{ij}^g \leq 0\}$, then

$$\text{pr}(Y_{ij} \leq g \mid z_{ij}, b_{ij}) = \text{pr}(Y_{ij}^g \leq 0 \mid z_{ij}, b_{ij}) = \text{pr}(\epsilon_{ij} \geq -\alpha_g - z_{ij}^T\beta - b_{ij}) = \exp\left(-e^{-\alpha_g - z_{ij}^T\beta + b_{ij}}\right)\,.$$

Therefore, conditionally, β is interpreted as the covariate effects on the mean of a latent outcome which follows a Gumbel distribution with the location parameter as $\alpha_g + z_{ij}^T\beta - b_{ij}$.

4.3 Ordinal Model Inference

4.3.1 Maximal Likelihood Estimation

MLE is applicable for this parametric model inference. The joint probability of the i^{th} cluster is expressed as a linear combination of probabilities having the following format:

$$\mathrm{pr}[Y_{ij} \leq g_{ij} \mid z_{ij}; j = 1, \ldots, n_i] = \mathrm{E}_{\boldsymbol{W}_i}\left[\exp\left(-\sum_{j=1}^{n_i} w_{ij}\exp\left(-z_{ij}^T\beta - \alpha_{g_{ij}}\right)\right)\right] .$$

By the Laplace transformation of multivariate exponential random variables, the above probability equals to

$$\mathrm{pr}[Y_{ij} \leq g_{ij} \mid z_{ij}; j = 1, \ldots, n_i] := \mathscr{L}(\boldsymbol{u}_i) = \mid I + \Gamma\mathrm{diag}(u_{i1}, \ldots, u_{in_i}) \mid^{-1} ,$$

where Γ is the component-wise square root of the frailty correlation matrix and $u_{ij} = \exp\left(-z_{ij}^T\beta - \alpha_{g_{ij}}\right)$. This form of joint probabilities makes MLE harder to find even with moderate number of observations. For an ordinal vector from three categories 1, 2, 3: $\boldsymbol{y}_i := (y_{i1} = 1, y_{i2} = 2, y_{i3} = 3, y_{i4} = 2)$, since there are three observations having values greater than one, I consider the following $8 = 2^3$ Boolean vectors:

$$\begin{aligned} &e_1 = (0,0,0,0), \quad e_2 = (0,1,0,0), \quad e_3 = (0,0,1,0), \quad e_4 = (0,0,0,1), \\ &e_5 = (0,0,1,1), \quad e_6 = (0,1,0,1), \quad e_7 = (0,1,1,0), \quad e_8 = (0,1,1,1) , \end{aligned}$$

and the joint probability of y_i is

$$\mathrm{pr}(\boldsymbol{y}_i = (1,2,3,2) \mid z_i) = \sum_{k=1}^{8}(-1)^{d_k}\mathrm{pr}\left(y \leq y_i - e_k \mid z_i\right), \quad d_k \text{ is the number of one's in } e_k ,$$

where " $\leq$ " and " $-$ " are component-wise operations.

MLE has two major drawbacks. First, consistent estimates require correct specification of the whole model, even when the marginal parameters are of primary interest. Calculation of the likelihood function involves computing $2^{\tilde{d}_i}$ terms for each cluster, where $\tilde{d}_i$ is the number of outcomes greater than one. Thus it is practically infeasible to apply to MLE except with small clusters.

4.3.2 Connections with Binary Outcomes

Following the binarization idea from Dale (1986) and the natural correspondence between the cumulative logit link and binary variables, I rewrite an ordinal observation (Y_{ij}, Z_{ij}) into a series of binary observations:

$$\begin{aligned}
\boldsymbol{Y}_{ij}^{*} &= (Y_{ij1}, \ldots, Y_{ijG})^{T} = (1\{Y_{ij} \leq 1\}, 1\{Y_{ij} \leq 2\}, \ldots, 1\{Y_{ij} \leq G\})^{T} , \\
\boldsymbol{Z}_{ij}^{*} &= \begin{pmatrix} Z_{ij1} \\ \vdots \\ Z_{ijG} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 & Z_{ij} \\ 0 & 1 & \ldots & 0 & Z_{ij} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & Z_{ij} \end{pmatrix} .
\end{aligned}$$

For the i^{th} cluster, its transformed observations $(\boldsymbol{Y}_{i}^{*}, \boldsymbol{Z}_{i}^{*})$ have the transformed data $(\boldsymbol{Y}_{ij}^{*}, \boldsymbol{Z}_{ij}^{*})$ of each observation put consecutively to each other, as if there were $G \times n_i$ observations.

4.3.3 Another MLE: Saturated Log-Linear Model

An alternative presentation of the likelihood is the log-linear model of the binarized data:

$$\mathrm{pr}(\boldsymbol{y}_{i}^{*}) = \exp\left(\Delta_i + \sum_{j=1}^{n_i \times G} \theta_{ij} y_{ij}^{*} + \sum_{j<k} \lambda_{ijk} y_{ij}^{*} y_{ik}^{*} + u_{i12,\ldots,n_i \times G} y_{i1}^{*} \ldots y_{i,n_i \times G}^{*} \right) , \tag{4.4}$$

where Δ_i is a normalizing factor and the other parameters (θ, λ, u) have conditional interpretations:

$$\begin{aligned}
\theta_{ij} &= \mathrm{logit}\left\{ \mathrm{pr}(y_{ij}^{*} \mid y_{ik}^{*} = 0, k \neq j) \right\} , \\
\lambda_{ijk} &= \log \mathrm{OR}\left\{ \mathrm{pr}(y_{ij}^{*}, y_{ik}^{*} \mid y_{il}^{*} = 0, l \neq j, k) \right\} , \\
u_{i123} &= \log \mathrm{OR}\left\{ \mathrm{pr}(y_{i1}^{*}, y_{i2}^{*} \mid y_{i3}^{*} = 1, y_{il}^{*} = 0, l > 3) \right\} - \log \mathrm{OR}\left\{ \mathrm{pr}(y_{i1}^{*}, y_{i2}^{*} \mid y_{i3}^{*} = 0, y_{il}^{*} = 0, l > 3) \right\} .
\end{aligned}$$

This joint distribution is very complex to work with. In the following I write down the log-linear model formula of my model to demonstrate the complexity. Suppose in the i^{th} cluster there are 2 observations coming from categories 1, 2 and 3:

$$\boldsymbol{y}_{i}^{*} = (y_{i1}^{*} = 1\{y_{i1} \leq 1\}, y_{i2}^{*} = 1\{y_{i1} \leq 2\}, y_{i3}^{*} = 1\{y_{i2} \leq 1\}, y_{i4}^{*} = 1\{y_{i2} \leq 2\}) ,$$

the saturated log-linear representation for my model is

$$\mathrm{pr}(\boldsymbol{y}_{i}^{*}) = \exp\left(u_i + u'_{i1} y_{i1}^{*} + u_{i2} y_{i2}^{*} + u'_{i3} y_{i3}^{*} + u_{i4} y_{i4}^{*} + u'_{i13} y_{i1}^{*} y_{i3}^{*} + u'_{i14} y_{i1}^{*} y_{i4}^{*} + u'_{i23} y_{i2}^{*} y_{i3}^{*} + u'_{i24} y_{i2}^{*} y_{i4}^{*} \right) . \tag{4.5}$$

Due to the inherent restriction between binarized outcomes, some canonical parameters merge together:

$$\begin{aligned}
u'_{i1} &= \theta_{i1} + \lambda_{i12} = \log\frac{\mathrm{pr}(y_{i1}=1, y_{i2}=3)}{\mathrm{pr}(y_{i1}=2, y_{i2}=3)}\,, \\
u_{i2} &= \theta_{i2} = \mathrm{logit}\left\{\mathrm{pr}(y^*_{i2} \mid y^*_{i1}=0, y^*_{i3}=0, y^*_{i4}=0)\right\} = \log\frac{\mathrm{pr}(y_{i1}=2, y_{i2}=3)}{\mathrm{pr}(y_{i1}=3, y_{i2}=3)}\,, \\
u'_{i3} &= \theta_{i3} + \lambda_{i34} = \log\frac{\mathrm{pr}(y_{i1}=3, y_{i2}=1)}{\mathrm{pr}(y_{i1}=3, y_{i2}=2)}\,, \\
u'_{i4} &= \theta_{i4} = \log\frac{\mathrm{pr}(y_{i1}=3, y_{i2}=2)}{\mathrm{pr}(y_{i1}=3, y_{i2}=3)}\,,
\end{aligned}$$

$$\begin{aligned}
u'_{i13} &= \lambda_{i13} + u_{i123} + u_{i134} + u_{i1234} = \log\frac{\mathrm{pr}(y_{i1}=1, y_{i2}=1)\mathrm{pr}(y_{i1}=2, y_{i2}=2)}{\mathrm{pr}(y_{i1}=1, y_{i2}=2)\mathrm{pr}(y_{i1}=2, y_{i2}=1)}\,, \\
u'_{i14} &= \lambda_{i14} + u_{i124} = \log\frac{\mathrm{pr}(y_{i1}=1, y_{i2}=2)\mathrm{pr}(y_{i1}=2, y_{i2}=3)}{\mathrm{pr}(y_{i1}=1, y_{i2}=3)\mathrm{pr}(y_{i1}=2, y_{i2}=2)}\,, \\
u'_{i23} &= \lambda_{i23} + u_{i234} = \log\frac{\mathrm{pr}(y_{i1}=2, y_{i2}=1)\mathrm{pr}(y_{i1}=3, y_{i2}=2)}{\mathrm{pr}(y_{i1}=2, y_{i2}=3)\mathrm{pr}(y_{i1}=3, y_{i2}=2)}\,, \\
u_{i24} &= \lambda_{i24} = \log\frac{\mathrm{pr}(y_{i1}=2, y_{i2}=2)\mathrm{pr}(y_{i1}=3, y_{i2}=3)}{\mathrm{pr}(y_{i1}=2, y_{i2}=3)\mathrm{pr}(y_{i1}=3, y_{i2}=2)}\,.
\end{aligned}$$

The last four canonical parameters are quite similar to the GCR's expression developed by Dale (1986). Canonical parameter number increases as cluster sizes increases. Canonical parameters are complicated to be expressed by the parameters from my model:

$$\begin{aligned}
u'_{i1} &= \log\frac{\mathrm{pr}(y_{i1}=1, y_{i2}=3)}{\mathrm{pr}(y_{i1}=1, y_{i2}=2)} \\
&= \log\frac{\mathrm{pr}(y_{i1}\le 1, y_{i2}\le 3) - \mathrm{pr}(y_{i1}\le 1, y_{i2}\le 2)}{\mathrm{pr}(y_{i1}\le 2, y_{i2}\le 2) - \mathrm{pr}(y_{i1}\le 1, y_{i2}\le 2) - \mathrm{pr}(y_{i1}\le 2, y_{i2}\le 1) + \mathrm{pr}(y_{i1}\le 1, y_{i2}\le 1)}\,.
\end{aligned}$$

But (4.5) implies likelihood of ordinal variables can be written in their binarized forms so in the following I discuss a technique originally developed to simplify the log-linear model of binary variables.

4.3.4 Quadratic Exponential Family

Researchers discussed model inference on marginal means as well as correlations, covariances, or odds ratios; however, the original log-linear model is too complicated and restrictive to work with. Zhao and Prentice (1990) proposed a simplified version, called the quadratic exponential model,

which makes a nice connection between canonical parameters and marginal parameters under interest. For the i^{th} cluster, the quadratic exponential log-likelihood of the clustered binary outcome $\boldsymbol{y}_i$ is

$$l_i := \log\left[\mathrm{pr}(\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i}))\right] = \boldsymbol{y}_i^T\boldsymbol{\theta}_i + \boldsymbol{v}_i^T\boldsymbol{\lambda}_i + c(\boldsymbol{y}_i) - \log(\Delta_i)\,, \quad \boldsymbol{v}_i = (y_{i1}y_{i2}, y_{i1}y_{i3}, \ldots, y_{i,n_i-1}y_{i,n_i})\,; \tag{4.6}$$

i.e. all higher order products in (4.4) are put into the "nuisance" part $c(\boldsymbol{y}_i)$, which is assumed to contain no parameter under interest.

Prentice and Zhao have shown there is an one-to-one mapping between $(\boldsymbol{\theta}_i, \boldsymbol{\lambda}_i)$ and the marginal parameters upto the second order, i.e., marginal mean parameters, marginal covariances, correlations or odds ratio parameters.

Here I review several critical steps in developing a set of estimating equations for estimating parameters. One wants to model the pairwise products: $\mathrm{E}(\boldsymbol{v}_i \mid \boldsymbol{Z}_i) := \boldsymbol{\eta}_i = h_1(\boldsymbol{Z}_i; \beta, \alpha)$ and the marginal means $\mathrm{E}(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i) := \boldsymbol{\mu}_i = h(\boldsymbol{Z}_i^T\beta)$. Separately taking derivatives of $\boldsymbol{\theta}_i$ and $\boldsymbol{\lambda}_i$ and taking expectations in (4.6) gives

$$\mathrm{E}\left(\boldsymbol{y}_i^T - \Delta_i^{-2}\frac{\partial\Delta_i}{\partial\boldsymbol{\theta}_i}\right) = 0\,, \quad \mathrm{E}\left(\boldsymbol{v}_i^T - \Delta_i^{-2}\frac{\partial\Delta_i}{\partial\boldsymbol{\lambda}_i}\right) = 0\,; \quad \text{i.e.} \quad \boldsymbol{\mu}_i = \Delta_i^{-2}\frac{\partial\Delta_i}{\partial\boldsymbol{\theta}_i}\,, \quad \boldsymbol{\eta}_i = \Delta_i^{-2}\frac{\partial\Delta_i}{\partial\boldsymbol{\lambda}_i}\,.$$

The above is true since

$$\boldsymbol{\mu}_i = \sum \boldsymbol{y}_i \exp\left\{\boldsymbol{y}_i^T\boldsymbol{\theta}_i + \boldsymbol{v}_i^T\boldsymbol{\lambda}_i + c(\boldsymbol{y}_i)\right\}\Delta_i^{-1}\,, \quad \boldsymbol{\eta}_i = \mathrm{E}(\boldsymbol{v}_i) = \sum \boldsymbol{v}_i \exp\left\{\boldsymbol{y}_i^T\boldsymbol{\theta}_i + \boldsymbol{v}_i^T\boldsymbol{\lambda}_i + c(\boldsymbol{y}_i)\right\}\Delta_i^{-1}\,.$$

Taking derivative of $\boldsymbol{\theta}_i$ in the first equation and taking derivative of $\boldsymbol{\lambda}_i$ in the second give

$$\begin{aligned}
\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\theta}_i} &= \sum \boldsymbol{y}_i\boldsymbol{y}_i^T \exp\left\{\boldsymbol{y}_i^T\boldsymbol{\theta}_i + \boldsymbol{v}_i^T\boldsymbol{\lambda}_i + c(\boldsymbol{y}_i)\right\}\Delta_i^{-1} - \sum \boldsymbol{y}_i \exp\left\{\boldsymbol{y}_i^T\boldsymbol{\theta}_i + \boldsymbol{v}_i^T\boldsymbol{\lambda}_i + c(\boldsymbol{y}_i)\right\}\Delta_i^{-2}\frac{\partial\Delta_i}{\partial\boldsymbol{\theta}_i} \\
&= \mathrm{E}[\boldsymbol{y}_i\boldsymbol{y}_i^T] - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T = \mathrm{cov}[\boldsymbol{y}_i] \\
\frac{\partial\boldsymbol{\eta}_i}{\partial\boldsymbol{\lambda}_i} &= \mathrm{cov}[\boldsymbol{v}_i] \\
\frac{\partial\boldsymbol{\eta}_i}{\partial\boldsymbol{\theta}_i} &= \sum \boldsymbol{y}_i^T\boldsymbol{v}_i \exp\left\{\boldsymbol{y}_i^T\boldsymbol{\theta}_i + \boldsymbol{v}_i^T\boldsymbol{\lambda}_i + c(\boldsymbol{y}_i)\right\}\Delta_i^{-1} - \sum \boldsymbol{v}_i \exp\left\{\boldsymbol{y}_i^T\boldsymbol{\theta}_i + \boldsymbol{v}_i^T\boldsymbol{\lambda}_i + c(\boldsymbol{y}_i)\right\}\Delta_i^{-2}\frac{\partial\Delta_i}{\partial\boldsymbol{\theta}_i} \\
&= \mathrm{cov}[\boldsymbol{y}_i, \boldsymbol{v}_i]\,.
\end{aligned}$$

Thus, the Jacobian matrix of the transformation between canonical parameters $(\boldsymbol{\theta}_i, \boldsymbol{\lambda}_i)$ and marginal parameters up to the second order, is in a nice form and the authors proposed the following estimat-

ing equation

$$\sum_{i=1}^{m}\begin{pmatrix} \partial h(\boldsymbol{Z}_i^T\beta)/\partial\beta & 0 \\ \partial h_1(\boldsymbol{Z}_i;\beta,\alpha)/\partial\beta & \partial h_1(\boldsymbol{Z}_i;\beta,\alpha)/\partial\alpha \end{pmatrix}^T \begin{pmatrix} \mathrm{cov}(\boldsymbol{y}_i) & \mathrm{cov}(\boldsymbol{y}_i,\boldsymbol{v}_i) \\ \mathrm{cov}(\boldsymbol{v}_i,\boldsymbol{y}_i) & \mathrm{cov}(\boldsymbol{v}_i) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y}_i-\boldsymbol{\mu}_i \\ \boldsymbol{v}_i-\boldsymbol{\eta}_i \end{pmatrix} = \mathbf{0}\,, \tag{4.7}$$

which is also the score equation of the *"likelihood"* in (4.6). A quotation mark was used to emphasize the quadratic exponential likelihood is not a real likelihood.

Since the parameters in my model can be grouped as marginal and correlations: η and ρ. Marginal means and pairwise products contain all parameters and thus the quadratic exponential distribution is a good candidate for my model inference.

4.3.5 A simplification: GEE1

(2.8) have several issues. First, most semi-parametric models only specify correlations up to the second order, and thus $\mathrm{cov}(\boldsymbol{v}_i,\boldsymbol{y}_i)$ and $\mathrm{cov}(\boldsymbol{v}_i)$ are intractable. Second, when the marginal regression parameters are under primary interest, (2.8) does not give consistent estimates when the correlation model is mis-specified. Besides, when taking the inverse of the covariance matrix, the computing burden is up to $O(n^6)$ and thus is not working quite well when cluster size is large. A set of simplified estimating equations called GEE1 was proposed:

$$\sum_{i=1}^{m}\left(\partial h(\boldsymbol{z}_i^T\beta)/\partial\beta\right)^T \mathrm{cov}^{-1}(\boldsymbol{y}_i)(\boldsymbol{y}_i-\boldsymbol{\mu}_i) \;=\; \mathbf{0}\,; \tag{4.8}$$

$$\sum_{i=1}^{m}\left(\partial h_1(\boldsymbol{z}_i;\beta,\alpha)/\partial\alpha\right)^T \mathrm{cov}^{-1}(\boldsymbol{v}_i)(\boldsymbol{v}_i-\boldsymbol{\eta}_i) \;=\; \mathbf{0}\,. \tag{4.9}$$

(4.8) is used for marginal parameter estimation here.

Concerning with partial model assumptions and the computing problem, a working version of $\mathrm{cov}(\boldsymbol{v}_i)$ is proposed by assuming there is no correlations among $\boldsymbol{v}_i$. Thus, the estimating equation of α can be written as

$$\sum_{i=1}^{m}\sum_{j<k}\left(\partial h_1(z_{ij},z_{ik};\beta,\alpha)/\partial\alpha\right)^T \frac{v_{ijk}-\eta_{ijk}}{\mathrm{var}(v_{ijk})} = \mathbf{0}\,. \tag{4.10}$$

(4.10) corresponds to the quasi-score equation for a new dataset containing independent observa-

tions

$$(Y_{11}Y_{12}, Z_{11}, Z_{12}) ,$$
$$(Y_{11}Y_{13}, Z_{11}, Z_{13}) ,$$
$$\dots ,$$
$$(Y_{m,n_m-1}Y_{m,n_m}, Z_{m,n_m-1}, Z_{m,n_m}) .$$

This is similar to the composite likelihood based on pairwise observations.

4.3.6 Generalizations

In (4.10), (4.9) is generalized into a composite score equation in which the component likelihoods are pairwise likelihoods. Researchers also directly worked on pairwise observations.

Carey et.al (1993) proposed the alternating logistic regression method: marginal parameters are estimated via (4.8) and correlation parameters, parametrized by pairwise odds ratio regression parameters α, are estimated by (GLM) logistic regressions. To be specific,

$$\text{logit}\,\text{pr}(Y_{ij} = 1 \mid Y_{ik} = y_{ik}, z_{ij}, z_{ik}) = \alpha y_{ik} + \log\left(\frac{\mu_{ij} - \mu_{ijk}}{1 - \mu_{ij} - \mu_{ik} - \mu_{ijk}}\right), \quad j \neq k ,$$

where $\mu_{ij} = \text{pr}(Y_{ij} = 1 \mid z_{ij})$ and $\mu_{ijk} = \text{pr}(y_{ij} = 1, y_{ik} = 1 \mid z_{ij}, z_{ik})$. This method is equivalent to maximizing a composite likelihood in α, whose components are conditional distributions $\text{pr}[Y_j \mid Y_k]$: $j < k$.

Kuk proposed to estimate correlation parameters via maximizing another composite likelihood whose components are the pairwise likelihoods.

Comparing these two methods, Kuk's method is more similar to (4.10). In binary data, using the pairwise products of binary data gives binary outcomes, which loses some information: observations $(Y_{11} = 0, Y_{12} = 0)$, $(Y_{11} = 0, Y_{12} = 1)$ and $(Y_{11} = 1, Y_{12} = 0)$ contribute equivalently to estimate correlation regression parameter. Thus, Prentice and Zhao chose to work with the centered version of the pairwise product: $(Y_{11} - \mu_{11})(Y_{12} - \mu_{12})$. In binary data, both the composite likelihood and the quasi-likelihood from (4.10) with centered products correspond to likelihoods of multinomial distributions. Their estimation efficiency depends on the specific parametric model and generally speaking, they are equivalent. However, when it comes to ordinal data, binarized outcome

is required to use the estimating equation from GEE1, which needs much more computation time as in Table 6.8. In this thesis, I choose to work with Kuk's method in binary and ordinal data, which uses composite likelihood of the original data, not binarizing the ordinal data.

4.3.7 Marginal Parameter Inference

In this subsection I formally introduce the inference methods for marginal parameters, denoted by $\eta := (\beta, \alpha_1, \ldots, \alpha_G)$. I use the simplified quadratic exponential score equation from GEE1 to estimate η. From a dataset containing m independent clusters, the estimating equation of η is

$$\sum_{i=1}^{m} \left(\partial \boldsymbol{\mu}_i^*(\eta)/\partial\eta\right)^T \left(V_{i11}^*\right)^{-1} (\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i^*(\eta)) = 0\ . \tag{4.11}$$

$V_{i11}^* = (v_{ist}^*)$ is a $G \times n_i$-by-$G \times n_i$ covariance matrix of the transformed binary outcome $\boldsymbol{Y}_i^*$. To be more specific, for $s = (j-1) \times G + g_1 \leq j \times G$, i.e., transformed observation for category g_1 of the j^{th} observation from the i^{th} cluster,

$$v_{iss}^* = \frac{e^{\alpha_{g_1} + z_{ij}^T \beta}}{\left(1 + e^{\alpha_{g_1} + z_{ij}^T \beta}\right)^2}\ .$$

In addition, for $g_2 > g_1$, let $t = (k-1) \times G + g_2 \leq k \times G$, i.e. transformed observation for category g_2 in the k^{th} observation from the i^{th} cluster.

When $j = k$; i.e. for the same observation, the corresponding entry is

$$v_{ist}^* = \frac{1}{1 + e^{-z_{ij}^T \beta - \alpha_{g_1}}} - \frac{1}{1 + e^{-z_{ij}^T \beta - \alpha_{g_1}}} \frac{1}{1 + e^{-z_{ij}^T \beta - \alpha_{g_2}}}\ .$$

When $j \neq k$; i.e. for different observations from the same cluster, the corresponding entry is

$$v_{ist}^* = \frac{1}{(1-\rho_{jk}) e^{-z_{ij}^T \beta - \alpha_{g_1} - z_{ik}^T \beta - \alpha_{g_2}} + e^{-z_{ij}^T \beta - \alpha_{g_1}} + e^{-z_{ik}^T \beta - \alpha_{g_2}} + 1} - \frac{1}{1 + e^{-z_{ij}^T \beta - \alpha_{g_1}}} \frac{1}{1 + e^{-z_{ik}^T \beta - \alpha_{g_2}}}\ ,$$

where ρ_{jk} is the correlation between frailties in observations j and k. In the case of an exchangeable correlation structure, ρ_{jk} identically equals to a scalar ρ. In the case of an auto-regressive correlation structure, ρ_{jk} is a function of a scalar ρ. For example, in longitudinal datasets, assuming there is an AR(1) correlation structure, the correlation between frailties W_{ij} and W_{ik} is

$$\rho_{jk} = \rho^{|\mathrm{Time}_{ij} - \mathrm{Time}_{ik}|}\ .$$

In more general correlation structures, such as the un-structured correlation structure, ρ_{ij}'s are functions of some parameter vector ρ.

As long as the marginal mean model is correctly specified, (4.11) gives a consistent estimate $\hat{\eta}$.

4.3.8 Correlation Parameter Inference

To estimate ρ, I consider a composite log-likelihood of the original ordinal outcomes. Denote

$$p_{ijkg_1g_2} = \begin{cases} \left\{(1-\rho_{jk})e^{-(z_{ij}+z_{ik})^T\beta-\alpha_{g_1}-\alpha_{g_2}} + e^{-z_{ij}^T\beta-\alpha_{g_1}} + e^{-z_{ik}^T\beta-\alpha_{g_2}} + 1\right\}^{-1} & \text{for } g_1, g_2 = 1, \ldots, G\ ; \\ \left\{e^{-z_{ij}^T\beta-\alpha_{g_1}} + 1\right\}^{-1} & \text{for } g_1 = 1, \ldots, G, g_2 = G+1\ ; \\ 1 & \text{for } g_1, g_2 = G+1\ . \end{cases}$$

The probability for a pair of observations $(Y_{ij} = y_1, Y_{ik} = y_2, Z_{ij} = z_1, Z_{ik} = z_2)$ is

$$\mathrm{pr}(Y_1 = y_1, Y_2 = y_2 \mid z_1, z_2) = \begin{cases} p_{ijky_1y_2} - p_{ijky_1(y_2-1)} - p_{ijk(y_1-1)y_2} + p_{ijk(y_1-1)(y_2-1)} & y_1 > 1, y_2 > 1\ ; \\ p_{ijky_1y_2} - p_{ijky_1(y_2-1)} & y_1 = 1, y_2 > 1\ ; \\ p_{ijky_1y_2} & y_1 = 1, y_2 = 1\ . \end{cases}$$

The composite log-likelihood I choose to maximize is

$$\sum_{j<k} l_{jk}(\boldsymbol{Z}, \boldsymbol{Y}; \eta, \rho) = \sum_{j<k} \log\left\{\mathrm{pr}(Y_j = y_j, Y_k = y_k \mid z_j, z_k)\right\}\ ;$$

i.e. with a dataset of m independent clusters, I solve for

$$\sum_{i=1}^{m}\sum_{j<k} \partial \log\left\{\mathrm{pr}(Y_{ij} = y_{ij}, Y_{ik} = y_{ik} \mid z_{ij}, z_{ik})\right\} / \partial\rho = 0\ . \tag{4.12}$$

4.3.9 Overall Inference Procedure

Similar to the binary model case, to estimate η and ρ jointly, my algorithm alternates between solving (4.11) with a fixed plug-in ρ from (4.12), and solving (4.12) with a fixed plug-in η from (4.11), until convergence, obtaining estimates $(\hat{\eta}_m, \hat{\rho}_m)$. I write m to indicate an estimate based on a dataset containing m independent clusters.

I suggest starting with solving for (4.11) fixing $\rho = 0$, i.e. GLM, and then estimating ρ from (4.12).

4.3.10 Discussion of My Inference Method

The binarization step is essential for applying the estimating equations derived from the quadratic exponential distribution. One may plug in the original ordinal data into the estimating equation (4.11):

$$\sum_{i=1}^{m} \left(\partial \boldsymbol{\mu}_i(\eta)/\partial\eta\right)^T \left(V_i'\right)^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\eta)) = 0\,. \tag{4.13}$$

where V_i' is the covariance matrix of the original ordinal outcome $\boldsymbol{Y}_i$ and

$$\mu_{ij}(\eta) = G + 1 - \sum_{g=1}^{G} \mathrm{pr}(Y_{ij} \le g \mid z_{ij})\,, \quad \boldsymbol{\mu}_i(\eta) = (\mu_{ij}(\eta))^T_{(j=1,\ldots,n_i)}\,.$$

It can be shown that (4.13) is not as efficient as (4.11). To demonstrate this point, I consider a cluster with two observations $(y_{i1}, y_{i2}, z_{i1}, z_{i2})$ and the outcomes come from three categories 1, 2 and 3.The joint log-likelihood can be expressed in several equivalent ways:

$$\begin{aligned}
\log\left(\mathrm{pr}[y_{i1}, y_{i2} \mid z_{i1}, z_{i2}]\right) = \quad & 1\{y_{i1} \le 1\}\theta_{11}(z_i) + 1\{y_{i2} \le 1\}\theta_{21}(z_i) + 1\{y_{i1} \le 2\}\theta_{12}(z_i) + 1\{y_{i2} \le 2\}\theta_{22}(z_i) \\
& +1\{y_{i1} \le 1\}1\{y_{i2} \le 1\}\lambda_{11}(z_i) + 1\{y_{i1} \le 1\}1\{y_{i2} \le 2\}\lambda_{12}(z_i) \\
& +1\{y_{i1} \le 2\}1\{y_{i2} \le 1\}\lambda_{21}(z_i) + 1\{y_{i1} \le 2\}1\{y_{i2} \le 2\}\lambda_{22}(z_i) \qquad (4.14) \\
= \quad & (3 - y_{i1})\frac{\theta_{11} + \theta_{12}}{2}(z_i) + 1\{y_{i1} = 2\}\frac{\theta_{12} - \theta_{11}}{2}(z_i) \\
& +(3 - y_{i2})\frac{\theta_{21} + \theta_{22}}{2}(z_i) + 1\{y_{i2} = 2\}\frac{\theta_{22} - \theta_{21}}{2}(z_i) + \ldots \qquad (4.15)
\end{aligned}$$

By the method of Zhao and Prentice, (4.11) was derived from presentation in (4.14) and I derive another approximate score equation from the presentation in (4.15):

$$\begin{pmatrix} \partial(\mathrm{E}[y_{i1} \mid z_{i1}])/\partial\eta \\ \partial(\mathrm{pr}[y_{i1} = 2 \mid z_{i1}])/\partial\eta \\ \partial(\mathrm{E}[y_{i2} \mid z_{i2}])/\partial\eta \\ \partial(\mathrm{pr}[y_{i2} = 2 \mid z_{i2}])/\partial\eta \end{pmatrix}^T \mathrm{cov}^{-1}\left[(y_{i1}, 1\{y_{i1} = 2\}, y_{i2}, 1\{y_{i2} = 2\})\right] \begin{pmatrix} y_{i1} - \mathrm{E}[y_{i1} \mid z_{i1}] \\ 1\{y_{i1} = 2\} - \mathrm{pr}[y_{i1} = 2 \mid z_{i1}] \\ y_{i2} - \mathrm{E}[y_{i2} \mid z_{i2}] \\ 1\{y_{i2} = 2\} - \mathrm{pr}[y_{i2} = 2 \mid z_{i2}] \end{pmatrix} = 0\,. \tag{4.16}$$

Compared to (4.16), (4.13) misses several components and thus uses less data. Figure 6.1 plots change in the values of estimating equations (4.11) and (4.13) with regards to different α values.

In Figures 6.1 and 6.2, the blue lines represent the change of (4.13) values is much more flat than the red one, indicating some information is lost by switching from (4.11) to (4.13).

Besides, inverse link function $\mathrm{E}[y_{i1} \mid z_{i1}]$ in (4.13) and (4.16) does not belong to the generalized linear family, making Newton's method for solving estimating equations harder.

I also consider estimating ρ from binarized outcomes, as in estimating marginal parameter η. I denote

$$\begin{aligned} p_{ijk}^{g_j,g_k} &:= \text{pr}[y_{ij} \le g_j, y_{ik} \le g_k \mid z_{ij}, z_{ik}] \\ &= \left\{(1-\rho_{jk})e^{-(z_{ij}+z_{ik})^T\beta-\alpha_{g_j}-\alpha_{g_k}} + e^{-z_{ij}^T\beta-\alpha_{g_j}} + e^{-z_{ik}^T\beta-\alpha_{g_k}} + 1\right\}^{-1} ; \\ p_{ij}^{g_j} := \text{pr}[y_{ij} \le g_j \mid z_{ij}] &= \left\{e^{-z_{ij}^T\beta-\alpha_{g_j}} + 1\right\}^{-1} ; \\ p_{ik}^{g_k} := \text{pr}[y_{ik} \le g_k \mid z_{ik}] &= \left\{e^{-z_{ik}^T\beta-\alpha_{g_k}} + 1\right\}^{-1} . \end{aligned}$$

The composite score equation for ρ inference is

$$\sum_{i=1}^{m} \partial\, l_i'(\boldsymbol{z}_i, \boldsymbol{y}_i; \eta, \rho)/\partial\rho = 0 \,, \tag{4.17}$$

where

$$\begin{aligned} & l_i'(\boldsymbol{z}_i, \boldsymbol{y}_i; \eta, \rho) \\ = \quad & \sum_{j<k}\sum_{g_j=1}^{G}\sum_{g_k=1}^{G} 1\{y_{ij} \le g_j\} \cdot 1\{y_{ik} \le g_k\} \log\left[p_{ijk}^{g_j,g_k}\right] + 1\{y_{ij} \le g_j\}(1 - 1\{y_{ik} \le g_k\}) \log\left[p_{ij}^{g_j} - p_{ijk}^{g_j,g_k}\right] \\ & + (1 - 1\{y_{ij} \le g_j\}) 1\{y_{ik} \le g_k\} \log\left[p_{ik}^{g_k} - p_{ijk}^{g_j,g_k}\right] \\ & + (1 - 1\{y_{ij} \le g_j\})(1 - 1\{y_{ik} \le g_k\}) \log\left[p_{ijk}^{g_j,g_k} - p_{ij}^{g_j} - p_{ik}^{g_k} + 1\right] . \end{aligned}$$

In simulations I compared the performance of these two different methods in Tables 6.6, 6.7 and 6.8. I observed these two methods have similar performance but the method based on (4.17) takes much more time.

Comparing estimating equation inference methods of correlation parameters from Carey et.al and Kuk, the latter is more general and more similar to the estimating equation in GEE1. Heagerty and Zeger compared the alternating logistic regression estimates to (4.9) estimates and the latter method has lower estimation efficiency. Thus I choose to use Kuk's method.

MLE inference has the highest estimation efficiency. However, it is over-complicated to work with and is not robust to model mis-specification. I choose to work with estimating equations. The proposed estimating equation for η is closely related to the MLE and thus has relatively small estimation efficiency loss. In the simulations I presented the average computing times of the two methods and compared their estimation efficiency by Monte Carlo mean squared errors.

4.3.11 Theorems Regarding Asymptotic Behaviour of Estimator

In this subsection I list several theorems regarding the asymptotic behaviour of the estimate $(\hat{\eta}_m, \hat{\rho}_m)$. Denote θ_0 to be true parameter, $D(\boldsymbol{Z}_i^*;\eta) = \partial\boldsymbol{\mu}_i^*(\eta)/\partial\eta$, $V(\boldsymbol{Z}_i^*;\theta) = V_i^*$ and $S(\boldsymbol{Y}_i^*, \boldsymbol{Z}_i^*;\eta) = (\boldsymbol{Y}_i^* - \boldsymbol{\mu}_i^*(\eta))$.

Theorem 4.3.1. *Assume conditions C1, C3 and C5 stated in Appendix A are satisfied, with a generalized C5:*

C^*.5 There is an unique root of η from its estimating equation, then the followings are true:

(a) the solution $\hat{\theta}_m := (\hat{\eta}_m, \hat{\rho}_m)$ from (4.11) *and* (4.12) *is consistent for $\theta_0 := (\eta_0, \rho_0)$;*

(b) $\sqrt{m}\left\{(\hat{\eta}_m - \eta_0)^T, (\hat{\rho}_m - \rho_0)^T\right\}^T$ converges weakly to a normal distribution of mean zero and variance matrix V given by

$$V = \{\mathrm{E}(B)\}^{-1}\{\mathrm{E}(C)\}\left\{\mathrm{E}(B)^T\right\}^{-1},$$

where

$$B = \begin{pmatrix} D(\boldsymbol{Z}^*;\eta_0)^T V^{-1}(\boldsymbol{Z}^*;\theta_0) D(\boldsymbol{Z}^*;\eta_0) & 0 \\ -\sum_{j<k} \frac{\partial^2 l_{jk}}{\partial\eta\partial\rho}(\boldsymbol{Y},\boldsymbol{Z};\theta)\mid_{\theta_0} & -\sum_{j<k} \frac{\partial^2 l_{jk}}{\partial\rho^2}(\boldsymbol{Y},\boldsymbol{Z};\theta)\mid_{\theta_0} \end{pmatrix},$$

$$C = \begin{pmatrix} D(\boldsymbol{Z}^*;\eta_0)^T V^{-1}(\boldsymbol{Z}^*;\theta_0) S(\boldsymbol{Y}^*,\boldsymbol{Z}^*;\eta_0) \\ \sum_{j<k} \frac{\partial l_{jk}}{\partial\rho}(\boldsymbol{Y},\boldsymbol{Z};\theta)\mid_{\theta_0} \end{pmatrix}^{\otimes 2}.$$

Its proof is quite similar to previous model proof and thus is omitted.

The next theorem discusses a mis-specified parametric model but a correct marginal mean model.

Theorem 4.3.2. *Suppose the marginal model in* (4.2) *is true but not the conditional model in* (4.3)*, and the other conditions in Theorem 4.3.1 are satisfied, then*

(a) the solution $\hat{\theta}_m$ is consistent for (η_0, ρ_1), and ρ_1 is derived from

$$\rho_1 = \arg\min_{\rho} KL_{\text{composite}}(L, L^*) := \arg\min_{\rho} P_0 \log\left(\frac{\prod_{j<k} L(Y_j, Y_k, Z_j, Z_k;\eta,\rho')}{\prod_{j<k} L^*(Y_j, Y_k, Z_j, Z_k;\eta,\rho)}\right),$$

where L denotes the likelihood of the true pairwise joint model and L^* for the mis-specified one; ρ' is some other parameters in the true model.

(b) $\sqrt{m}\left\{(\hat{\eta}_m - \eta_0)^T, (\hat{\rho}_m - \rho_1)^T\right\}^T$ *converges weakly to a normal distribution of mean zero and covariance matrix* W *given by*

$$W = \{\mathrm{E}(B_1)\}^{-1}\{\mathrm{E}(C_1)\}\left\{\mathrm{E}(B_1)^T\right\}^{-1} ,$$

where

$$B_1 = \begin{pmatrix} D(\boldsymbol{Z}^*;\eta_0)^T V^{-1}(\boldsymbol{Z}^*;\eta_0,\rho_1)D(\boldsymbol{Z}^*;\eta_0) & 0 \\ -\sum_{j<k}\frac{\partial^2 l^*}{\partial\eta\partial\rho}(\boldsymbol{Y},\boldsymbol{Z};\theta)\mid_{(\eta_0,\rho_1)} & -\sum_{j<k}\frac{\partial^2 l^*}{\partial\rho^2}(\boldsymbol{Y},\boldsymbol{Z};\theta)\mid_{(\eta_0,\rho_1)} \end{pmatrix} ,$$

$$C_1 = \begin{pmatrix} D(\boldsymbol{Z}^*;\eta_0)^T V^{-1}(\boldsymbol{Z}^*;\eta_0,\rho_1)S(\boldsymbol{Y}^*,\boldsymbol{Z}^*;\eta_0) \\ \sum_{j<k}\frac{\partial l^*}{\partial\rho}(\boldsymbol{Y},\boldsymbol{Z};\theta)\mid_{(\eta_0,\rho_1)} \end{pmatrix}^{\otimes 2} .$$

When the pairwise conditional model is correct, asymptotic covariance of $\sqrt{m}(\hat{\eta}_m - \eta_0)$ can be estimated by

$$\hat{V}_m^{\beta} := m\left(\sum_{i=1}^{m} D(\boldsymbol{Z}_i^*;\hat{\eta}_m)^T V^{-1}(\boldsymbol{Z}_i^*;\hat{\theta}_m)D(\boldsymbol{Z}_i^*;\hat{\eta}_m)\right)^{-1} .$$

Allowing for a potentially mis-specified parametric model, a robust estimate of the asymptotic covariance of $\sqrt{m}(\hat{\eta}_m - \eta_0)$ is in a sandwich form

$$\begin{aligned} \hat{V}_m^{\mathrm{robust}} &:= mA_m^{-1}B_mA_m^{-1} ; \\ A_m &:= \sum_{i=1}^{m} D(\boldsymbol{z}_i^*;\hat{\eta}_m)^T V^{-1}(\boldsymbol{z}_i^*;\hat{\theta}_m)D(\boldsymbol{z}_i^*;\hat{\eta}_m) , \\ B_m &:= \sum_{i=1}^{m} D(\boldsymbol{z}_i^*;\hat{\eta}_m)^T V^{-1}(\boldsymbol{z}_i^*;\hat{\theta}_m)S(\boldsymbol{y}_i^*,\boldsymbol{z}_i^*;\hat{\eta}_m)^{\otimes 2} . \end{aligned}$$

Chapter 5

MARGINALIZABLE FRAILTY MODEL

5.1 Overview

I present the formulation of the new frailty model, which is a generalization of the previous models. Inference procedure is also described and discussed briefly. In the end I present asymptotic theorems.

5.2 Model Formulation

In cluster i, given the frailty vector $\boldsymbol{W}_i$, which follows the standard multivariate exponential distribution, I assume the observations are independent and I assume the conditional hazard rate are independent with hazard rates

$$\lambda(t \mid \boldsymbol{W}_i = \boldsymbol{w}_i, Z_{ij} = z_{ij}) = \lambda_0(t) w_{ij} \exp\left(z_{ij}^T \beta\right), \quad t \in [0, \tau],$$

where $\lambda_0(\cdot)$ is a positive conditional baseline hazard rate function, β stands for the conditional log hazard rates ratio and τ is the study time.

I can show that

$$\begin{aligned}
& S(t \mid w_{i1}, Z_{i1} = z_{i1}) \\
= \quad & \int_0^\infty \cdots \int_0^\infty S(t \mid \boldsymbol{W}_i = \boldsymbol{w}_i, Z_{i1} = z_{i1}) f_{\boldsymbol{W}}(w_{i2}, \ldots, w_{in_i} \mid w_{i1}) dw_{i2} \cdots dw_{in_i} \\
= \quad & \int_0^\infty \cdots \int_0^\infty \exp\left(-w_{i1} \Lambda_0(t) e^{z_{i1}^T \beta}\right) f_{\boldsymbol{W}}(w_{i2}, \ldots, w_{in_i} \mid w_{i1}) dw_{i2} \cdots dw_{in_i} \\
= \quad & \exp\left(-w_{i1} \Lambda_0(t) e^{z_{i1}^T \beta}\right) \\
& \text{i.e. } \lambda(t \mid w_{ij}, Z_{ij} = z_{ij}) = \lambda_0(t) w_{ij} \exp\left(z_{ij}^T \beta\right)
\end{aligned}$$

To marginalize the conditional model, note that

$$S(t \mid Z_{ij} = z_{ij}) = \int_0^\infty \cdots \int_0^\infty S(t \mid z_{ij}, \boldsymbol{w}_i) f(\boldsymbol{w}_i) d\boldsymbol{w}_i = \left\{1 + \Lambda_0(t) \exp\left(z_{ij}^T \beta\right)\right\}^{-1},$$

i.e. odds of failure is

$$\frac{1 - S(t \mid z_{ij})}{S(t \mid z_{ij})} = \Lambda_0(t)\exp\left(z_{ij}^T\beta\right) \ . \tag{5.1}$$

5.2.1 Model Interpretation

β can be interpreted as the marginal proportional failure log odds ratio w.r.t. one unit change in the corresponding covariate.

Conditionally, β represents log hazard (rate) ratio w.r.t. one unit change in the corresponding covariate.

5.3 Model Inference

5.3.1 Composite Marginal Log-likelihood Contribution

I denote the parameters under interest by $\theta := (\beta, \rho, \Lambda)$. I consider Λ instead of λ, since the former can be estimated at the same rate as the finite-dimensional parameters (β, ρ). The true parameter is denoted by $\theta_0 := (\beta_0, \rho_0, \Lambda_0)$. I extend the parameter estimation method from Murphy et al. (1997) on independent data into my individual frailty model, maximizing the composite likelihood instead of the real likelihood non-parametrically.

To start with, some log-likelihood needed to be specified to work with. The following conditions are for the identifiability of my model and the construction of the marginal log-likelihood contribution:

C1. Given the total number of observations in the i^{th} cluster, the distribution of covariate vectors $\boldsymbol{Z}_i$ is independent of frailties and is non-informative; i.e. it does not contain θ.

C2. (Coarsening at random assumption) Conditioning on $(Z_{ij}, T_{ij}, \boldsymbol{W}_i)$, the hazard rate function of censoring time C_{ij} is only a function of covariates Z_{ij} and is non-informative.

C3. The number of observations inside the clusters are uniformly bounded from above by n_0.

C4. The true conditional baseline cumulative hazard function $\Lambda_0(t)$ is a strictly increasing function on $[0, \tau]$ and is continuously differentiable. In addition, $\Lambda_0(0) = 0$ and $\Lambda_0'(0) > 0$.

C5. Parameter spaces of β and ρ, denoted by $\mathcal{B}$ and $\mathcal{R}$, belong to some known convex and compact

subsets of $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$, respectively:

$$\begin{aligned}\mathcal{B} &:= \{\beta \in \mathbb{R}^{p_1} : ||\beta|| \leq B_0 \text{ for some finite constant } B_0\} \ , \\ \mathcal{R} &:= \{\rho \in \mathbb{R}^{p_2} : ||\rho|| \leq R_0 \text{ for some finite constant } R_0\} \ ,\end{aligned}$$

where $||\cdot||$ denotes the Euclidean norm; the true value (β_0, ρ_0) is not a boundary point of $\mathcal{B} \times \mathcal{R}$.

C6. The probability of covariate Z concentrating on a hyper-plane is zero. That is, $Z^T\beta = 0$ a.e. implies $\beta = 0$, and Z is bounded with probability one.

Conditions C1 and C3 guarantee that estimation can be carried out while ignoring the covariate and random cluster size distributions. The assumption in C2 follows from Nielsen et al. (1992) and Zeng et al. (2008) as a non-informative censoring condition. With these three conditions and under the same conditional independence assumptions in the other two models, the full likelihood of known frailties and non-missing failure and censoring times from a cluster is

$$\begin{aligned} & f(\boldsymbol{T}, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{W}, N) \\ = \quad & f(\boldsymbol{T}, \boldsymbol{C} \mid \boldsymbol{Z}, \boldsymbol{W}) \times f(\boldsymbol{Z}, \boldsymbol{W}, N) \\ = \quad & \left[\prod_{j=1}^{n} f_C(C_j \mid Z_j, \boldsymbol{W}, T_j) \times f_T(T_j \mid Z_j, \boldsymbol{W})\right] \times f(\boldsymbol{Z}_N) f(\boldsymbol{W}_N) f(N) \ . \end{aligned}$$

The full likelihood contribution of known frailties from a cluster is

$$\begin{aligned} & f(\boldsymbol{Y}, \boldsymbol{\Delta}, \boldsymbol{Z}, \boldsymbol{W}, N) \\ = \quad & \prod_{j=1}^{n} \left[\{f_T(Y_j \mid Z_j, \boldsymbol{W})\}^{\Delta_j} \{S_T(Y_j \mid Z_j, \boldsymbol{W})\}^{1-\Delta_j}\right] \ . \end{aligned}$$

Condition C3 removes the cases of ties and together with C5, leading to model identifiability. Condition C4 is a common technical condition for parameter spaces.

By conditions C1 and C3, I first remove factors involving distributions of covariates, censoring times and the random observation number from the full likelihood and integrated over frailties. This procedure gives the marginal log-likelihood contribution of observation $\boldsymbol{O}_i$ as:

$$\begin{aligned} & \log \int_{w_{in_i}} \cdots \int_{w_{i1}} \left\{\prod_{j=1}^{n_i} \left[f_T(T_{ij} = y_{ij} \mid z_{ij}, w_{ij})\right]^{\delta_{ij}} \left[\text{pr}(T_{ij} > y_{ij} \mid z_{ij}, w_{ij})\right]^{1-\delta_{ij}}\right\} \times f_{W_i}(w_i; \rho) dw_{i1} \cdots dw_{in_i} \\ = \quad & \sum_{j=1}^{n_i} \delta_{ij} \left[\log \lambda_0(y_{ij}) + z_{ij}^T \beta\right] \\ & + \log \int_{w_{in_i}} \cdots \int_{w_{i1}} \left\{\left[\prod_{j=1}^{n_i} w_{ij}^{\delta_{ij}}\right] \exp\left(-\sum_{j=1}^{n_i} w_{ij} \Lambda(y_{ij}) e^{z_{ij}^T \beta}\right)\right\} \times f_{\boldsymbol{W}_i}(\boldsymbol{w}_i; \rho) dw_{i1} \cdots dw_{in_i} \ , \end{aligned} \tag{5.2}$$

where f_T stands for the conditional density of failure time. This is equivalent to first integrating over frailties then removing irrelevant terms, as discussed by Nielsen et al. (1992) and Gill (1992).

Integration in (5.2) uses the Laplace transformation of frailties:

$$\mathscr{L}_{\boldsymbol{W}_i}(\boldsymbol{u}_i) := \mathrm{E}_{\boldsymbol{W}_i}\left[\exp\left(-\sum_{j=1}^{n_i} w_{ij}\Lambda(y_{ij})e^{z_{ij}^T\beta}\right)\right] = |I+\Gamma\mathrm{diag}(\Lambda(y_{i1})e^{z_{i1}^T\beta},\ldots,\Lambda(y_{in_i})e^{z_{in_i}^T\beta})|^{-1}\,,$$

where $\boldsymbol{u}_i := \left(\Lambda(y_{i1})e^{z_{i1}^T\beta},\ldots,\Lambda(y_{in_i})e^{z_{in_i}^T\beta}\right)$, and Γ is the element-wise square root of correlation matrix of frailties $R(\rho)$. Laplace transformation also yields quantities in the forms as

$$\begin{aligned}
\frac{\partial\mathscr{L}_{\boldsymbol{W}_i}(\boldsymbol{u}_i)}{\partial u_{ij}} &= \mathrm{E}_{\boldsymbol{W}_i}\left[u_{ij}\exp\left(-\sum_{j=1}^{n_i} w_{ij}\Lambda(y_{ij})e^{z_{ij}^T\beta}\right)\right]\,,\\
\frac{\partial^2\mathscr{L}_{\boldsymbol{W}_i}(\boldsymbol{u}_i)}{\partial u_{ij}\partial u_{ik}} &= \mathrm{E}_{\boldsymbol{W}_i}\left[u_{ij}u_{ik}\exp\left(-\sum_{j=1}^{n_i} w_{ij}\Lambda(y_{ij})e^{z_{ij}^T\beta}\right)\right]\,.
\end{aligned}$$

Thus, (5.2) is proportional to:

$$\prod_{j=1}^{n_i}\left\{\lambda(y_{ij})\exp\left(z_{ij}^T\beta\right)\right\}^{\delta_{ij}}\cdot\frac{\partial^{\tilde{d}_i}\mathscr{L}_{\boldsymbol{W}_i}(\boldsymbol{u}_i)}{\prod_{j=1}^{n_i}(\partial u_{ij})^{\delta_{ij}}},\quad \text{where } \tilde{d}_i := \sum_{j=1}^{n_i}\delta_{ij}\,,$$

where I take partial derivatives of this Laplace transformation $\mathscr{L}$ at values corresponding to failure events. Since $\mathscr{L}$ is a matrix determinant inverse, taking its derivative is quite complicated. Thus, I choose to work with a composite marginal log-likelihood contribution, which is a weighted summation of pairwise marginal log-likelihood contributions. This is equivalent to working with a mis-specified model under which correlations among three or more observations are ignored. I show the resulting estimator is still consistent in Appendix B, but the estimation efficiency is partially sacrificed since the true likelihood contribution is not used.

Here I explicitly write out a pairwise marginal log-likelihood contribution. Beforehand, I denote

$$\begin{aligned}
u(o_{ij};\beta,\Lambda) &= \Lambda(y_{ij})e^{z_{ij}^T\beta}\,,\\
v(o_{ij},o_{ik};\beta,\rho,\Lambda) &= (1-\rho_{jk})u(o_{ij};\beta,\Lambda)u(o_{ik};\beta,\Lambda)+u(o_{ij};\beta,\Lambda)+u(o_{ik};\beta,\Lambda)+1\,,\\
w(o_{ij},o_{ik};\beta,\rho,\Lambda) &= \delta_{ij}\delta_{ik}(1-\rho_{jk})^2u(o_{ij};\beta,\Lambda)u(o_{ik};\beta,\Lambda)\\
&\quad +\delta_{ij}(1-\rho_{jk})u(o_{ik};\beta,\Lambda)+\delta_{ik}(1-\rho_{jk})u(o_{ij};\beta,\Lambda)+1+\delta_{ij}\delta_{ik}\rho_{jk}\,,
\end{aligned}$$

where ρ_{jk} is the correlation between W_{ij} and W_{ik}. In the case of an exchangeable correlation structure, ρ_{jk}'s identically equal to a scalar parameter ρ. In general, ρ_{jk} is a function of ρ, which can be a vector.

The pairwise marginal log-likelihood contribution of correlated observations (O_{ij}, O_{ik}) is:

$$\begin{aligned} l(O_{ij}=o_{ij}, O_{ik}=o_{ik};\beta,\rho,\Lambda) &= \log w(o_{ij},o_{ik};\beta,\rho,\Lambda)+\delta_{ij}\log\lambda(y_{ij})+\delta_{ik}\log\lambda(y_{ik}) \\ &\quad +(\delta_{ij}z_{ij}+\delta_{ik}z_{ik})^T\beta-(1+\delta_{ij}+\delta_{ik})\log v(o_{ij},o_{ik};\beta,\rho,\Lambda)\ . \end{aligned} \tag{5.3}$$

The composite marginal log-likelihood contribution of $\boldsymbol{O}_i$ is

$$\begin{aligned} & clog(\boldsymbol{O}_i=\boldsymbol{o}_i;\beta,\rho,\Lambda) \\ = \quad & \frac{1}{n_i-1}\sum_{j<k} l(o_{ij},o_{ik};\beta,\rho,\Lambda) \\ = \quad & \sum_{j=1}^{n_i}\delta_{ij}\left\{\log\lambda(y_{ij})+z_{ij}^T\beta\right\}+\frac{1}{n_i-1}\sum_{j<k}\left\{\log w(o_{ij},o_{ik};\beta,\rho,\Lambda)-(1+\delta_{ij}+\delta_{ik})\log v(o_{ij},o_{ik};\beta,\rho,\Lambda)\right\}\ . \end{aligned}$$

Switching to a composite log-likelihood contribution, the high-dimensional integral in (5.2) was reduced to a set of double integrals. Comparing this quantity with the full log-likelihood contribution in (5.2), weights $1/(n_i-1)$ equate their first two terms.

Given a dataset of m independent clusters $(\boldsymbol{O}_1,\ldots,\boldsymbol{O}_m)$, I maximize the following empirical composite marginal log-likelihood contribution:

$$\mathbb{P}_m clog(O;\beta,\rho,\Lambda) := \frac{1}{m}\sum_{i=1}^{m} clog(o_i;\beta,\rho,\Lambda)\ . \tag{5.4}$$

An ordinary maximum likelihood estimator $\hat{\theta}$, in which $\hat{\Lambda}(t)$ is an absolutely continuous function, does not exist, due to the infinite-dimensional parameter $\Lambda(t)$. Therefore I narrow the estimator space of Λ_0 to be a function class containing càdlàg functions $\hat{\Lambda}$ and the derivative is

$$\hat{\lambda}(t) := \hat{\Lambda}(t)-\hat{\Lambda}(t-)\ .$$

In order to maximize (5.4), $\hat{\Lambda}$ shall be a non-decreasing step càdlàg function that only jumps at observed failure time points.

In summary, given a dataset of size m, I maximize (5.4) over the parameter space:

$$\begin{aligned} \Theta &:= \mathcal{B}\times\mathcal{R}\times\mathcal{L}\ , \qquad (5.5) \\ \mathcal{L} &:= \{\Lambda(\cdot):\ \text{a non-decreasing step càdlàg function in } [0,\tau] \text{ with jumps at observed failure time points} \\ &\qquad \text{and } \Lambda(0)=0\}\ . \end{aligned}$$

I denote the resulting non-parametric maximum composite likelihood estimate (NPMCLE) by $(\hat{\beta}_m, \hat{\rho}_m, \hat{\Lambda}_m)$.

5.3.2 EM Algorithm for NPMCLE of (β, Λ)

Treating frailties as missing data, for an observation, the composite complete log-likelihood contribution is

$$\begin{aligned} & \sum_{j=1}^{n_i}\sum_{k<j}\frac{1}{n_i-1}\left(\delta_{ij}\left\{\log\lambda(y_{ij})+\log(w_{ij})+z_{ij}^T\beta\right\}-e^{z_{ij}^T\beta}\Lambda(y_{ij})w_{ij}+\delta_{ik}\left\{\log\lambda(y_{ik})+\log(w_{ik})+z_{ik}^T\beta\right\}-e^{z_{ik}^T\beta}\Lambda(y_{ik})w_{ik}\right) \\ = & \sum_{j=1}^{n_i}\left(\delta_{ij}\left\{\log\lambda(y_{ij})+\log(w_{ij})+z_{ij}^T\beta\right\}-e^{z_{ij}^T\beta}\Lambda(y_{ij})w_{ij}\right) . \end{aligned} \tag{5.6}$$

Weights $1/(n_i-1)$ equate the composite complete log-likelihood contribution to the joint complete log-likelihood contribution inside the integral in (5.2).

1. E-step.

   According to Gao and Song (2011), the expectation of every pairwise complete log-likelihood contribution, conditioning on the data, is

$$\begin{aligned} & \mathrm{E}\left\{\left(\delta_{ij}\left\{\log\lambda(y_{ij})+\log(w_{ij})+z_{ij}^T\beta\right\}-e^{z_{ij}^T\beta}\Lambda(y_{ij})w_{ij}\right.\right. \\ & \left.\left.+\delta_{ik}\left\{\log\lambda(y_{ik})+\log(w_{ik})+z_{ik}^T\beta\right\}-e^{z_{ik}^T\beta}\Lambda(y_{ik})w_{ik}\right) \mid o_{ij}, o_{ik}\right\} . \end{aligned}$$

   One intuitive explanation is that in the composite likelihood, observation O_i is transformed into a new observation O_i^*: every component in O_i^* is a pair of original observations (O_{ij}, O_{ik}), $j<k$. And in O_i^*, these components are treated as if independent of each other.

   Since only w_{ij} and w_{ik} are involved with the parameters of interest, I only need to derive:

$$\mathrm{E}\left\{W_{ij} \mid o_{ij}, o_{ik}; \beta, \rho, \Lambda\right\}, \quad \mathrm{E}\left\{W_{ik} \mid o_{ij}, o_{ik}; \beta, \rho, \Lambda\right\}$$

   for every pair (j,k) in cluster i. To be specific, I consider four cases:

(a) $(\delta_{ij}, \delta_{ik}) = (1, 1)$

$$\begin{aligned}
& \mathrm{E}(W_{ij} \mid y_{ij}, \delta_{ij} = 1, z_{ij}, y_{ik}, \delta_{ik} = 1, z_{ik}; \beta, \rho, \Lambda) \\
= \quad & \left(\frac{\partial^3 \mathscr{L}(u_{ij}, u_{ik})}{\partial u_{ij}^2 \partial u_{ik}} \right) \Bigg/ \left(\frac{\partial^2 \mathscr{L}(u_{ij}, u_{ik})}{\partial u_{ij} \partial u_{ik}} \right) \\
= \quad & \frac{2\left((1-\rho_{jk})u(o_{ik};\beta,\Lambda)+1\right)}{1+u(o_{ij};\beta,\Lambda)+u(o_{ik};\beta,\Lambda)+(1-\rho_{jk})u(o_{ij};\beta,\Lambda)u(o_{ik};\beta,\Lambda)} \\
& \times \frac{\frac{3\left((1-\rho_{jk})u(o_{ik};\beta,\Lambda)+1\right)\left((1-\rho_{jk})u(o_{ij};\beta,\Lambda)+1\right)}{1+u(o_{ij};\beta,\Lambda)+u(o_{ik};\beta,\Lambda)+(1-\rho_{jk})u(o_{ij};\beta,\Lambda)u(o_{ik};\beta,\Lambda)} - 2(1-\rho_{jk})}{\frac{2\left((1-\rho_{jk})u(o_{ik};\beta,\Lambda)+1\right)\left((1-\rho_{jk})u(o_{ij};\beta,\Lambda)+1\right)}{1+u(o_{ij};\beta,\Lambda)+u(o_{ik};\beta,\Lambda)+(1-\rho_{jk})u(o_{ij};\beta,\Lambda)u(o_{ik};\beta,\Lambda)} - (1-\rho_{jk})} ;
\end{aligned}$$

(b) $(\delta_{ij}, \delta_{ik}) = (1, 0)$

$$\begin{aligned}
& \mathrm{E}(W_{ij} \mid y_{ij}, \delta_{ij} = 1, z_{ij}, y_{ik}, z_{ik}, \delta_{ik} = 0; \beta, \rho, \Lambda) \\
= \quad & \left(\frac{\partial^2 \mathscr{L}(u_{ij}, u_{ik})}{\partial u_{ij}^2} \right) \Bigg/ \left(\frac{\partial \mathscr{L}(u_{ij}, u_{ik})}{\partial u_{ij}} \right) \\
= \quad & \frac{2(1 + (1-\rho_{jk})u(o_{ik};\beta,\Lambda))}{1 + u(o_{ij};\beta,\Lambda) + u(o_{ik};\beta,\Lambda) + (1-\rho_{jk})u(o_{ij};\beta,\Lambda)u(o_{ik};\beta,\Lambda)} ;
\end{aligned}$$

(c) $(\delta_{ij}, \delta_{ik}) = (0, 1)$

$$\begin{aligned}
& \mathrm{E}(W_{ij} \mid y_{ij}, \delta_{ij} = 0, z_{ij}, y_{ik}, \delta_{ik} = 1, z_{ik}; \beta, \rho, \Lambda) \\
= \quad & \left(\frac{\partial^2 \mathscr{L}(u_{ij}, u_{ik})}{\partial u_{ij} \partial u_{ik}} \right) \Bigg/ \left(\frac{\partial \mathscr{L}(u_{ij}, u_{ik})}{\partial u_{ik}} \right) \\
= \quad & \frac{2(1 + (1-\rho_{jk})u(o_{ik};\beta,\Lambda))}{1 + u(o_{ij};\beta,\Lambda) + u(o_{ik};\beta,\Lambda) + (1-\rho_{jk})u(o_{ij};\beta,\Lambda)u(o_{ik};\beta,\Lambda)} - \frac{1-\rho_{jk}}{1 + (1-\rho_{jk})u(o_{ij};\beta,\Lambda)} ;
\end{aligned}$$

(d) $(\delta_{ij}, \delta_{ik}) = (0, 0)$

$$\begin{aligned}
& \mathrm{E}(W_{ij} \mid y_{ij}, \delta_{ij} = 0, z_{ij}, y_{ik}, \delta_{ik} = 0, z_{ik}; \beta, \rho, \Lambda) \\
= \quad & \left(\frac{\partial \mathscr{L}(u_{ij}, u_{ik})}{\partial u_{ij}} \right) \Big/ \mathscr{L}(u_{ij}, u_{ik}) \\
= \quad & \frac{1 + (1-\rho_{jk})u(o_{ik};\beta,\Lambda)}{1 + u(o_{ij};\beta,\Lambda) + u(o_{ik};\beta,\Lambda) + (1-\rho_{jk})u(o_{ij};\beta,\Lambda)u(o_{ik};\beta,\Lambda)} .
\end{aligned}$$

I denote $w_{ij}^{(r)}$ as the averaged expectation of the frailty conditional expectations in the r^{th} iteration E-step; i.e.

$$w_{ij}^{(r)} = \sum_{k \neq j} \frac{1}{n_i - 1} \mathrm{E}\left\{ W_{ij} \mid o_{ij}, o_{ik}; \hat{\beta}^{(r-1)}, \rho, \hat{\Lambda}^{(r-1)} \right\} ,$$

where $(\hat{\beta}^{(r-1)}, \hat{\Lambda}^{(r-1)})$ are estimates from the $(r-1)^{th}$ iteration's M-step and ρ is some fixed value plugged into the EM algorithm.

2. M-step.

   Since the complete composite log-likelihood contribution is exactly the complete joint log-likelihood contribution, the M-step inference is quite straightforward. In the r^{th} iteration's M-step, plugging the imputed frailty values into (5.6), the composite complete log-likelihood contribution becomes

$$\sum_{j=1}^{n_i}\left\{\delta_{ij}\left[\log\lambda(y_{ij})+z_{ij}^T\beta\right]-\left[\sum_{k\neq j}\frac{1}{n_i-1}\mathrm{E}\left\{W_{ij}\mid o_{ij},o_{ik}\right\}\right]e^{z_{ij}^T\beta}\Lambda(y_{ij})+\sum_{k\neq j}\frac{1}{n_i-1}\mathrm{E}\left[\log(W_{ij})\mid o_{ij},o_{ik}\right]\right\}.$$

   I update the estimates as $(\hat{\beta}^{(r)},\hat{\Lambda}^{(r)})$, where $\hat{\beta}^{(r)}$ solves

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}\left(z_{ij}-\frac{\sum_{k=1}^{m}\sum_{l=1}^{n_k}z_{kl}w_{kl}^{(r)}\exp(z_{kl}^T\beta)1\{y_{kl}\geq y_{ij}\}}{\sum_{k=1}^{m}\sum_{l=1}^{n_k}w_{kl}^{(r)}\exp(z_{kl}^T\beta)1\{y_{kl}\geq y_{ij}\}}\right)=0\,, \qquad (5.7)$$

   and $$\hat{\Lambda}^{(r)}(t)=\sum_{s\leq t}\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}1\{y_{ij}=s\}\delta_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}w_{ij}^{(r)}\exp(z_{ij}^T\hat{\beta}^{(r)})1\{y_{ij}\geq s\}}\,.$$

(5.7) is equivalent to a partial score equation offset by imputed w_{ij}'s from the E-step and the estimate of $\Lambda(t)$ is a Breslow-type estimator.

5.3.3 Estimating ρ from (5.4)

For some fixed value (β_1,Λ_1), I maximize (5.4) over ρ directly; i.e., I solve for ρ from

$$\sum_{i=1}^{m}\frac{1}{n_i-1}\sum_{j<k}\frac{\partial\rho_{jk}}{\partial\rho}\left(\frac{2(\rho_{jk}-1)\delta_{ij}\delta_{ik}u(o_{ij};\beta_1,\Lambda_1)u(o_{ik};\beta_1,\Lambda_1)-\delta_{ij}u(o_{ik};\beta_1,\Lambda_1)-\delta_{ik}u(o_{ij};\beta_1,\Lambda_1)+\delta_{ij}\delta_{ik}}{w(o_{ij},o_{ik};\beta_1,\rho,\Lambda_1)}\right.$$
$$\left.+(1+\delta_{ij}+\delta_{ik})\frac{u(o_{ij};\beta_1,\Lambda_1)u(o_{ik};\beta_1,\Lambda_1)}{v(o_{ij},o_{ik};\beta_1,\rho,\Lambda_1)}\right)=0\,.$$

5.3.4 Overall Inference Procedure

Direct maximization of (5.4) is computationally challenging due to unobservable frailties. By treating frailties as missing data, I use a generalized version of the EM algorithm for the composite likelihood, similar to Gao and Song (2011). However, the EM algorithm cannot directly estimate the correlation parameter and thus the whole algorithm is a hybrid.

Step 1 The algorithm starts from some value $\theta^{(0)} := (\beta^{(0)}, \rho^{(0)}, \Lambda^{(0)})$.

Step 2 This algorithm gives an estimate of (β, Λ), denoted as $(\beta^{(1)}, \Lambda^{(1)})$, from the EM algorithm, fixing ρ at $\rho^{(0)}$.

Step 3 This algorithm gives an estimate of ρ, denoted as $\rho^{(1)}$, by directly maximizing (5.4), fixing (β, Λ) at $(\beta^{(1)}, \Lambda^{(1)})$.

Step 4 This algorithm repeats Step 2 and Step 3 until estimates converge.

I suggest starting from $\rho^{(0)} = 0$ and carrying out Step 2 as in the standard proportional odds model for independent data.

5.3.5 Inference Method Discussion

The EM algorithm has been widely applied in shared frailty models, such as in the work by Klein (1992). Individual frailty models can flexibly model correlation structures but put several challenges in model inference. First, the observed likelihood involves high-dimensional integrations over individual frailties; thus, NPMLE is computationally infeasible. Second, the multivariate exponential distribution does not have a close form density, making the direct application of EM algorithm impossible: the full likelihood contribution cannot be written out.

Considering the first challenge, I maximize the composite log-likelihood contribution, which is the summation of all pairwise log-likelihood contributions. And the corresponding computation is relatively straightforward. As for the second challenge, I propose a hybrid EM algorithm on the composite likelihood contribution: in the EM algorithm part, I only solve for marginal regression parameters (β, Λ); thus, the explicit joint density of frailties is not needed. Correlation parameter was estimated by directly maximizing (5.4), which does not need the explicit density either.

Vu and Knuiman (2002) proposed a similar hybrid EM algorithm for a shared frailty model inference but they did not argue for convergence of their iterative algorithm. In the following I give a justification. My iterative inference procedure is guaranteed to at least converge to some local maximum of the empirical composite log-likelihood contribution. In Step 2 of the r^{th} iteration, by

the property of the EM algorithm ,

$$\mathbb{P}_m clog(O; \beta^{(r-1)}, \rho^{(r-1)}, \Lambda^{(r-1)}) \leq \mathbb{P}_m clog(O; \beta^{(r)}, \rho^{(r-1)}, \Lambda^{(r)})$$

is true. In Step 3 of the r^{th} iteration, since I maximize (5.4) over ρ with plug-in $(\beta^{(r)}, \Lambda^{(r)})$:

$$\mathbb{P}_m clog(O; \beta^{(r)}, \rho^{(r-1)}, \Lambda^{(r)}) \leq \mathbb{P}_m clog(O; \beta^{(r)}, \rho^{(r)}, \Lambda^{(r)})$$

is true. In summary,

$$\mathbb{P}_m clog(O; \beta^{(r-1)}, \rho^{(r-1)}, \Lambda^{(r-1)}) \leq \mathbb{P}_m clog(O; \beta^{(r)}, \rho^{(r)}, \Lambda^{(r)})$$

is true. Besides, since I can show the parameters are uniformly bounded (in Appendix B) and Z and Y are assumed to be uniformly bounded, $\mathbb{P}_m clog$ is an uniformly bounded function. Thus, the iterative algorithm will converge at some local maximum.

5.3.6 Asymptotic Theorems

Before presenting the results, I list additional technical assumptions needed for the theoretical results of the NPMCLE. Hereafter, $\tau < \infty$ denotes the endpoint time of the study.

C6. There exists some strictly positive constant c_0 such that

$$\text{pr}(C_{ij} \geq \tau \mid Z_{ij}) = \text{pr}(C_{ij} = \tau \mid Z_{ij}) \geq a_0 \quad \text{a.s. ;}$$

C7. The covariate Z is bounded.

Lemma 5.3.1. *The true parameter is identifiable from the composite marginal likelihood contribution. Furthermore, the composite Fisher information matrix, which is the average of Fisher information matrices from all pairwise observations, is invertible along any one-dimensional sub-model.*

Theorem 5.3.2. *NPMCLE $(\hat{\beta}_m, \hat{\rho}_m, \hat{\Lambda}_m)$ will converge uniformly to the true value $(\beta_0, \rho_0, \Lambda_0)$ as the number of independent clusters m goes to infinity, in the metric space $\mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \times l^\infty[0, \tau]$, where $l^\infty[0, \tau]$ is the linear space consisting of all the bounded functions on $[0, \tau]$ and is equipped with the total variation norm $|| \cdot ||_V$, defined as the maximum between the sup norm and the total variation of a function.*

Theorem 5.3.3. $\sqrt{m}(\hat{\beta}_m^T - \beta_0^T, \hat{\rho}_m^T - \rho_0^T, \hat{\Lambda}_m - \Lambda_0)^T$ *weakly converges to a mean zero Gaussian process in the same metric space as in the previous theorem.*

Given a dataset of m independent clusters, in the following I discuss how to estimate the asymptotic standard error of NPMCLE $(\hat{\beta}_m, \hat{\rho}_m)$ from the dataset. For the empirical composite marginal log-likelihood contribution in (5.4), its Hessian matrix was calculated by taking its second derivative at $(\beta, \rho, \Lambda(t_{(1)}), \ldots, \Lambda(t_{(Q)}))$, where $\{t_{(q)}\}_{q=1,\ldots,Q}$ is the set of ordered failure event times. I denote this matrix by H_m. The empirical composite score function $S_m := \sum_{i=1}^m S_{m,i}/m$ was derived by taking the first derivative of (5.4) at $(\beta, \rho, \Lambda(t_{(1)}), \ldots, \Lambda(t_{(Q)}))$, and I estimate the covariance of $\sqrt{m}S_m$ by $J_m := \sum_{i=1}^m S_{m,i}S_{m,i}^T/m$. For some arbitrary $h = (h_1, h_2, h_3)$ in which $(h_1, h_2) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ and $h_3 \in l^\infty(0, \tau]$, I denote h_m as the vector comprising of h_1, h_2 and $h_m = (h_1^T, h_2^T, h_3(t_{(1)}), \ldots, h_3(t_{(Q)}))^T$.

The following theorem finds an asymptotically consistent estimate of the NPMCLE covariance matrix.

Theorem 5.3.4. *Let* $V(h_1, h_2, h_3)$ *be the asymptotic covariance matrix of*

$$\sqrt{m}\left\{h_1^T(\hat{\beta}_m - \beta_0) + h_2^T(\hat{\rho}_m - \rho_0) + \int_0^\tau h_3(s)d[\hat{\Lambda}_m - \Lambda_0](s)\right\} .$$

Then $h_m^T H_m^{-1} J_m H_m^{-1} h_m \xrightarrow{p.} V(h_1, h_2, h_3)$ *uniformly for* (h_1, h_2, h_3) *such that*

$$||h_1|| \leq 1, \quad ||h_2|| \leq 1, \quad ||h_3||_V \leq 1 .$$

Proofs are provided in Appendix B.

To estimate the covariance matrix of $(\hat{\beta}_m, \hat{\rho}_m)$, I set

$$h_m^T = \begin{pmatrix} I & 0_{p_1+p_2} & \cdots & 0_{p_1+p_2} \end{pmatrix} , \tag{5.8}$$

in which I is a $(p_1 + p_2) \times (p_1 + p_2)$ identity matrix, and $0_{p_1+p_2}$ is a zero column vector of length $p_1 + p_2$ and Q such zero column vectors are put into h_m; i.e., $h_3(\cdot)$ is a zero function.

Chapter 6

NUMERIC STUDIES

6.1 Binary Model Numerical Studies

6.1.1 Simulation

I conducted simulation studies to evaluate the finite sample performance of my proposed estimators. In each simulation scenario, 1000 Monte Carlo datasets were generated. In each dataset, I generated 200 independent clusters.

Throughout this subsection, the marginal model was assumed to be

$$\mathrm{pr}(Y_{ij} = 1 \mid Z_{ij}) = \frac{1}{1 + \exp\left(-\beta_0 - \beta_1 Z_{ij1}\right)} , \tag{6.1}$$

where $\beta_0 = 1$ and $\beta_1 = -1.2$. A covariate Z_1 is included, which was a continuous normal random variable with mean zero and standard deviation 2.

In the first three simulation sets, model inference was carried out with correctly assumed parametric models: the conditional model in (3.2) and multivariate exponential distributed random effects. The first two simulation sets are of two-level clustering, in which cluster sizes varied from 5 to 7 with equal probabilities. I impose an exchangeable correlation structure and an AR(1) correlation structure. Exchangeable correlation structure is typically implemented to model correlations between individuals sampled from the same geographical region, hospital, etc; AR(1) correlation usually models longitudinal observations over time. The third simulation set is three-level clustering; there are 200 independent clusters, and inside each cluster there are several individuals and multiple observations are taken on every individual. I put 2 or 3 individuals into each cluster with probabilities $4/5$ and $1/5$ and generated 2 or 3 observations for every individual with probabilities $4/5$ and $1/5$. I imposed exchangeable correlation structure on the first level of clustering and the AR(1) structure onto the second. In Tables 6.1 and 6.2 I only listed model-based standard errors and 95% confidence interval coverage rates derived from (3.10), since their robust counterparts from (3.11) behaved quite similarly.

In the end of this subsection, a mis-specified joint model was also considered, in which I generated outcomes by some latent variable model. For each cluster i, I generated a uniform variable U_i and transformed it into a logistic distributed random variable $A_i := \log U_i - \log(1 - U_i)$; at the end, I simulated $(Y_{i1}, \ldots, Y_{in_i})$ by $Y_{ij} = 1\{Z_{ij}^T\beta + A_i > 0\}$. The marginal model in (6.1) is satisfied. For the proposed inference method, both the model-based and the robust standard errors and their respective 95% confidence interval coverage rates were presented in Table 5.3.

MLE was also implemented in my model, and I compared MLE to my estimators. MLE was carried out using the **optim** function in R throughout my simulations, provided gradients of the targeting function.

Table 6.1 lists the simulation results in the case of two-level clustering under an exchangeable correlation structure and an AR(1) correlation structure, respectively in (a) and (b). The estimation efficiency of my estimates, measured in Monte Carlo mean squared error (MSE), is quite close to the MLE's, but the proposed method takes much less computing time than MLE.

Table 6.2 lists simulation results for three-level clustering. When the correlation level is low, results from the two inference methods are pretty close.

Table 6.3 lists simulation results for the mis-specified model. As expected, MLE of β is biased while the proposed method gives consistent estimates of β, along with consistently estimated robust standard errors and correct confidence intervals.

In all simulation sets, my inference method has little estimation efficiency loss compared to MLE and MLE inference is much more computationally intensive.

6.1.2 Madras Longitudinal Schizophrenia Study

I further demonstrated the proposed method using the Madras longitudinal schizophrenia study from Thara et al. (1994), in which first-episode schizophrenics were followed for 10 years with the primary objective of characterizing the natural history of disease progression. The data contain several longitudinal binary outcome measurements indicating the presence of positive psychiatric symptoms over the time course: $t_{ij} = 0, \ldots, 11$ months during the first year following an initial hospitalization for 86 schizophrenia patients. The binary outcome Y_{ij} under interest is an indicator of whether or not a patient is observed to have thought disorders. Covariates include the time

variable t_{ij}, a binary indicator Z_{ij2} of whether or not a patient is younger than 20 at disease onset and gender Z_{ij3}: 0 for male and 1 for female. To assess the association between occurrence of thought disorders and the covariates, a marginal logistic regression model was proposed:

$$\operatorname{logit} \mathrm{E}(Y_{ij} \mid t_{ij}, Z_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 Z_{ij2} + \beta_3 Z_{ij3} \; ;$$

i.e. I assumed a linear trend of time in log odds ratio of thought disorders. This regression model is almost identical to the model from Heagerty (1999), except that I did not center the time covariate.

With the proposed method, I can answer whether the population-averaged probability of thought disorders differs over time, age-at-onset and gender subgroups. Both the proposed method and the MLE were used to analyze this dataset, assuming observations from the same patients are exchangeably and AR(1) correlated over time. Results were reported in Table 5.4. The results from different inference methods are pretty similar. Since this is a longitudinal dataset, the AR(1) correlation structure assumption should be more close to the real situation, and in the following I reported the results from the proposed inference method based on the AR(1) correlation assumption. Robust 95% confidence intervals were reported.

The estimated odds of thought disorder prevalence for a patient younger than 20 at the beginning of hospitalization is 50% higher (95% C.I.: 26% lower to 203% higher) than older patient, controlling for gender and observation time. The estimated odds of thought disorder prevalence for a female patient is 47% lower (95% CI: 72% lower to 1% higher) than a male patient, controlling for age at onset and observation time. The estimated odds of thought disorder decreases by 29% (95% CI: 35% to 23%) in one month during hospitalization, controlling for age at onset and gender. There is evidence of significant decrease in thought disorder occurrence as times passes in hospital.

The estimated correlations between random effects of the same patients over time is 0.97 (0.93, 1.00); I found a strong correlation within each patient, which is not surprising for longitudinal observations

6.1.3 British Social Attitudes Panel Survey.

To demonstrate the method for three-level clustering data, I analyzed the *British Social Attitudes Panel Survey* conducted from 1983 through 1986. In this survey, individuals were asked whether

they thought there should be no legal or governmental regulation on abortion (should permit=1, otherwise=0). This survey was carried out in 54 independent districts annually for four years among the same individuals. The dataset includes people who have completed all four surveys during the four years, adding up to 1,056 observations from 264 individuals in total. There are 54 districts, containing 3 to 40 individuals. Covariates can be categorized into three levels: the first level is a district-level covariate: the percent of protestants of each district; the second level includes individual-level demographic covariates, including social class (middle, upper and lower), gender (male and female) and religion (Protestant, Catholic, other and none); in the third level are three dummy variables for years $1984, 1985, 1986$. There are two covariates corresponding to protestant in the model, one on the district-level and the other on the individual-level. The inclusion of the two protestant variables is potentially of substantive interest by measuring the religious context impact on individual attitude in contrast to their own religious affiliation affect, as discussed in Heagerty and Zeger (2000a).

Table 6.6 lists the point estimates and 95% robust confidence intervals of odds ratio corresponding to the above covariates, from three methods. Method 1 is the proposed method assuming the frailty correlation structure within individuals is AR(1) and the correlation structure across individuals within a district is exchangeable. Method 2 is also the proposed method, but assuming both correlation structures within districts and individuals are exchangeable. Method 3 is GEE with an exchangeable working correlation matrix for observations within a district. This ignores the finer level of correlations within individuals over time. I did not compare the results with the MLE since the function **optim** in **R** failed to converge, due to large clusters.

Method 1 and Method 2 gave out roughly the same point estimates as GEE but with generally narrower 95% confidence intervals. The exceptions are categorical covariates representing religion contrasts between other religions and Protestants.

Results from Methods 1 and 2 are roughly the same, indicating the robustness of the proposed method with respect to different assumed correlation structures. Since observations are taken annually, their correlations are better described by an AR(1) correlation structure. In the following I reported results from Method 1.

The covariates I put into Method 1 decompose religion contrasts into within-district contrast and between-district contrast. The variable %Protestant is a district-level covariate and equals to the

sampled proportion of district Protestants. The estimated odds ratio is 2.17, 95% CI: (0.86, 5.52), indicating a non-significantly increasing trend of allowing abortions among individuals from districts of a higher level in Protestants, controlling for all the other covariates. The categorical covariate religion contrasting Catholic, other, none, to the reference Protestant group can be interpreted as comparing the propensity of allowing abortion among individuals of different religions who reside in districts of equal level in Protestants, controlling for year surveyed, social class and gender. Controlling for other covariates, compared to Protestants, non-significantly lower odds are observed among Catholics as of 0.67, 95% CI: (0.25, 1.80); non-significantly lower odds are observed among those of other religions with the ratio 0.52, 95% CI: (0.24, 1.11) and significantly higher odds are observed among those without any religions with odds ratio being 2.00, 95% CI: (1.21, 3.30). The propensity of allowing abortions among females is non-significantly lower than that among males, with an odds ratio as 0.72 95% CI: (0.48, 1.07). The odds ratio of allowing abortions from upper working class comparing to middle class is 0.76, 95% CI: (0.51, 1.14), and the odds ratio comparing lower working class to middle class is 0.80, 95% CI: (0.54, 1.19), controlling for all the other covariates. As for time trend in allowing for abortions propensity, there is a significant drop in Year 1984 compared to the previous year with odds 0.66, 95% CI: (0.49, 0.88), and there are non-significant increments in the following two years, compared to Year 1983.

6.2 Ordinal Model Numerical Studies

6.2.1 Simulation

I conducted simulation studies to evaluate the finite sample performance of my proposed estimators. In each simulation scenario, 1000 Monte Carlo datasets were generated. In each dataset, I generated 200 independent clusters.

In two-level clustering simulations, the outcomes come from five categories and the marginal model was

$$\mathrm{pr}(Y_{ij} \leq g \mid Z_{ij}) = \{1 + \exp(-\alpha_g - \beta_1 Z_{ij1} - \beta_2 Z_{ij2})\}^{-1} \ , \quad g = 1, 2, 3, 4 \ , \tag{6.2}$$

where $\beta_1 = 1$, $\beta_1 = 1.2$, $\alpha_1 = 0.1$, $\alpha_2 = 0.4$, $\alpha_3 = 0.7$ and $\alpha_4 = 1$. The first covariate Z_1 is a continuous normal random variable with mean zero and standard deviation 0.5; the second covariate

Z_2 is a linear combination of a Bernoulli variable Z_0 with mean 0.3 and Z_1: $Z_2 = Z_0 - 0.3 + 0.2 \times Z_1$. Cluster sizes varied from 7 to 9 with equal probabilities. Similar to the previous simulations, I imposed an exchangeable correlation structure and an AR(1) correlation structure. During inference, I assumed the correct joint model and only model-based standard errors and model-based 95% confidence interval coverage rates were listed since their robust counterparts behaved quite similarly. MLE is also applied, with **optim** function in R implemented. Note that the gradient matrix of the log-likelihood is not implemented into **optim** due to heavy computation. In the exchangeable correlation case in Table 6.6, MLE is carried out properly by **optim**, which is not the case of Table 6.9.

In the three-level clustering simulation, outcomes come from three categories and the marginal model is

$$\mathrm{pr}(Y_{ij} \leq g \mid Z_{ij}) = \{1 + \exp(-\alpha_g - \beta_1 Z_{ij1} - \beta_2 Z_{ij2})\}^{-1} \ , \quad g = 1, 2 \ ,$$

where $\beta_1 = 1$, $\beta_1 = 1.2$, $\alpha_1 = 0.1$ and $\alpha_2 = 0.4$. Covariates are identically generated as the previous simulations. I put 2, 3, or 4 clusters into each institution with equal probabilities and generated 2 to 4 observations in each cluster with equal probabilities. Both correlation structures were set to be exchangeable.

During model inference, I assumed the correct models. Results were listed in Tables 6.6 through 6.10. I only listed model-based standard errors and 95% confidence interval coverage rates, since their robust counterparts behaved quite similarly.

A mis-specified parametric joint model was also considered at the end, in which correlations were introduced via a latent variable model as in the binary case. For each cluster i, I generated a uniform variable U_i and transformed it into a logistic random variable $A_i = \log U_i - \log(1 - U_i)$; at the end, I simulated ordinal outcomes by $1\{Y_{ij} \leq g\} = 1\{Z_{ij}^T \beta + A_i + \alpha_g > 0\}$, which satisfies the marginal model in (6.2). For the proposed inference method, both the model-based and the robust standard errors and their respective 95% confidence interval coverage rates were presented.

Tables 6.6 and 6.9 list the simulation results in the case of two-level clustering under an exchangeable correlation structure and an AR(1) correlation structure, respectively. Under exchangeable correlation structures, estimation efficiency of my estimates, measured by mean squared error (MSE),is quite close to the MLE's. In both cases the proposed method takes much less computing

time than MLE.

Table 6.10 lists simulation results for three-level clustering. Results from the two inference methods are pretty close.

Table 6.11 lists simulation results for the mis-specified model. As expected, MLE of β and α is biased while the proposed method gives consistent estimates of β and α, along with consistently estimated robust standard errors.

6.2.2 Arthritis Study

Lipsitz et al. (1994) studied a randomized clinical trial on 302 patients with rheumatoid arthritis. The treatment arm uses a drug called auranofin and the control arm uses placebo. Before treatment assignment, age, gender and self-assessment score of rheumatoid arthritis of patients are recorded. At months 1,3 and 5 after treatment carried out, patients continue to self-assess the rheumatoid arthritis level. The score is a five-level ordinal response and a higher score corresponds to a better self-assessed health status. To evaluate the treatment efficacy on self-assessed rheumatoid arthritis level in this population, I fitted this marginal model:

$$\text{logit } \{\text{pr}[Y_{ij} \leq g \mid Z_{ij}]\} = \alpha_g + \beta_1 \text{Age}_{ij} + \beta_2 \text{Gender}_{ij} + \beta_3 \text{Time}_{ij} + \beta_4 \text{baseline self-assessment}_{ij} \, ;$$

i.e. baseline self-assessment score of rheumatoid arthritis was adjusted for as a continuous variable and time variable was also put in the model in the linear form. I assumed an exchangeable correlation structure. The results are listed in Table 6.12, which compares the proposed method to MLE. Robust 95% confidence intervals are reported.

The cumulative log odds ratio of auranofin treatment is -0.55 (-0.87, -0.24); this implies controlling for all the other covariates, the odds of self-assessing a score smaller than any fixed score g_0 for patients from the treatment arm is 42% (21%, 58%) lower than their counterparts in the control arm. There is statistically significant improvement in patient self-assessments by auranofin treatment.

It is also interesting that controlling for the other covariates, patients self-assessments of their status are becoming optimistic as time passes, where the odds of self-assessing a score smaller than any fixed score g_0 at a certain time is 7.7% (2.0%, 13%) times lower than one month ago during the study. And this trend is statistically significant. The correlation level between frailty is estimated to 0.84 (0.76, 0.92).

6.2.3 Television, School and Family Smoking Prevention and Cessation Project (TVSFP)

Television, School and Family Smoking Prevention and Cessation Project is a study designed to determine the efficacy of a school-based smoking prevention curriculum in conjunction with a television-based prevention program, to prevent the onset of smoking and to promote smoking cessation. The original study involved 6695 students in 47 schools in Southern California, by Flay et al. (1995). The dataset analyzed here consists of a subset of 1,600 seventh-grade students from 135 classes in 28 schools in Los Angeles. The outcome is measured by a tobacco and health knowledge scale (THKS), ranging from zero to seven, assessing a student's knowledge of associations between tobacco and health. A higher score means a better understanding of tobacco associations with health. This study uses a two-by-two factorial design, with four intervention conditions determined by the cross-classification of a classroom-based social-resistance curriculum (CC: coded 1 = yes, 0 = no) and a television-based prevention program (TV: coded 1 = yes. 0 = no). The baseline covariate is the baseline THKS. I fitted a marginal model:

$$\text{logit}\,\text{prob}[Y_{ij} \leq g \mid Z_{ij}] = \alpha_g + \beta_1 \text{CC}_{ij} + \beta_2 \text{TV}_{ij} + \beta_3 \text{CC}_{ij} \times \text{TV}_{ij} + \beta_4 \text{baseline}\,\text{THKS}_{ij}\,;$$

Table 6.13 lists the point estimates and 95% robust confidence intervals of odds ratio corresponding to the above covariates. I assumed correlation structures among schools and classes are both exchangeable. Likelihood inference is not applicable to this dataset since there are three clusters having more than 100 observations and R or MATLAB does not have enough memory to deal with such big clusters.

All the following covariate effects are evaluated in the form of cumulative log odds ratios. Here I listed the result from my proposed method, where the outcomes are from after-study THKS categories $g_0 = 0, \ldots, 7$.

I defined the reference group as the group without any interventions. Compared to the reference group, the cumulative log odds ratio of classroom-based smoking prevention group is -0.87 (-1.09, -0.66); this implies controlling for the baseline score, the odds of achieving a THKS lower than g_0 for students from the classroom-based smoking prevention arm is 58% (48%, 66%) lower than their counterparts in the reference group. There is statistically significant improvement in students THKS with classroom-based smoking prevention.

Compared to the reference group, the cumulative log odds ratio of Television-based prevention group is -0.23 (-0.62, 0.17); controlling for the baseline score, the odds of achieving a THKS lower than g_0 for students from the Television-based prevention arm is 21% lower than their counterparts in the reference group. However, the positive effect from Television-based prevention is not significant in this study.

Compared to the reference group, the cumulative log odds ratio of classroom-based smoking plus Television-based prevention group is -0.72 (-0.79 -0.66); this implies controlling for the baseline score, the odds of achieving a THKS lower than g_0 for students from the classroom-based smoking plus Television-based prevention arm is 51% (48%, 55%) lower than their counterparts in the reference group. There is statistically significant improvement in students THKS with classroom-based smoking prevention plus Television-based prevention.

Correlation of classes among the same school is small: 0.03 (0, 0.08), as compared to the correlation among students from the same class, 0.13 (0.04, 0.22). This is not surprising to see since social-resistance intervention is carried out in the unit of class so peer effects inside classes will affect the outcome.

6.3 Frailty Model Numerical Studies

6.3.1 Simulations

I conducted simulations to study the finite sample performance of the proposed hybrid EM algorithm under different censoring rates and correlation structures. Four simulation settings were considered, each based upon 1000 Monte Carlo datasets, in which each dataset contained 200 independent clusters, and cluster sizes varied from 5 to 7 with equal probabilities.

I put two covariates into the mean model: covariates Z_1 are normally distributed with mean 0 and standard deviation 0.5, and $Z_2 = 0.2 \times Z_1 + Z_0 - 0.3$, in which Z_0 is a Bernoulli variable being 1 with probability 0.3. Given the frailty W_{ij} and covariate $Z_{ij} = (Z_{ij1}, Z_{ij2})$, failure time T_{ij} was generated via:

$$S(T_{ij} = t \mid W_{ij} = w_{ij}, Z_{ij} = z_{ij}) = \exp\{-w_{ij}\Lambda_0(t)\exp(1.2 \times z_{ij1} + 2.5 \times z_{ij2})\} .$$

For each observation, its censoring time was the minimum between 10 and an exponential random

variable. This exponential random variable is i.i.d. across observations. Different means of exponential distributions were chosen to generate different censoring rates.

In the first two settings, I considered individual frailties being exchangeably correlated. In the first setting, censoring times were generated independently as the minimum of an exponential distribution with mean 3.64 and 10, resulting in a censoring rate around 40%. In the second setting, censoring times were generated in a similar way except now I adopted an exponential distribution with mean 0.59, resulting in a censoring rate around 75%. Results for finite-dimensional parameters are listed in Table 6.14. NPMLE results by ignoring any underlying correlations are also presented.

The AR(1) correlation structure has been widely used to model longitudinal data, such as in Liang and Zeger (1986); it is also widely used in spatial data analysis, as discussed by Gelfand and Vounatsou (2003) and Li and Lin (2006). I found it is also suitable for modeling correlation induced by common genetic factors in some family studies. For example, 100% of genetic material is shared by monozygotic twins, 50% is shared by parent-offspring pairs, 25% is shared by grandparent-offspring pairs, etc. In the last two settings with AR(1) correlation structures, the same pair of censoring rates were used. Results for finite-dimensional parameters were listed in Table 6.15. The standard error estimates were based upon Theorem 4.4.4 and (5.8).

When the censoring rate is lower, under both correlation structures, the biases are lower and the standard errors are smaller. Under the AR(1) correlation structure, estimates of ρ are slightly more biased than the exchangeable correlation structure case. However, the estimation performance of β estimates are similar. The empirical coverage rates of 95% confidence intervals are close to the nominal coverage rate. Comparing the NPMLE ignoring any underlying correlations to the NPMCLE proposed in this thesis, I found when the correlation level is larger, typically when $\rho \geq 0.5$, the Monte Carlo MSE is smaller in my NPMCLE estimators.

6.3.2 Real Data: Rats Study

In the Rats dataset from Mantel et al. (1977), three rats were chosen from each of 100 litters, one of which was treated with a drug while the other two served as controls. All mice were followed for tumor incidence in 2 years. A subject is censored if died from other causes.

I found my new frailty model is useful under this design where in each litter, some members

get the treatment while the others serve as controls. Conditioning on unobservable environmental or genetic factors, hazard ratio between individual rats in treatment and control groups is constantly proportional to each other over time. However, environmental or genetic factors could also affect survival distribution so marginally, this difference in hazard rate due to treatment is finally worn out as time passes. Thus it is reasonable to analyse the Rats dataset with this model. I ran the model with treatment indicator as the sole covariate, assuming an exchangeable correlation structure among rats from the same litter. Censoring rate of this study is approximately 75% .

I estimated the conditional hazard rate ratio comparing the treatment group to the control group as 2.56 (1.30, 5.02), which is also the marginal log failure odds ratio. Deterioration or tumorigenic effect of treatment on rat survival is statistically significant. I estimated the correlation between frailties to be 0.75 (0, 0.99). Correlation between individual frailties is high, indicating that there is strong litter effect such as common genetic factors.

6.3.3 Real Data: Lung Study

This is from the work in Loprinzi et al. (1994) and the study was carried out to determine whether descriptive information from a questionnaire could provide prognostic information that was independent from that already obtained by the patient's physician. In the dataset the failure event is death (time to death measured in days) and the censor event is lost-to-followup.

I assumed exchangeable correlation structures inside the institutions, putting variables: age (in years), gender and Karnofsky performance score (bad=0-good=100) rated by physicians into the model. I got the conditional hazard rate ratios as age: 1.030 (1.010, 1.050), sex: 0.455 (0.319, 0.648), Karnofsky: 0.973 (0.955, 0.992) and they are also the marginal log failure odds ratio. The correlation parameter was estimated to be 0.047 (0, 0.227).

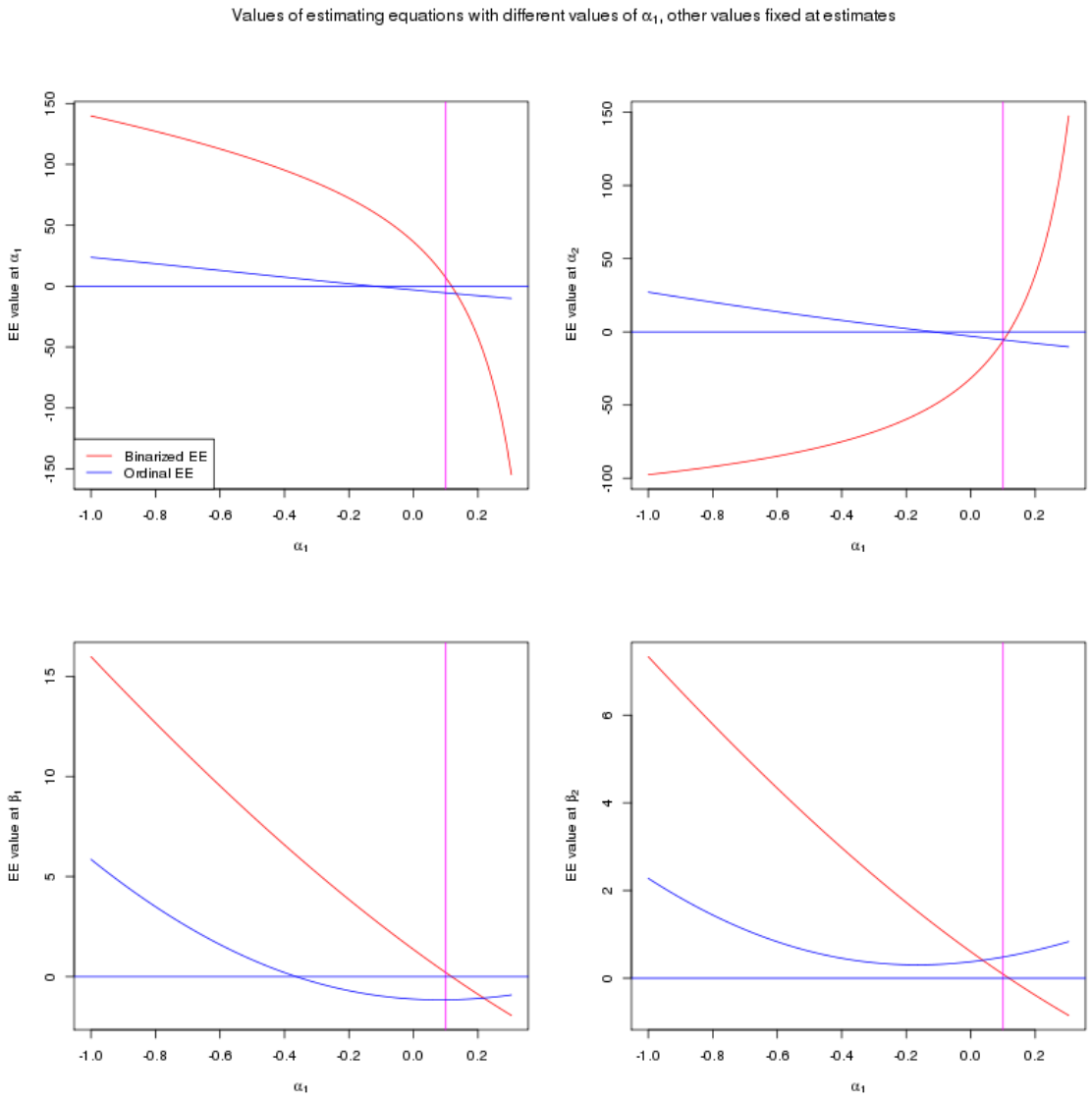

Figure 6.1: Fixing other parameters at estimated values, leaving α_1 varying. Red line is for (4.11) and blue line for (4.13).

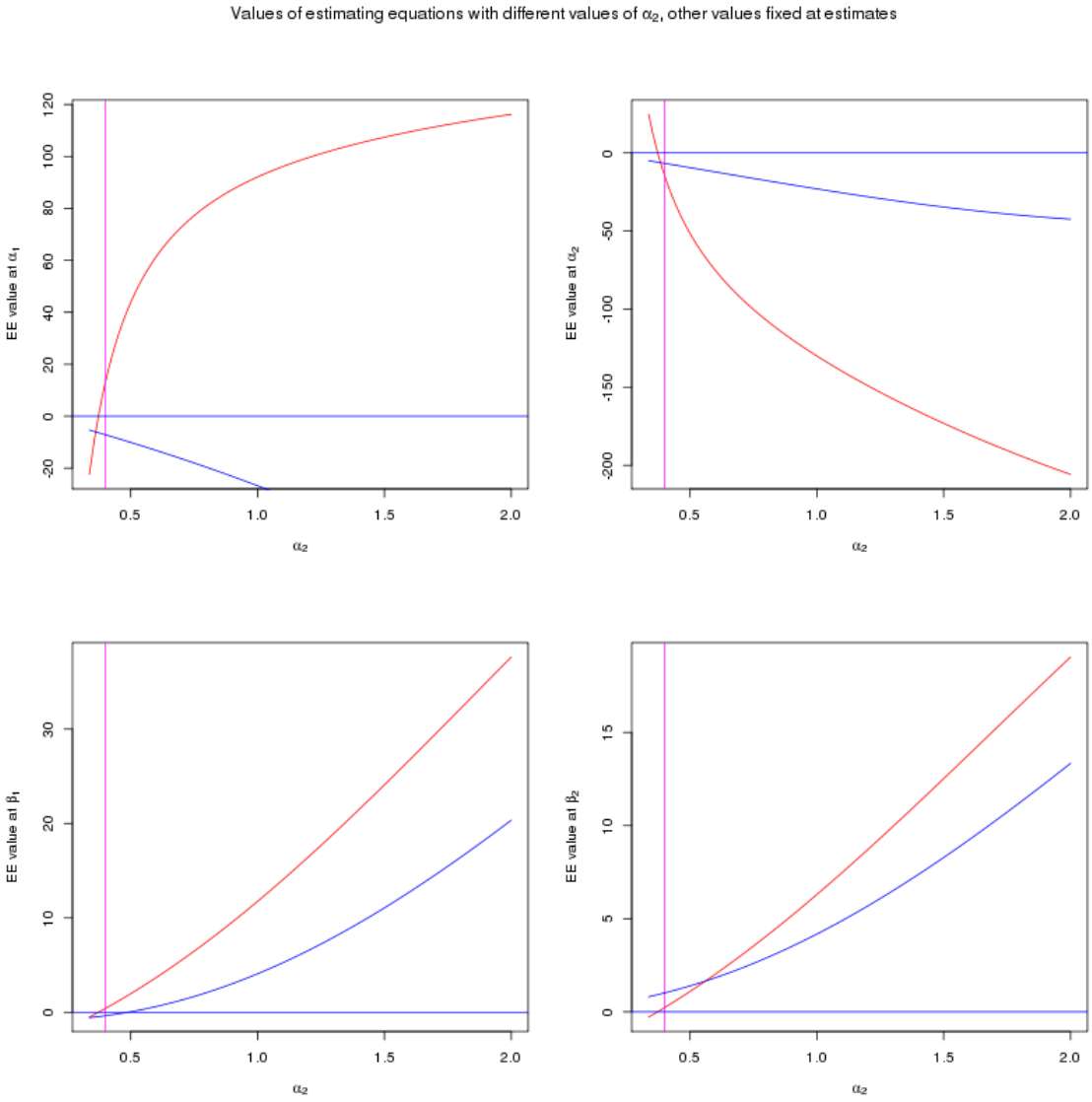

Figure 6.2: Fixing other parameters at estimated values, leaving α_2 varying. Red line is for (4.11) and blue line for (4.13). Data used for plotting comes from a Monte Carlo dataset following the working conditional model in (4.3).

Table 6.1: *Simulation results for estimating $(\beta_0, \beta_1, \rho_0)$ in two-level clustering in binary data, where ρ_0 is the correlation parameter of random effects. Bias represents the empirical bias, SSE represents the Monte Carlo standard error (s.e.), MSE is the mean squared error. SEE represents the averaged model-based s.e. estimates.*

(a) Two-level clustering, exchangeable correlation matrix.

		$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
		Our	MLE	Our	MLE	Our	MLE	Our	MLE	Our	MLE
β_0	Bias$\times 10^3$	-1	-1	-5	-5	-6	-6	-8	-9	-9	-8
	SEE$\times 10^3$	93	93	98	98	104	104	112	112	123	123
	SSE$\times 10^3$	92	92	103	103	106	106	114	113	121	119
	MSE$\times 10^3$	9	9	10	10	11	11	13	13	15	15
	C.I.	96%	96%	94%	94%	94%	94%	95%	94%	94%	95%
β_1	Bias$\times 10^3$	3	3	7	7	2	2	2	2	7	6
	SEE$\times 10^3$	71	71	71	71	71	71	72	72	74	67
	SSE$\times 10^3$	69	69	73	73	69	69	71	71	72	63
	MSE$\times 10^3$	5	5	5	5	5	5	5	5	6	6
	C.I.	96%	96%	95%	95%	96%	96%	95%	95%	97%	96%
ρ_0	Bias$\times 10^3$	6	2	-16	-16	-13	-12	-11	-9	-9	-10
	SEE$\times 10^3$	137	134	129	127	114	112	90	89	56	56
	SSE$\times 10^3$	99	86	128	119	118	115	92	88	59	50
	MSE$\times 10^3$	19	18	17	16	13	13	8	8	3	3
	C.I.	97%	96%	94%	96%	93%	94%	93%	94%	92%	92%
Time (in sec)		7	44	8	51	8	52	9	57	9	63

(b) Two-level clustering, AR(1) correlation matrix.

		$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
		Our	MLE	Our	MLE	Our	MLE	Our	MLE	Our	MLE
β_0	Bias$\times 10^3$	-4	-4	-9	-9	-7	-7	-9	-9	-8	-8
	SEE$\times 10^3$	92	92	94	94	98	98	104	104	117	117
	SSE$\times 10^3$	88	88	95	95	106	106	106	106	118	118
	MSE$\times 10^3$	8	8	9	9	10	10	11	11	14	14
	C.I.	96%	96%	94%	94%	94%	94%	95%	94%	95%	95%
β_1	Bias$\times 10^3$	6	6	11	11	6	6	9	9	6	6
	SEE$\times 10^3$	71	71	71	71	71	71	72	72	73	72
	SSE$\times 10^3$	72	72	74	74	72	72	73	73	71	70
	MSE$\times 10^3$	5	5	5	5	5	5	5	5	5	5
	C.I.	95%	95%	94%	94%	95%	95%	95%	95%	96%	97%
ρ_0	Bias$\times 10^3$	30	36	-24	-16	-21	-16	-16	-11	-7	-8
	SEE$\times 10^3$	201	201	181	180	142	139	93	91	42	38
	SSE$\times 10^3$	132	133	148	152	139	137	94	94	45	44
	MSE$\times 10^3$	42	42	33	32	21	20	9	8	2	2
	C.I.	92%	91%	95%	94%	94%	94%	94%	94%	93%	94%
Time (in sec)		7	80	9	66	9	60	9	58	9	57

Table 6.2: *Simulation results for estimating $(\beta_0, \beta_1, \rho_2, \rho_3)$ in a three-level clustering in binary data. I assumed the exchangeable correlation structure in the first level of clustering (larger clusters) and AR(1) on the second, and (ρ_2, ρ_3) represents the true correlations in the second and the third clustering levels. Bias, SSE, SEE, MSE represent the same quantities as in Table 6.1. 95% confidence interval coverage rates are presented, derived from model based s.e. estimates.*

	ρ_2	0.1				0.3			0.5		0.7
	ρ_3	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.1	0.3	0.1
β_0	Bias$\times 10^3$	-2.93	-1.78	-3.95	-1.09	-0.208	-0.814	-4.76	-6.19	-0.669	-5.28
	SEE$\times 10^3$	84	86	88.5	92.1	88.8	90.9	94.3	96.1	98.8	105
	SSE$\times 10^3$	86.4	86.4	87.1	91.4	90.9	90.3	94.3	97.2	103	106
	MSE$\times 10^3$	7.06	7.4	7.84	8.48	7.89	8.26	8.91	9.28	9.75	11
	C.I.	94%	95%	95%	96%	94%	95%	95%	94%	94%	95%
β_1	Bias$\times 10^3$	4.45	-0.251	4.34	1.73	-0.594	-1.52	3.24	3.01	0.53	6.15
	SEE$\times 10^3$	121	120	120	118	120	120	119	119	118	118
	SSE$\times 10^3$	123	127	122	125	118	122	123	121	120	118
	MSE$\times 10^3$	14.6	14.4	14.3	14	14.4	14.3	14.1	14.2	14	14
	C.I.	94.4%	92.7%	95.4%	94%	96%	95%	94%	95%	95%	96%
ρ_2	Bias$\times 10^3$	19	46.6	58.7	68.1	-40.4	-14.7	-6.3	-44	-17.2	-29.4
	SEE$\times 10^3$	146	147	154	165	135	137	146	115	119	92.4
	SSE$\times 10^3$	80.3	98.6	103	107	112	126	131	113	119	84.1
	MSE$\times 10^3$	21.6	23.6	27.1	32	19.8	19	21.4	15.2	14.5	9.4
	C.I.	99%	96%	95%	96%	99%	97%	96%	96%	95%	97%
ρ_3	Bias$\times 10^3$	59.1	-50	-104	-101	76.7	-12.7	-36	57	-16.6	28.6
	SEE$\times 10^3$	221	206	194	185	190	175	166	149	137	109
	SSE$\times 10^3$	115	141	151	135	121	151	157	111	133	84.8
	MSE$\times 10^3$	52.3	44.8	48.5	44.4	41.8	30.9	28.9	25.5	19.1	12.6
	C.I.	99%	98%	94%	94%	98%	97%	94%	97%	95%	97%
Time (in sec)		37.1	26.1	23.5	24.3	37.5	23.2	10.6	21.1	9.82	29.3

Table 6.3: *Simulation results for estimating (β_0, β_1) in a two-level clustering setting with a mis-specified joint model but a correct marginal model in binary data. Bias, SSE, SEE, MSE represent the same quantities as in Table 6.1. Two SEE's are presented, one is model-based while the other is robust. 95% confidence interval coverage rates are presented, derived by model based s.e. and robust s.e., respectively.*

Two-level clustering, mis-specified conditional model.														
Method	Bias $\times 10^3$		SEE $\times 10^3$		Robust SEE $\times 10^3$		SSE $\times 10^3$		MSE $\times 10^3$		95% C.I. coverage rate		Robust 95% C.I. coverage rate	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
Proposed	9	16	133	150	153	151	155	148	24	22	91.1%	95.5%	94.8%	95.9%
MLE	-308	201	144	164	-100	-100	159	153	121	64	42.0%	81.4%	-	-

Table 6.4: *Analysis of Madras longitudinal schizophrenia study. Covariate effects estimates and the corresponding 95% confidence intervals are listed in the scale of marginal odds ratioes. Different types of correlation structures are assumed over time.*

	Exchangeable		AR (1)	
Coefficients	$\exp(\beta)$	95% C.I.	$\exp(\beta)$	95% C.I.
		Likelihood		
Intercept	2.29	(1.44, 3.66)	2.31	(1.46, 3.67)
Time	0.70	(0.66, 0.75)	0.70	(0.66, 0.75)
Age > 20	-	-	-	-
Age ≤ 20	1.50	(0.83, 2.72)	1.28	(0.72, 2.28)
Male	-	-	-	-
Female	0.43	(0.24, 0.79)	0.47	(0.27, 0.82)
ρ	0.94	(0.84, 0.98)	0.95	(0.92,0.98)
		Proposed Method		
Intercept	2.41	(1.40, 4.10)	2.47	(1.45, 4.20)
Time	0.71	(0.65, 0.77)	0.71	(0.65, 0.77)
Age > 20	-	-	-	-
Age ≤ 20	1.60	(0.79, 3.32)	1.50	(0.74, 3.03)
Male	-	-	-	-
Female	0.53	(0.27, 1.01)	0.53	(0.28, 1.01)
ρ	0.92	(0.71, 0.98)	0.97	(0.93, 1.00)

Table 6.5: *Analysis of British Social Attitudes Panel Survey: years 1983-1986. Covariate effects estimates and the corresponding 95% confidence intervals are listed in the scale of marginal odds ratioes. Method 1 is my proposed method using the exchangeable-exchangle correlation structure assumption; Method 2 is my proposed method under the exchangeable-AR(1) correlation structure assumption; Method 3 is the GEE with exchangeable correlation structure, ignoring correlations in the same individual over years.*

	Method 1		Method 2		Method 3	
Coefficients	$\exp(\beta)$	95% C.I.	$\exp(\beta)$	95% C.I.	$\exp(\beta)$	95% C.I.
Intercept	0.61	(0.23, 1.63)	0.62	(0.24, 1.61)	0.74	(0.22, 2.43)
Year 1983	-	-	-	-	-	-
Year 1984	0.66	(0.49, 0.88)	0.65	(0.48, 0.88)	0.65	(0.47, 0.91)
Year 1985	1.06	(0.80, 1.41)	1.05	(0.78, 1.40)	1.04	(0.74, 1.46)
Year 1986	1.21	(0.91, 1.61)	1.20	(0.90, 1.61)	1.20	(0.88, 1.63)
Class: middle working	-	-	-	-	-	-
Class: upper working	0.76	(0.51, 1.14)	0.75	(0.50, 1.13)	0.72	(0.41, 1.24)
Class: lower working	0.80	(0.54, 1.19)	0.78	(0.52, 1.16)	0.66	(0.43, 1.02)
Gender: male	-	-	-	-	-	-
Gender: female	0.72	(0.48, 1.07)	0.72	(0.49, 1.07)	0.71	(0.45, 1.11)
Religion: protestant	-	-	-	-	-	-
Religion: catholic	0.67	(0.25, 1.80)	0.67	(0.26, 1.76)	0.76	(0.30, 1.91)
Religion: other	0.52	(0.24, 1.11)	0.52	(0.25, 1.08)	0.45	(0.23, 0.87)
Religion: none	2.00	(1.21, 3.30)	2.02	(1.23, 3.29)	2.12	(1.13, 3.97)
% protestant	2.17	(0.86, 5.52)	2.19	(0.88, 5.48)	1.94	(0.70, 5.41)

Table 6.6: *Simulation results of estimating (β_0, β_1, ρ) in two-level clustering with an exchangeable correlation structure in ordinal data. ρ_0 is the correlation parameter of random effects. Bias represents the empirical bias, SSE represents the Monte Carlo standard error, MSE is defined as the summation of square SSE and Bias. SEE represents the averaged model-based standard error estimate. 95% confidence interval coverage rates are presented in rows with name "C.I.", derived from model based s.e..*

		$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
		Our	MLE	Our	MLE	Our	MLE	Our	MLE	Our	MLE
	Bias$\times 10^3$	< 1	2	-2	1	< 1	1	-1	< 1	-2	2
	SEE$\times 10^3$	59	58	67	67	74	74	82	80	89	86
α_1	SSE$\times 10^3$	60	58	67	68	75	74	82	80	90	85
	MSE$\times 10^3$	4	3	5	5	6	5	7	6	8	7
	C.I.	94%	95%	95%	95%	95%	95%	95%	95%	95%	95%
	Bias$\times 10^3$	-1	3	-1	1	< 1	1	< 1	3	< 1	2
	SEE$\times 10^3$	60	59	67	67	74	74	81	80	88	86
α_2	SSE$\times 10^3$	60	59	70	69	74	74	80	81	88	84
	MSE$\times 10^3$	4	3	5	5	6	5	6	7	8	7
	C.I.	95%	96%	93%	94%	96%	96%	95%	95%	95%	95%
	Bias$\times 10^3$	< 1	2	< 1	3	< 1	1	< 1	5	1	2
	SEE$\times 10^3$	62	61	69	69	75	75	82	81	88	86
α_3	SSE$\times 10^3$	62	61	70	70	74	75	81	79	87	85
	MSE$\times 10^3$	4	4	5	5	6	6	7	6	8	7
	C.I.	95%	95%	94%	95%	96%	94%	96%	95%	95%	95%
	Bias$\times 10^3$	1	5	2	2	1	2	2	4	2	5
	SEE$\times 10^3$	64	63	71	71	77	77	83	82	89	88
α_4	SSE$\times 10^3$	64	63	72	73	76	80	82	81	88	87
	MSE$\times 10^3$	4	4	5	5	6	6	7	7	8	8
	C.I.	95%	96%	94%	95%	95%	94%	95%	95%	95%	95%
	Bias$\times 10^3$	1	1	-1	-2	-2	-2	1	-1	2	3
	SEE$\times 10^3$	82	82	81	81	80	79	78	77	77	74
β_1	SSE$\times 10^3$	85	81	84	82	83	82	80	78	78	76
	MSE$\times 10^3$	7	7	7	7	7	7	6	6	6	6
	C.I.	94%	96%	95%	94%	94%	94%	94%	94%	95%	95%
	Bias$\times 10^3$	6	4	3	9	4	9	9	9	8	8
	SEE$\times 10^3$	120	120	118	118	116	116	114	113	111	108
β_2	SSE$\times 10^3$	120	121	123	120	118	118	119	116	114	112
	MSE$\times 10^3$	14	15	15	14	14	14	14	14	13	13
	C.I.	95%	95%	94%	95%	95%	96%	94%	94%	95%	94%
	Bias$\times 10^3$	< 1	-25	-5	-3	-6	-5	-7	-8	-10	-6
	SEE$\times 10^3$	61	61	69	68	71	69	68	65	61	54
ρ	SSE$\times 10^3$	56	72	69	70	72	68	69	65	58	54
	MSE$\times 10^3$	3	6	5	5	5	5	5	4	3	3
	C.I.	98%	98%	94%	94%	94%	96%	94%	94%	97%	95%
Time (in sec)		27	565	32	753	30	867	32	1009	27	1096

Table 6.7: *Simulation results of estimating (β_0, β_1, ρ) in two-level clustering with an exchangeable correlation structure in ordinal data. ρ_0 is the correlation parameter of random effects. Composite likelihood inference on ρ is based on binarized outcomes. Bias represents the empirical bias, SSE represents the Monte Carlo standard error, MSE is defined as the summation of square SSE and Bias. SEE represents the averaged model-based standard error estimate. 95% confidence interval coverage rates are presented in rows with name "C.I.", derived from model based s.e..*

		$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.9$
α_1	Bias$\times 10^3$	< 1	-2	< 1	-1	-2
	SEE$\times 10^3$	59	67	74	82	89
	SSE$\times 10^3$	60	67	75	82	90
	MSE$\times 10^3$	4	5	6	7	8
	C.I.	94%	95%	95%	95%	96%
α_2	Bias$\times 10^3$	-1	-1	< 1	< 1	< 1
	SEE$\times 10^3$	60	67	74	81	88
	SSE$\times 10^3$	60	70	74	80	88
	MSE$\times 10^3$	4	5	6	6	8
	C.I.	95%	93%	96%	96%	95%
α_3	Bias$\times 10^3$	< 1	< 1	0	< 1	1
	SEE$\times 10^3$	62	69	75	82	88
	SSE$\times 10^3$	62	70	74	81	87
	MSE$\times 10^3$	4	5	6	7	8
	C.I.	95%	94%	96%	96%	95%
α_4	Bias$\times 10^3$	1	2	1	2	2
	SEE$\times 10^3$	64	71	77	83	89
	SSE$\times 10^3$	64	72	76	82	88
	MSE$\times 10^3$	4	5	6	7	8
	C.I.	95%	94%	95%	95%	95%
β_1	Bias$\times 10^3$	1	-1	-2	1	2
	SEE$\times 10^3$	82	81	80	78	77
	SSE$\times 10^3$	85	84	83	80	78
	MSE$\times 10^3$	7	7	7	6	6
	C.I.	94%	95%	94%	94%	95%
β_2	Bias$\times 10^3$	6	3	4	8	8
	SEE$\times 10^3$	120	118	116	114	111
	SSE$\times 10^3$	120	123	118	119	114
	MSE$\times 10^3$	14	15	14	14	13
	C.I.	95%	94%	95%	94%	94%
ρ	Bias$\times 10^3$	-2	-5	-6	-7	-9
	SEE$\times 10^3$	64	72	75	73	66
	SSE$\times 10^3$	61	72	75	74	66
	MSE$\times 10^3$	4	5	6	5	4
	C.I.	98%	94%	94%	94%	95%
Time (in sec)		273	486	422	468	281

Table 6.8: *Simulation results of estimating ρ in two-level clustering under an exchangeable correlation structure in ordinal data. Bias represents the empirical bias, SSE represents the Monte Carlo standard error, MSE is defined as the summation of square SSE and Bias, comparing between the composite inference on original ordinal data and the composite inference on binarized data.*

	Bias$\times 10^3$		SEE$\times 10^3$		SSE$\times 10^3$		MSE$\times 10^3$		95% C.I. coverage		Computation Time	
	O*	B*	O	B	O	B	O	B	O	B	O	B
$\rho_0 = 0.1$	< 1	-2	61	64	56	61	3	4	98%	98%	27	273
$\rho_0 = 0.3$	-5	-5	69	72	69	72	5	5	94%	94%	32	486
$\rho_0 = 0.5$	-6	-6	71	75	72	75	5	6	94%	94%	30	422
$\rho_0 = 0.7$	-7	-7	68	73	69	74	5	5	94%	94%	32	468
$\rho_0 = 0.9$	-10	-9	61	66	58	66	3	4	97%	95%	27	281

*: O stands for the composite likelihood using ordinal data and B stands for the composite likelihood using binarized data.

Table 6.9: *Simulation results of estimating (β_0, β_1, ρ) in two-level clustering with an AR(1) correlation structure in ordinal data. ρ_0 is the correlation parameter of random effects. Bias represents the empirical bias, SSE represents the Monte Carlo standard error, MSE is defined as the summation of square SSE and Bias. SEE represents the averaged model-based standard error estimate. 95% confidence interval coverage rates are presented in rows with name "C.I.", derived from model based s.e..*

		$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
		Our	MLE	Our	MLE	Our	MLE	Our	MLE	Our	MLE
	Bias$\times 10^3$	-1	-155	< 1	-165	-1	-173	< 1	-149	-1	12
	SEE$\times 10^3$	57	74	59	76	64	76	71	77	83	83
α_1	SSE$\times 10^3$	59	83	58	79	64	85	70	99	83	161
	MSE$\times 10^3$	3	31	3	34	4	37	5	32	7	26
	C.I.	94%	46%	94%	42%	95%	37%	94%	50%	95%	66%
	Bias$\times 10^3$	< 1	-117	-1	-125	1	-129	< 1	-103	-1	9
	SEE$\times 10^3$	57	74	60	75	64	76	71	76	82	84
α_2	SSE$\times 10^3$	60	81	61	80	64	85	71	99	83	132
	MSE$\times 10^3$	4	20	4	22	4	24	5	21	7	17
	C.I.	94%	62%	94%	59%	95%	58%	95%	66%	95%	78%
	Bias$\times 10^3$	< 1	-84	< 1	-91	2	-92	2	-64	2	10
	SEE$\times 10^3$	59	75	62	76	66	76	72	77	83	85
α_3	SSE$\times 10^3$	61	79	64	85	65	87	72	99	82	111
	MSE$\times 10^3$	4	13	4	15	4	16	5	14	7	12
	C.I.	94%	78%	94%	74%	95%	73%	95%	78%	95%	86%
	Bias$\times 10^3$	1	-58	1	-62	3	-63	2	-33	3	8
	SEE$\times 10^3$	62	76	64	77	68	78	74	78	84	87
α_4	SSE$\times 10^3$	63	79	66	85	66	87	75	101	83	99
	MSE$\times 10^3$	4	10	4	11	4	12	6	11	7	10
	C.I.	95%	88%	95%	84%	96%	84%	95%	85%	95%	90%
	Bias$\times 10^3$	2	-200	-2	-209	1	-200	-2	-182	-1	-107
	SEE$\times 10^3$	82	78	82	77	81	77	80	77	77	71
β_1	SSE$\times 10^3$	82	88	84	76	80	77	81	79	83	85
	MSE$\times 10^3$	7	48	7	50	6	46	7	39	7	19
	C.I.	95%	25%	94%	22%	96%	28%	94%	34%	94%	65%
	Bias$\times 10^3$	4	-200	3	-221	4	-205	5	-181	5	-106
	SEE$\times 10^3$	120	115	119	114	118	114	116	115	113	106
β_2	SSE$\times 10^3$	124	125	121	120	118	115	118	124	114	131
	MSE$\times 10^3$	15	56	15	63	14	55	14	48	13	28
	C.I.	94%	55%	95%	49%	95%	56%	94%	61%	95%	77%
ρ	Bias$\times 10^3$	12	607	-8	450	-10	262	-11	79	-5	45
	Time	42	775	23	866	23	890	23	996	23	951

Table 6.10: *Simulation results of estimating (β_0, β_1, ρ) in three-level clustering with exchangeable correlation structures on both levels in ordinal data. ρ_2 and ρ_2 are the correlation parameter of random effects on the first and second levels respectively. Bias represents the empirical bias, SSE represents the Monte Carlo standard error, MSE is defined as the summation of square SSE and Bias. SEE represents the averaged model-based standard error estimate. 95% confidence interval coverage rates are presented in rows with name "C.I.", derived from model based s.e..*

	ρ_1	0.1				0.3			0.5		0.7
	ρ_2	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.1	0.3	0.1
	Bias$\times 10^3$	1	1	< 1	-1	1	1	1	1	< 1	2
	SEE$\times 10^3$	58	60	63	66	66	68	71	74	76	82
α_1	SSE$\times 10^3$	56	60	63	65	66	68	71	75	77	83
	MSE$\times 10^3$	3	4	4	4	4	5	5	6	6	7
	C.I.	95%	95%	95%	96%	95%	95%	94%	94%	95%	95%
	Bias$\times 10^3$	1	2	2	1	1	2	1	1	< 1	3
	SEE$\times 10^3$	59	61	63	66	66	69	71	74	76	81
α_2	SSE$\times 10^3$	58	60	62	65	65	69	71	75	74	83
	MSE$\times 10^3$	3	4	4	4	4	5	5	6	6	7
	C.I.	95%	94%	95%	95%	95%	94%	95%	94%	95%	95%
	Bias$\times 10^3$	3	1	< 1	1	1	1	1	2	1	1
	SEE$\times 10^3$	81	81	80	79	80	80	79	79	79	78
β_1	SSE$\times 10^3$	82	81	80	79	83	80	80	79	76	77
	MSE$\times 10^3$	7	7	6	6	7	6	6	6	6	6
	C.I.	95%	95%	95%	95%	95%	96%	95%	96%	96%	95%
	Bias$\times 10^3$	6	1	4	4	7	5	2	9	7	11
	SEE$\times 10^3$	116	115	114	112	115	114	112	113	112	111
β_2	SSE$\times 10^3$	116	117	117	112	113	110	112	111	113	113
	MSE$\times 10^3$	13	14	14	12	13	12	13	12	13	13
	C.I.	94%	94%	94%	96%	95%	96%	95%	95%	95%	95%
	Bias$\times 10^3$	-3	4	5	6	-11	-8	-9	-10	-9	-7
	SEE$\times 10^3$	68	73	81	88	76	82	88	78	84	75
ρ_2	SSE$\times 10^3$	57	64	65	70	75	81	91	78	82	73
	MSE$\times 10^3$	3	4	4	5	6	7	8	6	7	5
	C.I.	99%	98%	99%	98%	96%	94%	94%	94%	96%	95%
	Bias$\times 10^3$	21	-10	-5	-3	19	2	1	13	3	7
	SEE$\times 10^3$	111	111	110	109	99	100	101	86	88	72
ρ_3	SSE$\times 10^3$	81	106	105	105	76	98	104	72	91	64
	MSE$\times 10^3$	7	11	11	11	6	10	11	5	8	4
	C.I.	99%	96%	97%	97%	98%	96%	94%	97%	93%	98%
	Time	80	60	73	87	79	37	37	60	42	57

Table 6.11: *Simulation results of estimating (β_0, β_1, ρ) in two-level clustering with a misspecified conditional model in ordinal data. Bias represents the empirical bias, SSE represents the Monte Carlo standard error, MSE is defined as the summation of square SSE and Bias. SEE represents the averaged model-based standard error estimate. robust SEE is the averaged robust standard error estimate. 95% confidence interval coverage rates are presented in rows names "C.I.", derived from model based s.e. and robust s.e..*

	Bias$\times 10^3$		SEE$\times 10^3$		Robust SEE$\times 10^3$		SSE$\times 10^3$		MSE$\times 10^3$		C.I.		robust C.I.	
	Our	MLE	Our	MLE	Our	MLE	Our	MLE	Our	MLE	Our	MLE	Our	MLE
α_1	3	-1233	74	124	129	-	131	247	17	1581	72%	0%	95%	-
α_2	5	-975	75	116	131	-	134	243	18	1009	73%	0%	95%	-
α_3	6	-717	76	110	137	-	140	236	20	570	71%	3%	94%	-
α_4	7	-455	78	106	145	-	150	233	22	261	70%	13%	94%	-
β_1	5	670	76	80	89	-	93	134	7	467	92%	0%	94%	-
β_2	3	789	110	117	118	-	117	167	12	651	95%	0%	95%	-

Table 6.12: *Analysis of the Arthritis study. Covariate effect estimates and corresponding 95% confidence intervals are presented in the marginal cumulative odds ratio form. Exchangeable correlation strcture was assumed.*

	Proposed		MLE	
Coefficients	$\exp(\beta)$	95% C.I.	$\exp(\beta)$	95% C.I.
Category 1 intercept	-0.26	(-1.55, 1.03)	-0.49	(-1.48, 0.5)
Category 2 intercept	1.88	(0.61, 3.15)	1.66	(0.68, 2.64)
Category 3 intercept	3.84	(2.53, 5.15)	3.77	(2.77, 4.78)
Category 4 intercept	6.06	(4.66, 7.45)	6.09	(5.03, 7.16)
Category 4 intercept	-	-	–	-
Baseline score	-0.89	(-1.11, -0.67)	-0.88	(-1.05, -0.71)
Age	0.01	(0, 0.03)	0.02	(0, 0.03)
Gender: female	-	-	–	-
Gender: male	-0.21	(-0.57, 0.14)	-0.28	(-0.57, 0)
Treatment	-0.55	(-0.88, -0.22)	-0.61	(-0.84, -0.37)
Time (in months)	-0.08	(-0.14, -0.03)	-0.10	(-0.16, -0.03)
Correlation	0.84	(0.76, 0.92)	0.97	(0.56,1.00)

Table 6.13: *Analysis of Television, School and Family Smoking Prevention and Cessation Project. Covariate effect estimates and corresponding 95% confidence intervals are presented in the marginal cumulative odds ratio form.*

	Proposed Method	
Coefficients	β	95% C.I.
Category 0 intercept	-1.55	(-1.84, -1.26)
Category 1 intercept	0.39	(0.11, 0.67)
Category 2 intercept	1.61	(1.31, 1.92)
Category 3 intercept	2.77	(2.44, 3.11)
Category 4 intercept	4.17	(3.8, 4.55)
Category 5 intercept	5.85	(5.34, 6.36)
Category 6 intercept	7.97	(6.85, 9.08)
Category 7 intercept	-	-
Baseline THKS	-0.42	(-0.50, -0.34)
Classroom-based social-resistance curriculum (CC)	-0.87	(-1.09, -0.66)
Television-based prevention (TV)	-0.23	(-0.62, 0.17)
Intervention interaction (CC$\times$TV)	0.37	(-0.16, 0.90)
School level correlation	0.03	(0, 0.08)
Classroom level correlation	0.13	(0.04, 0.22)

Table 6.14: *Simulation results of estimating (β_0, β_1, ρ) in different scenarios. ρ_0 is the true correlation parameter of frailties, under an exchangeable correlation structure. Bias represents the empirical bias. SEE represents the averaged model-based standard error estimates. SSE represents the Monte Carlo standard error, MSE is the summation of squared SSE and squared Bias. "C" and "I" represents NPMLE results using composite pairwise likelihood and likelihood assuming indepdent observations. Coverage probabilities for 95% confidence intervals are presented.*

(a) Two-level clustering, 40% censoring rate.

		$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
		C	I	C	I	C	I	C	I	C	I
β_0	Bias$\times 10^3$	-3	< 1	-1	-1	5	2	2	< 1	1	-4
	SEE$\times 10^3$	90	90	90	90	89	91	87	91	84	93
	SSE$\times 10^3$	94	91	92	91	89	94	86	94	86	95
	MSE$\times 10^3$	8	8	8	8	8	8	8	8	7	9
	C.I.	94%	95%	93%	95%	95%	94%	95%	94%	94%	94%
β_1	Bias$\times 10^3$	-4	-6	-4	-4	-3	-6	< 1	-5	-3	-5
	SEE$\times 10^3$	136	136	135	136	135	137	133	138	130	142
	SSE$\times 10^3$	139	137	133	136	133	135	134	140	133	145
	MSE$\times 10^3$	18	18	18	18	18	19	18	19	17	20
	C.I.	94%	95%	95%	95%	95%	96%	95%	95%	94%	94%
ρ	Bias$\times 10^3$	2	-	-5	-	-5	-	-3	-	-4	-
	SEE$\times 10^3$	74	-	78	-	75	-	64	-	44	-
	SSE$\times 10^3$	68	-	80	-	74	-	65	-	43	-
	MSE$\times 10^3$	5	-	6	-	5	-	4	-	2	-
	C.I.	97%	-	95%	-	94%	-	93%	-	93%	-

(b) Two-level clustering, 75% censoring rate.

		$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
		C	I	C	I	C	I	C	I	C	I
β_0	Bias$\times 10^3$	7	1	3	-4	2	-1	-2	1	2	-3
	SEE$\times 10^3$	121	121	121	121	119	121	119	121	117	122
	SSE$\times 10^3$	128	121	126	119	121	122	120	124	122	122
	MSE$\times 10^3$	15	15	15	15	14	15	14	15	14	15
	C.I.	93%	95%	95%	95%	94%	94%	96%	95%	94%	95%
β_1	Bias$\times 10^3$	-3	< 1	< 1	< 1	1	-7	2	-2	2	2
	SEE$\times 10^3$	172	172	172	172	170	172	170	173	169	175
	SSE$\times 10^3$	174	182	174	175	175	168	172	172	171	181
	MSE$\times 10^3$	30	29	29	30	29	30	29	30	29	31
	C.I.	96%	93%	95%	94%	94%	96%	95%	95%	96%	95%
ρ	Bias$\times 10^3$	13	-	-10	-	-13	-	-8	-	-15	-
	SEE$\times 10^3$	136	-	138	-	136	-	125	-	105	-
	SSE$\times 10^3$	111	-	139	-	136	-	127	-	98	-
	MSE$\times 10^3$	13	-	19	-	19	-	16	-	10	-
	C.I.	96%	-	95%	-	94%	-	93%	-	96%	-

Table 6.15: *Simulation results of estimating (β_0, β_1, ρ) in different scenarios. ρ_0 is the true correlation parameter of frailties, under an AR(1) correlation structure. Bias represents the empirical bias. SEE represents the averaged model-based standard error estimates. SSE represents the Monte Carlo standard error, MSE is the summation of squared SSE and squared Bias. "C" and "I" represents NPMLE results using composite pairwise likelihood and likelihood assuming indepdent observations. Coverage probabilities for 95% confidence intervals are presented.*

(a) Two-level clustering, 40% censoring rate.

		$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
		C	I	C	I	C	I	C	I	C	I
β_0	Bias$\times 10^3$	< 1	< 1	2	-3	-2	-3	1	-4	2	-3
	SEE$\times 10^3$	90	90	90	90	89	90	88	91	85	92
	SSE$\times 10^3$	91	95	90	90	90	91	89	94	86	94
	MSE$\times 10^3$	8	8	8	8	8	8	8	8	7	8
	C.I.	95%	94%	94%	94%	94%	94%	94%	94%	96%	94%
β_1	Bias$\times 10^3$	-2	-4	-3	-8	-5	-5	-2	-6	< 1	-8
	SEE$\times 10^3$	136	135	136	136	135	136	134	137	131	140
	SSE$\times 10^3$	139	139	138	142	133	142	136	139	132	140
	MSE$\times 10^3$	19	18	18	18	18	19	18	19	17	20
	C.I.	95%	94%	94%	94%	95%	93%	94%	94%	94%	96%
ρ	Bias$\times 10^3$	7	-	-5	-	-7	-	-6	-	-3	-
	SEE$\times 10^3$	116	-	102	-	83	-	60	-	31	-
	SSE$\times 10^3$	94	-	102	-	83	-	61	-	30	-
	MSE$\times 10^3$	9	-	11	-	7	-	4	-	1	-
	C.I.	96%	-	95%	-	95%	-	94%	-	93%	-

(b) Two-level clustering, 75% censoring rate.

		$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
		C	I	C	I	C	I	C	I	C	I
β_0	Bias$\times 10^3$	1	-1	-3	-4	-5	-6	-6	-6	-1	-6
	SEE$\times 10^3$	121	121	120	121	120	121	120	121	118	121
	SSE$\times 10^3$	121	123	122	123	121	123	124	125	120	123
	MSE$\times 10^3$	15	15	15	15	14	15	14	15	14	15
	C.I.	95%	95%	95%	94%	94%	95%	94%	94%	94%	95%
β_1	Bias$\times 10^3$	1	1	-2	3	1	2	-6	-5	-2	2
	SEE$\times 10^3$	172	172	172	172	172	172	171	173	170	174
	SSE$\times 10^3$	178	176	180	181	175	180	172	173	173	176
	MSE$\times 10^3$	29	30	29	30	29	30	29	30	29	30
	C.I.	94%	95%	95%	95%	94%	93%	95%	95%	95%	95%
ρ	Bias$\times 10^3$	34	-	-12	-	-31	-	-21	-	-10	-
	SEE$\times 10^3$	206	-	193	-	159	-	117	-	63	-
	SSE$\times 10^3$	147	-	178	-	162	-	125	-	65	-
	MSE$\times 10^3$	23	-	32	-	27	-	16	-	4	-
	C.I.	92%	-	93%	-	93%	-	94%	-	95%	-

Chapter 7

CONCLUDING REMARKS AND DISCUSSIONS

7.1 Potential Generalizations of the Models

Suppose frailties are exponentially distributed with a variance γ^{-2}, the new model for clustered binary data is still marginalizable:

$$\mathrm{pr}(Y_{ij}=1 \mid Z_{ij}=z_{ij}) = \frac{1}{\gamma}\int_0^\infty \exp\left(-w_{ij}e^{-z_{ij}^T\beta} - \frac{w_{ij}}{\gamma}\right) dw_{ij} = \frac{1}{\gamma}\frac{1}{e^{-z_{ij}^T\beta}+\frac{1}{\gamma}} = \frac{1}{e^{-z_{ij}^T\beta+log\gamma}+1}\,.$$

$\log(\gamma)$ is aliased with the intercept in the marginal probability, implying γ is not identifiable marginally. Moreover, γ is not identifiable in the joint likelihood.

To derive joint likelihood, I consider this conditional probability first:

$$\begin{aligned}
& f(Y_1 \mid Z_1, Z_2, \ldots, Z_n, W_1, W_2, \ldots, W_n) \\
= \quad & \int_{Y_n} \cdots \int_{Y_2} f(Y_1, Y_2, \ldots, Y_n \mid Z_1, Z_2, \ldots, Z_n, W_1, W_2, \ldots, W_n) dY_2, \ldots, dY_n \quad (7.1) \\
= \quad & \int_{Y_n} \cdots \int_{Y_2} \prod_{j=1}^{n} f(Y_j \mid Z_j, W_j) dY_2, \ldots, dY_n \quad (7.2) \\
= \quad & f(Y_1 \mid Z_1, W_1)\,.
\end{aligned}$$

where the equality between (7.1) and (7.2) is due to (3.3). Thus, the joint probability is

$$\begin{aligned}
& \mathrm{pr}(Y_{i1}=0, Y_{i2}=1, \ldots, Y_{in_i}=1 \mid \boldsymbol{Z}_i = \boldsymbol{z}_i) \\
= \quad & \mathrm{pr}(Y_{i2}=1, \ldots, Y_{in_i}=1 \mid \boldsymbol{z}_i) - \mathrm{pr}(Y_{i1}=1, \ldots, Y_{in_i}=1 \mid \boldsymbol{z}_i) \\
= \quad & |I + \Gamma_{-1}\mathrm{diag}(\gamma e^{-z_{i2}^T\beta}, \ldots, \gamma e^{-z_{in_i}^T\beta})|^{-1} - |I + \Gamma \mathrm{diag}(\gamma e^{-z_{i1}^T\beta}, \ldots, \gamma e^{-z_{in_i}^T\beta})|^{-1}\,,
\end{aligned}$$

where Γ_{-1} is the element-wise square root of the correlation matrix among $(W_2, \ldots, W_{n_i})$ and Γ is the element-wise square root of the correlation matrix among $(W_1, \ldots, W_{n_i})$. Therefore, $\log(\gamma)$ merges with the intercept in joint probabilities as well, and the variance of the random effect cannot be separately estimated from the intercept.

In view of this identifiability problem, I standardize the random effect distribution having variance one.

The proposed method for marginal inference is unaffected when the frailty distribution is covariate-dependent. Suppose conditional on a covariate Z_{ij}, W_{ij} is exponential distributed with mean $e^{Z_{ij}^T\gamma}$, and a frailty vector W_i conditional $\boldsymbol{Z}_i$ has a correlation matrix R. Consider the re-scaled frailty vector $\tilde{W}_i := (e^{-Z_{i1}^T\gamma}W_{i1}, \ldots, e^{-Z_{in_i}^T\gamma}W_{in_i})$, which is multivariate exponential with mean one and has the same correlation matrix R, the conditional probability of the binary outcome follows from (3.2), i.e.

$$\mathrm{pr}(Y_{ij}=1 \mid Z_{ij}=z_{ij}, w_{ij}) = \exp\left(-w_{ij}e^{-z_{ij}^T\beta}\right) = \exp\left(-\tilde{w}_{ij}e^{-z_{ij}^T\tilde{\beta}}\right), \quad \text{where } \tilde{\beta}=\beta-\gamma\ .$$

The marginal probability (3.4) becomes

$$\mathrm{pr}(Y_{ij}=1 \mid Z_{ij}=z_{ij}) = \frac{e^{z_{ij}^T\tilde{\beta}}}{1+e^{z_{ij}^T\tilde{\beta}}}\ .$$

For marginal regression parameter inference, $\tilde{\beta}$ is estimated as if the random effects were covariate independent.

As for the new ordinal data model and the new frailty model, arguments are very similar and thus omitted.

7.2 Discussions of the New Models: Correlation Modeling

7.2.1 Limited Correlation Level

In conventional linear and logistic mixed effects models, the within-cluster correlation is controlled by the variance of some shared random effects, such as the model in Ten Have (1996) and the shared Gamma frailty model. While the variance of frailties is standardized in my models, flexible within-cluster correlation structure modeling can be done by correlated random effects, when the correlation level is not high.

I examine the covariance between any two correlated binary observations:

$$\begin{aligned}
\mathrm{cov}(Y_{ij}, Y_{ik}) &= \left[(1-\rho_{jk})e^{(z_{ij}+z_{ik})^T\beta} + e^{z_{ij}^T\beta} + e^{z_{ik}^T\beta} + 1\right]^{-1} - \left(1+e^{z_{ij}^T\beta}\right)^{-1}\left(1+e^{z_{ik}^T\beta}\right)^{-1} \\
&= \begin{cases} \left[e^{z_{ij}^T\beta} + e^{z_{ik}^T\beta} + 1\right]^{-1} - \left(1+e^{z_{ij}^T\beta}\right)^{-1}\left(1+e^{z_{ik}^T\beta}\right)^{-1} & \rho_{jk}=1 \\ 0 & \rho_{jk}=0 \end{cases}
\end{aligned}$$

where $\rho_{jk} := \text{cor}(W_{ij}, W_{ik})$, $i = 1, \ldots, m$.

When $\rho_{jk} = 0$, these two observations are independent of each other; when $\rho_{jk} = 1$, they share a random intercept. In theory, the largest covariance achievable by a pair of binary variables (Y_{ij}, Y_{ik}) with marginal means $(1/(1 + e^{-z_{ij}^T\beta}), 1/(1 + e^{-z_{ik}^T\beta}))$ is

$$\text{cov}(1-Y_{ij}, 1-Y_{ik}) = \left(1 + e^{z_{ij}^T\beta}\right)^{-1} \wedge \left(1 + e^{z_{ik}^T\beta}\right)^{-1} - \left(1 + e^{z_{ij}^T\beta}\right)^{-1} \left(1 + e^{z_{ik}^T\beta}\right)^{-1}; \quad \text{i.e. Fréchet bound.}$$

Even at $\rho_{jk} = 1$, the upper limit of Fréchet bound is not achievable in my model for binary data.

The correlation between survival times from the new frailty model is also limited. To evaluate correlation levels, I consider the conditional hazard ratio, proposed by Clayton (1978b)

$$\theta(t_1, t_2) := \frac{\lambda_{T_1|T_2}(t_1 \mid T_2 = t_2)}{\lambda_{T_1|T_2}(t_1 \mid T_2 > t_2)} ,$$

which can be viewed as some generalized odds ratio.

In the shared Gamma frailty model, frailties follow the Gamma distribution of mean one and unknown variance $1/\gamma$ to be estimated:

$$\theta(t_1, t_2) = (\gamma + 1)/\gamma .$$

This measure is time-invariant; ranging in $(1, +\infty)$; this measure increases as γ decreases.

In the shared positive stable frailty model, frailties follow the positive stable distribution of some parameter $\alpha \in (0, 1]$ to be estimated:

$$\theta(t_1, t_2) = 1 + \frac{1-\alpha}{\alpha} \left(\Lambda_0(t_1)e^{z_1^T\beta} + \Lambda_0(t_2)e^{z_2^T\beta}\right)^{-\alpha} .$$

It implies the correlation diminishes as time passes. This measure ranges in $(1, +\infty)$ and the correlation increases as $\alpha \to 0$.

In my frailty model, frailties follow the multivariate Exponential distribution of mean one and correlation matrix $R(\rho)$:

$$\theta(t_1, t_2) = 1 + \frac{\rho}{[(1-\rho)\Lambda_0(t_1)e^{z_1^T\beta} + 1][(1-\rho)\Lambda_0(t_2)e^{z_2^T\beta} + 1]} .$$

The correlation also diminishes as time passes. This measure ranges in $(1, 2)$ and the correlation increases as $\rho \to 1$. Thus, the most correlated case of this frailty model is equivalent to the shared Gamma frailty model in which the frailty follows the standard Gamma distribution.

7.2.2 Detecting Correlations

One interesting hypothesis regarding these new models is: $\rho = 0$; i.e. observations are independent. Since in my models, ρ is restricted to be non-negative, 0 is a boundary point. Thus, only score test based on composite likelihood works in this case. Future research can be done.

7.2.3 Alternative Modeling of Correlations

Heagerty and Zeger (1996) modeled binary variable correlations by their pairwise odds ratios, which, compared to correlation ρ modeling in this thesis, are unconstrained and covariate effects on pairwise odds ratio are readily intepretable. However, ρ can model correlations between three or even more observations while pairwise odds ratio only focuses on paired observations; generalizing pairwise odds ratio is possible but requires a balanced design. And in modeling complex correlation structures pairwise odds, ratios are not as flexible as random effects correlations.

7.2.4 Remarks

I can choose not to fix frailties having variance one and instead incorporating individual Gamma frailties. However, the marginal model is no longer conveniently interpretable and there is an identifiability problem, as discussed by Coull et al. (2006).

7.3 Model Inference Characteristics

All inference methods have the same characteristic. Taking the inference method for binary data model for example, I split the parameters into two sets: marginal parameters β and correlation parameters ρ. For each set, I propose a separate inference method and the overall inference is carried out by:

1. starting with some initial value of marginal parameter $\beta^{(0)}$ and correlation parameter $\rho^{(0)}$,

2. doing inference on β with plug-in and fixed $\rho^{(0)}$, getting $\beta^{(1)}$,

3. doing inference on ρ with plug-in and fixed $\beta^{(1)}$, getting $\rho^{(1)}$,

4. repeating Steps 2 and 3 until the L^{th} iteration such that $|| \beta^{(L)} - \beta^{(L-1)} ||< \epsilon$ and $|| \rho^{(L)} - \rho^{(L-1)} ||< \epsilon$, where ϵ is a pre-specified small positive number, usually set up as 10^{-5};

giving the estimates $(\beta^{(L)}, \rho^{(L)})$.

I suggest starting with the GLM estimates of β, denoted by $\beta^{(0)}$, and carrying out Step 3 to solve for $\rho^{(0)}$. A heuristic argument for this suggestion is available in the end of Appendix A.

7.4 Model applicability to longitudinal data

Pepe and Anderson (1994) pointed out several marginal model inference methods, such as GEE, GEE1 and GEE2, need the following assumption when modeling longitudinal data:

$$\mathrm{E}[Y_{ij} \mid \boldsymbol{Z}_i] = \mathrm{E}[Y_{ij} \mid Z_{ij}] \,. \tag{7.3}$$

And in this thesis, I also build my models on this assumption:

$$\begin{aligned}
& f(Y_1 \mid Z_1, Z_2, \ldots, Z_n) \\
= \quad & \int_{W_n} \cdots \int_{W_1} f(Y_1 \mid Z_1, Z_2, \ldots, Z_n, W_1, W_2, \ldots, W_n) f(W_1, \ldots, W_n) dW_1, \ldots, dW_n \\
= \quad & \int_{W_n} \cdots \int_{W_1} f(Y_1 \mid Z_1, W_1) f(W_1, \ldots, W_n) dW_1, \ldots, dW_n \\
= \quad & \int_{W_1} f(Y_1 \mid Z_1, W_1) f(W_1) dW_1 \\
= \quad & f(Y_1 \mid Z_1) \,.
\end{aligned}$$

Here I show my models for clustered binary and ordinal data satisfy this assumption. I denote the full data of a cluster as

$$(Y_1, Y_2, \ldots, Y_n),\ (Z_1, Z_2, \ldots, Z_n),\ (W_1, W_2, \ldots, W_n),\ .$$

So my parametric models for clustered binary and ordinal data satisfies the assumptions in (7.3).

7.5 Conclusions

In this thesis I introduce a set of new marginalizable mixed effects model for analysing clustered binary, ordinal and survival data.

As for the binary data and ordinal data models, a working generalized linear mixed effects model and a multivariate Gumbel random intercept distribution were proposed, which respectively yield a logistic regression model or a cumulative logit model that have population-level interpretations. Unlike most marginal models which only model the first and perhaps the second moments, I have come up with a parametric model, which guarantees there is always a real joint distribution for the marginal logistic regression model and parameters being estimated always exist. In contrast, one criticism of GEE is that there may not be any multivariate distribution with a correlation structure being equivalent to the working correlation structure of GEE. My proposed inference yields consistent estimates of marginal regression parameters even under mis-specified model, along with consistent estimates of standard deviations of the estimates.

For the new frailty model, a multivariate exponential distribution for individual frailties was proposed, which yields a marginal proportional odds model that has a population-level interpretation, and the model also allows flexible correlation structures among observations. For model inference, I maximize a composite marginal log-likelihood contribution. I do not choose to maximize the joint marginal log-likelihood contribution in the end, due to the computation complexity with large clusters and long computation in the simulations. I neither apply to the penalized log-likelihood method, because the density of a multivariate exponential random variable is intractable and a density is needed to construct a penalized term. The estimation efficiency is not optimal among all asymptotically linear estimators. However, by only specifying pairwise joint distribution, the inference method has some level of robustness in return, as pointed out by Varin et al. (2011).

The marginalization property is based on a standard exponential frailty assumption, which can be viewed as a special case of the Gamma frailty models considered in Coull et al. (2006). Exponential distributed frailties should not be considered as a limitation, since

1. a marginal interpretation is often desirable in practice;

2. an exponential distributed frailty is equivalent to a Gumbel random intercept which has physical interpretations. Gumbel distribution can model the distribution of maximum of normal or exponential type random variables, so Gumbel random intercept is reasonable when scientists believe there are many latent cluster effects and the maximum dominates the others; i.e. the random effect can be modeled as the maximum of many latent cluster effects;

3. the proposed robust estimation procedure would yield consistent estimates for marginal parameters even when the multivariate exponential frailty distribution or the conditional mean model is mis-specified;

4. marginal inference is un-affected when frailty distribution is covariate dependent.

BIBLIOGRAPHY

Albert, P. S. (2000). A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics*, 56(2):602–608.

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(2):pp. 203–210.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, 81(4):pp. 767–775.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2(2):273–277.

Bishop, Y., Feinberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis*. MIT Press.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):pp. 9–25.

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

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(4):pp. 851–860.

Chaganty, N. R. and Joe, H. (2006). Range of correlation matrices for dependent bernoulli random variables. *Biometrika*, 93(1):197–206.

Clayton, D. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society. Series A (General)*, 148(2):pp. 82–117.

Clayton, D. G. (1978a). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):pp. 141–151.

Clayton, D. G. (1978b). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):pp. 141–151.

Conaway, M. (1990). A random effects model for binary data. *Biometrics*, pages 317–328.

Coull, B. A., Houseman, E. A., and Betensky, R. A. (2006). A computationally tractable multivariate random effects model for clustered binary data. *Biometrika*, 93(3):pp. 587–599.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):pp. 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):pp. 269–276.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(1):pp. 1–39.

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, 42(4):pp. 909–917.

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):pp. 1–38.

Diggle, P., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall: London.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80(1):pp. 141–151.

Flay, B., Miller, T., Hedeker, D., Siddiqui, O., Britton, C., Brannon, B., Johnson, C., Hansen, W., Sussman, S., and Dent, C. (1995). The television, school, and family smoking prevention and cessation project: Viii. student outcomes and mediating variables. *Preventive Medicine*, 24(1):29 – 40.

Furman, E. and Landsman, Z. (2005). Risk capital decomposition for a multivariate dependent gamma portfolio. *Insurance: Mathematics and Economics*, 37(3):635–649.

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

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15.

Geman, S. and Geman, D. (1993). Stochastic relaxation, gibbs distributions and the bayesian restoration of images*. *Journal of Applied Statistics*, 20(5-6):25–62.

Gill, R. D. (1992). Marginal partial likelihood. *Scandinavian Journal of Statistics*, 19(2):pp. 133–137.

Gray, S. M. and Ron, B. (2000). Multidimensional longitudinal data: Estimating a treatment effect from continuous, discrete, or time-to-event response variables. *Journal of the American Statistical Association*, 95(450):pp. 396–406.

Hartzel, J., Agresti, A., and Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, 1(2):81–102.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):pp. 320–338.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, 55(3):688–698.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*, 58(2):pp. 342–351.

Heagerty, P. J. and Zeger, S. L. (1996). Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association*, 91(435):pp. 1024–1036.

Heagerty, P. J. and Zeger, S. L. (1998). Lorelogram: A regression approach to exploring dependence in longitudinal categorical responses. *Journal of the American Statistical Association*, 93(441):pp. 150–162.

Heagerty, P. J. and Zeger, S. L. (2000a). Marginalized multilevel models and likelihood inference. *Statistical Science*, 15(1):pp. 1–19.

Heagerty, P. J. and Zeger, S. L. (2000b). Multivariate continuation ratio models: Connections and caveats. *Biometrics*, 56(3):719–732.

Henderson, R. and Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika*, 90(2):pp. 355–366.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71(1):pp. 75–83.

Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika*, 73(3):pp. 671–678.

Klein, J. P. (1992). Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, 48(3):pp. 795–806.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer: New York.

Krishnamoorthy, A. S. and Parthasarathy, M. (1951). A multivariate gamma-type distribution. *The Annals of Mathematical Statistics*, 22(4):pp. 549–557.

Kuk, A. Y. C. (2007). A hybrid pairwise likelihood method. *Biometrika*, 94(4):939–952.

Lang, J. B. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, 89(426):pp. 625–632.

Li, Y. and Lin, X. (2006). Semiparametric normal transformation models for spatially correlated survival data. *Journal of the American Statistical Association*, 101(474):pp. 591–603.

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

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

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine*, 13(21):2233–2247.

Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):711–730.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80:pp. 221239.

Lipsitz, S. R., Kim, K., and Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, 13(11):1149–1163.

Liu, D., Kalbfleisch, J. D., and Schaubel, D. E. (2011). A positive stable frailty model for clustered failure time data with covariate-dependent frailty. *Biometrics*, 67(1):8–17.

Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., Bartel, J., Law, M., Bateman, M., and Klatt, N. E. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. north central cancer treatment group. *Journal of Clinical Oncology*, 12(3):601–7.

Mantel, N., Bohidar, N. R., and Ciminera, J. L. (1977). Mantel-haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information. *Cancer Research*, 37(11):3863–3868.

McCulloch, C. E., Searle, S. R., and M., N. J. (2008). *Generalized, linear, and mixed models*. Wiley.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate plackett distribution. *Journal of the American Statistical Association*, 89(426):pp. 633–644.

Murphy, S. A., Rossini, A. J., and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):pp. 968–976.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450):pp. 449–465.

Neuhaus, J., Kalbfleisch, J., and Hauck, W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review/Revue Internationale de Statistique*, pages 25–35.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, 19(1):pp. 25–43.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(3):pp. 414–422.

O'Brien, S. M. and Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics*, 60(3):pp. 739–746.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics*, 26(1):pp. 183–214.

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

Parzen, M., Ghosh, S., Lipsitz, S., Debajyoti, S., Fitzmaurice, G. M., Mallick, B. K., and Ibrahim, J. G. (2011). A generalized linear mixed model for longitudinal binary data with a marginal logit link function. *Ann Appl Stat*, 5(1):pp. 449467.

Patefield, W. M. (1977). On the maximized likelihood function. *Sankhy: The Indian Journal of Statistics, Series B (1960-2002)*, 39(1):pp. 92–96.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):pp. 545–554.

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(4):939–951.

Petersen, J. H. (1998). An additive frailty model for correlated life times. *Biometrics*, 54(2):pp. 646–661.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44(4):pp. 1033–1048.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47(3):pp. 825–839.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):pp. 1016–1022.

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

Ten Have, T. R. (1996). A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics*, 52(2):pp. 473–491.

Thara, R., Henrietta, M., Joseph, A., Rajkumar, S., and Eaton, W. W. (1994). Ten-year course of schizophreniathe madras longitudinal study. *Acta Psychiatrica Scandinavica*, 90(5):329–336.

Tsutakawa, R. K. (1988). Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association*, 83(401):pp. 37–42.

Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, 19(24):3309–3324.

van der Vaart, A. W. (1995). Efficiency. of infinite dimensional m- estimators. *Statistica Neerlandica*, 49(1):9–30.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

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

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):pp. 439–454.

Vu, H. T. and Knuiman, M. W. (2002). A hybrid ml-em algorithm for calculation of maximum likelihood estimates in semiparametric shared frailty models. *Computational Statistics & Data Analysis*, 40(1):173 – 187.

Wang, Z. and Louis, T. A. (2004). Marginalized binary mixed-effects models with covariate-dependent random effects and likelihood inference. *Biometrics*, 60(4):pp. 884–891.

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

Wolfinger, R. and O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48(3-4):233–243.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American Statistical Association*, 86(413):pp. 79–86.

Zeng, D., Lin, D., and Lin, X. (2008). Semiparametric transformation models with random effects for clustered failure time data. *Statistica Sinica*, pages pp. 355–377.

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

Appendix A

PROOF OF THEOREM 3.5.1

Here I list the conditions of Theorem 3.5.1 and proved it. The proofs of Theorem 3.5.2, Theorem 4.3.1 and Theorem 4.3.2 are quite similar and are omitted.

$$\text{I define } \Psi(\theta) = \begin{pmatrix} \Psi_1(\theta) \\ \Psi_2(\theta) \end{pmatrix} = \begin{pmatrix} \mathrm{E}\left\{ f_1(\boldsymbol{Z}, \boldsymbol{Y}; \theta) \right\} \\ \mathrm{E}\left\{ f_2(\boldsymbol{Z}, \boldsymbol{Y}; \theta) \right\} \end{pmatrix},$$

$$\text{and } \Psi_m(\theta) = \begin{pmatrix} \Psi_{1,m}(\theta) \\ \Psi_{2,m}(\theta) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{m} f_1(\boldsymbol{Z}_i, \boldsymbol{Y}_i; \theta) \\ \sum_{i=1}^{m} f_2(\boldsymbol{Z}_i, \boldsymbol{Y}_i; \theta) \end{pmatrix},$$

where f_1 and f_2 correspond to estimating equations in (3.7) and (3.9):

$$\begin{cases} f_1(\boldsymbol{Z}_i, \boldsymbol{Y}_i; \theta) := D(\boldsymbol{Z}_i; \beta)^T V^{-1}(\boldsymbol{Z}_i; \theta) S(\boldsymbol{Z}_i, \boldsymbol{Y}_i; \beta) \, , \\ f_2(\boldsymbol{Z}_i, \boldsymbol{Y}_i; \theta) := \sum_{j<k} \frac{\partial l_{jk}}{\partial \rho}(\boldsymbol{Z}_i, \boldsymbol{Y}_i; \theta) \, . \end{cases}$$

Theorem 3.5.1 is true under the following conditions:

C.1 Observations from different clusters are independent and identically distributed.

C.2 The number of observations per cluster is uniformly bounded.

C.3 The parameter space Θ is a convex and compact subset of $\mathbb{R}^p$ and the true value of parameter, θ_0, is not a boundary point of Θ.

C.4 The probability of covariate Z concentrating on a hyper-plane is zero. That is, $Z^T\beta = 0$ a.e. implies $\beta = 0$, and Z is bounded with probability one.

C.5 There is an unique root of β from $\Psi_1(\beta, \rho) = 0$ for all ρ.

C.6 The joint distribution is correctly specified.

Proof. First I point out that even though different clusters may contain different numbers of observations, joint observations from a cluster can still be viewed as i.i.d.

Each cluster in theory contains infinite subjects and are denoted by $(Z(\cdot), Y(\cdot), W(\cdot))$: $\cdot$ varies

with different subjects. The observed data from a cluster is a deterministic projection of $(Z(\cdot), Y(\cdot), W(\cdot))$. Assuming the stochastic process $(Z(\cdot), Y(\cdot), W(\cdot))$ are i.i.d. and the projection procedure is also i.i.d., I conclude observations from different clusters are i.i.d.. P_0 is denoted as the joint distribution.

Second, I argue that ρ_0 is the unique solution to $\Psi_2(\theta) = 0$ at $\beta = \beta_0$. This can be shown by the composite Kullback-Leibler divergence defined below.

The composite likelihood of the i^{th} cluster is $\prod_{j<k} L_{jk}(\boldsymbol{Z}_i, \boldsymbol{Y}_i; \theta)$.

The composite Kullback-Leibler divergence is defined as

$$\begin{aligned} & KL_{\text{composite}}(L_0, L_1) \\ := \quad & P_0 \log \left(\frac{\prod_{j<k} L_0(Z_j, Z_k, Y_j, Y_k; \beta_0, \rho_0)}{\prod_{j<k} L_1(Z_j, Z_k, Y_j, Y_k; \beta_0, \rho_1)} \right) \\ = \quad & P_0 \left[\sum_{j<k} \log \left(\frac{L_0(Z_j, Z_k, Y_j, Y_k; \beta_0, \rho_0)}{L_1(Z_j, Z_k, Y_j, Y_k; \beta_0, \rho_1)} \right) \right] > 0 \, . \end{aligned}$$

For more details, please refer to Lindsay (1988). The last strict inequality is due to Jensen's Inequality and the fact that $L_1 = L_0$ if and only if $\rho_1 = \rho_0$.

Thus ρ_0 is the unique value maximizing composite likelihood expectation with plug-in β_0. Since the model is smooth in parameters, $\Psi_2(\theta) = 0$ uniquely at $\rho = \rho_0$ when β is fixed at β_0.

Next, consider an index set $\mathcal{H} := \{h \in \mathbb{R}^p : ||h|| \leq 1; \}$ in which $||\cdot||$ is the Euclidean norm and p is the dimension of parameter $\theta = (\beta, \rho)$. Suppose $(\boldsymbol{Z}, \boldsymbol{Y})$ has distribution P_0. Then the function class:

$$\mathcal{F}_0 := \left\{ h^T (f_1(\boldsymbol{Z}, \boldsymbol{Y}; \theta), f_2(\boldsymbol{Z}, \boldsymbol{Y}; \theta)) : \theta \in \Theta, h \in \mathcal{H} \right\}$$

is P_0-Donsker.

Here is the argument. For an arbitrary pair of functions from $\mathcal{F}_0$:

$$\begin{aligned} & |h_1^T (f_1(\boldsymbol{Z}, \boldsymbol{Y}; \theta_1), f_2(\boldsymbol{Z}, \boldsymbol{Y}; \theta_1)) - h_2^T (f_1(\boldsymbol{Z}, \boldsymbol{Y}; \theta_2), f_2(\boldsymbol{Z}, \boldsymbol{Y}; \theta_2))| \\ \leq \quad & C_0 ||\theta_1 - \theta_2|| \cdot ||h_1 - h_2|| \, . \end{aligned}$$

Since everything in $h^T (f_1(\boldsymbol{Z}, \boldsymbol{Y}; \theta), f_2(\boldsymbol{Z}, \boldsymbol{Y}; \theta))$ is continuous in θ so Mean Value Theorem can be used; C_0 is some finite number by conditions C.2 and C.3. Since $\theta_1, \theta_2 \in \Theta$ and Θ is a com-

pact subset of Euclidean space, the number of brackets needed to cover $\mathcal{F}_0$ satisfies P_0-Donsker requirement, according to van der Vaart and Wellner (1996), page 129.

Now I can claim

$$\sup_{\theta\in\Theta, h\in\mathcal{H}} |h^T\Psi_m(\theta) - h^T\Psi(\theta)| \to 0,$$

implying that

$$\begin{aligned} & \sup_{h\in\mathcal{H}} |h^T \left[\Psi_m(\hat{\theta}_m) - \Psi(\hat{\theta}_m)\right]| \to 0 \ , \\ \text{i.e.} \qquad & \sup_{h\in\mathcal{H}} |h^T\Psi(\hat{\theta}_m)| \to 0 \ ; \quad \text{thus, } ||\Psi(\hat{\theta}_m)|| \to 0 \ . \end{aligned}$$

$(\Psi_1(\theta), \Psi_2(\theta)) = \mathbf{0}$ has an unique root at θ_0 implies $||\Psi(\hat{\theta}_m)|| \to ||\Psi(\theta_0)||$. Suppose $\hat{\theta}_m$ does not converge to θ_0, then there exists a sub-sequence $\{\hat{\theta}_{m'}\}$ of $\{\hat{\theta}_m\}$ such that $\hat{\theta}_{m'}$ converge to $\theta_1 (\neq \theta_0)$. Since $(\Psi_1(\theta), \Psi_2(\theta))$ are continuous in θ, $\Psi(\hat{\theta}_{m'})$ converges to $\Psi(\theta_1) \neq 0$ and thus contradicts. I have shown $\hat{\theta}_m \xrightarrow{p} \theta_0$, where the convergence in probability comes from the Glivenko-Cantelli class definition.

□

The proof of the weak convergence of $\sqrt{m}\left\{(\hat{\beta}_m - \beta_0)^T, (\hat{\rho}_m - \rho_0)^T\right\}^T$ makes use of Theorem 3.3.1 of van der Vaart and Wellner (1996), which is stated as the following.

Suppose there are two random mappings Ψ_m and Ψ such that $\Psi(\beta_0, \rho_0) = 0$ for some interior point $(\beta_0, \rho_0) \in \Theta$, $\Psi_m(\beta_m, \rho_m) \xrightarrow{p} 0$ for some random sequence $(\beta_m, \rho_m) \subset \Theta$, and assume the followings are true:

P.1 (β_m, ρ_m) is consistent for (β_0, ρ_0);

P.2 $\sqrt{m}\left(\Psi_m - \Psi\right)(\beta_0, \rho_0)$ converges in distribution to a tight random element Z;

P.3

$$\begin{aligned} & \sqrt{m}\left(\Psi_m - \Psi\right)(\beta_m, \rho_m) - \sqrt{m}\left(\Psi_m - \Psi\right)(\beta_0, \rho_0) \\ = \quad & o_p\left(1 + \sqrt{m}||\beta_m - \beta_0|| + \sqrt{m}||\rho_m - \rho_0||\right) \ ; \end{aligned}$$

P.4 $\Psi(\beta, \rho)$ is Fréchet differentiable at (β_0, ρ_0);

P.5 The derivative of $\Psi(\beta, \rho)$ at (β_0, ρ_0), denoted by $\dot{\Psi}(\beta_0, \rho_0)$, is continuously invertible.

Then

$$\sqrt{m}\left\{(\hat{\beta}_m - \beta_0)^T, (\hat{\rho}_m - \rho_0)^T\right\}^T \xrightarrow{d} -\dot{\Psi}(\beta_0, \rho_0)^{-1}(Z) \ .$$

Proof. Condition P.1 has been verified.

Since $\mathcal{F}_0$ is P_0-Donsker, condition P.2 is verified.

By P_0-Donsker preservation theorem 2.10.3 in van der Vaart and Wellner (1996), this function class

$$\left\{h^T\left[\{(f_1(\beta,\rho),f_2(\beta,\rho))-(f_1(\beta_0,\rho_0),f_2(\beta_0,\rho_0))\}\right]:(\beta,\rho)\in\Theta,h\in\mathcal{H}\right\}$$

is P_0-Donsker as well.

$$\begin{aligned}&\sup_{h\in\mathcal{H}}P_0\left(h^T\left[\{(f_1(\beta,\rho),f_2(\beta,\rho))-(f_1(\beta_0,\rho_0),f_2(\beta_0,\rho_0))\}\right]\right)^2\\ \leq\quad & P_0\left(C_0||\theta_0-\theta||\right)^2\to 0\quad \text{ as } ||(\beta,\rho)-(\beta_0,\rho_0)||\to 0\ .\end{aligned}$$

Therefore, according to Lemma 3.3.5 of van der Vaart and Wellner (1996), P.3 holds.

As for P.4, since Ψ is continuously differentiable in parameters, it is trivial to verify that $-E(B)$ is its Fréchet derivative at (β_0,ρ_0). Due to model identifiability, $E(B)$ is a negative definite matrix and thus continuously invertible. Therefore, P.5 is also satisfied. □

NOTE: I assume unique root exists in C.5. Yet generally this is not true globally. However, with a proper value plugged into the estimating equation to start the solving procedure, final solution from this estimating equation is consistent. That is to say, the estimating equation does not have a global positive/negative definite derivative matrix but has a local positive/negative definite derivative matrix. This can be shown as the following. Plugging $\hat{\beta}$, which is a consistent estimate of β, into the estimating equation gives

$$\sum_{i=1}^{m}D(\boldsymbol{Z}_i;\hat{\beta})V(\boldsymbol{Z}_i;\hat{\beta},\rho)(\boldsymbol{Y}_i-h(\boldsymbol{Z}_i^T\hat{\beta}))=\mathbf{0}\ .$$

Taking the derivative of β gives

$$\sum_{i=1}^{m}\left\{-D(\boldsymbol{Z}_i;\hat{\beta})V(\boldsymbol{Z}_i;\hat{\beta},\rho)D(\boldsymbol{Z}_i;\hat{\beta})+\left[\partial D(\boldsymbol{Z}_i;\beta)V(\boldsymbol{Z}_i;\beta,\rho)/\partial\beta|_{\hat{\beta}}\right](\boldsymbol{Y}_i-h(\boldsymbol{Z}_i^T\hat{\beta}))\right\}\ .$$

Since $\hat{\beta}$ is a consistent estimate of β, the second term in the above equation is negligible compared to the first term.

Appendix B

PROOF OF FRAILTY MODEL THEOREMS

In this appendix, I outline the proofs of the lemma and theorems for the new frailty model, under the following conditions:

C1. The covariate $\boldsymbol{Z}_i$ is independent of frailties $\boldsymbol{W}_i$ and its distribution is non-informative; i.e. it does not contain θ.

C2. (Coarsening at random assumption) Conditioning on $(\boldsymbol{Z}_i, \boldsymbol{T}_i, \boldsymbol{W}_i)$, the hazard rate function of censoring time C_{ij} is a bounded function of covariates $\boldsymbol{Z}_i$ and is non-informative.

C3. The true conditional baseline cumulative hazard function $\Lambda_0(t)$ is a strictly increasing function on $[0, \tau]$ and is continuously differentiable. In addition, $\Lambda_0(0) = 0$ and $\Lambda_0'(0) > 0$.

C4. Parameter spaces of β and ρ, denoted by $\mathcal{B}$ and $\mathcal{R}$, belong to some known convex and compact subsets of $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$, respectively:

$$\begin{aligned} \mathcal{B} &:= \{\beta \in \mathbb{R}^{p_1} : ||\beta|| \leq B_0 \text{ for some finite constant } B_0\} \ , \\ \mathcal{R} &:= \{\rho \in \mathbb{R}^{p_2} : ||\rho|| \leq R_0 \text{ for some finite constant } R_0\} \ , \end{aligned}$$

where $||\cdot||$ denotes the Euclidean norm; the true value (β_0, ρ_0) is not a boundary point of $\mathcal{B} \times \mathcal{R}$.

C5. The probability of covariate Z concentrating on a hyper-plane is zero. That is, $Z^T\beta = 0$ a.e. implies $\beta = 0$, and Z is bounded with probability one.

C6. There exists some strictly positive constant a_0 such that

$$\text{pr}(C_{ij} \geq \tau \mid \boldsymbol{Z}_i) = \text{pr}(C_{ij} = \tau \mid \boldsymbol{Z}_i) \geq a_0 \quad \text{a.s.} ;$$

C7. The covariate Z is bounded.

I denote the whole observations from each cluster by $\boldsymbol{O}$. Let $\mathbb{P}_m$ be the empirical measure of m i.i.d. cluster observations: $\boldsymbol{O}_1, \ldots, \boldsymbol{O}_m$; denote P_0 as the expectation of cluster observations. That is, for any measurable function $g(\boldsymbol{O})$, I define

$$\mathbb{P}_m[g(\boldsymbol{O})] = \frac{1}{m}\sum_{i=1}^{m} g(\boldsymbol{O}_i); \quad P_0[g(\boldsymbol{O})] = \text{E}[g(\boldsymbol{O})] \ .$$

Even though different clusters may contain different numbers of observations, I can still view joint observations from each cluster as i.i.d. samples, by the same reasoning from Appendix A.

B.1 Lemma 5.2.1

Proof. Suppose there is a pair of parameter values $(\Lambda_0, \beta_0, \rho_0)$ and $(\Lambda_1, \beta_1, \rho_1)$ such that the marginal likelihood contributions for any pair of dependent observations, (O_j, O_k), $j \neq k$ under these two sets of parameters are identical.

Parameters Λ_0 and Λ_1 come from the function class $\mathcal{L}$

$$\mathcal{L} = \{\Lambda(\cdot) : \text{ a non-decreasing step càdlàg function in } [0,\tau] \text{ with jumps at observed failure time points and } \Lambda(0) = 0\} .$$

For a pair of observations $(y_j, \delta_j = 1, z_j)$ and $(y_k, \delta_k = 1, z_k)$, taking ratio of their marginal likelihood contributions under these two sets of parameters gives

$$\begin{aligned}
&\frac{L(y_j, y_k, z_j, z_k, \delta_j = \delta_k = 1; \beta_0, \rho_0, \Lambda_0)}{L(y_j, y_k, z_j, z_k, \delta_j = \delta_k = 1; \beta_1, \rho_1, \Lambda_1)} \\
= \; &\frac{(1-\rho_{0jk})^2 \Lambda_0(y_j) e^{z_j^T\beta_0} \Lambda_0(y_k) e^{z_k^T\beta_0} + (1-\rho_{0jk}) \Lambda_0(y_j) e^{z_j^T\beta_0} + (1-\rho_{0jk}) \Lambda_0(y_k) e^{z_k^T\beta_0} + 1}{(1-\rho_{1jk})^2 \Lambda_1(y_j) e^{z_j^T\beta_1} \Lambda_1(y_k) e^{z_k^T\beta_1} + (1-\rho_{1jk}) \Lambda_1(y_j) e^{z_j^T\beta_1} + (1-\rho_{1jk}) \Lambda_1(y_k) e^{z_k^T\beta_1} + 1} \\
&\cdot \left(\frac{(1-\rho_{1jk}) \Lambda_1(y_j) e^{z_j^T\beta_1} \Lambda_1(y_k) e^{z_k^T\beta_1} + \Lambda_1(y_j) e^{z_j^T\beta_1} + \Lambda_1(y_k) e^{z_k^T\beta_1} + 1}{(1-\rho_{0jk}) \Lambda_0(y_j) e^{z_j^T\beta_0} \Lambda_0(y_k) e^{z_k^T\beta_0} + \Lambda_0(y_j) e^{z_j^T\beta_0} + \Lambda_0(y_k) e^{z_k^T\beta_0} + 1} \right)^3 \\
&\cdot \frac{d\Lambda_0(y_j) d\Lambda_0(y_k)}{d\Lambda_1(y_j) d\Lambda_1(y_k)} \exp\left((z_j + z_k)^T (\beta_0 - \beta_1) \right) = 1 . \qquad \text{(B.1)}
\end{aligned}$$

This is true for any observations (y_j, z_j, y_k, z_k) such that $\mathrm{pr}(C_j \geq y_j, C_k \geq y_k \mid \boldsymbol{z}) > 0$. I consider two monotone decreasing sequences $\{y_{jr} : r = 1, 2, \ldots\}$ and $\{y_{kr} : r = 1, 2, \ldots\}$ such that

$$y_{jr} \downarrow 0, \quad y_{kr} \downarrow 0 \quad \text{as } r \to \infty .$$

(B.1) holds a.e. for every $\boldsymbol{z}$ from the set

$$A_q := \{\boldsymbol{z} : \mathrm{pr}\,[C_j \geq y_{jr}, C_k \geq y_{kr} \mid \boldsymbol{Z} = \boldsymbol{z}] > 0\} .$$

As $r \to \infty$, $A_r \uparrow A := \{\boldsymbol{z} : \mathrm{pr}\,[C_j \geq 0, C_k \geq 0 \mid \boldsymbol{Z} = \boldsymbol{z}] > 0\}$, which has probability one under $F_{\boldsymbol{Z}}$ by conditions C2 and C3. Thus

$$\lim_{q\to\infty} \left(\frac{L(y_{jq}, y_{kq}, z_{jq}, z_{kq}, \delta_j = \delta_k = 1; \Lambda_0, \beta_0, \rho_0)}{L(y_{jq}, y_{kq}, z_{jq}, z_{kq}, \delta_j = \delta_k = 1; \Lambda_1, \beta_1, \rho_1)} \right) = \left(\frac{d\Lambda_0(0)}{d\Lambda_1(0)} \right)^2 \exp\left\{ (z_j + z_k)^T (\beta_0 - \beta_1) \right\} = 1 .$$

By condition C5, $\beta_1 = \beta_0$. Consequently, $d\Lambda_1(0) = d\Lambda_0(0)$.

In the following I show $\Lambda_1(t) = \Lambda_0(t)$ for $\forall t \in [0, \tau]$.

Since each pair of the marginal likelihood contribution is identical, the marginal likelihood contribution for a single observation, which is an integration of the former one, should also be identical:

$$\frac{\mathrm{pr}(y_j, z_j, \delta_j = 1; \Lambda_0, \beta_0, \rho_0)}{\mathrm{pr}(y_j, z_j, \delta_j = 1; \Lambda_1, \beta_0, \rho_1)} = \frac{d\Lambda_0(y_j)}{\left(\Lambda_0(y_j)e^{z_j^T\beta_0} + 1\right)^2} \Bigg/ \frac{d\Lambda_1(y_j)}{\left(\Lambda_1(y_j)e^{z_j^T\beta_0} + 1\right)^2} = 1 \ .$$

Integrating from 0 to arbitrary $t \in (0, \tau]$ gives

$$\frac{1}{\Lambda_0(t)e^{z_j^T\beta_0} + 1} = \frac{1}{\Lambda_1(t)e^{z_j^T\beta_0} + 1} \ .$$

Thus $\Lambda_1(\cdot) = \Lambda_0(\cdot)$ on $[0, \tau]$. Since $\beta_1 = \beta_0$ and $\Lambda_1(\cdot) = \Lambda_0(\cdot)$, it is trivial to show $\rho_0 = \rho_1$.

Next, I connect the above conclusion to the composite Kullback-Leibler Distance, which is defined as the following

$$\begin{aligned}
& P_0\left[clog(O; \beta_0, \rho_0, \Lambda_0) - clog(O; \beta_1, \rho_1, \Lambda_1)\right] \\
= \quad & \mathrm{E}_n\left\{\frac{1}{n-1}\sum_{j<k} P_0\left[\log L(O_j, O_k; \beta_0, \rho_0, \Lambda_0) - \log L(O_j, O_k; \beta_1, \rho_1, \Lambda_1)\right] \mid n\right\} \quad \text{(B.2)} \\
\geq \quad & -\mathrm{E}_n\left\{\frac{1}{n-1}\sum_{j<k} \log \mathrm{E}_O\left[\frac{L(O_j, O_k; \beta_1, \rho_1, \Lambda_1)}{L(O_j, O_k; \beta_0, \rho_0, \Lambda_0)} \mid n\right]\right\} = 0 \ . \quad \text{(B.3)}
\end{aligned}$$

E_n in (B.2) and (B.3) denotes the expectation with respect to the distribution of the random cluster size n. Equality in (B.3) holds if and only if $(\beta_0, \rho_0, \Lambda_0) = (\beta_1, \rho_1, \Lambda_1)$. I have shown that similar to the Kullback-Leibler Distance, the composite Kullback-Leibler Distance is always non-negative and equals to zero if and only if at identical parameter sets.

As for the second part of the lemma, I show it by contradiction.

Suppose there exists some one-dimensional sub-model passing through the true parameters that has a singular composite Fisher information matrix; i.e., the weighted average of all Fisher information matrices corresponding to different pairs of observations. I denote this one-dimensional sub-model by $(\beta_0 + \epsilon h_1, \rho_0 + \epsilon h_2, \Lambda_0 + \epsilon \int h_3 d\Lambda_0)$, where $h := (h_1, h_2, h_3) \in \mathcal{H}$ such that

$$\mathcal{H} := \left\{(h_1, h_2, h_3) : h_1 \in \mathbb{R}^{d_1}, h_2 \in \mathbb{R}^{d_2}, h_3(t) \text{ is a càdlàg function on } [0, \tau]\right\} \ ,$$

equipped with the norm $||h||_{\mathcal{H}} := ||h_1|| + ||h_2|| + ||h_3||_V$.

In my case, the composite Fisher information matrix is a weighted summation of the Fisher information matrices corresponding to pairwise observations. So a singular composite Fisher information matrix implies every Fisher information matrices of pairwise observations is singular. In this submodel, correlated observations j and k have the score function:

$$\begin{aligned} S(O_j, O_k; \beta_0, \rho_0, \Lambda_0, h_1, h_2, h_3) &:= h_1^T clog_\beta(O_j, O_k; \beta_0, \rho_0, \Lambda_0) + h_{2jk} clog_{\rho_{jk}}(O_j, O_k; \beta_0, \rho_0, \Lambda_0) \\ &\quad + clog_\Lambda(O_j, O_k; \beta_0, \rho_0, \Lambda_0) \left[\int h_3 d\Lambda_0 \right]; \end{aligned} \tag{B.4}$$

in which

$$\begin{aligned} & h_1^T clog_\beta(O_j, O_k; \beta, \rho, \Lambda) \\ = \quad & h_1^T \left\{ \int_0^\tau Z_j dN_j(s) + \int_0^\tau Z_k dN_k(s) - \int_0^\tau A(u, O_j, O_k; \beta, \rho, \Lambda) d\Lambda(u) \right\}, \\ & h_{2jk} clog_{\rho_{jk}}(O_j, O_k; \beta, \rho, \Lambda) \\ = \quad & h_{2jk} \left\{ \frac{-2(1-\rho_{jk})\Delta_j \Delta_k e^{Z_j^T \beta} e^{Z_k^T \beta} \Lambda(Y_j)\Lambda(Y_k) - \Delta_k e^{Z_j^T \beta} \Lambda(Y_j) - \Delta_j e^{Z_k^T \beta} \Lambda(Y_k) + \Delta_j \Delta_k}{(1-\rho_{jk})^2 \Delta_j \Delta_k e^{Z_j^T \beta} e^{Z_k^T \beta} \Lambda(Y_j)\Lambda(Y_k) + (1-\rho_{jk})\Delta_k e^{Z_j^T \beta} \Lambda(Y_j) + (1-\rho_{jk})\Delta_j e^{Z_k^T \beta} \Lambda(Y_k) + 1 + \Delta_j \Delta_k \rho_{jk}} \right. \\ & \left. + (1 + \Delta_j + \Delta_k) \frac{e^{Z_j^T \beta} e^{Z_k^T \beta} \Lambda(Y_j)\Lambda(Y_k)}{(1-\rho_{jk}) e^{Z_j^T \beta} e^{Z_k^T \beta} \Lambda(Y_j)\Lambda(Y_k) + e^{Z_j^T \beta} \Lambda(Y_j) + e^{Z_k^T \beta} \Lambda(Y_k) + 1} \right\}, \\ & clog_\Lambda(O_j, O_k; \beta, \rho, \Lambda) \left[\int h_3 d\Lambda \right] \\ = \quad & \left(\int_0^\tau h_3(s) dN_j(s) + \int_0^\tau h_3(s) dN_k(s) \right) - \int_0^\tau D(u, O_j, O_k; \beta, \rho, \Lambda) h_3(u) d\Lambda(u), \end{aligned}$$

and I denote

$$\begin{aligned} u(o_j; \beta, \Lambda) &= \Lambda(y_j) e^{z_j^T \beta}, \\ v(o_j, o_k; \beta, \rho, \Lambda) &= (1-\rho_{jk}) u(o_j; \beta, \Lambda) u(o_k; \beta, \Lambda) + u(o_j; \beta, \Lambda) + u(o_k; \beta, \Lambda) + 1, \\ w(o_j, o_k; \beta, \rho, \Lambda) &= \delta_j \delta_k (1-\rho_{jk})^2 u(o_j; \beta, \Lambda) u(o_k; \beta, \Lambda) \\ &\quad + \delta_j (1-\rho_{jk}) u(o_k; \beta, \Lambda) + \delta_k (1-\rho_{jk}) u(o_j; \beta, \Lambda) + 1 + \delta_j \delta_k \rho_{jk}, \end{aligned}$$

$$
\begin{aligned}
& D(u, o_j, o_k; \beta, \rho, \Lambda) \\
:= \quad & (1+\delta_j+\delta_k)\frac{e^{z_j^T\beta}1\{y_j \geq u\}\left[1+(1-\rho_{jk})e^{z_k^T\beta}\Lambda(y_k)\right] + e^{z_k^T\beta}1\{y_k \geq u\}\left[1+(1-\rho_{jk})e^{z_j^T\beta}\Lambda(y_j)\right]}{v(o_j, o_k; \beta, \Lambda, \rho)} \\
& -\frac{\delta_k(1-\rho_{jk})e^{z_j^T\beta}1\{y_j \geq u\}\left[\delta_j(1-\rho_{jk})e^{Z_k^T\beta}\Lambda(Y_k)+1\right]}{w(o_j, o_k; \beta, \Lambda, \rho)} \\
& -\frac{\delta_j(1-\rho_{jk})e^{z_k^T\beta}1\{y_k \geq u\}\left[\delta_k(1-\rho_{jk})e^{z_j^T\beta}\Lambda(y_j)+1\right]}{w(o_j, o_k; \beta, \Lambda, \rho)}, \\
& A(u, o_j, o_k; \beta, \rho, \Lambda) \\
= \quad & z_j e^{z_j^T\beta}1\{y_j \geq u\}\left\{(1+\delta_j+\delta_k)\frac{1+(1-\rho_{jk})e^{z_k^T\beta}\Lambda(y_k)}{v(o_j, o_k; \beta, \Lambda, \rho)} - \frac{\delta_k(1-\rho_{jk})\left[\delta_j(1-\rho_{jk})e^{z_k^T\beta}\Lambda(y_k)+1\right]}{w(o_j, o_k; \beta, \Lambda, \rho)}\right\} \\
& + z_k e^{z_k^T\beta}1\{y_k \geq u\}\left\{(1+\delta_j+\delta_k)\frac{1+(1-\rho_{jk})e^{z_j^T\beta}\Lambda(y_j)}{v(o_j, o_k; \beta, \Lambda, \rho)} - \frac{\delta_j(1-\rho_{jk})\left[\delta_k(1-\rho_{jk})e^{z_j^T\beta}\Lambda(y_j)+1\right]}{w(o_j, o_k; \beta, \Lambda, \rho)}\right\}.
\end{aligned}
$$

The Fisher information matrix for such a sub-model is scalar, and equals to

$$\mathrm{E}[S(O_j, O_k; \beta_0, \rho_0, \Lambda_0)^2] .$$

If it is singular, then $S(O_j, O_k; \beta_0, \rho_0, \Lambda_0, h_1, h_2, h_3) = 0$ a.s.. Like in the previous part, I consider sequences of paired observations $\{y_{jr}, \delta_j = 1, z_j, y_{kr}, \delta_k = 1, z_k : r = 1, 2, \ldots\}$ such that

$$y_{jr} \downarrow 0 \quad \text{as } r \to \infty; \quad y_{kr} \downarrow 0 \quad \text{as } r \to \infty .$$

These paired observations are plausible under my model with covariates $\boldsymbol{z}$ such that $\mathrm{pr}(C_{jr} \geq y_j, C_{kr} \geq y_k \mid \boldsymbol{z}) > 0$. As $r \to \infty$, covariates $\boldsymbol{z}$ satisfying this condition span the whole covariate space. I notice this fact as well:

$$\lim_{r\to\infty} \mid \int_0^\tau D(u, o_{jr}, o_{kr}; \beta_0, \rho_0, \Lambda_0) h_3(u) d\Lambda_0(u) \mid \leq \lim_{r\to\infty} [y_{jq} \vee y_{kq}] \cdot M_3 \cdot \Lambda_0(y_{jq} \vee y_{kq}) \to 0 ,$$

where $M_3 < \infty$. Plugging paired observations (o_{jr}, o_{kr}) into (B.4) and taking its limit in r give

$$h_1^T (z_j + z_k) + h_{2jk}\frac{1}{1+\rho_{jk}} + 2h_3(0) = 0 . \tag{B.5}$$

Thus $h_1 = 0$, using the same argument in the first part.

I consider another pair of observations $o_j = (y_j = \tau, \delta_j = 0, z_j)$ and $o_k = (y_k = \tau, \delta_k = 0, z_k)$. According to condition C6, any covariate $\boldsymbol{z}$ will satisfy $\mathrm{pr}(C_j \geq \tau, C_k \geq \tau \mid \boldsymbol{z}) > 0$. Then I write

(B.4) as

$$h_{2jk}\frac{e^{z_j^T\beta_0}e^{z_k^T\beta_0}\Lambda_0(\tau)\Lambda_0(\tau)}{(1-\rho_{jk})e^{z_j^T\beta_0}e^{z_k^T\beta_0}\Lambda_0(\tau)\Lambda_0(\tau)+e^{z_j^T\beta_0}\Lambda_0(\tau)+e^{z_k^T\beta_0}\Lambda_0(\tau)+1}$$
$$-\frac{\left[1+(1-\rho_{jk})e^{z_k^T\beta_0}\Lambda_0(\tau)\right]e^{z_j^T\beta_0}+\left[1+(1-\rho_{jk})e^{z_j^T\beta_0}\Lambda_0(\tau)\right]e^{z_k^T\beta_0}}{(1-\rho_{jk})e^{z_j^T\beta_0}e^{z_k^T\beta_0}\Lambda_0(\tau)\Lambda_0(\tau)+e^{z_j^T\beta_0}\Lambda_0(\tau)+e^{z_k^T\beta_0}\Lambda_0(\tau)+1}\int_0^\tau h_3(u)d\Lambda_0(u)=0\ ;$$
i.e. $h_{2jk}e^{z_j^T\beta_0}e^{z_k^T\beta_0}\Lambda_0(\tau)\Lambda_0(\tau)$
$$-\left(\left[1+(1-\rho_{jk})e^{z_k^T\beta_0}\Lambda_0(\tau)\right]e^{z_j^T\beta_0}+\left[1+(1-\rho_{jk})e^{z_j^T\beta_0}\Lambda_0(\tau)\right]e^{z_k^T\beta_0}\right)\int_0^\tau h_3(u)d\Lambda_0(u)=0\ . \tag{B.6}$$

I plug $z_j'=z_j+(\log 2)/\beta$ into (B.6):

$$2h_{2jk}e^{z_j^T\beta}e^{z_k^T\beta}\Lambda_0(\tau)\Lambda_0(\tau)$$
$$-\left(2\left[1+(1-\rho_{jk})e^{z_k^T\beta}\Lambda_0(\tau)\right]e^{z_j^T\beta}+\left[1+2(1-\rho_{jk})e^{z_j^T\beta}\Lambda_0(\tau)\right]e^{z_k^T\beta}\right)\int_0^\tau h_3(u)d\Lambda(u)=0\ .$$
If $h_{2jk}\neq 0$ or $\displaystyle\int_0^\tau h_3(u)d\Lambda_0(u)\neq 0$, then
$$1+2(1-\rho_{jk})e^{z_j^T\beta}\Lambda_0(\tau)=2+2(1-\rho_{jk})e^{z_j^T\beta}\Lambda_0(\tau)\ .$$

Contradiction is achieved. I have shown $h_{2jk}=0$ and thus $h_2=0$; consequently, $h_3(0)=0$ by (B.5).

Since I have shown h_1 and h_2 are both zero vectors, I can claim that

$$\int_0^\tau h_3(u)d\left(N_j(u)+N_k(u)\right)=\int_0^\tau D(u,O_j,O_k;\beta_0,\rho_0,\Lambda_0)h_3(u)d\Lambda_0(u)\ ,$$

for arbitrary pairs of observations.

I consider another sequence of paired observations $\{y_j,\delta_j=1,z_j,y_{kr},\delta_k=1,z_k:r=1,2,\ldots\}$ such that $y_{kr}\downarrow 0$. I denote the set $A_r:=\{\boldsymbol{z}:\mathrm{pr}(C_j\geq y_j,C_k\geq y_{kr}\mid \boldsymbol{Z}=\boldsymbol{z})>0\}$ and according to conditions C2 and C3, as $r\to\infty$, $A_r\to A$ such that A is the whole space of covariate $\boldsymbol{Z}$. Plugging $\{y_j,\delta_j=1,z_j,y_{kr},\delta_k=1,z_k:r=1,2,\ldots\}$ into (B.4) and taking the limit in r give

$$h_3(y_j)=\left(3\frac{e^{z_j^T\beta_0}}{u(o_j;\beta_0,\Lambda_0)+1}-\frac{(1-\rho_{jk})e^{z_j^T\beta_0}}{(1-\rho_{jk})u(o_j;\beta_0,\Lambda_0)+1+\rho_{jk}}\right)\int_0^{y_j}h_3(u)d\Lambda_0(u)\ .$$

The following quantity

$$\begin{aligned} & 3\frac{e^{z_j^T\beta_0}}{u(o_j;\beta_0,\Lambda_0)+1} - \frac{(1-\rho_{jk})e^{z_j^T\beta_0}}{(1-\rho_{jk})u(o_j;\beta_0,\Lambda_0)+1+\rho_{jk}} \\ = \quad & e^{z_j^T\beta_0}\frac{2(1-\rho_{jk})u(o_j;\beta_0,\Lambda_0)+2+4\rho_{jk}}{[u(o_j;\beta_0,\Lambda_0)+1][(1-\rho_{jk})u(o_j;\beta_0,\Lambda_0)+1+\rho_{jk}]} \end{aligned}$$

is always positive and varies arbitrarily with different values of z_j. Thus, $h_3(y_j) = 0$ for $\forall y_j \in [0,\tau]$.

□

Theorem 5.2.2

Consistency of the NPMCLE can be demonstrated by first showing $\hat{\Lambda}_m(\tau)$ is uniformly bounded a.s.. Then by Helly's selection lemma and the compactness of $\mathcal{B}\times\mathcal{R}$, for every subsequence of NPMCLE denoted as $\{\hat{\theta}_n\} = \{(\hat{\beta}_n,\hat{\rho}_n,\hat{\Lambda}_n)\}$, there exists a subsequence $\{\hat{\theta}_{n'}\}$ such that $\hat{\theta}_{n'} \to \theta^* := (\beta^*,\rho^*,\Lambda^*)$, which is an inner point of parameter space Θ. This convergence is point-wise but I should strengthen it into uniform convergence. The whole proof will be complete if I can show $\theta^* = \theta_0$. However, I cannot write out Λ^* explicitly. I switch to an intermediate function sequence $\{\tilde{\Lambda}_{n'}\}$ which converges to Λ_0 uniformly on $[0,\tau]$. In the following I present several key steps of the proof, following the structure from Murphy et al. (1997).

Proof. To show the uniform boundedness of $\{\hat{\Lambda}_m(\tau)\}$, I compare the values of the empirical composite marginal log-likelihood contribution evaluated at the NPMCLE $(\hat{\beta}_m,\hat{\rho}_m,\hat{\Lambda}_m)$ and another set of parameters. I show if $\{\hat{\Lambda}_m(\tau)\}$ is not uniformly bounded, there is a subsequence of the empirical composite marginal log-likelihood contributions evaluated at the NPMCLE going to negative infinity as $m\to\infty$.

1. I construct a step function $\bar{\Lambda}_m$:

$$\begin{aligned} & \bar{\Lambda}_m(t) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}1\{y_{ij}\le t\}, \quad t\in[0,\tau]\,. \\ \text{Consequently,} \quad & \bar{\Lambda}_m(t) = O(1)\,, \qquad \Delta\bar{\Lambda}_m(t) = O(1/m)\,, \\ \text{and} \quad & \mathbb{P}_m clog(O;\beta_0,\rho_0,\bar{\Lambda}_m) = O(1) + \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}\log(1/m)\,. \end{aligned} \tag{B.7}$$

2. I find the empirical composite marginal log-likelihood contribution evaluated at the NPMCLE as

$$\begin{aligned} & \mathbb{P}_m \left\{ \frac{1}{n-1} \sum_{j<k} l(O_j, O_k; \hat{\beta}_m, \hat{\rho}_m, \hat{\Lambda}_m) \right\} \\ = \quad & O(1) + \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \log \Delta \hat{\Lambda}_m(y_{ij}) - \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (1+\delta_{ij}) \log \left(\hat{\Lambda}_m(y_{ij}) + 1 \right) . \end{aligned} \tag{B.8}$$

3. Assuming $\{\hat{\Lambda}_m(\tau)\}$ is not uniformly bounded, I establish a contradiction.
Considering a partition $\tau = s_0 > s_1 > \ldots > s_N > s_{N+1} = 0$, I find the difference between (B.8) and (B.7) as the following:

$$\begin{aligned} & \text{(B.8)} - \text{(B.7)} \qquad\qquad \text{(B.9)} \\ & = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \log \left(m \Delta \hat{\Lambda}_m(y_{ij}) \right) - \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (1+\delta_{ij}) \log \left(\hat{\Lambda}_m(y_{ij}) + 1 \right) + O(1) \\ = \quad & \sum_{q=0}^{N} \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} 1\left\{ y_{ij} \in [s_{q+1}, s_q) \right\} \log \left(m \Delta \hat{\Lambda}_m(y_{ij}) \right) - \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\left\{ y_{ij} = \tau \right\} (1+\delta_{ij}) \log \left(\hat{\Lambda}_m(\tau) + 1 \right) \\ - \quad & \sum_{q=0}^{N} \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\left\{ y_{ij} \in [s_{q+1}, s_q) \right\} (1+\delta_{ij}) \log \left(\hat{\Lambda}_m(y_{ij}) + 1 \right) + O(1) \\ \leq \quad & \sum_{q=0}^{N} \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\left\{ y_{ij} \in [s_{q+1}, s_q) \right\} \delta_{ij} \log \left(m \Delta \hat{\Lambda}_m(y_{ij}) \right) - \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\left\{ y_{ij} = \tau \right\} (1+\delta_{ij}) \log \left(\hat{\Lambda}_m(\tau) + 1 \right) \\ - \quad & \sum_{q=0}^{N} \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\left\{ y_{ij} \in [s_{q+1}, s_q) \right\} (1+\delta_{ij}) \log \left(\hat{\Lambda}_m(s_{q+1}) + 1 \right) + O(1) . \end{aligned} \tag{B.10}$$

Since $\log(x)$ is a concave function, by Jensen's Inequality,

$$\begin{aligned} & \frac{\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} 1\left\{ y_{ij} \in [s_{q+1}, s_q) \right\} \log \left(m \Delta \hat{\Lambda}_m(y_{ij}) \right)}{\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} 1\left\{ y_{ij} \in [s_{q+1}, s_q) \right\}} \\ \leq \quad & \log \left(m \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} 1\left\{ y_{ij} \in [s_{q+1}, s_q) \right\} \Delta \hat{\Lambda}_m(y_{ij})}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} 1\left\{ y_{ij} \in [s_{q+1}, s_q) \right\}} \right) . \end{aligned}$$

Thus

$$
\begin{aligned}
& \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}1\left\{y_{ij}\in[s_{q+1},s_q)\right\}\log\left(m\Delta\hat{\Lambda}_m(y_{ij})\right) \\
\leq\quad & \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}1\left\{y_{ij}\in[s_{q+1},s_q)\right\}\times\log\left(m\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}1\left\{y_{ij}\in[s_{q+1},s_q)\right\}\Delta\hat{\Lambda}_m(y_{ij})}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}1\left\{y_{ij}\in[s_{q+1},s_q)\right\}}\right) \\
=\quad & \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}1\left\{y_{ij}\in[s_{q+1},s_q)\right\}\times\left\{\log\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}1\left\{y_{ij}\in[s_{q+1},s_q)\right\}\Delta\hat{\Lambda}_m(y_{ij})\right)\right. \\
& \left.-\log\left(\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}1\left\{y_{ij}\in[s_{q+1},s_q)\right\}\right]/m\right)\right\} \\
\leq\quad & O(1)+\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}1\left\{y_{ij}\in[s_{q+1},s_q)\right\}\times\log\hat{\Lambda}_m(s_q)\,,
\end{aligned}
$$

since

$$
\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}1\left\{y_{ij}\in[s_{q+1},s_q)\right\}\Delta\hat{\Lambda}_m(y_{ij})\leq\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}1\left\{y_{ij}\in[0,s_q]\right\}\Delta\hat{\Lambda}_m(y_{ij})=\hat{\Lambda}_m(s_q)
$$

and $\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}1\left\{y_{ij}\in[s_{q+1},s_q)\right\}/m=O(1)$.

Then the right side of (B.10) is bounded from above by

$$
\begin{aligned}
& -\sum_{q=0}^{N-1}\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[(1+\delta_{ij})1\left\{y_{ij}\in[s_{q+1},s_q)\right\}-\delta_{ij}1\left\{y_{ij}\in[s_{q+2},s_{q+1})\right\}\right]\log\left(\hat{\Lambda}_m(s_{q+1})+1\right) \\
& -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[1\left\{y_{ij}=\tau\right\}(1+\delta_{ij})-1\left\{y_{ij}\in[s_1,\tau)\right\}\delta_{ij}\right]\log\left(\hat{\Lambda}(\tau)+1\right) \\
& -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}1\{y_{ij}\in[0,s_N)\}(1+\delta_{ij})\log\left(\hat{\Lambda}_m(0)+1\right)+O(1)\,. \qquad \text{(B.11)}
\end{aligned}
$$

I choose a partition from τ to 0. First, I found some $s_1\in[0,\tau)$ such that

$$
\frac{1}{2}\mathrm{E}\left\{\sum_{j=1}^{n_i}1\left\{Y_{ij}=\tau\right\}\right\}=\frac{1}{2}\mathrm{E}\left\{\sum_{j=1}^{n_i}(1+\Delta_{ij})1\left\{Y_{ij}=\tau\right\}\right\}>\mathrm{E}\left\{\sum_{j=1}^{n_i}\Delta_{ij}1\left\{Y_{ij}\in[s_1,\tau)\right\}\right\}.
$$

By conditions C3 and C6, such an s_1 exists.

Define a constant $\epsilon \in (0,1)$ such that

$$\frac{\epsilon}{1-\epsilon} < \frac{\mathrm{E}\left\{\sum_{j=1}^{n_i} 1\left\{Y_{ij} \in [s_1, s_0)\right\}\right\}}{\mathrm{E}\left\{\sum_{j=1}^{n_i} \Delta_{ij} 1\left\{Y_{ij} \in [0, s_0)\right\}\right\}} .$$

If $s_1 > 0$, I choose $s_2 = 0 \vee s$ such that s is the minimum value less than s_1 satisfying:

$$(1-\epsilon)\mathrm{E}\left\{\sum_{j=1}^{n_i}(1+\Delta_{ij})1\left\{Y_{ij} \in [s_1, s_0)\right\}\right\} \geq \mathrm{E}\left\{\sum_{j=1}^{n_i}\Delta_{ij}1\left\{Y_{ij} \in [s, s_1)\right\}\right\} .$$

Clearly, s_2 exists. The process can continue so that I obtain a sequence $\tau = s_0 > s_1 > \ldots \geq 0$ such that

$$\begin{aligned}
\frac{1}{2}\mathrm{E}\left\{\sum_{j=1}^{n_i}(1+\Delta_{ij})1\left\{Y_{ij} = \tau\right\}\right\} &> \mathrm{E}\left\{\sum_{j=1}^{n_i}\Delta_{ij}1\left\{Y_{ij} \in [s_1, \tau)\right\}\right\} , \\
(1-\epsilon)\mathrm{E}\left\{\sum_{j=1}^{n_i}(1+\Delta_{ij})1\left\{Y_{ij} \in [s_q, s_{q-1})\right\}\right\} &\geq \mathrm{E}\left\{\sum_{j=1}^{n_i}\Delta_{ij}1\left\{Y_{ij} \in [s_{q+1}, s_q)\right\}\right\}, \quad q \geq 1 .
\end{aligned}$$

I claim that such a sequence cannot be infinite; i.e. there exists a finite N such that $s_{N+1} = 0$. Suppose suppose $s_q \to s^* \geq 0$ then by the definition of s_q, it holds that

$$(1-\epsilon)\mathrm{E}\left\{\sum_{j=1}^{n_i}(1+\Delta_{ij})1\left\{Y_{ij} \in [s_q, s_{q-1})\right\}\right\} = \mathrm{E}\left\{\sum_{j=1}^{n_i}\Delta_{ij}1\left\{Y_{ij} \in [s_{q+1}, s_q)\right\}\right\}, \quad q \geq 1 ,$$

since failure is possible at any moment and thus the right side is continuous in s_{q+1}. Summing over $q = 1, 2, \ldots$ gives

$$(1-\epsilon)\mathrm{E}\left\{\sum_{j=1}^{n_i}(1+\Delta_{ij})1\left\{Y_{ij} \in [s^*, \tau)\right\}\right\} = \mathrm{E}\left\{\sum_{j=1}^{n_i}\Delta_{ij}1\left\{Y_{ij} \in [s^*, s_1)\right\}\right\} .$$

I remove $(1-\epsilon)\mathrm{E}\left\{\sum_{j=1}^{n_i}\Delta_{ij}1\left\{Y_{ij} \in [s^*, \tau)\right\}\right\}$ from the left,
$(1-\epsilon)\mathrm{E}\left\{\sum_{j=1}^{n_i}\Delta_{ij}1\left\{Y_{ij} \in [s^*, s_1)\right\}\right\}$ from the right and got

$$(1-\epsilon)\mathrm{E}\left\{\sum_{j=1}^{n_i}1\left\{Y_{ij} \in [s^*, \tau)\right\}\right\} \leq \epsilon\mathrm{E}\left\{\sum_{j=1}^{n_i}\Delta_{ij}1\left\{Y_{ij} \in [s^*, s_1)\right\}\right\} .$$

And thus

$$\frac{\epsilon}{1-\epsilon} \geq \frac{\mathrm{E}\left\{\sum_{j=1}^{n_i} 1\left\{Y_{ij} \in [s^*, \tau)\right\}\right\}}{\mathrm{E}\left\{\sum_{j=1}^{n_i} \Delta_{ij} 1\left\{Y_{ij} \in [s^*, s_1)\right\}\right\}} > \frac{\mathrm{E}\left\{\sum_{j=1}^{n_i} 1\left\{Y_{ij} \in [s_1, \tau)\right\}\right\}}{\mathrm{E}\left\{\sum_{j=1}^{n_i} \Delta_{ij} 1\left\{Y_{ij} \in [0, \tau)\right\}\right\}} > \frac{\epsilon}{1-\epsilon} ;$$

a contradiction is achieved. I have shown there exists a finite N such that $s_{N+1} = 0$.

If $\{\hat{\Lambda}_m\}$ is not uniformly bounded, then there exists a subsequence $\{\hat{\Lambda}_{m'}\}$ such that $\hat{\Lambda}_{m'}(\tau) \to \infty$ as $m' \to \infty$. If this is true, (B.11) will go to negative infinity, contradicting the definition of the NPMCLE.

Therefore, $\hat{\Lambda}_m(\tau)$ is uniformly bounded.

I rewrite the parameter space for NPMCLE inference:

$$\Theta := \mathcal{B} \times \mathcal{R} \times \mathcal{L} , \tag{B.12}$$

$$\mathcal{L} := \{\Lambda(\cdot) : \text{ non-decreasing step càdlàg function in } [0, \tau] \text{ with jumps at observed failure time points and } \Lambda(0) = 0, \Lambda(\tau) = V_0 < \infty\} .$$

I rewrite $\hat{\Lambda}_m$ as

$$\hat{\Lambda}_m(t) = \int_0^t \frac{1}{W_m(u; \hat{\beta}_m, \hat{\rho}_m, \hat{\Lambda}_m)} dG_m(u) ,$$

where

$$\begin{aligned}
W(u, O; \beta, \rho, \Lambda) &= \frac{1}{n-1} \sum_{k<j} D(u, O_j, O_k; \beta, \rho, \Lambda) , \\
W_m(u; \beta, \rho, \Lambda) &= \mathbb{P}_m \left[W(u, O; \beta, \rho, \Lambda)\right] , \quad W_0(u; \beta, \rho, \Lambda) = P_0 \left[W(u, O; \beta, \rho, \Lambda)\right] , \\
G(t) &= \sum_{j=1}^{n} \Delta_j 1\{Y_j \leq t\} , \\
G_m(t) &= \mathbb{P}_m \left[\sum_{j=1}^{n} \Delta_j 1\{Y_j \leq t\}\right] , \quad G_0(t) = P_0 \left[\sum_{j=1}^{n} \Delta_j 1\{Y_j \leq t\}\right] .
\end{aligned}$$

The above identity of $\hat{\Lambda}_m$ is derived from the score function of an one-dimensional sub-model passing through parameters $(\hat{\beta}_m, \hat{\rho}_m, \hat{\Lambda}_m)$ in the "direction" of h_t:

$$\hat{\Lambda}_m^{\epsilon}(u) = \int_0^u (1 + \epsilon h_t(s)) d\hat{\Lambda}_m(s) ;$$

i.e. taking the derivative of the empirical composite marginal log-likelihood contribution under $(\hat{\beta}_m, \hat{\rho}_m, \hat{\Lambda}_m^{\epsilon})$ of ϵ at $\epsilon = 0$. Setting $h_t(s) = 1\{s \leq t\}$ and letting t vary arbitrarily on $[0, \tau]$, the identity is found. The intermediate term $\tilde{\Lambda}_m$ is defined

$$\tilde{\Lambda}_m(t) := \int_0^t \frac{1}{W_m(u; \beta_0, \rho_0, \Lambda_0)} dG_m(u) \, .$$

Trivially, I can write

$$\Lambda_0(t) := \int_0^t \frac{1}{W_0(u; \beta_0, \rho_0, \Lambda_0)} dG_0(u) \, .$$

In Appendix C, I verify that the following two function classes indexed by t and (u, θ) respectively

$$\begin{aligned} \mathcal{F}_1 &:= \left\{ f_t(O) := \int_0^t g(s) dG(s) : g \text{ is a càdlàg function on } [0,\tau] \text{ and } ||g||_V \leq M_1 < \infty \right\} \\ \mathcal{W} &:= \{ W(u, O; \beta, \rho, \Lambda) : u \in [0, \tau], (\beta, \rho, \Lambda) \in \Theta \} \end{aligned}$$

are P_0-Donsker. Thus $\tilde{\Lambda}_m(t)$ uniformly converges to $\Lambda_0(t)$:

$$\begin{aligned} & \left|\left| \tilde{\Lambda}_m(\cdot) - \Lambda_0(\cdot) \right|\right|_\infty = \left|\left| \int_0^{\cdot} \left(\frac{1}{W_m(u; \beta_0, \rho_0, \Lambda_0)} dG_m(u) - \frac{1}{W_0(u; \beta_0, \rho_0, \Lambda_0)} dG_0(u) \right) \right|\right|_\infty \\ \leq \quad & \left|\left| \int_0^{\cdot} \left(\frac{1}{W_m(u; \beta_0, \rho_0, \Lambda_0)} - \frac{1}{W_0(u; \beta_0, \rho_0, \Lambda_0)} \right) dG_0(u) \right|\right|_\infty \\ & + \left|\left| \int_0^{\cdot} \frac{1}{W_m(u; \beta_0, \rho_0, \Lambda_0)} \left(dG_m(u) - dG_0(u) \right) \right|\right|_\infty \\ \leq \quad & \sup_{u \in [0,\tau]} \left| \frac{1}{W_m(u; \beta_0, \rho_0, \Lambda_0)} - \frac{1}{W_0(u; \beta_0, \rho_0, \Lambda_0)} \right| \int_0^{\tau} dG_0(u) \\ & + \left|\left| \int_0^{\cdot} \frac{1}{W_m(u; \beta_0, \rho_0, \Lambda_0)} \left(dG_m(u) - dG_0(u) \right) \right|\right|_\infty \, . \end{aligned}$$

I do not apply Helly's Selection Lemma directly to the subsequence of NPMCLE $\{\hat{\Lambda}_n\}$. Instead, I define a point-wise converging sub-subsequence $\{n'\}$ of subsequence $\{n\}$ as

$$\hat{\beta}_{n'} \to \beta^*, \quad \hat{\rho}_{n'} \to \rho^*, \quad W_{n'}(\cdot; \hat{\beta}_{n'}, \hat{\rho}_{n'}, \hat{\Lambda}_{n'}) \to W^*(\cdot) \text{ pointwise.}$$

I define $\Lambda^*(t) := \int_0^t \frac{1}{W^*(u)} dG_0(u)$. By the Dominated Convergence Theorem, $\hat{\Lambda}_{n'}$ converges to Λ^*

uniformly since:

$$\begin{aligned}
\left|\left|\hat{\Lambda}_{n'} - \Lambda^*\right|\right|_\infty &= \left|\left|\int_0^\cdot \frac{1}{W_{n'}(u;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'})} dG_{n'}(u) - \int_0^\cdot \frac{1}{W^*(u)} dG_0(u)\right|\right|_\infty \\
&\leq \left|\left|\int_0^\cdot \frac{1}{W_{n'}(u;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'})} (dG_{n'}(u) - dG_0(u))\right|\right|_\infty \\
&\quad + \left|\left|\int_0^\cdot \left(\frac{1}{W_{n'}(u;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'})} - \frac{1}{W^*(u)}\right) dG_0(u)\right|\right|_\infty \\
&\leq \left|\left|\int_0^\cdot \frac{1}{W_{n'}(u;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'})} (dG_{n'}(u) - dG_0(u))\right|\right|_\infty \\
&\quad + \int_0^\tau \left|\frac{1}{W_{n'}(u;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'})} - \frac{1}{W^*(u)}\right| dG_0(u) \\
&\to 0 . \qquad \text{(B.13)}
\end{aligned}$$

By definitions of the four sets of parameters (sequences), these two quantities $d\hat{\Lambda}_{n'}(t)/d\tilde{\Lambda}_{n'}(t)$ and $d\Lambda^*(t)/d\Lambda_0(t)$ are also defined. By definitions, I write

$$\hat{\Lambda}_{n'}(t) = \int_0^t \frac{W_{n'}(u;\beta_0,\rho_0,\Lambda_0)}{W_{n'}(u;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'})} d\tilde{\Lambda}_{n'}(u) . \qquad \text{(B.14)}$$

Since

1) $W_{n'}(u;\beta,\rho,\Lambda) \to W_0(u;\beta,\rho,\Lambda)$ uniformly in $[0,\tau]\times\Theta$;

2) $W_0(\cdot;\beta,\rho,\Lambda)$ is continuous in β,ρ,Λ;

3) $W_0(\cdot;\cdot,\cdot,\cdot)$ is uniformly bounded in some strictly positive interval.

$$\begin{aligned}
&\left|\left|W_{n'}(u;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'}) - W_0(u;\beta^*,\rho^*,\Lambda^*)\right|\right|_\infty \\
\leq\ &\left|\left|W_{n'}(u;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'}) - W_0(u;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'})\right|\right|_\infty + \left|\left|W_0(u;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'}) - W_0(u;\beta^*,\rho^*,\Lambda^*)\right|\right|_\infty .
\end{aligned}$$

The first term converges to 0 as $n' \to \infty$. As for the second term, the Mean Value Theorem gives

$$\begin{aligned}
&\left|\left|W(u;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'}) - W(u;\beta^*,\rho^*,\Lambda^*)\right|\right|_\infty \\
\leq\ &L_1||\hat{\beta}_{n'} - \beta^*|| + L_2||\hat{\rho}_{n'} - \rho^*|| + L_3\left|\left|\hat{\Lambda}_{n'} - \Lambda^*\right|\right|_\infty \to 0 ,
\end{aligned}$$

since all first derivatives in $W(u;\beta,\rho,\Lambda)$ of (β,ρ,Λ) are uniformly bounded.

In Lin et al. (2000) Lemma 1, the authors claimed that: let $f_{n'}$, $g_{n'}$ be two sequences of bounded functions such that for some constant τ:

(a) $\sup_{0\le t\le\tau} |f_{n'}(t) - f(t)| \to 0$ where f is continuous on $[0,\tau]$;

(b) $\{g_{n'}\}$ is monotone on $[0,\tau]$;

(c) $\sup_{0\le t\le\tau} |g_{n'}(t) - g(t)| \to 0$ for some bounded function g;

Then

$$\begin{aligned}\sup_{0\le t\le\tau}\left|\int_0^t f_{n'}(s)dg_{n'}(s) - \int_0^t f(s)dg(s)\right| &\to 0\,,\\ \sup_{0\le t\le\tau}\left|\int_0^t g_{n'}(s)df_{n'}(s) - \int_0^t g(s)df(s)\right| &\to 0\,.\end{aligned}$$

In (B.14) $\{\tilde\Lambda_{n'}\}$ serves the character of $\{g_{n'}\}$ and $\frac{W_{n'}(u;\beta_0,\rho_0,\Lambda_0)}{W_{n'}(u;\hat\beta_{n'},\hat\rho_{n'},\hat\Lambda_{n'})}$ serves the character of $\{f_{n'}\}$ since

$$\begin{aligned}&\left|\left|\frac{W_{n'}(u;\beta_0,\rho_0,\Lambda_0)}{W_{n'}(u;\hat\beta_{n'},\hat\rho_{n'},\hat\Lambda_{n'})} - \frac{W_0(u;\beta_0,\rho_0,\Lambda_0)}{W_0(u;\beta^*,\rho^*,\Lambda^*)}\right|\right|_\infty\\ \le\;& C\left|\left|W_{n'}(u;\beta_0,\rho_0,\Lambda_0)W_0(u;\beta^*,\rho^*,\Lambda^*) - W_{n'}(u;\hat\beta_{n'},\hat\rho_{n'},\hat\Lambda_{n'})W_0(u;\beta_0,\rho_0,\Lambda_0)\right|\right|_\infty\\ =\;& C\,||W_{n'}(u;\beta_0,\rho_0,\Lambda_0)W_0(u;\beta^*,\rho^*,\Lambda^*) - W_0(u;\beta_0,\rho_0,\Lambda_0)W_0(u;\beta^*,\rho^*,\Lambda^*)||_\infty\\ &+C\left|\left|W_{n'}(u;\hat\beta_{n'},\hat\rho_{n'},\hat\Lambda_{n'})W_0(u;\beta_0,\rho_0,\Lambda_0) - W_0(u;\beta^*,\rho^*,\Lambda^*)W_0(u;\beta_0,\rho_0,\Lambda_0)\right|\right|_\infty\\ \to\;& 0\,.\end{aligned}$$

I take limits of both sides in (B.14), and uniformly for $t\in[0,\tau]$, it is true that

$$\Lambda^*(t) = \int_0^t \frac{W_0(u;\beta_0,\rho_0,\Lambda_0)}{W_0(u;\beta^*,\rho^*,\Lambda^*)}d\Lambda_0(u)\,,$$

which gives this equality:

$$\frac{d\Lambda^*(u)}{d\Lambda_0(u)} = \frac{W_0(u;\beta_0,\rho_0,\Lambda_0)}{W_0(u;\beta^*,\rho^*,\Lambda^*)}\,,$$

implying uniformly on $u\in[0,\tau]$: $d\hat\Lambda_{n'}(t)/d\tilde\Lambda_{n'}(t) \to d\Lambda^*(t)/d\Lambda_0(t)$.

Since $\{clog(O;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'}) - clog(O;\beta_0,\rho_0,\tilde{\Lambda}_{n'})\}$ is shown to be P_0-Glivenko-Cantelli in Appendix C, I come up with

$$\mathbb{P}_{n'}\{clog(O;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'}) - clog(O;\beta_0,\rho_0,\tilde{\Lambda}_{n'})\} \geq 0 \ ,$$

$$\text{implying} \qquad P_0\{clog(O;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'}) - clog(O;\beta_0,\rho_0,\tilde{\Lambda}_{n'})\} \geq -o(1) \ .$$

Together with the above proved (uniform) convergence sequences:

$$\hat{\beta}_{n'} \to \beta^*, \quad \hat{\rho}_{n'} \to \rho^*, \quad \hat{\Lambda}_{n'} \to \Lambda^*, \quad d\hat{\Lambda}_{n'}/d\tilde{\Lambda}_{n'} \to d\Lambda^*/d\Lambda_0 \ ;$$

$$\text{I get} \qquad P_0\{clog(O;\hat{\beta}_{n'},\hat{\rho}_{n'},\hat{\Lambda}_{n'}) - clog(O;\beta_0,\rho_0,\tilde{\Lambda}_{n'})\} \to P_0\{clog(O;\beta^*,\rho^*,\Lambda^*) - clog(O;\beta_0,\rho_0,\Lambda_0)\} \ .$$

By model identifiability with regards to the composite Kullback-Leibler distance,

$$\beta^* = \beta_0, \quad \rho^* = \rho_0, \quad \Lambda^* = \Lambda_0 \ .$$

Therefore, consistency is proved. □

Theorem 5.2.3

Since this frailty model has an infinite dimensional parameter Λ, I need to take care of score and information operator calculations to prove the NPMCLE weak convergence. To facilitate the development of these operators, I use the Banach space from **Lemma 5.2.1** to index the infinite-dimensional parameters:

$$\mathcal{H} := \Big\{(h_1,h_2,h_3) : h_1 \in \mathbb{R}^{d_1}, h_2 \in \mathbb{R}^{d_2}, h_3(t) \text{ is a càdlàg function on } [0,\tau]\Big\} \ ,$$

$$\text{equipped with the norm } ||h||_{\mathcal{H}} := ||h_1|| + ||h_2|| + ||h_3||_V \ .$$

I define subspaces of $\mathcal{H}$ as

$$\mathcal{H}_p := \{h \in \mathcal{H} : ||h||_{\mathcal{H}} \leq p\}, \quad \forall p > 0 \ ,$$

and the inequality will be strict if $p = \infty$.

$\mathcal{H}_1$ is sufficient to extract all components of θ, if θ is viewed as an element of $l^\infty(\mathcal{H})$, i.e. bounded functions defined on $\mathcal{H}$ such that $\theta(h) := h_1^T\beta + h_2^T\rho + \int_0^\tau h_3(u)d\Lambda(u)$. For example, $\tilde{h}_1 = ((1,0,\ldots,0)^T,(0,\ldots,0)^T,0)$ extracts out the first component of β,

$\tilde{h}_2 = ((0,\ldots,0)^T,(0,1,0,\ldots,0)^T,0)$ extracts out the second component of ρ and $\tilde{h}_3 = ((0,\ldots,0)^T,(0,\ldots,0)^T,1\{(\cdot) \leq u\})$ extracts $\Lambda(u)$. I also view $(\beta-\beta_0,\rho-\rho_0,\Lambda-\Lambda_0)$ as bounded functions on $\mathcal{H}_1$ by defining its value at (h_1,h_2,h_3)

$$(\beta-\beta_0)^T h_1 + (\rho-\rho_0)^T h_2 + \int_0^\tau h_3(u)d(\Lambda-\Lambda_0)(u) \ .$$

In the following I discuss the weak uniform convergence of $\sqrt{m}(\hat{\beta}_m-\beta_0,\hat{\rho}_m-\rho_0,\hat{\Lambda}_m-\Lambda_0)$ in the metric space $l^\infty(\mathcal{H}_1)$, which has the following norm:

$$\text{for some } f \in l^\infty(\mathcal{H}_1): \quad ||f|| = \sup_{h\in\mathcal{H}_1}|f(h)| \ .$$

I use Theorem 3.3.1 in van der Vaart and Wellner (1996) to show the weak convergence. This theorem is stated as the followings.

Suppose there are two random mappings Ψ_m and Ψ, to be defined later, such that $\Psi(\beta_0,\rho_0,\Lambda_0)=0$ for some interior point $(\beta_0,\rho_0,\Lambda_0)\in\Theta$, $\Psi_m(\hat{\beta}_m,\hat{\rho}_m,\hat{\Lambda}_m)\xrightarrow{P}0$ for some random sequence $(\hat{\beta}_m,\hat{\rho}_m,\hat{\Lambda}_m)\subset\Theta$, and the followings are true:

P.1 $(\hat{\beta}_m,\hat{\rho}_m,\hat{\Lambda}_m)$ is consistent for $(\beta_0,\rho_0,\Lambda_0)$;

P.2 $\sqrt{m}\left(\Psi_m-\Psi\right)(\beta_0,\rho_0,\Lambda_0)$ converges in distribution to a tight random element Z;

P.3

$$\begin{aligned} & \sqrt{m}\left(\Psi_m-\Psi\right)(\hat{\beta}_m,\hat{\rho}_m,\hat{\Lambda}_m) - \sqrt{m}\left(\Psi_m-\Psi\right)(\beta_0,\rho_0,\Lambda_0) \\ = \quad & o_p\left(1+\sqrt{m}||\hat{\beta}_m-\beta_0|| + \sqrt{m}||\hat{\rho}_m-\rho_0|| + \sqrt{m}||\hat{\Lambda}_m-\Lambda_0||_\infty\right) \ ; \end{aligned}$$

P.4 $\Psi(\beta,\rho,\Lambda)$ is Fréchet differentiable at $(\beta_0,\rho_0,\Lambda_0)$;

P.5 The derivative of $\Psi(\beta,\rho,\Lambda)$ in (β,ρ,Λ) at $(\beta_0,\rho_0,\Lambda_0)$, denoted by $\dot{\Psi}(\beta_0,\rho_0,\Lambda_0)$, is continuously invertible.

Then

$$\sqrt{m}(\hat{\beta}_m^T-\beta_0^T,\hat{\rho}_m^T-\rho_0^T,\hat{\Lambda}_m-\Lambda_0)\xrightarrow{d.} -\dot{\Psi}(\beta_0,\rho_0,\Lambda_0)^{-1}(Z) \ .$$

Proof. In the following I show that conditions P.1∼P.5 are satisfied in my model.

I define two random maps Ψ_m and Ψ.

For this purpose, I define a neighbourhood of the true parameter $(\beta_0,\rho_0,\Lambda_0)$, denoted by U, which is a subset of Θ:

$$U := \left\{(\beta,\rho,\Lambda): ||\beta-\beta_0|| + ||\rho-\rho_0|| + \sup_{t\in[0,\tau]}|\Lambda(t)-\Lambda_0(t)| < \epsilon_0\right\} \ ,$$

for a very small fixed constant $\epsilon_0 > 0$. Clearly, when the sample size m is large enough, $(\hat{\beta}_m, \hat{\rho}_m, \hat{\Lambda}_m)$ belongs to U with probability approaching one.

I note Ψ_m and Ψ are maps from U to $l^\infty(\mathcal{H}_1)$ such that $l^\infty(\mathcal{H}_1)$ consists of all the bounded functions on $\mathcal{H}_1$:

$$\begin{aligned}
\Psi_m(\beta,\rho,\Lambda)[h_1,h_2,h_3] &= \frac{d}{d\epsilon}\mathbb{P}_m clog(O;\beta+\epsilon h_1,\rho+\epsilon h_2,\Lambda+\epsilon\int h_3 d\Lambda)\Big|_{\epsilon=0} = \mathbb{P}_m V(O;\beta,\rho,\Lambda)(h)\ ,\\
\Psi(\beta,\rho,\Lambda)[h_1,h_2,h_3] &= \frac{d}{d\epsilon}P_0 clog(O;\beta+\epsilon h_1,\rho+\epsilon h_2,\Lambda+\epsilon\int h_3 d\Lambda)\Big|_{\epsilon=0} = P_0 V(O;\beta,\rho,\Lambda)(h)\ ,\\
\text{where } V(O;\beta,\rho,\Lambda)[h] &:= \frac{d}{d\epsilon}clog(O;\beta+\epsilon h_1,\rho+\epsilon h_2,\Lambda+\epsilon\int h_3 d\Lambda)\Big|_{\epsilon=0}\\
&:= \left\{h_1^T clog_\beta(O;\beta,\rho,\Lambda) + h_2^T clog_\rho(O;\beta,\rho,\Lambda) + clog_\Lambda(O;\beta,\rho,\Lambda)\left[\int h_3 d\Lambda\right]\right\}\ .
\end{aligned}$$

$clog_\beta(O;\beta,\rho,\Lambda)$ is the composite score for β, $clog_\rho(O;\beta,\rho,\Lambda)$ is the composite score for ρ and $clog_\Lambda(O;\beta,\rho,\Lambda)\left[\int h_3 d\Lambda\right]$ is the composite score for Λ from the submodel $\Lambda+\epsilon\int h_3 d\Lambda$: ϵ is the parameter. These quantities were derived by using the fact that the integral and differentiation signs are exchangeable. It is trivial that

$$\Psi_m(\hat{\beta}_m,\hat{\rho}_m,\hat{\Lambda}_m) = 0; \quad \Psi(\beta_0,\rho_0,\Lambda_0) = 0\ .$$

Condition P.1 has been shown.

To show the weak convergence in P.2 of the theorem, I want to verify the function class

$$\{V(O;\beta,\rho,\Lambda)(h) : (h_1,h_2,h_3)\in\mathcal{H}_1, (\beta,\rho,\Lambda)\in U\}$$

is P_0-Donsker. This procedure is quite similar to the two classes considered in Appendix C and thus omitted.

To verify P.3, by the P_0-Donsker preservation theorem,

$$\{V(O;\beta,\rho,\Lambda)(h) - V(O;\beta_0,\rho_0,\Lambda_0)(h) : (\beta,\rho,\Lambda)\in U, h\in\mathcal{H}_1\}$$

is P_0-Donsker as well. I claim

$$\begin{aligned}
&\sup_{h\in\mathcal{H}_1} P_0\left[V(O;\beta,\rho,\Lambda)(h) - V(O;\beta_0,\rho_0,\Lambda_0)(h)\right]^2\\
\leq\ & P_0\left[M_1||\beta_0-\beta|| + M_2||\rho_0-\rho|| + M_3||\Lambda_0-\Lambda||_\infty\right]^2 \qquad \text{(B.15)}\\
\to\ & 0 \quad \text{as } ||(\beta,\rho,\Lambda)-(\beta_0,\rho_0,\Lambda_0)||_\infty \to 0\ ,
\end{aligned}$$

since everything in $V(O;\beta,\rho,\Lambda)(h)$ is continuous with regards to $(\beta,\rho,\Lambda(Y_1),\ldots,\Lambda(Y_n))$ so the Mean Value Theorem can be applied. Because all random variables are uniformly bounded, there exists finite constants (M_1,M_2,M_3).

Therefore, according to Lemma 3.3.5 from van der Vaart and Wellner (1996), P.3 holds.

To verify the Fréchet differentiability of the composite score function, I first consider the Găteaux derivative of Ψ at $(\beta_0,\rho_0,\Lambda_0)$, denoted by $\dot{\Psi}$, which is a map from the set $\dot{U} \equiv \{(\beta-\beta_0,\rho-\rho_0,\Lambda-\Lambda_0):(\beta,\rho,\Lambda)\in U\}$ to $l^\infty(\mathcal{H}_1)$; i.e. $l^\infty(\mathcal{H}_1)\mapsto l^\infty(\mathcal{H}_1)$.

Straightforward calculations yield that

$$
\begin{aligned}
& \dot{\Psi}(\beta-\beta_0,\rho-\rho_0,\Lambda-\Lambda_0)[h_1,h_2,\int h_3 d\Lambda_0] \\
= \quad & (\beta-\beta_0)^T\mathcal{T}_{1,\theta_0}(h_1,h_2,h_3)+(\rho-\rho_0)^T\mathcal{T}_{2,\theta_0}(h_1,h_2,h_3)+\int_0^\tau \mathcal{T}_{3,\theta_0}(h_1,h_2,h_3)d(\Lambda-\Lambda_0) \\
= \quad & (\beta-\beta_0)^T\left[\mathcal{T}_{1,\beta,\theta_0}(h_1)+\mathcal{T}_{1,\rho,\theta_0}(h_2)+\mathcal{T}_{1,\Lambda,\theta_0}(h_3)\right] \\
& +(\rho-\rho_0)^T\left[\mathcal{T}_{2,\beta,\theta_0}(h_1)+\mathcal{T}_{2,\rho,\theta_0}(h_2)+\mathcal{T}_{2,\Lambda,\theta_0}(h_3)\right] \\
& +\int_0^\tau\left[\mathcal{T}_{3,\beta,\theta_0}(h_1)+\mathcal{T}_{3,\rho,\theta_0}(h_2)+\mathcal{T}_{3,\Lambda,\theta_0}(h_3)\right]d(\Lambda-\Lambda_0)\ .
\end{aligned}
$$

The operator $\mathcal{T}_{\theta_0}: lin\ \mathcal{H}_1\mapsto lin\ \mathcal{H}_1$ can be written as

$$
\mathcal{T}_{\theta_0}(h)=\begin{pmatrix}\mathcal{T}_{1,\beta,\theta_0} & \mathcal{T}_{1,\rho,\theta_0} & \mathcal{T}_{1,\Lambda,\theta_0}\\ \mathcal{T}_{2,\beta,\theta_0} & \mathcal{T}_{2,\rho,\theta_0} & \mathcal{T}_{2,\Lambda,\theta_0}\\ \mathcal{T}_{3,\beta,\theta_0} & \mathcal{T}_{3,\rho,\theta_0} & \mathcal{T}_{3,\Lambda,\theta_0}\end{pmatrix}\begin{pmatrix}h_1\\h_2\\h_3\end{pmatrix}\ .
$$

Since integral under P_0 and differentiation are exchangeable, I get

$$
\begin{aligned}
\mathcal{T}_{1,\beta,\theta_0}(h_1) &= h_1^T P_0 clog_{\beta\beta}(O;\beta_0,\rho_0,\Lambda_0)\\
\mathcal{T}_{1,\rho,\theta_0}(h_2) &= h_2^T P_0 clog_{\beta\rho}(O;\beta_0,\rho_0,\Lambda_0)\\
\mathcal{T}_{1,\Lambda,\theta_0}(h_3) &= P_0\int_0^\tau C_\beta(u,O;\beta_0,\rho_0,\Lambda_0)h_3(u)d\Lambda_0(u)\\
\mathcal{T}_{2,\beta,\theta_0}(h_1) &= h_1^T P_0 clog_{\rho\beta}(O;\beta_0,\rho_0,\Lambda_0)\\
\mathcal{T}_{2,\rho,\theta_0}(h_2) &= h_2^T P_0 clog_{\rho\rho}(O;\beta_0,\rho_0,\Lambda_0)\\
\mathcal{T}_{2,\Lambda,\theta_0}(h_3) &= P_0\int_0^\tau C_\rho(u,O;\beta_0,\rho_0,\Lambda_0)h_3(u)d\Lambda_0(u)\\
\mathcal{T}_{3,\beta,\theta_0}(h_1)(t) &= P_0\left[\sum_{j=1}^n q_j(O;\beta_0,\rho_0,\Lambda_0)1\{Y_j\geq t\}\right]^T h_1
\end{aligned}
$$

$$\mathcal{T}_{3,\rho,\theta_0}(h_2)(t) = P_0\left[\sum_{j=1}^{n}\tilde{q}_j(O;\beta_0,\rho_0,\Lambda_0)1\{Y_j\geq t\}\right]^T h_2$$

$$\mathcal{T}_{3,\Lambda,\theta_0}(h_3)(t) = P_0\left(\int_0^\tau \sum_{j=1}^{n} B_j(u,O;\beta_0,\rho_0,\Lambda_0)1\{Y_j\geq t\}h_3(u)d\Lambda_0(u)\right),$$

where I define

$$\begin{aligned}
C_\beta(u,O;\beta_0,\rho_0,\Lambda_0) &:= -\bigtriangledown_\beta W(u,O;\beta_0,\rho_0,\Lambda_0)\\
C_\rho(u,O;\beta_0,\rho_0,\Lambda_0) &:= -\bigtriangledown_\rho W(u,O;\beta_0,\rho_0,\Lambda_0)\\
q_j(O;\beta_0,\rho_0,\Lambda_0) &:= \frac{\partial clog_\beta(O;\beta_0,\rho_0,\Lambda_0)}{\partial\Lambda_0(Y_j)}\\
\tilde{q}_j(O;\beta_0,\rho_0,\Lambda_0) &:= \frac{\partial clog_\rho(O;\beta_0,\rho_0,\Lambda_0)}{\partial\Lambda_0(Y_j)}\\
B_j(u,O;\beta_0,\rho_0,\Lambda_0) &:= -\partial\left(\frac{1}{n-1}\sum_{k\neq j}\left[(1+\Delta_j+\Delta_k)\frac{e^{Z_j^T\beta_0}1\{Y_j\geq u\}\left[1+(1-\rho_{jk})e^{Z_k^T\beta_0}\Lambda_0(Y_k)\right]}{v(O_j,O_k;\beta_0,\Lambda_0,\rho_0)}\right.\right.\\
&\quad\left.\left.-\frac{\Delta_k(1-\rho_{jk})e^{Z_j^T\beta_0}1\{Y_j\geq u\}\left[\Delta_j(1-\rho_{jk})e^{Z_k^T\beta_0}\Lambda_0(Y_k)+1\right]}{w(O_j,O_k;\beta_0,\Lambda_0,\rho_0)}\right]\right)\Big/\partial\Lambda_0(Y_j).
\end{aligned}$$

In the following I verify the above derivative is indeed the Fréchet derivative. I work on this quantity, which is a map from $l^\infty(\mathcal{H}_1)$ to $l^\infty(\mathcal{H}_1)$:

$$\begin{aligned}
&\sup_{||h_3||_V\leq 1}\left|P_0\left\{-\int_0^\tau W(u,O;\beta,\rho_0,\Lambda_0)h_3(u)d\Lambda_0(u)+\int_0^\tau W(u,O;\beta_0,\rho_0,\Lambda_0)h_3(u)d\Lambda_0(u)\right\}\right.\\
&\left.-(\beta-\beta_0)^TP_0\left\{\int_0^\tau C_\beta(u,O;\beta_0,\rho_0,\Lambda_0)h_3(u)d\Lambda_0(u)\right\}\right|/||\beta-\beta_0||\\
=\;&\frac{\sup_{||h_3||_V\leq 1}\left|(\beta-\beta_0)^T\left[\int_0^\tau P_0\left\{C_\beta(u,O;\beta^*,\rho_0,\Lambda_0)h_3(u)d\Lambda_0(u)-\int_0^\tau C_\beta(u,O;\beta_0,\rho_0,\Lambda_0)h_3(u)d\Lambda_0(u)\right\}\right]\right|}{||\beta-\beta_0||}\\
\leq\;&\sup_{||h_3||_V\leq 1}\left|\int_0^\tau P_0\left\{C_\beta(u,O;\beta^*,\rho_0,\Lambda_0)h_3(u)d\Lambda_0(u)-\int_0^\tau C_\beta(u,O;\beta_0,\rho_0,\Lambda_0)h_3(u)d\Lambda_0(u)\right\}\right|, \qquad \text{(B.16)}
\end{aligned}$$

such that $||\beta^*-\beta_0||\leq||\beta-\beta_0||$; the last inequality is Cauchy-Schwartz inequality. As $||\beta-\beta_0||\to 0$, due to the smoothness of $C_\beta(u,O;\beta,\rho_0,\Lambda_0)$ in β and since Λ_0 and h_3 are bounded functions on $[0,\tau]$, quantity in (B.16) goes to zero.

Similarly I get

$$\lim_{\rho\to\rho_0}\sup_{||h_3||_V\leq 1}\left|P_0\left\{-\int_0^\tau W(u,O;\beta_0,\rho\Lambda_0)h_3(u)d\Lambda_0(u)+\int_0^\tau W(u,O;\beta_0,\rho_0,\Lambda_0)h_3(u)d\Lambda_0(u)\right\}\right.$$
$$\left.-(\rho-\rho_0)^T P_0\left\{\int_0^\tau C_\rho(u,O;\beta_0,\rho_0,\Lambda_0)h_3(u)d\Lambda_0(u)\right\}\right|/||\rho-\rho_0||\to 0\,.$$

I define $q_j(O;\beta_0,\rho_0,\Lambda_0)=\partial clog_\beta(O;\beta_0,\rho_0,\Lambda_0)/\partial\Lambda_0(Y_j)$ and considered this quantity

$$\begin{aligned}
&\sup_{||h_1||\leq 1}\left|h_1^T\left\{P_0\left[clog_\beta(O;\beta_0,\rho_0,\Lambda)-clog_\beta(O;\beta_0,\rho_0,\Lambda_0)\right]\right.\right.\\
&\left.\left.-P_0\left[\sum_{j=1}^n q_j(O;\beta_0,\rho_0,\Lambda_0)\int_0^\tau 1\{Y_j\geq t\}d(\Lambda-\Lambda_0)(t)\right]\right\}\right|/\sup_{||h_3||_V\leq 1}|\int_0^\tau h_3(u)d(\Lambda-\Lambda_0)(u)|\\
\leq\quad&\left|\left|P_0\left\{\sum_{j=1}^n q_j(O;\beta_0,\rho_0,\Lambda^*)\int_0^\tau 1\{Y_j\geq t\}d(\Lambda-\Lambda_0)(t)\right]\right.\right.\\
&\left.\left.-P_0\left[\sum_{j=1}^n q_j(O;\beta_0,\rho_0,\Lambda_0)\int_0^\tau 1\{Y_j\geq t\}d(\Lambda-\Lambda_0)(t)\right\}\right|\right|/||\Lambda-\Lambda_0||_\infty\\
=\quad&\frac{\left|\left|P_0\left[\sum_{j=1}^n(q_j(O;\beta_0,\rho_0,\Lambda^*)-q_j(O;\beta_0,\rho_0,\Lambda_0))\int_0^\tau 1\{Y_j\geq t\}d(\Lambda-\Lambda_0)(t)\right]\right|\right|}{||\Lambda-\Lambda_0||_\infty}\\
\leq\quad&\left|\left|P_0\left[\sum_{j=1}^n(q_j(O;\beta_0,\rho_0,\Lambda^*)-q_j(O;\beta_0,\rho_0,\Lambda_0))\right]\right|\right|\,,
\end{aligned}$$

due to the smoothness of q_j in Λ_0. It will converge to zero as $||\Lambda-\Lambda_0||_\infty\to 0$.

Similarly,

$$\begin{aligned}
&\sup_{||h_2||\leq 1}\left|h_2^T\left\{P_0\left[clog_\rho(O;\beta_0,\rho_0,\Lambda)-clog_\rho(O;\beta_0,\rho_0,\Lambda_0)\right]\right.\right.\\
&\left.\left.-P_0\left[\sum_{j=1}^n\tilde{q}_j(O;\beta_0,\rho_0,\Lambda_0)\int_0^\tau 1\{Y_j\geq t\}d(\Lambda-\Lambda_0)(t)\right]\right\}\right|/||\Lambda-\Lambda_0||_\infty\to 0
\end{aligned}$$

as $||\Lambda-\Lambda_0||_\infty\to 0$.

In the end, I consider this quantity

$$
\begin{aligned}
& \sup_{||h_3||_V \le 1} \Bigg| - \int_0^\tau W(u;\beta_0,\rho_0,\Lambda) h_3(u) d\Lambda(u) + \int_0^\tau W(u;\beta_0,\rho_0,\Lambda_0) h_3(u) d\Lambda(u) \\
& \qquad - P_0 \int_0^\tau \int_0^\tau \sum_{j=1}^n B_j(u;\beta_0,\rho_0,\Lambda_0) 1\{Y_j \ge t\} d(\Lambda - \Lambda_0)(t) h_3(u) d\Lambda_0(u) \Bigg| \\
= & \sup_{||h_3||_V \le 1} \Bigg| P_0 \left\{ \int_0^\tau \sum_{j=1}^n B_j(u;\beta_0,\rho_0,\Lambda^*)(\Lambda(Y_j) - \Lambda_0(Y_j)) h_3(u) d\Lambda_0(u) \right\} \\
& \qquad - P_0 \left\{ \int_0^\tau \int_0^\tau \sum_{j=1}^n B_j(u;\beta_0,\rho_0,\Lambda_0) 1\{Y_j \ge t\} d(\Lambda - \Lambda_0)(t) h_3(u) d\Lambda_0(u) \right\} \Bigg| \\
= & \sup_{||h_3||_V \le 1} \left| P_0 \left\{ \int_0^\tau \sum_{j=1}^n \left[B_j(u;\beta_0,\rho_0,\Lambda^*) - B_j(u;\beta_0,\rho_0,\Lambda_0) \right] (\Lambda - \Lambda_0)(Y_j) h_3(u) d\Lambda_0(u) \right\} \right| \\
\le & \sup_{||h_3||_V \le 1} P_0 \left\{ \int_0^\tau \sum_{j=1}^n \left| B_j(u;\beta_0,\rho_0,\Lambda^*) - B_j(u;\beta_0,\rho_0,\Lambda_0) \right| h_3(u) d\Lambda_0(u) \right\} \times ||\Lambda - \Lambda_0||_\infty \, .
\end{aligned}
$$

Thus, I have

$$
\begin{aligned}
& \sup_{||h_2|| \le 1} \Bigg| - \int_0^\tau W(u;\beta_0,\rho_0,\Lambda) h_3(u) d\Lambda(u) + \int_0^\tau W(u;\beta_0,\rho_0,\Lambda_0) h_3(u) d\Lambda(u) \\
& - \left(P_0 \int_0^\tau \int_0^\tau \sum_{j=1}^n B_j(u;\beta_0,\rho_0,\Lambda_0) 1\{Y_j \ge t\} d(\Lambda - \Lambda_0)(t) h_3(u) d\Lambda_0(u) \right) \Bigg| \Bigg/ ||\Lambda - \Lambda_0||_\infty
\end{aligned}
$$

converging to 0 as $||\Lambda - \Lambda_0||_\infty \to 0$.

As a summary, I have verified the G$\check{a}$teaux derivative is the Fréchet derivative of $\Psi(\beta,\rho,\Lambda)[h_1,h_2,\int h_3 d\Lambda_0]$ at $(\beta_0,\rho_0,\Lambda_0)$ by parts.

Before showing $\dot{\Psi}$ is continuously invertiable, in the following I first show $\mathcal{T}_{\theta_0}(h)$ is invertible and then I show it is also a Fredholm operator. I note that

$$
\begin{aligned}
& \mathcal{T}_{\theta_0}(h) \\
\propto \quad & -\mathrm{E} \left[\frac{1}{n-1} \sum_{j=1}^n \sum_{j<k} \left\{ h_1^T clog_\beta(O_j,O_k;\beta_0,\rho_0,\Lambda_0) + h_2^T clog_\rho(O_j,O_k;\beta_0,\rho_0,\Lambda_0) \right. \right. \\
& \qquad \left. \left. + clog_\Lambda(O_j,O_k;\beta_0,\rho_0,\Lambda_0)[\int h_3 d\Lambda_0] \right\}^2 \right] \\
= \quad & 0 \, ;
\end{aligned}
$$

i.e. all pairwise score functions will be zero a.s. for the one-dimensional sub-model defined in the direction of h. By **Lemma 5.2.1**, it implies $h = 0$ and thus $\mathcal{T}_{\theta_0}(h)$ is invertible.

To show $\mathcal{T}_{\theta_0}(h)$ is a Fredholm operator, I define

$$
\begin{aligned}
A(h) &:= \begin{pmatrix} P_0 clog_{\beta,\beta}(O;\beta_0,\rho,\Lambda_0) & P_0 clog_{\beta,\rho}(O;\beta_0,\rho,\Lambda_0) & 0 \\ P_0 clog_{\rho,\beta}(O;\beta_0,\rho,\Lambda_0) & P_0 clog_{\beta,\beta}(O;\beta_0,\rho,\Lambda_0) & 0 \\ 0 & 0 & -P_0 W(t;\beta_0,\rho,\Lambda_0) \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix} \\
&:= \begin{pmatrix} A_0 & 0 \\ 0 & -P_0 W(t;\beta_0,\rho,\Lambda_0) \end{pmatrix} \begin{pmatrix} (h_1,h_2)^T \\ h_3 \end{pmatrix} .
\end{aligned}
$$

$A(h)$ is a continuously invertible operator trivially:

$$
A^{-1}(h) = \begin{pmatrix} A_0^{-1} & 0 \\ 0 & -\frac{1}{P_0 W(t;\beta_0,\rho,\Lambda_0)} \end{pmatrix} \begin{pmatrix} (h_1,h_2)^T \\ h_3 \end{pmatrix}
$$

In the following I show the remaining part $K(h) := \mathcal{T}_{\theta_0}(h) - A(h)$ is a compact operator. I write out $K(h)$ explicitly

$$
K(h) = K_1(h) + K_2(h) + K_3(h) ,
$$

where

$$
\begin{aligned}
K_1(h) &= P_0 \int_0^\tau C_\beta(u,O;\beta_0,\rho_0,\Lambda_0) h_3(u) d\Lambda_0(u) + P_0 \int_0^\tau C_\rho(u,O;\beta_0,\rho_0,\Lambda_0) h_3(u) d\Lambda_0(u) , \\
K_2(h) &= P_0 \left[\sum_{j=1}^n q_j(O;\beta_0,\rho_0,\Lambda_0) 1\{Y_j \geq t\} \right]^T h_1 + P_0 \left[\sum_{j=1}^n \tilde{q}_j(O;\beta_0,\rho_0,\Lambda_0) 1\{Y_j \geq t\} \right]^T h_2 , \\
K_3(h) &= P_0 \left(\int_0^\tau \sum_{j=1}^n B_j(u,O;\beta_0,\rho_0,\Lambda_0) 1\{Y_j \geq t\} h_3(u) d\Lambda_0(u) \right) .
\end{aligned}
$$

$K_1(h)$ and $K_2(h)$ are bounded linear operators with finite-dimensional range and thus are compact, as discussed in Murphy et al. (1997).

For $K_3(h)$, I consider a sequence of indexing elements $\{h_{1n}, h_{2n}, h_{3n}\} \subset \mathcal{H}_1$. I write every h_{3n}

in the form as

$$\begin{aligned} h_{3n}(t) &= h_{3n}^{+}(t) - h_{3n}^{-}(t) \\ \text{where } h_{3n}^{+}(t) &= \begin{cases} h_{3n}(t) & \text{if } h_{3n}(t) \geq 0 \\ 0 & \text{otherwise} \end{cases} , \\ \text{and } h_{3n}^{-}(t) &= \begin{cases} -h_{3n}(t) & \text{if } h_{3n}(t) < 0 \\ 0 & \text{otherwise} \end{cases} . \end{aligned}$$

By Helly's Selection Lemma, there exists a subsequence of $\{h_{3n}\}$: $\{h_{3n_z}\}$ such that $h_{3n_z}^{+} \to g_{03}^{+}$ and $h_{3n_z}^{-} \to g_{03}^{-}$ point-wise, where $||g_{03}^{+}||_V \leq 1$ and $||g_{03}^{-}||_V \leq 1$. By the Dominant Convergence Theorem

$$||K_3(h_{3n_z}^{+}) - K_3(g_{03}^{+})||_V = ||K_3(h_{3n_z}^{+} - g_{03}^{+})||_V \leq C \int_0^{\tau} \left| h_{3n_z}^{+} - g_{03}^{+} \right| (u) d\Lambda_0(u) \to 0 .$$

Thus

$$||K_3(h_{3n_z}) - K_3(g_{03})||_V \to 0 .$$

Therefore I have shown there exists a subsequence and an element $g_0 \in \mathcal{H}_1$ such that

$$||K(h_{n_z}) - K(g_0)||_{\mathcal{H}_1} \to 0 .$$

As a summary, I have shown the operator $\mathcal{T}_{\theta_0} : \mathcal{H}_1 \mapsto \mathcal{H}_1$ is a Fredholm operator.

I write the mapping $\dot{\Psi}$ as

$$\begin{aligned} & \dot{\Psi}(\beta - \beta_0, \rho - \rho_0, \Lambda - \Lambda_0)[h_1, h_2, \int h_3 d\Lambda_0] \\ = \quad & (\beta - \beta_0)^T \mathcal{T}_1(h_1, h_2, h_3) + (\rho - \rho_0)^T \mathcal{T}_2(h_1, h_2, h_3) + \int_0^{\tau} \mathcal{T}_3(h_1, h_2, h_3)(u) d(\Lambda - \Lambda_0)(u) . \end{aligned}$$

Due to the property of parameter space, it is trivial that $\dot{\Psi}$ is continuously invertible.

Therefore, conditions P.1$\sim$P.5 are satisfied and Theorem 2 holds. □

Appendix C

ADDITIONAL PROOFS FOR APPENDIX B

Here I show two classes of functions in Appendix B are P_0-Donsker and the third class of functions is P_0-Glivenko-Cantelli.

C.0.1 The first class of functions

Remember I define

$$G(\boldsymbol{O};t) = \sum_{j=1}^{n} \Delta_j 1\{Y_j \leq t\} \ .$$

Then following class of functions, indexed by $t \in [0,\tau]$:

$$\mathcal{F}_2 := \left\{ f_2(\boldsymbol{O};t) := \int_0^t g(s) dG(\boldsymbol{O};s) : g \text{ is a càdlàg function on } [0,\tau] \text{ and } ||g||_V \leq M_1 < \infty, \boldsymbol{O} \sim P_0 \right\}$$

is P_0-Donsker.

Proof. I consider another function class which is also indexed by t

$$\mathcal{F}_0 := \left\{ f_0(\boldsymbol{O};t) := \int_0^t g(s) dG(\boldsymbol{O};s) : g \text{ is monotone on } [0,\tau] \text{ and } ||g||_V \leq M_1 < \infty, \boldsymbol{O} \sim P_0 \right\} \ ,$$

which can be rewritten as

$$\begin{aligned} \mathcal{F}_0 = \quad & \Bigg\{ f_0(\boldsymbol{O};t) := \sum_{j=1}^{n} g(Y_j) 1\{Y_j \leq t\} \Delta_j : t \in [0,\tau], \\ & \quad g \text{ is a monotone function on } [0,\tau] \text{ and } ||g||_V \leq M_1 < \infty, \boldsymbol{O} \sim P_0 \} \ . \end{aligned}$$

For a single observation O_j, consider the function class:

$$\begin{aligned} \mathcal{F}_1 = \quad & \{ f_1(O_j;t) := g(O_j) 1\{O_j \leq t\} \Delta_j : t \in [0,\tau], \\ & \quad g \text{ is a monotone function on } [0,\tau] \text{ and } ||g||_V \leq M_1 < \infty \} \ . \end{aligned}$$

For some fixed O_j, $f_1(O_j;t)$ is a monotone function in t. According to Exercise 3 on page 165 in van der Vaart and Wellner (1996), logarithm of the corresponding number of brackets is in a polynomial order and thus $\mathcal{F}_1$ is P_0-Donsker. That is to say, for any $\epsilon > 0$, denote $[f_1(O_j;t_s^L), f_1(O_j;t_s^U)]$, $s = 1,\ldots,N_{j,\epsilon}$ as the set of brackets covering $\mathcal{F}_1$ such that

$$||f_1(O_j;t_s^L) - f_1(O_j;t_s^U)||_{L_2(P_0)} \leq \epsilon/n_0$$

and $\log(N_{j,\epsilon})$ is in the order $(O(1)/\epsilon)^{M_\epsilon}$, M_ϵ is some finite number.

I propose a new set of brackets whose boundary point/function corresponding parameter values are the union of $\{t_{js}^L, t_{js}^U : s = 1,\ldots,N_{j,\epsilon},\ j = 1,\ldots,n\}$ and denoted them by

$$[f_0(O;t_q^L), f_0(O;t_q^U)]\ ; \quad q = 1,\ldots,N_\epsilon \sim (O(1)/\epsilon)^{M_\epsilon}\ .$$

Thus, I get

$$\begin{aligned} & f_0(O;t_q^L) \leq f_0(O;t') \leq f_0(O;t_q^U) \\ & ||f_0(O;t_q^L) - f_0(O;t_q^U)||_{L_2(P_0)} \leq \sum_{j=1}^{n} ||f_1(O_j;t_{js_j}^L) - f_1(O_j;t_{js_j}^U)||_{L_2(P_0)} \leq \epsilon\ . \end{aligned}$$

Since there is also a square integrable envelope function for $\mathcal{F}_0$, $\mathcal{F}_0$ is P_0-Donsker. Since every element in $\mathcal{F}_2$ can be expressed as a summation of two elements in $\mathcal{F}_0$, $\mathcal{F}_2$ is also P_0-Donsker by the preservation theorem. □

C.0.2 The second class of functions

The function class:

$$\mathcal{W} = \{W(u,\boldsymbol{O};\beta,\rho,\Lambda) : \quad \boldsymbol{O} \sim P_0; \quad u \in [0,\tau], \quad (\beta,\rho,\Lambda) \in \Theta\}$$

is P_0-Donsker, where

$$\begin{aligned} W(u,\boldsymbol{O};\beta,\rho,\Lambda) \quad = \quad & \sum_{j=1}^{n} \left\{ \frac{1}{n-1} \sum_{k\neq j} \left[(1+\Delta_j+\Delta_k) \frac{e^{Z_j^T\beta} 1\{Y_j \geq u\} \left[1 + (1-\rho_{jk}) e^{Z_k^T\beta} \Lambda(Y_k)\right]}{v(O_j,O_k;\beta,\Lambda,\rho)} \right.\right. \\ & \left.\left. - \frac{\Delta_k (1-\rho_{jk}) e^{Z_j^T\beta} 1\{Y_j \geq u\} \left[\Delta_j (1-\rho_{jk}) e^{Z_k^T\beta} \Lambda(Y_k) + 1\right]}{w(O_j,O_k;\beta,\Lambda,\rho)} \right] \right\}, \quad \boldsymbol{O} \sim P_0\ . \end{aligned}$$

Proof. For a pair of elements from $\mathcal{W}$, given a sample $\boldsymbol{O}$, their absolute difference is bounded by

$$|W(u_1, \boldsymbol{O}; \beta_2, \rho_2, \Lambda_2) - W(u_1, \boldsymbol{O}; \beta_1, \rho_1, \Lambda_1)| \leq A_0 \left\{ ||\beta_1 - \beta_2|| + ||\rho_1 - \rho_2|| + \sum_{j=1}^{n} |\Lambda_1(Y_j) - \Lambda_2(Y_j)| \right\} .$$

The above bound is achieved by noting $W(u, \boldsymbol{O}; \beta, \rho, \Lambda)$ is absolute continuous in $(\beta, \rho, \Lambda(Y_1), \ldots, \Lambda(Y_n))$ and every element in $W(u, \boldsymbol{O}; \beta, \rho, \Lambda)$ is uniformly bounded; thus $W(u, \boldsymbol{O}; \beta, \rho, \Lambda)$ is Lipschitz continuous in $(\beta, \rho, \Lambda(Y_1), \ldots, \Lambda(Y_n))$. Define a function

$$h(\boldsymbol{O}; \Lambda_1, \Lambda_2) = \sum_{j=1}^{n} |\Lambda_2(Y_j) - \Lambda_1(Y_j)| .$$

The right side of the above inequality can be rewritten as
$A_0 \{||\beta_1 - \beta_2|| + ||\rho_1 - \rho_2|| + h(\boldsymbol{O}; \Lambda_1, \Lambda_2)\}$.

By Theorem 2.7.5 in van der Vaart and Wellner (1996), the number of brackets of

$$\{ \quad \Lambda(\cdot) : \text{ non-decreasing step function in } [0, \tau] \text{ with jumps at the observed failure times and } \Lambda(0) = 0, \Lambda(\tau) \leq C\}$$

is in the order of $\exp(O(1)/\epsilon)$, under probability measure P_0. Further due to the compactness in the finite-dimensional part of the parameter, for $\forall \epsilon > 0$, there exists a finite bracket interval $[\beta_s^L, \beta_s^U] \times [\rho_s^L, \rho_s^U] \times [\Lambda_s^L, \Lambda_s^U]$, $s = 1, \ldots, N_\epsilon$ covering Θ, where $N_\epsilon \sim \exp(O(1)/\epsilon)$ such that for arbitrary $\theta' = (\beta', \rho', \Lambda') \in \Theta$, there is some s such that

$$\begin{aligned}
&\beta_s^L \leq \beta' \leq \beta_s^U, \quad \rho_s^L \leq \rho' \leq \rho_s^U, \quad \Lambda_s^L(\cdot) \leq \Lambda'(\cdot) \leq \Lambda_s^U(\cdot); \\
&||\beta_s^L - \beta_s^U|| < \frac{\epsilon}{12 n_0 A_0}, \quad ||\rho_s^L - \rho_s^U|| < \frac{\epsilon}{12 n_0 A_0}, \quad ||\Lambda_s^L(Y) - \Lambda_s^U(Y)||_{L_2(P_0)} < \frac{\epsilon}{12 n_0 A_0} ; \\
\text{i.e.} \quad &||\beta_s^L - \beta'|| \leq ||\beta_s^L - \beta_s^U|| < \frac{\epsilon}{12 n_0 A_0}, \quad ||\rho_s^L - \rho'|| \leq ||\rho_s^L - \rho_s^U|| < \frac{\epsilon}{12 n_0 A_0}, \\
&||h(\boldsymbol{O}; \Lambda_s^L, \Lambda')||_{L_2(P_0)} \leq ||h(\boldsymbol{O}; \Lambda_s^L, \Lambda_s^U)||_{L_2(P_0)} < \frac{\epsilon}{12 A_0} .
\end{aligned}$$

I fix bracketing set and considered the function classes, $s = 1, \ldots, N_\epsilon$:

$$\mathcal{W}_{\epsilon,s} := \left\{ W(u, \boldsymbol{O}; \beta_s^L, \rho_s^L, \Lambda_s^L) : u \in [0, \tau] \right\} .$$

By Exercise 3 on page 165 in van der Vaart and Wellner (1996), I denote the number of brackets for $\mathcal{W}_{\epsilon,s}$ by $[u_{\epsilon,s,t}^L, u_{\epsilon,s,t}^U]$, $t = 1, \ldots N_\epsilon^{\epsilon,s}$, where $\log(N_\epsilon^{\epsilon,s})$ is in a polynomial order. That is to say, for

arbitrary $\forall u \in [0,\tau]$, there is some $t = 1,\ldots,N_\epsilon^{\epsilon,s}$ such that

$$W(u^L_{\epsilon,s,t},\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s) \le W(u,\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s) \le W(u^U_{\epsilon,s,t},\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s)\,,$$
$$||W(u^L_{\epsilon,s,t},\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s) - W(u^U_{\epsilon,s,t},\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s)||_{L_2(P_0)} < \frac{\epsilon}{2}\,.$$

For an arbitrary function $W(u',\boldsymbol{O};\beta',\rho',\Lambda') \in \mathcal{W}$, it is contained in bracket

$$\left[W(u^L_{\epsilon,s,t},\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s) - A_0\left\{||\beta^U_s-\beta^L_s|| + ||\rho^U_s-\rho^L_s|| + h(\boldsymbol{O};\Lambda^L_s,\Lambda^U_s)\right\},\right.$$
$$\left. W(u^U_{\epsilon,s,t},\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s) + A_0\left\{||\beta^U_s-\beta^L_s|| + ||\rho^U_s-\rho^L_s|| + h(\boldsymbol{O};\Lambda^L_s,\Lambda^U_s)\right\}\right]\,,$$

such that the distance between the boundary functions:

$$\begin{aligned}
& ||W(u^U_{\epsilon,s,t},\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s) - W(u^L_{\epsilon,s,t},\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s) + 2A_0\left\{||\beta^U_s-\beta^L_s|| + ||\rho^U_s-\rho^L_s|| + h(\boldsymbol{O};\Lambda^L_s,\Lambda^U_s)\right\}||_{L_2(P_0)} \\
= \quad & ||W(u^U_{\epsilon,s,t},\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s) - W(u^L_{\epsilon,s,t},\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s) + 2A_0\left\{||\beta^U_s-\beta^L_s|| + ||\rho^U_s-\rho^L_s|| + h(\boldsymbol{O};\Lambda^L_s,\Lambda^U_s)\right\}||_{L_2(P_0)} \\
\le \quad & ||W(u^L_{\epsilon,s,t},\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s) - W(u^U_{\epsilon,s,t},\boldsymbol{O};\beta^L_s,\rho^L_s,\Lambda^L_s)||_{L_2(P_0)} \\
& +2A_0||\beta^L_s-\beta^U_s|| + 2A_0||\rho^L_s-\rho^U_s|| + 2A_0||h(\boldsymbol{O};\Lambda^L_s,\Lambda^U_s)||_{L_2(P_0)} \\
\le \quad & \frac{\epsilon}{2}+\frac{\epsilon}{6}+\frac{\epsilon}{6}+\frac{\epsilon}{6} < \epsilon\,.
\end{aligned}$$

Since it is straightforward to show the number of brackets is in the order $\exp(O(1)/\epsilon)$, $\mathcal{W}$ is P_0-Donsker by definition. □

C.0.3 The third class of functions

The function class indexed by the estimator sequence $\left\{\left(\hat\beta_{n'},\hat\rho_{n'},\hat\Lambda_{n'}\right),\left(\beta_0,\rho_0,\tilde\Lambda_{n'}\right)\right\}$:

$$\{clog(\boldsymbol{O};\hat\beta_{n'},\hat\rho_{n'},\hat\Lambda_{n'}) - clog(\boldsymbol{O};\beta_0,\rho_0,\tilde\Lambda_{n'});\ \boldsymbol{O}\sim P_0\}$$

is P_0-Glivenko-Cantelli.

Proof. First, I want to show this function class is contained by another function class $\mathcal{G}_0$ defined by:

$$\begin{aligned}
\mathcal{G}_0 \quad = \quad & \{f(\boldsymbol{O};\beta_1,\rho_1,\Lambda_1,\beta_2,\rho_2,\Lambda_2) = clog(\boldsymbol{O};\beta_1,\rho_1,\Lambda_1) - clog(\boldsymbol{O};\beta_2,\rho_2,\Lambda_2) : \\
& (\beta_1,\rho_1,\Lambda_1),(\beta_2,\rho_2,\Lambda_2)\in\Theta, y\mapsto\frac{\Delta\Lambda_1}{\Delta\Lambda_2}(u)\in[m_1,M_1] \text{ and is } BV_{M_2}; \boldsymbol{O}\sim P_0\}\,.
\end{aligned}$$

This relationship is true

$$\frac{\Delta\hat\Lambda_{n'}}{\Delta\tilde\Lambda_{n'}}(u) = \frac{W_{n'}(u;\beta_0,\rho_0,\Lambda_0)}{W_{n'}(u;\hat\beta_{n'},\hat\rho_{n'},\hat\Lambda_{n'})}\,.$$

I consider the partition at observed failure event time points corresponding to the dataset generating $(\hat{\beta}_{n'}, \hat{\rho}_{n'}, \hat{\Lambda}_{n'})$: $t_1 < t_2 < \ldots < t_Q$, and set $t_0 = 0$, $t_{Q+1} = \tau$; I write the total variation of $\frac{\Delta\hat{\Lambda}_{n'}}{\Delta\tilde{\Lambda}_{n'}}$ as:

$$\begin{aligned}
& \sum_{q=0}^{Q} \left| \frac{W_{n'}(t_{q+1}; \beta_0, \rho_0, \Lambda_0)}{W_{n'}(t_{q+1}; \hat{\beta}_{n'}, \hat{\rho}_{n'}, \hat{\Lambda}_{n'})} - \frac{W_{n'}(t_q; \beta_0, \rho_0, \Lambda_0)}{W_{n'}(t_q; \hat{\beta}_{n'}, \hat{\rho}_{n'}, \hat{\Lambda}_{n'})} \right| \\
= & \sum_{q=0}^{Q} \frac{\left| W_{n'}(t_{q+1}; \beta_0, \rho_0, \Lambda_0) W_{n'}(t_q; \hat{\beta}_{n'}, \hat{\rho}_{n'}, \hat{\Lambda}_{n'}) - W_{n'}(t_q; \beta_0, \rho_0, \Lambda_0) W_{n'}(t_{q+1}; \hat{\beta}_{n'}, \hat{\rho}_{n'}, \hat{\Lambda}_{n'}) \right|}{W_{n'}(t_{q+1}; \hat{\beta}_{n'}, \hat{\rho}_{n'}, \hat{\Lambda}_{n'}) \cdot W_{n'}(t_q; \hat{\beta}_{n'}, \hat{\rho}_{n'}, \hat{\Lambda}_{n'})} \\
\leq & \sum_{q=0}^{Q} \frac{\left| W_{n'}(t_{q+1}; \beta_0, \rho_0, \Lambda_0) W_{n'}(t_q; \hat{\beta}_{n'}, \hat{\rho}_{n'}, \hat{\Lambda}_{n'}) - W_{n'}(t_q; \beta_0, \rho_0, \Lambda_0) W_{n'}(t_{q+1}; \hat{\beta}_{n'}, \hat{\rho}_{n'}, \hat{\Lambda}_{n'}) \right|}{m_2^2}
\end{aligned} \tag{C.1}$$

where m_2 denotes the lower bound of $W(\cdot; \cdot, \cdot, \cdot)$.

For some $0 \leq q < Q$, suppose $1\{\delta_{ij} y_{ij} \in (t_q, t_{q+1}]\} = 1$, then

$$\begin{aligned}
& W_m(t_q; \beta, \rho, \Lambda) - W_m(t_{q+1}; \beta, \rho, \Lambda) \\
= & \frac{1}{m} \sum_{i=1}^{m} \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \sum_{k \neq j} \left\{ (2 + \delta_{ik}) \frac{e^{z_{ij}^T \beta} \left[1 + (1 - \rho_{jk}) e^{z_{ik}^T \beta} \Lambda(y_k) \right]}{v(X_{ij}, X_{ik}; \beta, \Lambda, \rho)} \right. \\
& \left. - \frac{\delta_{ik}(1 - \rho_{jk}) e^{z_{ij}^T \beta} \left[(1 - \rho_{jk}) e^{z_{ik}^T \beta} \Lambda(y_{ik}) + 1 \right]}{w(X_{ij}, X_{ik}; \beta, \Lambda, \rho)} \right\} \\
\leq & \frac{1}{m_2} M_2
\end{aligned}$$

uniformly for (β, ρ, Λ) varying over Θ; i.e.

$$\{\hat{\Lambda}_{n'}, \tilde{\Lambda}_{n'}\} \subset \left\{ (\Lambda_1, \Lambda_2) : \Lambda_1 \in \mathcal{L}, \Lambda_2 \in \mathcal{L}, y \mapsto \frac{\Delta\Lambda_1}{\Delta\Lambda_2}(u) \in [m_1, M_1] \text{ and is } BV_{M_2} \right\} .$$

If can show $\mathcal{G}_0$ is P_0-Glivenko-Cantelli, the proof is completed. In the following I show $\mathcal{G}_0$ is

P_0-Glivenko-Cantelli. I write functions from $\mathcal{G}_0$ as

$$
\begin{aligned}
& clog(\boldsymbol{O};\beta_1,\rho_1,\Lambda_1) - clog(\boldsymbol{O};\beta_2,\rho_2,\Lambda_2) \\
= \quad & \sum_{j=1}^{n}\Delta_j\left\{log\frac{\Delta\Lambda_1(Y_j)}{\Delta\Lambda_2(Y_j)}\right\} + \sum_{j=1}^{n} Z_j^T(\beta_1-\beta_2) \\
& -\frac{1}{n-1}\sum_{j<k}\left[(1+\Delta_j+\Delta_k)log\frac{(1-\rho_{1jk})e^{Z_j^T\beta_1}e^{Z_k^T\beta_1}\Lambda_1(Y_j)\Lambda_1(Y_k)+e^{Z_j^T\beta_1}\Lambda_1(Y_j)+e^{Z_k^T\beta_1}\Lambda_1(Y_k)+1}{(1-\rho_{2jk})e^{Z_j^T\beta_2}e^{Z_k^T\beta_2}\Lambda_2(Y_j)\Lambda_2(Y_k)+e^{Z_j^T\beta_2}\Lambda_2(Y_j)+e^{Z_k^T\beta_2}\Lambda_2(Y_k)+1}\right] \\
& +\frac{1}{n-1}\sum_{j<k}\Big(\Delta_j\Delta_k(1-\rho_{1jk})^2e^{Z_j^T\beta_1}e^{Z_k^T\beta_1}\Lambda_1(Y_j)\Lambda_1(Y_k)+\Delta_k(1-\rho_{1jk})e^{Z_j^T\beta_1}\Lambda_1(Y_j) \\
& \quad +\Delta_j(1-\rho_{1jk})e^{Z_k^T\beta_1}\Lambda_1(Y_k)+1+\Delta_j\Delta_k\rho_{1jk}\Big)\Big/ \\
& \quad \Big(\Delta_j\Delta_k(1-\rho_{2jk})^2e^{Z_j^T\beta_2}e^{Z_k^T\beta_2}\Lambda_2(Y_j)\Lambda_2(Y_k)+\Delta_k(1-\rho_{2jk})e^{Z_j^T\beta_2}\Lambda_2(Y_j) \\
& \quad +\Delta_j(1-\rho_{2jk})e^{Z_k^T\beta_2}\Lambda_2(Y_k)+1+\Delta_j\Delta_k\rho_{2jk}\Big)\ .
\end{aligned}
$$

The first term in the above is P_0-Glivenko-Cantelli with a similar argument as for $\mathcal{F}_2$. The remaining terms form a Lipschitz function of $(\beta_1,\rho_1,\Lambda_1(Y_1),\ldots,\Lambda_1(Y_n),\beta_2,\rho_2,\Lambda_2(Y_1),\ldots,\Lambda_2(Y_n))$ and similar to $\mathcal{W}$ argument, they also form a P_0-Glivenko-Cantelli class of functions. By addition preservation Corollary 9.27 in Kosorok (2008), I have shown $\mathcal{G}_0$ is P_0-Glivenko-Cantelli. □