Marginal Models for Modeling Clustered Failure Time Data

by

Yi Niu

A thesis submitted to the
Department of Mathematics and Statistics
in conformity with the requirements
for the degree of

Doctor of Philosophy

Queen's University
Kingston, Ontario, Canada
January 2013

Dedicated to my grandma, Qiue Yang and Qiang Zhang

Abstract

Clustered failure time data often arise in biomedical and clinical studies where potential correlation among survival times is induced in a cluster. In this thesis, we develop a class of marginal models for right censored clustered failure time data and propose a novel generalized estimating equation approach in a likelihood-based context. We first investigate a semiparametric proportional hazards model for clustered survival data and derive the large sample properties of the regression estimators. The finite sample studies demonstrate that the good applicability of the proposed method as well as the substantial efficiency improvement in comparison with the existing marginal model for clustered survival data.

Another important feature of failure time data we will consider in this thesis is a possible fraction of cured subjects. To accommodate the potential cure fraction, we consider a proportional hazards mixture cure model for clustered survival data with long-term survivors and develop a set of estimating equations by incorporating working correlation matrices in an EM algorithm. The dependence among the cure statuses and among the survival times of uncured patients within clusters are modeled by working correlation matrices in the estimating equations. For the parametric proportional hazards mixture cure model, we show that the estimators of the regression parameters and the parameter in the baseline hazard function are consistent

and asymptotically normal with a sandwich covariance matrix that can be consistently estimated. A numerical study presents that the proposed estimation method is comparable with the existing parametric marginal method.

We also extend the proposed generalized estimating equation approach to a semiparametric proportional hazards mixture cure model where the baseline survival function is nonparametrically specified. A bootstrap method is used to obtain the variances of the estimates. The proposed method is evaluated by a simulation study from which we observe a noticeable efficiency gain of the proposed method over the existing semiparametric marginal method for clustered failure time data with a cure fraction.

Acknowledgments

First and foremost, I would like to thank my supervisor, Professor Yingwei Peng, for providing me the precious opportunity and quality environment to carry out my graduate studies. From the first email he replied for my application to the last revision of my Ph.D. thesis, from weekly group meetings to casual individual discussions, I can always feel his enthusiasm and patience. Without his excellent guidance, constant encouragement and continuous support, this thesis would not have been possible.

I would like to thank my thesis committee members, Professor Wenqing He, Professor Dongsheng Tu, Professor Wenyu Jiang, Professor Fady Alajaji, Professor Robin Hutchinson, for their careful reading of this thesis. Their insightful comments and suggestions lead to great improvement of my work.

I want to express my sincere gratitude to Professor Dongsheng Tu for generously sharing his experience and knowledge during my graduate studies. I would like to thank Professor Alan Ableson, Professor Wenyu Jiang, and Mr. Charles Molson for their dedications in helping me improve my teaching skills. I want to thank Professor Agnes Herzberg and Professor Chunfang Devon Lin for their encouragement and invaluable advices in my research.

I am deeply grateful to Ms. Jennifer Read, the Graduate Program Assistant in the Department of Mathematics and Statistics, for her consistent support not only to me but also to all graduate students in almost all aspects of our life at Queen's

University.

I would like to thank the Department of Mathematics and Statistics and the Cancer Research Institute at Queen's University. Special thanks to Dr. Jianfeng Xu, Dr. Shan Jiang, Weidong Kong, Qun Miao, Brenda Bass for their help and encouragement. Thanks to all my friends, all of you make my life at Kingston more enjoyable and colorful.

Finally, I want to thank my beloved wife, Hui, for her love so that I am not alone in this long journey of graduate studies. I want to express my deepest appreciation to our parents and grandparents for their love, kindness, encouragement, patience and support. One of the best experiences I have had during the Ph.D. period was the birth of my daughter, Jessie, whose wonderful smile is an amazing power for me to complete this thesis.

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work and that any ideas or quotations from the work of other people, published or unpublished, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Clustered Failure Time Data

Clustered failure time data are frequently observed in biomedical and epidemiologic research. For example, times to occurrence of blindness of two eyes from the same patient with diabetic retinopathy are possibly correlated, ages at diagnosis of breast cancer from female siblings may be associated due to similar genetic structures, or failure times of patients from the same community may be related because of shared environments or treatment resources. Therefore, it is important to take the correlation into account when analyzing clustered failure times.

In some cancer studies, a fraction of patients may respond favorably to the treatment and have long-term censored survival times. They are often considered cured in the sense that they will not experience relapse/death due to the cancer even after an extended follow-up. For example, in breast cancer study (Farewell, 1986) the Kaplan-Meier survival curves from three treatment groups level off to nonzero proportions and a number of long-term censored observations appear at the tail of these curves. In a head and neck cancer study (Taylor, 1995) only between 5 and 50% of patients experienced local recurrences whereas the remaining patients were free of symptoms of

the cancer at the end of the sufficiently long observation period. Due to the existence of long-term survivors in some cancer studies, the use of standard survival models, which assume that all subjects would eventually experience the event of interest, will not be appropriate for the analysis of the failure time data with a cure proportion.

1.2 Motivating Examples

1.2.1 Diabetic Retinopathy Study

The well-known Diabetic Retinopathy Study (Diabetic Retinopathy Study Research Group, 1981) was conducted to assess the effectiveness of laser photocoagulation

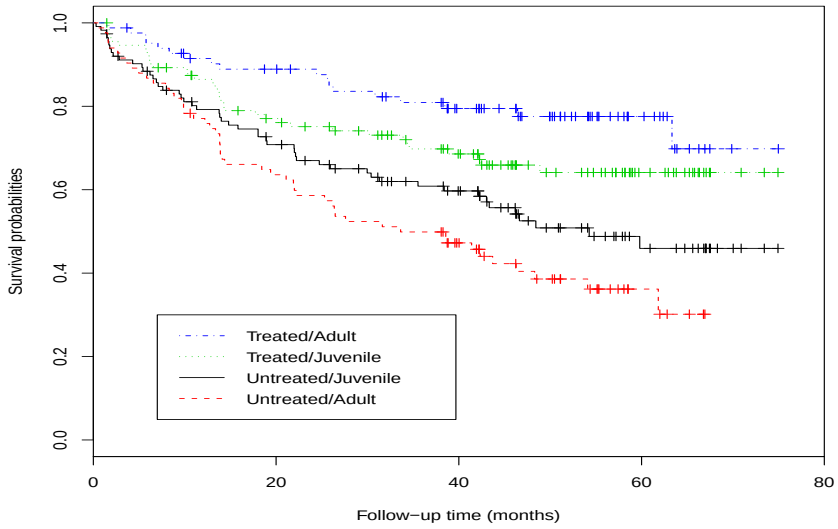

Figure 1.1: Kaplan-Meier survival curves for the time to blindness stratified by treatment and type of diabetes.

in delaying visual loss among patients with diabetic retinopathy. There were 1742 patients entered this study between 1972 to 1975. One eye of each patient was randomly selected to receive the laser treatment while the other eye was observed without treatment. The endpoint used to assess the treatment effect is the time (in

months) to the first occurrence of visual acuity less than 5/200. Besides the effects of treatment, the types of diabetes as well as the interaction between them (Figure 1.1), we are also interested in the potential dependence between a patients's two eyes which form a cluster. This data set will be analyzed in Chapter 3.

1.2.2 The Study of Infections in Kidney Patients

We consider a data set on the recurrence times (in days) of infections, at the point of insertion of the catheter, for 38 kidney patients using the same type of portable

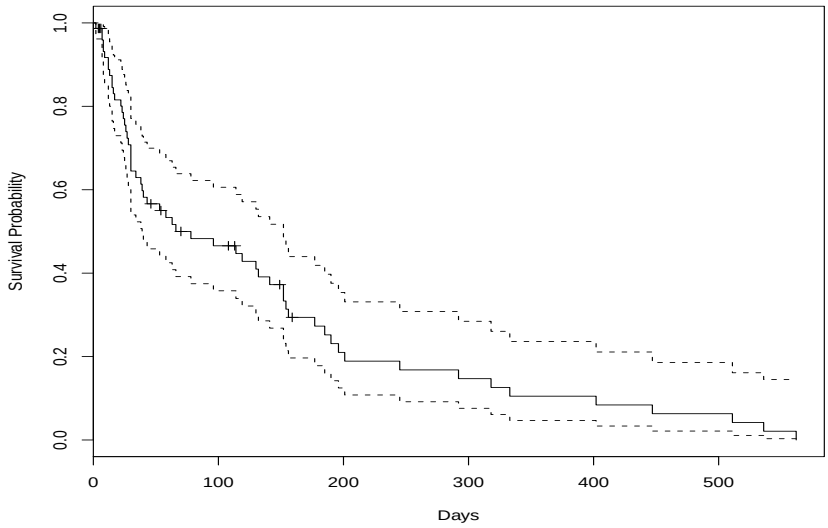

Figure 1.2: Kaplan-Meier survival curve and its pointwise 95% confidence interval for the kidney data.

dialysis equipment (McGilchrist and Aisbett, 1991). Two recurrence times and the corresponding censoring indicators were recorded for each patient. As introduced by Cleves et al. (2008), the first recurrence time to infection is measured when a catheter is inserted. The second recurrence time to infection is measured as time elapsed between the second insertion and the second infection or censoring. The primary interest of the study is to assess the factors such as age, gender and the

type of kidney disease to the development of infections. We plot the Kaplan-Meier survival curve based on 76 observations. Figure 1.2 shows that the patients experience the infection given sufficient follow-up time. Meanwhile, the correlation between the recurrence times within each patient is of interest. This data set will be analyzed in Chapter 3.

1.2.3 Smoking Cessation Study

We consider a data set from a smoking cessation study (Banerjee and Carlin, 2004). The original data consist of 223 people enrolled in the study between November 1986 and February 1989 from 51 zip codes in the southeastern corner of Minnesota in the United States. In this study, smokers were randomly assigned to one of two treatment

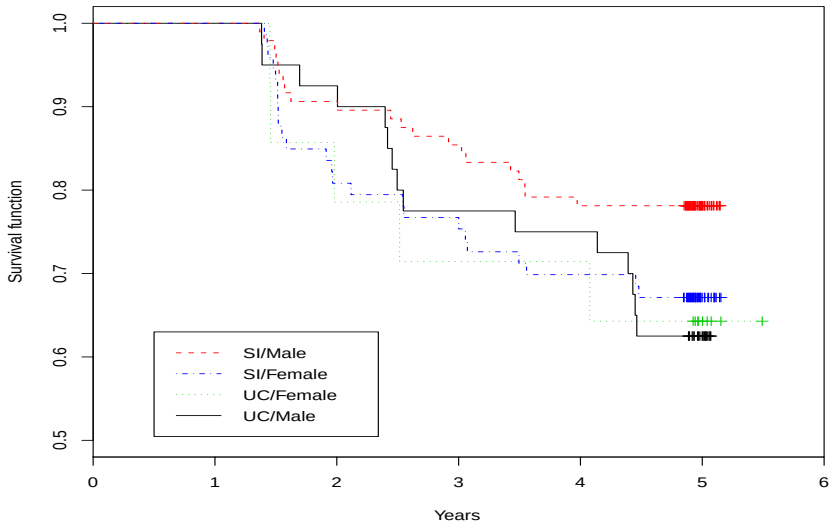

Figure 1.3: Kaplan-Meier survival curves of smoking cessation data stratified by intervention type and sex.

groups: smoking intervention (SI) group or usual care (UC) group. The survival time is defined as the time (in years) required for a failed quitter to resume smoking. The people residing in the area with the same zip code form a cluster and may be spatially

correlated due to the shared environment. Also the data reveals (Murray et al. 1998) that many former smokers have successfully given up smoking. Therefore, a cure fraction exists in this data set. We plot the Kaplan-Meier survival curves by sex and intervention type in Figure 1.3. This data set will be analyzed in Chapter 4.

1.2.4 Multi-Center Clinical Trial of Tonsil Carcinoma

We consider a data set from a tonsil cancer clinical trial study conducted by the Radiation Therapy Oncology Group in the United States. The survival time is defined as the time (in days) from diagnosis to death. In this study, patients in one institution were randomly assigned to one of two treatment groups: radiation therapy alone or radiation therapy together with a chemotherapeutic agent. A part of the data from the study is available in Kalbfleisch and Prentice (2002). We plot the Kaplan-Meier

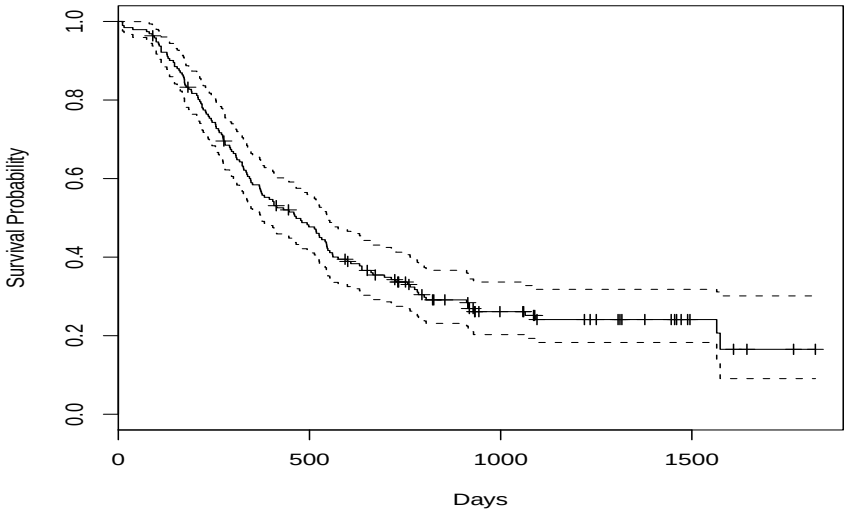

Figure 1.4: Kaplan-Meier survival curve and its pointwise 95% confidence interval for the tonsil data.

survival curve and its pointwise 95% confidence interval in Figure 1.4 and observe that the curve levels off at about 0.18, which suggests that a cure fraction may be

present in this data and a cure model should be considered. Another important feature of this data is that the patients are clustered by institutions in this study. The shared environment and the treatment facilities in one institution may induce correlation among the cure statuses and among the failure times of uncured patients in one institution. Therefore, it is important that both the cure fraction and the cluster effect are considered in the model for the data. This data set will be analyzed in Chapter 5.

1.2.5 Bone Marrow Transplantation Data

We consider the bone marrow transplantation data (Klein and Moeschberger, 2003). This multi-center acute leukemia study consists of 137 patients with acute myelocytic

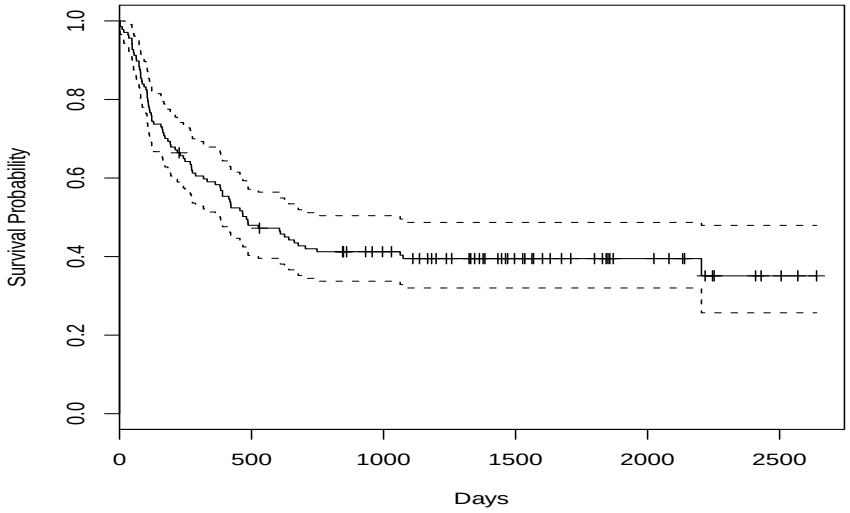

Figure 1.5: Kaplan-Meier survival curve and its pointwise 95% confidence interval for the leukemia data.

leukemia (AML) or acute lymphoblastic leukemia (ALL) aged 7 to 52 from March 1, 1984 to June 30, 1989 at four institutions. The failure time on study is defined as time (in days) to relapse or death. The Kaplan-Meier survival curve (Figure 1.5)

suggests that the existence of a cure proportion in acute leukemia patients and a cure model should be applied to the data. In addition, the patients are clustered by four institutions which may induce correlation among the cure statuses and among the failure times of uncured patients. Therefore, both the cure fraction and the cluster effect should be considered. This data set will be analyzed in Chapter 5.

1.3 Organization of This Thesis

The objective in this thesis is to develop new marginal models for analyzing clustered failure time data with/without a cure fraction to improve the estimation efficiency. In Chapter 2, two important models including the Cox proportional hazards model and proportional hazards mixture cure model are presented under correlation structures. Then we review random effects models and marginal models for clustered failure time data with/without a cure proportion. The generalized estimating equations, the EM algorithm, and the ES algorithm are introduced at the end of this chapter.

The Cox proportional hazards model is considered as a standard model for investigating the classical clustered failure time data which assume that all subjects would eventually experience the event of interest. In Chapter 3, we revisit the marginal method developed by Segal and Neuhaus (1993) for classical clustered failure time data and propose an unbiased weighted estimating function for regression parameters in a semiparametric proportional hazards model.

When there exists a fraction of cured subjects in the clustered survival data, the marginal mixture cure model has been received much attention. In Chapter 4, we propose a new generalized estimating equation approach to modeling the clustered survival data with a cure fraction through a marginal parametric proportional hazards mixture cure model.

In Chapter 5, we consider a semiparametric marginal proportional hazards mixture cure model for clustered failure time data with a cure fraction. A set of generalized estimating equations are proposed for the regression parameters. We briefly discuss the iterative algorithm used for solving the equations.

Chapter 6 presents a summary of this dissertation contributions and future research directions.

Chapter 2

Literature Review of Relevant Models and Methods

In the first two sections of this chapter, we present two important models including the Cox proportional hazards model and the mixture cure models that will be extended in the following chapters. Based on these two models, in Section 2.3, we review the random effects models and marginal models which are commonly used to handle the potential correlation within clustered survival times. In Section 2.4, we describe the generalized estimating equation approach which is an extension of generalized linear models by explicitly incorporating the correlation structure in the estimation procedure. The EM algorithm is introduced in Section 2.5. A review of the ES algorithm which is a combination of the EM algorithm and the generalized estimating equation approach will be given in Section 2.6.

Throughout this thesis, we assume that there are n_i individuals in the ith ($i = 1, \cdots, K$) cluster, and K clusters in total. The total number of observations is $N = \sum_{i=1}^{K} n_i$. The function $I(A) = 1$ if A is true and 0 otherwise. Let $\tilde{T}_{ij}$ and C_{ij} be the failure and censoring times for the jth subject in the ith cluster ($j = 1, \cdots, n_i, i = 1, \cdots, K$) where $\tilde{T}_{ij} < \infty$. Let $T_{ij} = \min(\tilde{T}_{ij}, C_{ij})$ be the observed failure time and

$\delta_{ij} = I(\tilde{T}_{ij} \leq C_{ij})$ be the right censoring indicator. The censoring mechanism is assumed non-informative. That is, the censoring time is statistically independent of the failure time given observed covariates. X_{ij} is a vector of time independent covariates that may have effect on the failure time distribution. Given X_{ij}, we assume that $\tilde{T}_{ij}$ and $\tilde{T}_{ij'}$ are correlated in a cluster if $j \neq j'$. However, $\tilde{T}_{ij}$ and $\tilde{T}_{i'j'}$ are independent if $i \neq i'$.

Although it is common to assume that all subjects would eventually experience the event of interest, in some social and biomedical studies (see Examples 1.2.3, 1.2.4, and 1.2.5), a certain fraction of the population may never experience a particular type of failure and is often considered as cured. Graphically, these fractions are often characterized by the survival curves being leveled off at nonzero probabilities. Here we let Y_{ij} denote the cure status of subject j in cluster i, that is, $Y_{ij} = 1$ if the subject is uncured (susceptible) and 0 otherwise.

Similar to the definition of $\tilde{T}_{ij}$, we let $\tilde{T}^*_{ij}$ be the failure time for the jth subject which may be cured in the ith cluster ($j = 1, \cdots, n_i, i = 1, \cdots, K$) where $\tilde{T}^*_{ij} \leq \infty$. Therefore, the modeling of cure rate is a decomposition of the failure time, i.e., $\tilde{T}^*_{ij} = Y_{ij}\tilde{T}_{ij} + (1 - Y_{ij})\infty$ where $\tilde{T}^*_{ij} = \tilde{T}_{ij}$ denotes the failure time of a susceptible subject and $\tilde{T}^*_{ij} = \infty$ denotes the event that the individual will not experience relapse or death from the cause of interest. The observed failure time is $T^*_{ij} = \min(\tilde{T}^*_{ij}, C_{ij})$. Let $1-\pi(Z_{ij})$ denote the cure probability for the jth individual in the ith cluster. We define X_{ij} and Z_{ij} as two vectors of time independent covariates (these two vectors may share some covariates) that may have effects respectively on the failure time distribution of uncured subjects and the cure probability. It is obvious that if $\delta_{ij} = 1$, then $Y_{ij} = 1$. However, if $\delta_{ij} = 0$, the value of Y_{ij} is unknown and Y_{ij} is a latent variable. We further assume that given X_{ij} and Z_{ij}, $\tilde{T}^*_{ij}|\{Y_{ij} = 1\}$ and $\tilde{T}^*_{ij'}|\{Y_{ij'} = 1\}$, and Y_{ij} and $Y_{ij'}$ are correlated respectively in a cluster if $j \neq j'$. However, $\tilde{T}^*_{ij}|\{Y_{ij} = 1\}$

and $\tilde{T}^*_{i'j'}|\{Y_{i'j'}=1\}$, and Y_{ij} and $Y_{i'j'}$ are respectively independent if $i \neq i'$. Note that if all subjects are assumed to experience the event of interest, the latent variable Y_{ij} is known and equal to 1 and the cure probability is 0.

Let $0 < \tau_1 < \tau_2 < \cdots < \tau_k < \infty$ denote the k distinct ordered event times. Let d_s be the number of deaths at τ_s and $\mathcal{D}_s$ be the set of all individuals who die at time τ_s, i.e., $\mathcal{D}_s = \{(i,j) : T_{ij} = \tau_s, \delta_{ij} = 1; j = 1, \cdots, n_i, i = 1, \cdots, K\}$. The risk set $\mathcal{R}_s$ is defined by $\mathcal{R}_s = \{(i,j) : T_{ij} \geq \tau_s; j = 1, \cdots, n_i, i = 1, \cdots, K\}$, i.e., the set of individuals alive and uncensored just prior to τ_s. Let $\mathcal{E}_s$ be the set of individuals with censoring times in $[\tau_s, \tau_{s+1}), s = 0, \cdots, k$, where $\tau_0 = 0$ and $\tau_{k+1} = \infty$, i.e., $\mathcal{E}_s = \{(i,j) : \tau_s \leq T_{ij} < \tau_{s+1}, \delta_{ij} = 0; j = 1, \cdots, n_i, i = 1, \cdots, K\}$.

2.1 Cox Proportional Hazards Model

The Cox proportional hazards model (Cox, 1972) is a multiplicative hazards model as well as a semiparametric model because a parametric form is assumed only for the covariate effect and the baseline hazard rate is treated nonparametrically. Given the classical clustered survival data $O = \{(T_{ij}, \delta_{ij}, X_{ij}), j = 1, \cdots, n_i, i = 1, \cdots, K\}$, the hazard rate $\lambda(t; X_{ij})$ for the jth individual in the ith cluster is defined by

$$\lambda(t; X_{ij}) = \lambda_0(t) \exp(\beta' X_{ij}), \tag{2.1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function and β is a $p_X \times 1$ unknown parameter vector for X_{ij}.

It is straightforward to extend the partial likelihood proposed by Cox (1975) to the clustered failure time data under the assumption that the survival times are

independent of each other (Wei et al., 1989; Lee et al.,1992; Lin, 1994). That is

$$\mathcal{L}(\beta) = \prod_{s=1}^{k} \frac{\exp(\beta' X_s)}{\{\sum_{(i,j)\in\mathcal{R}_s} \exp(\beta' X_{ij})\}^{d_s}}, \tag{2.2}$$

where X_s is the sum of the vectors X_{ij} over all individuals who die at time τ_s, i.e., $X_s = \sum_{(i,j)\in\mathcal{D}_s} X_{ij}$. Let $\hat{\beta}$ be the value that maximizes $\mathcal{L}(\beta)$, then the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ could be estimated by

$$\hat{\Lambda}_0(t) = \sum_{s:\tau_s \le t} \frac{d_s}{\sum_{(i,j)\in\mathcal{R}_s} \exp(\hat{\beta}' X_{ij})}, \tag{2.3}$$

which is the Breslow's estimator. We will apply the proportional hazards model for classical clustered survival data in Chapter 3.

2.2 Mixture Cure Models

Mixture cure models (Boag, 1949; Berkson and Gage, 1952) postulating a subpopulation of cured patients are intriguing from both the biological and statistical viewpoints. The models are composed by the probability of being a long-term survivor plus the probability of a death which occurs after time t (Farewell, 1982). They are often employed to handle the survival data which may contain a cure proportion. We let $S(t; X_{ij}, Z_{ij})$ and $S_u(t; X_{ij})$ denote the marginal survival functions of $\tilde{T}^*_{ij}$ and $\tilde{T}^*_{ij}|\{Y_{ij} = 1\}$, respectively. We say that the marginal survival function $S(t; X_{ij}, Z_{ij})$ is from a mixture cure model if

$$S(t; X_{ij}, Z_{ij}) = P(\tilde{T}^*_{ij} > t; X_{ij}, Z_{ij}) = 1 - \pi(Z_{ij}) + \pi(Z_{ij}) S_u(t; X_{ij}), \tag{2.4}$$

where the uncure rate $\pi(Z_{ij})$ is considered as the logistic form, i.e.,

$$\pi(Z_{ij}) = P(Y_{ij} = 1; Z_{ij}) = \frac{\exp(\gamma' Z_{ij})}{1 + \exp(\gamma' Z_{ij})} \tag{2.5}$$

and γ is unknown regression parameter for Z_{ij}. Other link functions such as complementary log-log, i.e., $\pi(Z_{ij}) = \exp(-e^{\gamma' Z_{ij}})$ or probit function, i.e., $\pi(Z_{ij}) = \Phi(\gamma' Z_{ij})$ where Φ is the cumulative distribution function of the standard normal distribution may also be applied to describe the effects of covariate Z_{ij} on uncure rate $\pi(Z_{ij})$ (Peng, 2003).

As we discussed in Section 2.1.2, the proportional hazards model is popular in the analysis of classical clustered survival time data. Similarly, this model can be used to describe the survival function of uncured patients, i.e., $S_u(t; X_{ij})$, in the mixture cure model. Therefore, one can assume

$$S_u(t; X_{ij}) = P(\tilde{T}^*_{ij} > t | Y_{ij} = 1; X_{ij}) = S_{u0}(t; \alpha)^{\exp(\beta' X_{ij})}, \tag{2.6}$$

where $S_{u0}(t; \alpha)$ is the baseline survival function of $\tilde{T}^*_{ij} | \{Y_{ij} = 1\}$ when $X_{ij} = 0$ and α denotes a set of unknown parameters in the baseline distribution. Here β is a set of unknown regression parameters for X_{ij}. Let $\theta^* = (\gamma, \beta, \alpha)$. An alternative to the commonly-used proportional hazards model for the uncured patients is accelerated failure time model, i.e., $S_u(t; X_{ij}) = S_{u0}(te^{\beta' X_{ij}})$ where the baseline survival function could be an extended generalized gamma distribution (Yamaguchi, 1992), a generalized F distribution (Peng et al., 1998), or a nonparametric form (Li and Taylor, 2002; Zhang and Peng, 2007). Also a transformation model which accommodates the proportional hazards model and proportional odds model was considered by Lu and Ying (2004) to model $S_u(t; X_{ij})$.

The mixture cure model (2.4) composed by the logistic model (2.5) and the proportional hazards model (2.6) is called the proportional hazards mixture cure model. We will focus on this model for clustered survival data with a cure fraction in Chapters 4 and 5. Under the complete clustered survival data $O_c^* = \{(T_{ij}^*, \delta_{ij}, X_{ij}, Z_{ij}, Y_{ij}), j = 1, \cdots, n_i, i = 1, \cdots, K\}$ and the independent observation assumption, an adjusted likelihood function for β (Peng and Dear, 2000; Sy and Taylor, 2000) could be written as

$$\mathcal{L}^*(\beta; g^{(m)}) = \prod_{s=1}^{k} \frac{\exp(\beta' X_s)}{\{\sum_{(i,j)\in\mathcal{R}_s} g_{ij}^{(m)} \exp(\beta' X_{ij})\}^{d_s}}, \tag{2.7}$$

where

$$\begin{aligned} g_{ij}^{(m)} &= E(Y_{ij}|\theta^{*(m)}, O^*) \\ &= \left[\delta_{ij} + \frac{(1-\delta_{ij})\pi(Z_{ij}) S_{u0}(t_{ij}^*)^{\exp(\beta' X_{ij})}}{1 - \pi(Z_{ij}) + \pi(Z_{ij}) S_{u0}(t_{ij}^*)^{\exp(\beta' X_{ij})}}\right]_{\theta=\theta^{*(m)}} \end{aligned} \tag{2.8}$$

where $\theta^{*(m)}$ is the current estimate of θ^* at the mth iteration of the EM algorithm and $O^* = \{(T_{ij}^*, \delta_{ij}, X_{ij}, Z_{ij}), j = 1, \cdots, n_i, i = 1, \cdots, K\}$ is the observed cluster failure time data.

Following the discussion of Kalbfleisch and Prentice (2002, P.115) for the proportional hazards model, the contribution of the likelihood of an individual who fails at τ_s is $S(\tau_s^-) - S(\tau_s)$ and the contribution of a censored observation at time t is $S(t)$. Then the likelihood function for α, the parameters in the baseline distribution, can

be rewritten as

$$\begin{aligned} L(\alpha; y_{ij}) &= \prod_{s=0}^{k}\{ \prod_{(i,j)\in\mathcal{D}_s} [S_u(\tau_s^-; X_{ij}) - S_u(\tau_s; X_{ij})] \prod_{(i,j)\in\mathcal{E}_s} S_u(t; X_{ij})^{y_{ij}}\} \\ &= \prod_{s=0}^{k}[\prod_{(i,j)\in\mathcal{D}_s} \{\lambda_u(\tau_s; X_{ij}) S_{u0}(\tau_s^-)^{\exp(\beta' X_{ij})}\} \prod_{(i,j)\in\mathcal{E}_s} S_{u0}(\tau_s)^{y_{ij}\exp(\beta' X_{ij})}]. \end{aligned}$$

Furthermore, a discrete proportional hazards model is assumed such that $S_{u0}(t)$ has the product-limit form $S_{u0}(t) = \prod_{s:\tau_s\le t} \alpha_s$. The α's are nonnegative parameters at each of the k distinct event times with $\alpha_0 = 1$ and $0 \le \alpha_s \le 1$. $S_{u0}(\tau_s^-) = \prod_{l=1}^{s-1} \alpha_l$ and $S_{u0}(\tau_s) = \prod_{l=1}^{s} \alpha_l$. $\lambda_u(\tau_s; X_{ij}) = 1 - \alpha_s^{\exp(\beta' X_{ij})}$ is the hazard function given X_{ij}. Rearranging terms, we obtain

$$\begin{aligned} L(\alpha; y_{ij}) &= \prod_{s=0}^{k}[\prod_{(i,j)\in\mathcal{D}_s} \{(1 - \alpha_s^{\exp(\beta' X_{ij})}) \prod_{l=0}^{s-1} \alpha_l^{\exp(\beta' X_{ij})}\} \prod_{(i,j)\in\mathcal{E}_s} \{\prod_{l=0}^{s} \alpha_l^{y_{ij}\exp(\beta' X_{ij})}\}] \\ &= \prod_{s=1}^{k}[\prod_{(i,j)\in\mathcal{D}_s} (1 - \alpha_s^{\exp(\beta' X_{ij})})] \\ &\qquad \times \prod_{s=1}^{k}[\prod_{(i,j)\in\mathcal{D}_s} \{\prod_{l=0}^{s-1} \alpha_l^{\exp(\beta' X_{ij})}\} \prod_{(i,j)\in\mathcal{E}_s} \{\prod_{l=0}^{s} \alpha_l^{y_{ij}\exp(\beta' X_{ij})}\}] \\ &= \prod_{s=1}^{k}[\prod_{(i,j)\in\mathcal{D}_s} (1 - \alpha_s^{\exp(\beta' X_{ij})})] \prod_{s=1}^{k}[\prod_{(i,j)\in\mathcal{R}_s - \mathcal{D}_s} \alpha_s^{y_{ij}\exp(\beta' X_{ij})}] \\ &= \prod_{s=1}^{k}\{ \prod_{(i,j)\in\mathcal{D}_s} (1 - \alpha_s{}^{\exp(\beta' X_{ij})}) \prod_{(i,j)\in\mathcal{R}_s - \mathcal{D}_s} \alpha_s^{y_{ij}\exp(\beta' X_{ij})}\}. \end{aligned}$$

After taking derivatives of $E_Y(L(\alpha; y_{ij}))$ with respect to α_s, we obtain the estimating equations for each α_s given β and γ. That is

$$\sum_{(i,j)\in\mathcal{D}_s} \frac{e^{\beta' X_{ij}}}{1 - \alpha_s^{\exp(\beta' X_{ij})}} = \sum_{(i,j)\in\mathcal{R}_s} g_{ij}^{(m)} \exp(\beta' X_{ij}), \quad s = 1, \cdots, k. \tag{2.9}$$

These equations are similar to (5) in Sy and Taylor (2000) but with clustered data settings. Therefore, an approximate estimator for the nonparametric baseline cumulative hazard function $\hat{\Lambda}_{u0}^{(m)}(t)$ can be obtained (Peng and Dear, 2000) by

$$\hat{\Lambda}_{u0}^{(m)}(t) = \sum_{s:\tau_s \leq t} \frac{d_s}{\sum_{(i,j)\in\mathcal{R}_s} g_{ij}^{(m)} \exp(\beta' X_{ij})}. \tag{2.10}$$

Note that if all patients are uncured, then $g_{ij}^{(m)} \equiv 1$, the likelihood function (2.7) for β reduces to the partial likelihood function (2.2) in the proportional hazards model, and the estimating function (2.10) reduces to (2.3).

2.3 Existing Methods for Clustered Failure Time Data

To appropriately account for the correlation in a cluster, the two most studied approaches are random effects models and marginal models. Random effects models (frailty models, cluster-specific models, conditional models, or multilevel models) explicitly formulate the underlying dependence via a cluster specific variable known as the random effect representing the heterogeneity in each cluster. Marginal (population-averaged) models focus on the population average on the margins of the joint distribution of data from one cluster, and the correlation is often treated as a nuisance parameter to reduce the dependence of marginal models on the specification of unobservable correlation structures of clustered data.

These two models with applications to the proportional hazards model and the proportional hazards mixture cure model for modeling clustered survival data have received much attention for the last decades. For classical clustered survival data, the commonly used proportional hazards frailty model is the so-called shared frailty model (Klein and Moeschberger, 2003). That is, the hazard rate for the jth subject

in the ith cluster is of the multiplicative form

$$\lambda_{ij}(t) = \lambda_0(t) u_i \exp(\beta' X_{ij}), j = 1, \cdots, n_i, i = 1, \cdots, K,$$

where $\lambda_0(t)$ is an unspecified and arbitrary baseline hazard rate and u_i is an unobservable random effect (frailty) shared by subjects in the ith cluster. Therefore, the frailty u_i induces the dependence among the failure times in the ith cluster. Note that the survival times in the ith group are independent of each other given u_i and covariates. Usually we assume that the u_i's are an independent and identically distributed sample from a distribution with mean 1 and some unknown variance.

If $g(u)$ is the density function of the distribution of u_i, then the joint unconditional survival function of the failure times in group i is

$$\begin{aligned} S(t_{i1}, \cdots, t_{in_i}) &= P(\tilde{T}_{i1} > t_{i1}, \cdots, \tilde{T}_{in_i} > t_{in_i}) \\ &= \int \exp\left(-u \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\beta' X_{ij})\right) g(u) du. \end{aligned}$$

The frequently used distributions for frailty are the gamma distribution (Clayton, 1978; Clayton and Cuzick, 1985), the inverse Gaussian distribution (Hougaard, 1986a), the positive stable distribution (Hougaard, 1986b), and the log normal distribution (McGilchrist and Aisbett, 1991). Excellent discussions on the proportional hazards frailty model can be found in Hougaard (1995, 2000) and Therneau and Grambsch (2000).

For clustered survival data with a cure fraction, Yau and Ng (2001) considered the proportional hazards mixture cure model by using two independent normal random effects to characterize the correlation among cure statuses and the correlation among the failure times of uncured patients in a cluster. They proposed a best linear

unbiased prediction (BLUP) method to estimate the parameters in the model. Lai and Yau (2008) extended this method by allowing dependent random effects and a nonparametric baseline distribution in the model. Peng and Taylor (2011) considered maximum likelihood estimation for the mixture cure model with random effects. Their method provides flexibility in specifying distribution for the random effects and is computationally intensive because of the numerical integration involved in the method.

Chatterjee and Shih (2001) also extended the univariate mixture cure models to bivariate survival data. They modeled the correlation among the cure statuses and the failure times of uncured subjects in a familial cluster in a breast cancer study using pairwise odds ratios and a copula model respectively, and proposed a quasi-likelihood method to estimate the parameters in the model. Wienke et al. (2003) considered a full likelihood method with a similar model for bivariate data. Both methods do not consider covariate effects and the estimation methods become infeasible when cluster size is large.

As an alternative method to the random effects models, the marginal models take a population-average approach to model the marginal mean while treating the correlations as nuisance parameters. The proportional hazards model has been investigated extensively for correlated failure time data with the marginal method. Wei et al. (1989) introduced a marginal proportional hazards model for the multivariate failure time observations with respect to different types of failures. Huster et al. (1989) proposed a parametric marginal proportional hazards model for modeling paired survival data. Lee et al. (1992) analyzed the clustered survival data with the common baseline hazard function and showed that the regression parameters are consistent and asymptotically normal. Liang et al. (1993), Lin (1994), Spiekerman and Lin (1998), and Clegg et al. (1999) independently proposed a marginal mixed baseline hazards

model where the baseline hazards combine both common and distinguishable baselines. Yang and Ying (2001) introduced parametric models for ordered event times with proper joint density functions and marginal proportional hazards. Chen et al. (2010) analyzed marginal proportional hazards model based on a linear combination of martingale residuals.

To further improve the estimation efficiency, an estimating equation approach has been investigated by Segal and Neuhaus (1993), Cai and Prentice (1995, 1997), Prentice and Hsu (1997), and Gray and Li (2002), among others. This method clearly specifies working correlation structures in the estimating equations to accommodate the dependence of failure times in each cluster. Specifically, Segal and Neuhaus (1993) developed a synthesis of the Poisson regression model and generalized estimating equations based on a parametric proportional hazards model for multivariate survival data. Cai and Prentice (1995, 1997) derived a weighted partial likelihood estimating equation based on a counting process approach for correlated failure time data. They developed the asymptotic distribution for the hazard ratio parameter estimates with different nonparametric baseline specifications. Prentice and Hsu (1997) extended Cai and Prentice (1995) by developing joint estimating equations for hazard ratio and pairwise dependence parameters. Gray and Li (2002) considered the optimal selection of weights in martingale estimating equations for clustered failure time data based on the marginal proportional hazards model.

For the marginal method in the analysis of clustered failure time data with a cure fraction, Peng et al. (2007) proposed a semiparametric marginal proportional hazards mixture cure model to analyze survival data from a multi-institutional study of tonsil cancer and provided robust variance estimates of parameters. Yu and Peng (2008) also considered a marginal mixture cure model with Weibull baseline distribution for a smoking cessation study and provided jackknife variance estimates of the parameters

in the model. Chen and Lu (2012) further extended the work of Peng et al. (2007) by considering a transformation model for uncured patients. All these marginal mixture cure models are robust to misspecification of the correlation structure. However, when the correlation is of interest and there is partial information available for the correlation structure, an efficiency loss may be incurred in using the marginal method for the clustered failure time data with a survival proportion.

Therefore, parallel to the generalized estimating equations approach in the marginal proportional hazards model, we are interested in developing a marginal method that accommodates the correlation in clustered failure time data with a cure fraction in the proportional hazards mixture cure model to improve the estimation efficiency.

2.4 Generalized Estimating Equations

Generalized estimating equations (GEEs) approach is originally proposed for the situation where it is reasonable to assume that the marginal mean response conforms to a generalized linear models (GLMs) (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989; Dobson, 2002). To handle non-normal longitudinal data, Liang and Zeger (1986) introduced a working correlation matrix with a set of nuisance parameters to avoid the specification of correlation between measurements within clusters. When the primary interest is on the marginal regression parameters and the dependence among observations in a cluster is nuisance, GEEs provide a useful approach in the analysis of correlated outcomes.

Let $\tilde{T}_i = (\tilde{T}_{i1}, \cdots, \tilde{T}_{in_i})'$ be the $n_i \times 1$ vector of outcome values and $X_i = (X_{i1}, \cdots, X_{in_i})'$ be the $n_i \times p$ covariate matrix for the ith subject $(i = 1, \cdots, K)$. Here $\tilde{T}_{ij} \in R$. Liang and Zeger (1986) assumed that the observations from the distinct subjects are

independent and the marginal density of $\tilde{T}_{ij}$ is from an exponential family

$$f(\tilde{t}_{ij}) = \exp[\{\tilde{t}_{ij}\theta_{ij} - a(\theta_{ij}) + b(\tilde{t}_{ij})\}/\phi],$$

where $\theta_{ij} = h(\eta_{ij})$, $\eta_{ij} = \beta' X_{ij}$, and moreover

$$E(\tilde{T}_{ij}) = \frac{da(\theta_{ij})}{d\theta_{ij}} = a'(\theta_{ij}), \qquad Var(\tilde{T}_{ij}) = \frac{d^2a(\theta_{ij})}{d\theta_{ij}^2}\phi = a''(\theta_{ij})\phi,$$

where the $p \times 1$ vector β are regression parameters which embody the relationship between the responses and the covariates and ϕ is a scale (dispersion) parameter. Traditionally, we could consider $f(\tilde{t}_{ij})$ as an exponential family which includes Gaussian, binomial, gamma, inverse Gaussian, Poisson, geometric, and negative binomial distributions.

Under the working assumption that the repeated observations from a subject are independent of one another, Liang and Zeger (1986) derived the independent estimating equations (IEEs), i.e.,

$$U_I(\beta) = \sum_{i=1}^{K} X_i^T \Delta_i S_i = 0, \tag{2.11}$$

where $\Delta_i = \text{diag}\{\delta\theta_{ij}/\delta\eta_{ij}\}$ is an $n_i \times n_i$ matrix and $S_i = \tilde{T}_i - a_i'(\theta)$ is of order $n_i \times 1$ for the ith subject. When the marginal model is correctly specified, under mild regularity conditions, the solution of equations (2.11), $\hat{\beta}_I$, is a consistent estimate of β and $\text{var}(\hat{\beta}_I)$ can be consistently estimated by

$$\{\sum_{i=1}^{K} X_i^T \Delta_i A_i \Delta_i X_i\}^{-1}\{\sum_{i=1}^{K} X_i^T \Delta_i S_i S_i^T \Delta_i X_i\}\{\sum_{i=1}^{K} X_i^T \Delta_i A_i \Delta_i X_i\}^{-1}|_{\hat{\beta}_I},$$

where A_i is a diagonal matrix of order $n_i \times n_i$ with elements $a''(\theta_{ij})$. They also showed

that $\hat{\beta}_I$ are reasonable efficient for a few simple designs such as the true correlation is moderate or the variation of cluster sizes is small.

However, the use of independent working correlation structure may result in a notable loss of efficiency when, for example, the response correlation coefficient is large, or variation in cluster sizes is large. To accommodate the within-cluster dependence and improve the efficiency, the diagonal covariance matrix Δ_i in the score equations (2.11) is replaced by a 'working' covariance matrix of order $n_i \times n_i$

$$V_i(\rho) = A_i^{\frac{1}{2}} R(\rho) A_i^{\frac{1}{2}} \phi, \tag{2.12}$$

which will be $\text{cov}(\tilde{T}_i)$ if $R(\rho)$ is the true correlation matrix. Here ρ is a set of parameters that fully characterizes the working correlation matrix $R(\rho)$. Therefore, the modified score equations, i.e., GEEs, are defined by Liang and Zeger (1986) as

$$U_G(\beta) = \sum_{i=1}^{K} D_i^T V_i^{-1}(\rho) S_i = 0, \tag{2.13}$$

where D_i is the matrix of derivatives of $a'(\theta)$ with respect to the regression parameters β, i.e., $D_i = \delta\{a_i'(\theta)\}/\delta\beta = A_i \Delta_i X_i$. If $R(\rho)$ is specified as an identity matrix, equations (2.13) reduce to the IEEs (2.11). Let $\hat{\beta}_G$ and $\hat{V}_G$ denote the regression estimates and the corresponding variance estimates from (2.13). As in the independence case, the consistency of $\hat{\beta}_G$ and $\hat{V}_G$ depend only on the correct specification of the mean structure, not on the correct choice of $R(\rho)$ and the estimators for ρ and ϕ as long as they are $K^{\frac{1}{2}}$-consistent.

To obtain $\hat{\beta}_G$, the authors suggested an iteration between the Fisher scoring method for β and the moment method for ρ and ϕ. Given the current estimates

$\hat{\rho}$ and $\hat{\phi}$,

$$\begin{aligned}\hat{\beta}_{j+1} &= \hat{\beta}_j - [E(\frac{\partial U_G(\beta)}{\partial \beta})]^{-1} U_G(\beta)|_{\hat{\beta}_j} \\ &= \hat{\beta}_j + \{\sum_{i=1}^{K} D_i^T(\hat{\beta}_j)\tilde{V}_i^{-1}(\hat{\beta}_j)D_i(\hat{\beta}_j)\}^{-1}\{\sum_{i=1}^{K} D_i^T(\hat{\beta}_j)\tilde{V}_i^{-1}(\hat{\beta}_j)S_i(\hat{\beta}_j)\},\end{aligned}$$

where $\tilde{V}_i(\beta) = V_i[\beta, \hat{\rho}\{\beta, \hat{\phi}(\beta)\}]$. That is,

$$\hat{\beta}_{j+1} = \{D^T\tilde{V}^{-1}D\}^{-1}D^T\tilde{V}^{-1}Z|_{\hat{\beta}_j},$$

where $D = (D_1^T(\beta), \cdots, D_K^T(\beta))^T$ is a matrix of order $N \times p$, $S = (S_1^T(\beta), \cdots, S_K^T(\beta))^T$ is of order $N \times 1$ and $\tilde{V} = \text{diag}\{\tilde{V}_1(\beta), \cdots, \tilde{V}_K(\beta)\}$ which is a block diagonal matrix of order $N \times N$. Vector $Z = D\beta + S$ is of order $N \times 1$.

At each iteration step, the scale parameter ϕ can be estimated by the moment method

$$\hat{\phi} = \sum_{i=1}^{K}\sum_{j=1}^{n_i} \hat{r}_{ij}^2/(N-p)\,, \tag{2.14}$$

where the current Pearson residual $\hat{r}_{ij}$ is $\{\tilde{T}_{ij} - a'(\hat{\theta}_{ij})\}/\{a''(\hat{\theta}_{ij})\}^{\frac{1}{2}}$. The estimate of ρ varies based on the different choices of $R(\rho)$. There are a variety of common correlation structures such as an independent working correlation matrix where $R(\rho)$ is an identity matrix; an exchangeable (equicorrelated, compound symmetric) one where $\text{corr}(\tilde{T}_{ij}, \tilde{T}_{ij'}) = \rho$ for all $j \neq j'$; a first-order autoregressive (AR-1) one where $\text{corr}(\tilde{T}_{ij}, \tilde{T}_{ij'}) = \rho^{|j-j'|}$; and an unstructured one where no restriction for correlation but $n_i(n_i-1)/2$ correlation parameters are required. For instance, when $R(\rho)$ has

an exchangeable correlation structure, given ϕ, ρ can be estimated by

$$\hat{\rho} = \phi^{-1} \sum_{i=1}^{K} \sum_{j>j'} \hat{r}_{ij}\hat{r}_{ij'} / \{\sum_{i=1}^{K} \frac{1}{2} n_i(n_i - 1) - p\}. \tag{2.15}$$

The variance of $\hat{\beta}_G$ can be consistently estimated by

$$\{\sum_{i=1}^{K} D_i^T V_i^{-1} D_i\}^{-1} \{\sum_{i=1}^{K} D_i^T V_i^{-1} S_i S_i^T V_i^{-1} D_i\} \{\sum_{i=1}^{K} D_i^T V_i^{-1} D_i\}^{-1}|_{(\hat{\beta}_G, \hat{\phi}, \hat{\rho})} .$$

Generally, GEEs is a marginal approach since the underlying GLMs involve regression models defining the mean of the marginal distribution. As we introduced above, Liang and Zeger (1986) gave an algorithm for estimating both β and ρ, as well as established the asymptotic multivariate normal distribution for the regression parameters given the consistent estimates of the correlation and scale parameters. The fact that the asymptotic distribution is independent of a specific estimator of ρ allows for robustness to misspecification of the working correlation matrix. Qu et al. (2000) utilized quadratic inference functions which avoid direct estimation of the correlation parameters. Their method guarantees that the estimator of correlation always exists and hence solve the issues raised by Crowder (1995) where the estimator of ρ does not exist in some simple cases of misspecification which results in inconsistency. Alternatively, Stoner and Leroux (2002) proposed an optimal (in terms of estimation efficiency) combination of estimating equations approach to model the correlation structure of the observations in a more efficient manner.

2.5 The EM Algorithm

The expectation maximization (EM) algorithm (Dempster et al., 1977) is a popular method for maximum likelihood estimation in incomplete-data problems. The EM

algorithm estimates the parameters of a model iteratively with some initial values. Specifically, each iteration consists of an expectation (E) step, which calculates the expected value of the full likelihood function with respect to the unobservable variables using the current estimates of the parameters, and a maximization (M) step, which estimates the parameters by maximizing the expected value of the full likelihood function derived in the E-step. The EM algorithm is easy to implement in many applications because of the numerical stability. However, one drawback of the EM algorithm is that it does not produce valid standard errors directly. To address this issue, Louis (1982) used the complete log-likelihood to derive the observed information matrix. That is, let $S(y, \theta)$ and $B(y, \theta)$ be the gradient and the negative of the associated second derivative matrices based on the completely log-likelihood separately, then the observed information $I(\theta)$ could be represented by

$$I(\theta) = E_\theta\{B(Y, \theta)|X \in R\} - E_\theta\{S(Y, \theta)S^T(Y, \theta)|Y \in R\},$$

where the first term on the right hand side can be viewed as the complete information, and the second term can be viewed as the missing information (Meng and Rubin, 1991).

Other methods such as using numerical differentiation to obtain the standard errors in the EM algorithm were also investigated by Meng and Rubin (1991) and Jamshidian and Jennrich (2000). Their methods are especially useful when the analytic calculation of derivatives was cumbersome or impossible.

2.6 The ES Algorithm

As an extension of the EM algorithm, Rosen et al. (2000) proposed an Expectation-Solution (ES) algorithm for mixtures of the generalized linear models where the GEEs

are embedded into the M-step of the EM algorithm to account for the correlation among the responses. They defined the marginal probability density of $\tilde{T}_{ij}^*$ (here $\tilde{T}_{ij}^* \in R$) by the mixture density function

$$p(\tilde{t}_{ij}^*|X_{ij}, \theta_s) = \sum_{r=1}^{I^*} p(e_r|X_{ij}, \gamma) p(\tilde{t}_{ij}^*|X_{ij}, \beta_r, \phi_r),$$

where θ_s denotes all the parameters; $\beta_1, \cdots, \beta_{I^*}, \phi_1, \cdots, \phi_{I^*}$, and $\gamma = (\gamma_1, \cdots, \gamma_{I^*-1})$, e_r is an $I^* \times 1$ vector with 1 at the rth position and 0's elsewhere. For each covariate X_{ij} in the ith cluster, the response $\tilde{T}_{ij}^*$ is generated from the rth subprocess (component) $p(\tilde{t}_{ij}^*|X_{ij}, \beta_r, \phi_r)$ with probability $p(e_r|X_{ij}, \gamma_r) = \pi_r^{(ij)}$, $r = 1, \cdots, I^*$, $j = 1, \cdots, n_i, i = 1, \cdots, K$. The weights $\pi_r^{(ij)}$ depend on the covariates and are expressed in a multinomial logit form. Specifically,

$$\pi_r^{(ij)} = \frac{\exp(\gamma_r' X_{ij})}{\sum_{d=1}^{I^*} \exp(\gamma_d' X_{ij})}$$

for each r. For the jth observation in the ith cluster, $\sum_{r=1}^{I^*} \pi_r^{(ij)} = 1$. $p(\tilde{t}_{ij}^*|X_{ij}, \beta_r, \phi_r)$ is assumed to be a member of the exponential family. $\pi_r^{(ij)}$ and $p(\tilde{t}_{ij}^*|X_{ij}, \beta_r, \phi_r)$ may share some covariates. Obviously, the expectation of $\tilde{T}_{ij}^*$ varies with different component in the mixture models. Given covariates X_{ij}, $\mu_r^{(ij)} = E(\tilde{T}_{ij}^*)$ which is described by a function of parameter vector β_r from the rth component, i.e., $\mu_r^{(ij)} = h(\beta_r' X_{ij})$ and h is a link function.

Suppose that the observed data $O_s = \{(\tilde{T}_{ij}^*, X_{ij}), j = 1, \cdots, n_i, i = 1, \cdots, K\}$ are independent and there are no censored observations, the observed full likelihood is given by

$$L(\theta_s|O_s) = \prod_{i=1}^{K} \prod_{j=1}^{n_i} \sum_{r=1}^{I^*} p(e_r|X_{ij}, \gamma_r) p(\tilde{t}_{ij}^*|X_{ij}, \beta_r, \phi_r). \tag{2.16}$$

When unobserved indicator vectors $y^{(ij)} = (y_1^{(ij)}, \cdots, y_{I^*}^{(ij)})$ are added to the observations O_s, i.e., $O'_s = \{(\tilde{T}^*_{ij}, X_{ij}, y^{(ij)}), j = 1, \cdots, n_i, i = 1, \cdots, K\}$, the complete full likelihood, an augmented version of the observed one, can be written as

$$\begin{aligned} L(\theta_s|O'_s) &= \prod_{i=1}^{K}\prod_{j=1}^{n_i}\prod_{r=1}^{I^*}\{p(e_r|X_{ij},\gamma_r)\}^{y_r^{(ij)}}\{p(\tilde{t}^*_{ij}|X_{ij},\beta_r,\phi_r)\}^{y_r^{(ij)}} \\ &= \prod_{i=1}^{K}\prod_{j=1}^{n_i}\prod_{r=1}^{I^*}\{\pi_r^{(ij)}p(\tilde{t}^*_{ij}|X_{ij},\beta_r,\phi_r)\}^{y_r^{(ij)}}. \end{aligned} \tag{2.17}$$

The authors used the EM algorithm to derive the maximum likelihood estimate $\hat{\theta}_s$. The E-step consists of calculating the expectation of the log form of (2.17) with respect to $y_r^{(ij)}$ given the current estimate of $\hat{\theta}_s$ and the complete data O'_s. That is,

$$\begin{aligned} l(\theta_s|O'_s) &= E(\log L(\theta_s|O'_s)) \\ &= \sum_{i=1}^{K}\sum_{j=1}^{n_i}\sum_{r=1}^{I^*} g_r^{(ij)}\{\log\pi_r^{(ij)} + \log p(\tilde{t}^*_{ij}|X_{ij},\beta_r,\phi_r)\}, \end{aligned} \tag{2.18}$$

where

$$\begin{aligned} g_r^{(ij)} &= E(y_r^{(ij)}|O_s;\theta_s) = P(y_r^{(ij)}=1|O_s;\theta_s) \\ &= \frac{p(\tilde{t}^*_{ij}|e_r,X_{ij};\theta)p(e_r|X_{ij};\theta_s)}{\sum_{r=1}^{I^*}p(\tilde{t}^*_{ij}|e_r,X_{ij};\theta_s)p(e_r|X_{ij};\theta_s)} \\ &= \frac{\pi_r^{(ij)}p(\tilde{t}^*_{ij}|X_{ij},\beta_r,\phi_r)}{\sum_{l=1}^{I^*}\pi_l^{(ij)}p(\tilde{t}^*_{ij}|X_{ij},\beta_l,\phi_l)}, \end{aligned} \tag{2.19}$$

which is the success probability of Bernoulli random variable $y_r^{(ij)}$. Furthermore, $l(\theta_s|O'_s)$ can be separated into I^*+1 log likelihood functions including

$$l_\gamma = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\sum_{r=1}^{I^*} g_r^{(ij)}\log\pi_r^{(ij)}, \tag{2.20}$$

and

$$l_{\beta_r,\phi_r} = \sum_{i=1}^{K}\sum_{j=1}^{n_i} g_r^{(ij)} \log p(\tilde{t}_{ij}^*|X_{ij},\beta_r,\phi_r) \quad (r=1,\cdots,I^*). \tag{2.21}$$

The M-step consists of maximizing (2.20) with respect to γ_r and maximizing (2.21) with respect to β_r and ϕ_r for the fixed $\{g_r^{(ij)}\}_{i=1}^n$. To estimate γ, the Newton-Raphson method can be applied based on (2.20). The authors proposed I^* generalized linear models with observed data O_s to fit β_r. Let $\mu_r^{(i)} = (\mu_r^{(i1)},\cdots,\mu_r^{(in_i)})^T$, the I^* systems of score equations based on (2.21) are

$$\sum_{i=1}^{K} D_r^{(i)T}(V_r^{(i)})^{-1}G_r^{(i)}(\tilde{t}_i^* - \mu_r^{(i)}) = 0 \quad (r=1,\cdots,I^*), \tag{2.22}$$

where $G_r^{(i)} = \mathrm{diag}(g_r^{(i1)},\cdots,g_r^{(in_i)})$, $V_r^{(i)} = \mathrm{diag}(\phi_r v(\mu_r^{(i1)}),\cdots,\phi_r v(\mu_r^{(in_i)}))$, and $D_r^{(i)} = \partial\mu_r^{(i)}/\partial\beta_r'$. $\mathrm{var}_r(\tilde{t}_{ij}^*) = \phi_r v(\mu_r^{(ij)})$. For instance, $v(\mu_r^{(ij)}) = \mu_r^{(ij)}(1-\mu_r^{(ij)})$ and $\phi_r = 1$ for Bernoulli outcome data, $v(\mu_r^{(ij)}) = \mu_r^{(ij)}$ and $\phi_r = 1$ for Poisson outcome data. Function (2.20) can be considered as a log likelihood for generalized Bernoulli outcome data where $\sum_{r=1}^{I^*} g_r^{(ij)} = 1$ for given (i,j).

To capture the correlation among the observations from the mixture model, the authors incorporated the working correlation matrices in (2.22). The GEEs with respect to β_r are

$$\sum_{i=1}^{K} D_r^{(i)T}(V^{(i)}(\rho_r))^{-1}G_r^{(i)}(\tilde{t}_i^* - \mu_r^{(i)}) = 0 \quad (r=1,\cdots,I^*), \tag{2.23}$$

where $V^{(i)}(\rho_r) = \phi_r(A_r^{(i)})^{1/2}R_i(\rho_r)(A_r^{(i)})^{1/2}$, $A_r^{(i)} = \mathrm{diag}(v(\mu_r{}^{(i1)}),\cdots,v(\mu_r^{(in_i)}))$, $R_i(\rho_r)$ is a working correlation matrix depending on the rth component's association parameters ρ_r which is a d_r-dimensional vector.

Rosen et al. (2000) further showed that the estimating functions with respect to γ and β are unbiased. Based on Carroll et al. (1995), the solutions of unbiased estimating equations are consistent and asymptotically normally distributed as the sample size $K \to \infty$. Under certain regularity conditions (Gallant and White, 1988), the estimated asymptotic variances of $\hat{\theta}_s$ are obtained by

$$\hat{\text{var}}(\hat{\theta}_s) = \{\hat{\mathcal{F}}^{-1}\}\hat{\mathcal{V}}\{\hat{\mathcal{F}}^{-1}\}^T, \tag{2.24}$$

where

$$\hat{\mathcal{F}} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\sum_{\zeta=1}^{I^*} \nabla s_{ij}(e_\zeta; \hat{\theta}_s, \hat{\theta}_s), \qquad \hat{\mathcal{V}} = \sum_{i=1}^{K}\left\{\sum_{j=1}^{n_i}\sum_{\zeta=1}^{I^*} s_{ij}(e_\zeta; \hat{\theta}_s, \hat{\theta}_s)\right\}^{\otimes^2},$$

where $v^{\otimes^2} \equiv vv^T$ for a general column vector v, $[\nabla v]_{kl} = \frac{\partial (v)_k}{\partial (\theta_s)_l}, k, l = 1, \cdots, d_{\theta_s}$, and $d_{\theta_s} = \dim(\theta_s) = I\dim(\beta_r) + I\dim(\phi_r) + (I^* - 1)\dim(\gamma_r) = Id_\beta + Id_\phi + (I^* - 1)d_\gamma$. $s_{ij}(e_\zeta; \hat{\theta}_s, \hat{\theta}_s) = q_{ij}(\tilde{t}^*_{ij}, e_\zeta; \hat{\theta}_s) g_\zeta^{(ij)}$ for $j = 1, \cdots, n_i$ and $i = 1, \cdots, K$. $q_{ij}(\tilde{t}^*_{ij}, e_\zeta; \theta_s)$ is a $d_{\theta_s} \times 1$ vector for each $\zeta = 1, \cdots, I^*, j = 1, \cdots, n_i$ and $i = 1, \cdots, K$. The first (Id_β) components are

$$\sum_{u=1}^{n_i} [D_\zeta^{(i)}]_{ul} [V^{(i)}(\rho_\zeta)^{-1}]_{uj} \delta^*_{\zeta r} (\tilde{t}^*_{ij} - \mu_\zeta^{(ij)})$$

for $r = 1, \cdots, I^*$ and $l = 1, \cdots, d_\beta$. Note that $\delta^*_{\zeta r} = 1$ when $\zeta = r$ and 0 elsewhere. The first (Id_β) components are corresponding to (2.23). That is, given i, r and l,

$$\sum_{j=1}^{n_i}\sum_{\zeta=1}^{I^*}\sum_{u=1}^{n_i} [D_\zeta^{(i)}]_{ul} [V^{(i)}(\rho_\zeta)^{-1}]_{uj} \delta^*_{\zeta r} (\tilde{t}^*_{ij} - \mu_\zeta^{(ij)}) = D_r^{(i)T} (V^{(i)}(\rho_r))^{-1} (\tilde{t}^*_i - \mu_r^{(i)}) \,.$$

The middle I^* components are

$$\{(\tilde{t}^*_{ij} - \mu^{(ij)}_\zeta)^2 - \phi_\zeta v(\mu^{(ij)}_\zeta)\}\delta^*_{\zeta r}$$

for $r = 1, \cdots, I^*$. The last $(I^* - 1)d_\gamma$ components are

$$(\pi^{(ij)}_\zeta)^{-1} \left\{ \frac{\partial \pi^{(ij)}_\zeta}{\partial (\gamma_r)_k} \right\}$$

for $r = 1, \cdots, I^*, k = 1, \cdots, d_\gamma$.

Chapter 3

Marginal Proportional Hazards Model

3.1 Introduction

In this chapter, we consider the marginal method developed by Segal and Neuhaus (1993) for the classical clustered failure time data without a cured fraction. We observe in numerical studies that, when correlation exists within clusters, the estimating function proposed by Segal and Neuhaus (1993) for hazards ratio regression parameters is biased. Therefore, the estimates from the existing estimating equation are biased and the variance estimates are unstable. To address this issue, we propose an unbiased weighted estimating function and show that the estimators based on the proposed estimating equation are consistent and asymptotically normal. A consistent estimator of the covariance matrix for regression parameters is also provided. We will demonstrate via a simulation study that the proposed estimating equation approach produces unbiased regression estimators as well as improves the estimation efficiency compared to the existing marginal methods.

The rest of the chapter is organized as follows. In Section 3.2, we propose an unbiased weighted estimating function for the hazard ratio parameters based on the marginal semiparametric proportional hazards model. The asymptotic properties of

the estimators and the variance estimates are obtained in Section 3.3. We perform a simulation study in Section 3.4 to evaluate the performance of the proposed estimating equation, and apply this approach in the analysis of Diabetic Retinopathy Study and in the study of Infection in Kidney Patients in Section 3.5. Finally, we provide conclusions on the proposed model and estimation method in Section 3.6.

3.2 Model and Estimating Equation

The marginal survival function of $\tilde{T}_{ij}$ is assumed to follow the proportional hazards model, i.e.,

$$S(t; X_{ij}) = S_0(t)^{\exp(\beta' X_{ij})}, \tag{3.1}$$

where $S_0(t)$ is the baseline survival function of $\tilde{T}_{ij}$ when $X_{ij} = 0$, and has the product-limit form as we defined in Section 2.2, i.e., $S_0(t) = \prod_{s:\tau_s \le t} \alpha_s$ where $0 \le \alpha_s \le 1$. If we ignore the correlation within clusters, the unknown parameters $\theta = (\beta, \alpha)$ in the model could be estimated based on a log-likelihood function with the observations $O = \{(T_{ij}, \delta_{ij}, X_{ij}), j = 1, \cdots, n_i, i = 1, \cdots, K\}$. That is,

$$\begin{aligned}
& l(\theta; O) = \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} f(t_{ij}; X_{ij})^{\delta_{ij}} S(t_{ij}; X_{ij})^{1-\delta_{ij}} \\
= \; & \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} \{\Lambda_0(t_{ij}) \exp(\beta' X_{ij})\}^{\delta_{ij}} \exp\{-\Lambda_0(t_{ij}) \exp(\beta' X_{ij})\} \\
& + \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} \left(\frac{\lambda_0(t_{ij})}{\Lambda_0(t_{ij})} \right)^{\delta_{ij}} \qquad (3.2) \\
= \; & \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} \left[\{\exp(\beta' X_{ij})\}^{\kappa_{ij}} \exp\{-\exp(\beta' X_{ij})\} \right]^{\Lambda_0(t_{ij})} \\
& + \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} (\lambda_0(t_{ij}))^{\delta_{ij}} , \qquad (3.3)
\end{aligned}$$

where $f(t_{ij}; X_{ij})$ is the density function of $S(t_{ij}; X_{ij})$, and $\lambda_0(t_{ij})$ and $\Lambda_0(t_{ij})$ are the hazard and cumulative hazard functions corresponding to $S_0(t_{ij})$. Here $\kappa_{ij} = \delta_{ij}/\Lambda_0(t_{ij})$.

Based on (3.2), Segal and Neuhaus (1993) proposed an estimating function (denoted by U_{SN}) for the regression parameters β. That is

$$\begin{aligned} U_{SN} &= \sum_{i=1}^{K} (U_{SN})_i \\ &= \sum_{i=1}^{K} \{\frac{\partial \mu_{SN}(X_i)}{\partial \beta}\}^T \{A_i^{1/2} Q_i(\rho_{SN}) A_i^{1/2} \phi_{SN}\}^{-1} \{\delta_i - \mu_{SN}(X_i)\}, \end{aligned} \quad (3.4)$$

where $\mu_{SN}(X_i) = \{\mu_{SN}(X_{i1}), \cdots, \mu_{SN}(X_{in_i})\}^T$ with $\mu_{SN}(X_{ij}) = \Lambda_0(t_{ij}) \exp(\beta' X_{ij})$, $A_i = \text{diag}\{\mu_{SN}(X_i)\}$, $\delta_i = (\delta_{i1}, \cdots, \delta_{in_i})^T$, $Q_i(\rho_{SN})$ is the working correlation matrix, ρ_{SN} is a group of unknown parameters in the correlation matrix, and ϕ_{SN} is an unknown scale parameter. The Newton-Raphson method can be used to solve the equation $U_{SN} = 0$ to obtain the estimate of the regression parameter vector β. We let $\hat{\beta}_{SN}$ denote the solution of $U_{SN} = 0$. As pointed out by Segal and Neuhaus (1993), the robust variance estimates are obtained from $(-\frac{\partial U_{SN}}{\partial \beta})^{-1} (\sum_{i=1}^{K} \{U_{SN}\}_i \{U_{SN}\}_i) (-\frac{\partial U_{SN}}{\partial \beta})^{-T}$. Different from the Poisson likelihood mentioned above, Lee et al. (1992), based on the partial likelihood, proposed robust sandwich variance estimates for the regression parameter β without specifying dependence structure within clusters.

As we discussed in Section 2.3, one attractive property of the GEEs method is that the estimation efficiency may be improved by using the working correlation matrix. However, the estimating function (3.4) displays considerable biases that may lead to biased estimate of β. This can be seen from a numerical study based on 1000 data sets generated from the model (3.1) (details for data generation are given in Section 3.4). By plotting the 1000 values of function (3.4) given the true

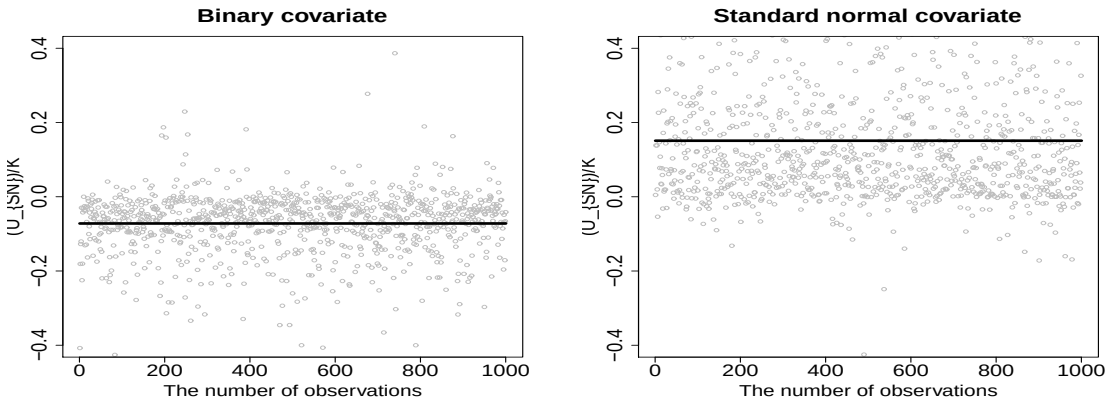

Figure 3.1: The average (black line) of 1000 values of function U_{SN}/K based on the correlated failure time (Kendall's tau=0.8) with binary covariate (left) and standard normal covariate (right). The regression parameter $\beta = \log(2)$ and the baseline survival function follows exponential distribution with parameter $\alpha = 2$. Here $K = 40$.

parameter settings, we observe that the average value of function U_{SN} is -0.072 for the binary covariate and 0.151 for the continuous covariate when the Kendall's tau equals 0.8 (Figure 3.1). Both considerably deviate from 0. Here Kendall's tau is used to measure the strength of association among observations in one cluster. The larger value of Kendall's tau represents the stronger correlation within clusters. Figure 3.1 shows that the estimating function U_{SN} is biased when the correlation within a cluster exists. Consequently, the estimators based on estimating equation $U_{SN} = 0$ are biased and the corresponding variance estimates are unstable as we can observe that in the simulation study in Section 3.4.

To address this issue, we propose a weighted estimating function for β based on the log-likelihood (3.3), i.e.,

$$\begin{aligned} U_{New} &= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(U_{New})_{ij} = \sum_{i=1}^{K}(U_{New})_i \\ &= \sum_{i=1}^{K}\{\frac{\partial \mu_{New}(X_i)}{\partial \beta}\}^T\{B_i^{1/2}Q_i(\rho_{New})B_i^{1/2}\phi_{New}\}^{-1}W_i\{\kappa_i - \mu_{New}(X_i)\}, \quad (3.5) \end{aligned}$$

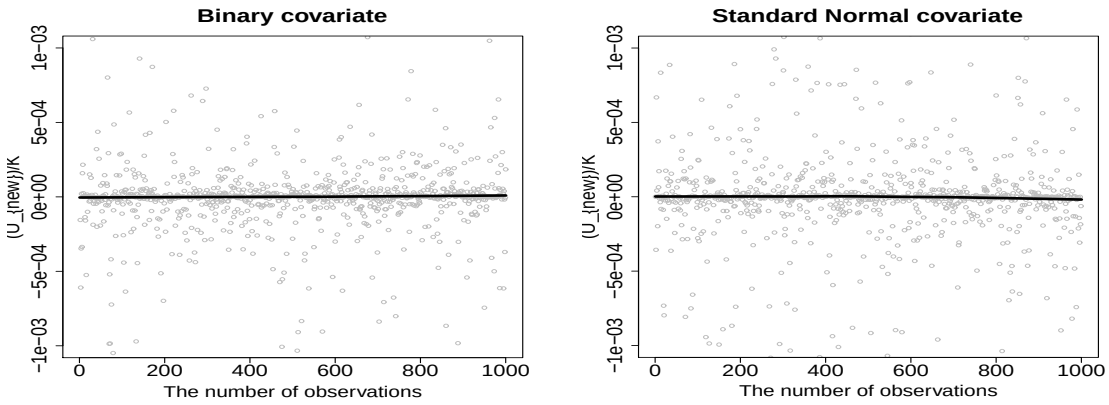

Figure 3.2: The average (black line) of 1000 values of function U_{New}/K based on the correlated failure time (Kendall's tau=0.8) with binary covariate (left) and standard normal covariate (right). The regression parameter $\beta = \log(2)$ and the baseline survival function follows exponential distribution with parameter $\alpha = 2$. Here $K = 40$.

where $\mu_{New}(X_i) = \{\mu_{New}(X_{i1}), \cdots, \mu_{New}(X_{in_i})\}^T$ with $\mu_{New}(X_{ij}) = \exp(\beta' X_{ij})$, $B_i = \text{diag}\{\mu_{New}(X_i)\}$, $\kappa_i = (\kappa_{i1}, \cdots, \kappa_{in_i})^T$, $W_i = \text{diag}(\Lambda_0(t_{i1}), \cdots, \Lambda_0(t_{in_i}))$, $Q_i(\rho_{New})$ is the working correlation matrix, and ρ_{New} is a group of unknown parameters in the matrix that needs to be estimated. Similar to function (3.4), the scale parameter ϕ_{New} is incorporated in the estimating function (3.5) to accommodate the over- or under-dispersion.

Based on the same data set used in Figure 3.1, we plot the 1000 values of function (3.5) and observe that the average value of U_{New} is much closer to zero than that from (3.4) for both binary covariate (about -3.178e-05) and continuous covariate (about 3.024e-04) (Figure 3.2). That is, empirically, the function (3.5) tends to be unbiased. We will show the unbiasedness of the proposed weighted estimating function (3.5) in Section 3.3. Therefore, by letting $U_{New} = 0$, we establish an unbiased weighted estimating equation for the hazard ratio parameters in (3.1). We let $\hat{\beta}_{New}$ denote the solution of equation $U_{New} = 0$.

Different from the parametric baseline specified by Segal and Neuhaus (1993), we

estimate the baseline survival function $S_0(t)$ by using the nonparametric maximum likelihood estimator (Kalbfleisch and Prentice, 2002). As we discussed in Section 2.1.2, a suitable initial value for $S_0(t)$ could be chosen as

$$\hat{S}_0(t) = \exp(-\sum_{s:\tau_s \leq t} \frac{d_s}{\sum_{(i,j)\in R_s} \exp(\beta' X_{ij})}). \tag{3.6}$$

based on the nonparametric estimate of $\Lambda_0(t)$.

To obtain $\hat{\beta}_{New}$, we suggest a dual iteration algorithm as follows:

1. Set initial values for β_{New} and calculate $\hat{S}_0(t)$ based on (3.6).

2. Given $\hat{S}_0(t)$, calculate the updated estimate of β_{New} using Newton-Raphson method, i.e.,

   (a) Given current estimates of ρ_{New} and ϕ_{New}, calculate the updated estimate of β_{New} from (3.5).

   (b) Given the estimate of β_{New}, calculate the standardized Pearson residuals $\hat{r}_{ij} = \{\kappa_{ij} - \mu_{New}(X_{ij})\}/\{\mu_{New}(X_{ij})\}^{\frac{1}{2}}$.

   (c) Use the residuals $\hat{r}_{ij}$ to estimate $\hat{\rho}_{New}$ and $\hat{\phi}_{New}$.

   (d) Repeat steps (a), (b), and (c) until convergence.

3. Given $\hat{\beta}_{New}$, update $\hat{S}_0(t)$.

4. Repeat steps 2 and 3 until convergence.

Here we consider an exchangeable correlation structure for $Q_i(\rho_{New})$ as it is often used for clustered data. Such a correlation structure was also considered in U_{SN} by Segal and Neuhaus (1993). As we discussed in Section 2.3, following the formulas (2.14) and (2.15), the correlation parameter ρ_{New} can be estimated from the standardized

Pearson residuals. That is

$$\hat{\rho}_{New} = \hat{\phi}_{New}^{-1} \sum_{i=1}^{K} \sum_{j>j'} \hat{r}_{ij}\hat{r}_{ij'} / \{\sum_{i=1}^{K} \frac{1}{2} n_i(n_i - 1) - p_X\}, \tag{3.7}$$

where $\hat{\phi}_{New} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \hat{r}_{ij}^2/(N - p_X)$ and p_X is the length of covariate X_{ij}.

3.3 Asymptotic Results and Variance Estimation

3.3.1 Asymptotic Properties of $\hat{\beta}_{New}$

Theorem 3.1. *Let β_0 be the true parameter of β and $\Psi_K(\beta) = \frac{U_{New}}{K} = \frac{1}{K}\sum_{i=1}^{K}(U_{New})_i$. Under the following set of conditions (Yuan and Jennrich, 1998; Liang and Zeger, 1986)*

A1. *$\Psi_K(\beta_0) \to 0$ with probability 1;*

A2. *There is a neighborhood N of β_0 on which with probability one all $\Psi_K(\beta)$ are continuously differentiable and the Jacobians $\frac{\partial \Psi_K(\beta)}{\partial \beta}$ converge uniformly to a nonstochastic limit which is nonsingular at β_0;*

A3. *$K^{1/2}\Psi_K(\beta_0) \to N(0, \mathcal{V})$ in distribution where $\mathcal{V} = \lim_{K\to\infty}(\sum_{i=1}^{K} E((U_{New})_i^T (U_{New})_i)/K)$;*

A4. *$\partial\Psi_K(\beta_0)/\partial\rho_{New} \to 0$ with probability 1;*

A5. *$K^{1/2}(\hat{\phi}_{New} - \phi_{New}) = O_p(1)$ given β;*

A6. *$K^{1/2}(\hat{\rho}_{New} - \rho_{New}) = O_p(1)$ given β and ϕ_{New};*

A7. *$|\partial\hat{\rho}(\beta, \phi_{New})/\partial\phi_{New}|$ is bounded by a function $H(T, \beta)$ which is $O_p(1)$,*

the estimator $\hat{\beta}_{New}$ solving (3.5) is a consistent estimator of β_0. Also $K^{1/2}(\hat{\beta}_{New} - \beta_0)$ is asymptotically normally distributed with mean vector $\mathbf{0}$ and with variance

matrix $\Sigma = \mathcal{A}^{-1}(\beta_0)\mathcal{V}(\beta_0)\mathcal{A}^{-T}(\beta_0)$ *where* $\mathcal{A}(\beta_0) = -\partial U_{New}(\beta_0)/\partial\beta_0$ *and* $\mathcal{V}(\beta_0) = \sum_{i=1}^{K}\{U_{New}(\beta_0)\}_i\{U_{New}(\beta_0)\}_i^T$. *Moreover,* Σ *can be consistently estimated by* $\hat{\Sigma} = \mathcal{A}^{-1}(\hat{\beta}_{New})\mathcal{V}(\hat{\beta}_{New})\mathcal{A}^{-T}(\hat{\beta}_{New})$.

Proof. To show the large sample properties of $\hat{\beta}_{New}$, we first consider the unbiasedness of $\{(U_{New})_{ij}|c_{ij},\theta\}$. Let $\mu_{ij} = (\mu_{New})_{ij}$ and $F_{ij}(t)$ be the distribution function of $\tilde{T}_{ij}$. Then we have

$$
\begin{aligned}
& E\{(U_{New})_{ij}|c_{ij},\theta\} \\
= & \int \sum_{l=1}^{n_i} \phi_{New}^{-1} X_{il}\mu_{il}^{1/2}\tilde{Q}_i(\rho_{New})_{lj}\mu_{ij}^{-1/2}\Lambda_0(t)(\kappa_{ij}-\mu_{ij})dF_{ij}(t|c_{ij},\theta) \\
= & \{\phi_{New}^{-1}\sum_{l=1}^{n_i} X_{il}\mu_{il}^{1/2}\tilde{Q}_i(\rho_{New})_{lj}\mu_{ij}^{-1/2}\}\int \Lambda_0(t)(\kappa_{ij}-\mu_{ij})dF_{ij}(t|c_{ij},\theta) \\
= & M_{ij}\int X_{ij}(\delta_{ij}-\Lambda_0(t)\mu_{ij})dF_{ij}(t|c_{ij},\theta) \\
= & M_{ij}\int \frac{\partial\mu_{ij}}{\partial\beta}\frac{1}{\mu_{ij}}(\delta_{ij}-\Lambda_0(t)\mu_{ij})dF_{ij}(t|c_{ij},\theta) \\
= & M_{ij}\int \frac{\partial}{\partial\beta}\log[\{\lambda_0(t)\mu_{ij}\}^{\delta_{ij}}\exp(-\Lambda_0(t)\mu_{ij})]dF_{ij}(t|c_{ij},\theta) \\
= & M_{ij}\int \frac{\partial}{\partial\beta}\log[\lambda^{\delta_{ij}}(t;X_{ij})\exp\{-\Lambda(t;X_{ij})\}]dF_{ij}(t|c_{ij},\theta) \\
= & M_{ij}\int \frac{\partial}{\partial\beta}[\log\{f_{ij}^{\delta_{ij}}(t|c_{ij},\theta)S_{ij}^{1-\delta_{ij}}(t|c_{ij},\theta)\}]dF_{ij}(t|c_{ij},\theta) \\
= & M_{ij}\{\int \delta_{ij}\frac{\partial}{\partial\beta}\log f_{ij}(t|c_{ij},\theta)dF_{ij}(t|c_{ij},\theta) \\
& + \int (1-\delta_{ij})\frac{\partial}{\partial\beta}\log S_{ij}(t|c_{ij},\theta)dF_{ij}(t|c_{ij},\theta)\} \\
= & M_{ij}\{\int_0^{c_{ij}} \frac{1}{f_{ij}(t|c_{ij},\theta)}\frac{\partial f_{ij}(t|c_{ij},\theta)}{\partial\beta}dF_{ij}(t|c_{ij},\theta) \\
& + \int_{c_{ij}}^{\infty}\frac{1}{S_{ij}(t|c_{ij},\theta)}\frac{\partial S_{ij}(t|c_{ij},\theta)}{\partial\beta}dF_{ij}(t|c_{ij},\theta)\} \\
= & M_{ij}\{\int_0^{c_{ij}}\frac{\partial f_{ij}(t|c_{ij},\theta)}{\partial\beta}dt + \int_{c_{ij}}^{\infty}\frac{\partial S_{ij}(t|c_{ij},\theta)}{\partial\beta}d\Lambda_{ij}(t|c_{ij},\theta)\} \\
= & M_{ij}\frac{\partial}{\partial\beta}\{F_{ij}(c_{ij}|c_{ij},\theta)+S_{ij}(c_{ij}|c_{ij},\theta)\} = \pi_{ij}M_{ij}\frac{\partial}{\partial\beta}(1) = \mathbf{0},
\end{aligned}
$$

where $M_{ij} = \{\phi_{New}^{-1}\sum_{l=1}^{n_i} X_{il}\mu_{il}^{1/2}\tilde{Q}_i(\rho_{New})_{lj}\mu_{ij}^{-1/2}/X_{ij}\}$ and $\tilde{Q}_i(\rho_{New}) = Q_i^{-1}(\rho_{New})$ and $Q_i(\rho_{New})$ is an $n_i \times n_i$ working correlation matrix for $i = 1, \cdots, K$. Therefore, $E\{(U_{New})_{ij}|\theta\} = E(E\{(U_{New})_{ij}|c_{ij}, \theta\}) = \mathbf{0}$. That is, U_{New} is an unbiased estimating function.

To prove the asymptotic normality of $\hat{\beta}_{New}$, we follow the similar discussion in Liang and Zeger (1986). For simplicity, we let $\rho_{New}^*(\beta_0) = \hat{\rho}(\beta_0, \hat{\phi}_{New}(\beta_0))$ and $U = U_{New}$. By Taylor expansion, we have

$$K^{1/2}(\hat{\beta}_{New} - \beta_0) \approx [\sum_{i=1}^{K} -\frac{\delta}{\delta\beta_0}U_i(\beta_0, \rho_{New}^*(\beta_0))/K]^{-1}[\sum_{i=1}^{K} U_i(\beta_0, \rho_{New}^*(\beta_0))/K^{1/2}],$$

where

$$\begin{aligned}\frac{\delta}{\delta\beta_0}U_i(\beta_0, \rho_{New}^*(\beta_0)) &= \frac{\partial}{\partial\beta_0}U_i(\beta_0, \rho_{New}^*(\beta_0)) + \frac{\partial U_i(\beta_0, \rho_{New}^*(\beta_0))}{\partial\rho_{New}^*(\beta_0)}\frac{\partial\rho_{New}^*(\beta_0)}{\partial\beta_0} \\ &= A_i^* + B_i^*C^*.\end{aligned}$$

Next we let β_0 be fixed and Taylor expansion gives

$$\begin{aligned}\frac{\sum_{i=1}^{K} U_i(\beta_0, \rho_{New}^*(\beta_0))}{K^{1/2}} &= \frac{\sum_{i=1}^{K} U_i(\beta_0, \rho_{New})}{K^{1/2}} \\ &\quad + \frac{\partial/\partial\rho\sum_{i=1}^{K} U_i(\beta_0, \rho_{New})}{K}K^{1/2}(\rho_{New}^*(\beta_0) - \rho_{New}) + o_p(1) \\ &= A^{**} + B^{**}C^{**} + o_p(1),\end{aligned}$$

where $B^{**} = o_p(1)$ by (A4) and

$$\begin{aligned}C^{**} &= K^{1/2}[\hat{\rho}(\beta_0, \hat{\phi}_{New}(\beta_0)) - \hat{\rho}(\beta_0, \phi_{New}) + \hat{\rho}(\beta_0, \phi_{New}) - \rho_{New}] \\ &= K^{1/2}[\frac{\partial\hat{\rho}(\beta_0, \phi_{New}^*)}{\partial\phi_{New}}(\hat{\phi}_{New} - \phi_{New}) + \hat{\rho}(\beta_0, \phi_{New}) - \rho_{New}] = O_p(1)\end{aligned}$$

by (A5), (A6) and (A7). Therefore, $K^{-1/2}\sum_{i=1}^{K} U_i(\beta_0, \rho_{New}^*(\beta_0))$ is asymptotically

equivalent to A^{**}. From (A3), we know that the asymptotic distribution of A^{**} is multivariate Gaussian with zero mean and covariance matrix $\lim_{K\to\infty}(\sum_{i=1}^{K} E((U_{New})_i^T (U_{New})_i)/K)$. It is easy to see that $\sum_{i=1}^{K} B_i^*/K = o_p(1)$ and $C^* = O_p(1)$. Therefore, $\sum_{i=1}^{K} \frac{\delta}{\delta\beta_0} U_i(\beta_0, \rho^*_{New}(\beta_0))/K$ is asymptotically equivalent to $\sum_{i=1}^{K} A_i^*/K$ which converges to $\frac{1}{K}\partial U_{New}(\beta_0)/\partial\beta_0$ as $K \to \infty$. By Slutsky's Theorem, as $K \to \infty$, we have

$$K^{1/2}(\hat{\beta}_{New} - \beta_0) \to N(\mathbf{0}, \Sigma) \quad \text{in distribution,}$$

where $\Sigma = \mathcal{A}^{-1}(\beta_0)\mathcal{V}(\beta_0)\mathcal{A}^{-T}(\beta_0)$, $\mathcal{A}(\beta_0) = -\partial U_{New}(\beta_0)/\partial\beta_0$ and $\mathcal{V}(\beta_0) = \sum_{i=1}^{K} \{U_{New}(\beta_0)\}_i\{U_{New}(\beta_0)\}_i^T$. Furthermore, we have

$$\hat{\beta}_{New} - \beta_0 = \frac{\Sigma}{K^{1/2}}(K^{1/2}\frac{\hat{\beta}_{New} - \beta_0}{\Sigma}) \to \lim_{K\to\infty}(\frac{\Sigma}{K^{1/2}})Z = 0,$$

where $Z \sim N(0,1)$. So $\hat{\beta}_{New} - \beta_0 \to 0$ in distribution by Slutsky's Theorem. We know that convergence in distribution to a point is equivalent to convergence in probability, so $\hat{\beta}_{New}$ is a consistent estimator of β_0.

3.3.2 Variance Estimation for $\hat{\beta}_{New}$ and $\hat{\rho}_{New}$

The estimating functions of β and α are

$$U_{New} = \sum_{i=1}^{K}(U_{New})_i = \sum_{i=1}^{K}\{\frac{\partial\mu(X_i)}{\partial\beta}\}^T\{B_i^{1/2}Q_i(\rho_{New})B_i^{1/2}\phi_{New}\}^{-1}W_i\{\kappa_i - \mu(X_i)\},$$

and

$$U_\alpha = \sum_{\{i,j\}\in D_k} \frac{\exp(\beta' X_{ij})}{1 - \exp(-\alpha_k \exp(\beta' X_{ij}))} - \sum_{(i,j)\in R_k} \exp(\beta' X_{ij}).$$

Therefore, the derivatives of the estimating functions U_{New} and U_α with respect to $\theta = (\beta, \alpha)$ have four elements including $U_{\beta\beta}$, $U_{\beta\alpha_k}$, $U_{\alpha_k\alpha_{k'}}$ and $U_{\alpha_k\beta}$. Specifically,

$$U_{\beta\beta} = \frac{\partial U_{New}}{\partial \beta} = \sum_{i=1}^{K} A_i^{(\beta)}[B_i^{(\beta)} W_i C_i^{(\beta)} - D_i^{(\beta)} W_i E_i^{(\beta)}],$$

where

$$A_i^{(\beta)} = (X_{i1\nu}, X_{i2\nu}, \cdots, X_{in_i\nu})_{1\times n_i},$$

$$B_i^{(\beta)} = \begin{pmatrix} B_{i11}^{(\beta)} & \cdots & B_{i1n_i}^{(\beta)} \\ \vdots & \ddots & \vdots \\ B_{in_i1}^{(\beta)} & \cdots & B_{in_in_i}^{(\beta)} \end{pmatrix}_{n_i\times n_i},$$

where

$$B_{imn}^{(\beta)} = \frac{1}{2\phi_{New}}(X_{im\omega} - X_{in\omega})(\mu_{im})^{1/2}(\mu_{in})^{-1/2}\tilde{Q}_i(\rho_{New})_{mn}$$

for $m \neq n$, otherwise $B_{imm}^{(\beta)} = 0$, $m, n = 1, 2, \cdots, n_i$.

$$C_i^{(\beta)} = (\kappa_{i1} - \mu_{i1}, \kappa_{i2} - \mu_{i2}, \cdots, \kappa_{in_i} - \mu_{in_i})^T,$$

$$D_i^{(\beta)} = \begin{pmatrix} D_{i11}^{(\beta)} & \cdots & D_{i1n_i}^{(\beta)} \\ \vdots & \ddots & \vdots \\ D_{in_i1}^{(\beta)} & \cdots & D_{in_in_i}^{(\beta)} \end{pmatrix}_{n_i\times n_i},$$

where

$$D_{imn}^{(\beta)} = \phi_2^{-1}(\mu_{im})^{1/2}(\mu_{in})^{-1/2}\tilde{Q}_i(\rho_{New})_{mn}$$

for $m \neq n$, otherwise $D^{(\beta)}_{imm} = \tilde{Q}_i(\rho_{New})_{mm}$.

$$E^{(\beta)}_i = (X_{i1\omega}\mu_{i1}, X_{i2\omega}\mu_{i2}, \cdots, X_{in_i\omega}\mu_{in_i})^T,$$

and $W_i = \text{diag}(\Lambda_0(t_{i1}), \cdots, \Lambda_0(t_{in_i}))$.

$$U_{\alpha_k\alpha_{k'}} = -\sum_{(i,j)\in D_k} \frac{\exp(2\beta' X_{ij} - \alpha_k \exp(\beta' X_{ij}))}{(1-\exp(-\alpha_k \exp(\beta' X_{ij})))^2}, \text{if } k' = k, \text{ 0 otherwise,}$$

$$\begin{aligned} U_{\alpha_k\beta} = & \sum_{(i,j)\in D_k} \frac{\exp(\beta' X_{ij})(1-\exp(-\alpha_k \exp(\beta' X_{ij})))}{(1-\exp(-\alpha_k \exp(\beta' X_{ij})))^2} X_{ij} \\ & - \sum_{(i,j)\in D_k} \frac{-\alpha_k \exp(\beta' X_{ij}) \exp(-\alpha_k \exp(\beta' X_{ij})))}{(1-\exp(-\alpha_k \exp(\beta' X_{ij})))^2} X_{ij} \\ & - \sum_{(i,j)\in R_k} \exp(\beta' X_{ij}) X_{ij}, \end{aligned}$$

$$U_{\beta\alpha_k} = -\sum_{(i,j):t_{ij}\geq\tau_k} (\sum_{l=1}^{n_i} X_{il}\mu_{il}^{1/2}\mu_{ij}^{-1/2}\tilde{Q}_i(\rho_{New})_{lj}) \exp(\beta' X_{ij}).$$

We consider a bootstrap method (Efron and Tibshirani, 1993; Monaco et al., 2005) to obtain the variance estimates of the correlation coefficient $\hat{\rho}_{New}$. The bootstrap sample is obtained from sampling clusters with replacement and is fitted with the proposed method to obtain the parameter estimates, denoted as $\hat{\rho}_b$. The variance estimates can be estimated by $\sum_{i=1}^{B}(\hat{\rho}_b^{(i)} - \sum_{j=1}^{B}\hat{\rho}_b^{(j)}/B)^2/(B-1)$, where B is the number of bootstrap samples for each simulated data set.

3.4 A Simulation Study

We conduct a simulation study to investigate the performances of the proposed method (denoted as New in the following tables) and to compare the results with those from Lee et al. (1992) (denoted as LWA in the following tables) and Segal and Neuhaus (1993) (denoted as SN in the following tables). The classical clustered survival data in the simulation study are generated from the proportional hazards model (3.1) with the exchangeable correlation structure. In particular, we consider a single binary covariate (mean=0.5) with value 0 for a control group and 1 for a treatment group, and a single continuous covariate with values generated from the standard normal distribution. Note that Segal and Neuhaus (1993) only considered the standard normal covariate in their simulation study. The effect of the covariate on $S(t; X_{ij})$ is specified by $\beta = \log(2)$ and the baseline distribution is the exponential distribution, i.e., $S_0(t; \alpha) = e^{-\alpha t}$ where $\alpha = 2$.

For each data set, we consider the following pairs of the number of clusters and the cluster size: (40,10), (80,5), and (200,2). For each cluster in a data set, the correlated failure times are generated by using the Clayton copula model (Clayton, 1978),

$$\begin{aligned} &P(\tilde{T}_{i1} > t_{i1}, \cdots, \tilde{T}_{in_i} > t_{in_i} | X_{ij}, j = 1, \cdots, n_i) \\ &= \{\sum_{j=1}^{n_i} S(t_{ij}; X_{ij})^{-\xi} - n_i + 1\}^{-1/\xi}, \end{aligned}$$

where $S(t_{ij}; X_{ij})$ is the marginal survival function given by (3.1). The value of ξ measures the degree of dependence among the failure times within cluster i and it relates to Kendall's tau by $\tau = \xi/(\xi + 2)$. We set $\xi = 8, 2$, and 0. The corresponding values of τ are 0.8, 0.5 and 0 respectively, and the larger value the stronger correlation of the failure times. When $\xi = 0$ or $\tau = 0$, it implies the independence among

the failure times. The censoring times are non-informative and generated from the uniform distribution in $(0, c)$ with c chosen to produce about 10%, 30%, and 50% censoring.

For each setting above, we generate 1000 data sets and fit each data set with the marginal model using the proposed estimating function (3.5) with an exchangeable working correlation matrix. As a comparison, we also estimate the parameters in the marginal model using the robust sandwich methods in Lee et al. (1992), i.e., estimating function (3.4) or (3.5) with an identity working correlation matrix, and in Segal and Neuhaus (1993), i.e., estimating function (3.4) with exchangeable working correlation matrix. The biases, empirical variances (Var), the averages of estimated variances (Var*), and the coverage probabilities (CP) of 95% confidence intervals of the parameter estimates based on the above three methods are reported in Tables 3.1, 3.2, and 3.3 for different censoring rates.

The simulation results indicate that the proposed estimation method outperforms the existing estimation methods. The average estimated variances of the regression parameters from the proposed method are close to their empirical variances, and the 95% confidence interval coverage rates are satisfactory and close to the nominal level. When the failure times within a cluster are correlated, the variances from the proposed method are consistently smaller than those from the Lee et al.'s method (1992). We notice that Segal et al.'s method (1993) also improves the estimation efficiency when the correlation exists within clusters. However, the biases of $\hat{\beta}_{SN}$ based on the estimating equation $U_{SN} = 0$ are obvious and significantly affect the coverage probability, especially when the correlation is strong and the cluster size is large. When the correlation reduces to zero, the empirical variances based on the proposed method and Lee et al.'s method are comparable.

To further evaluate the efficiency gains from the proposed method, we calculate

Table 3.1: Bias, empirical variance (Var), average of estimated variance (Var*), coverage percentage (CP) of 95% confidence intervals of the estimate of β with censoring rate equals 0.5.

n_i/K	40/10			80/5			200/2		
	LWA	SN	New	LWA	SN	New	LWA	SN	New
binary covariate									
$\tau = 0.8$									
Bias	0.021	-0.153	0.019	0.005	-0.147	0.004	0.002	-0.096	0.001
Var	0.032	0.026	0.021	0.027	0.017	0.016	0.022	0.016	0.018
Var*	0.030	0.012	0.019	0.025	0.009	0.015	0.022	0.008	0.018
CP	94.8	60.2	93.8	93.9	57.4	93.8	94.2	73.3	94.4
$\tau = 0.5$									
Bias	0.011	-0.535	0.011	0.011	-0.428	0.010	0.000	-0.224	0.000
Var	0.023	0.022	0.018	0.024	0.020	0.019	0.020	0.019	0.019
Var*	0.024	0.017	0.018	0.023	0.013	0.018	0.022	0.010	0.020
CP	94.3	2.6	95.4	94.5	6.3	95.5	95.7	41.4	95.6
$\tau = 0$									
Bias	-0.002	-0.540	-0.003	0.002	-0.297	0.003	0.008	-0.085	0.008
Var	0.023	0.038	0.023	0.020	0.038	0.020	0.020	0.029	0.021
Var*	0.021	0.013	0.021	0.021	0.010	0.021	0.021	0.009	0.021
CP	92.8	4.3	92.6	95.5	26.3	95.7	95.9	66.3	95.4
continuous covariate									
$\tau = 0.8$									
Bias	0.015	0.104	0.016	0.009	0.088	0.009	0.006	0.057	0.005
Var	0.014	0.010	0.011	0.010	0.007	0.007	0.007	0.005	0.006
Var*	0.013	0.004	0.011	0.010	0.003	0.007	0.007	0.004	0.006
CP	95.1	57.1	94.9	94.2	58.9	94.5	94.3	75.8	94.2
$\tau = 0.5$									
Bias	0.019	0.034	0.017	0.003	0.022	0.003	0.008	0.016	0.007
Var	0.010	0.007	0.008	0.007	0.005	0.006	0.006	0.005	0.006
Var*	0.009	0.005	0.007	0.007	0.005	0.006	0.006	0.005	0.006
CP	92.9	87.4	92.3	94.3	90.5	93.6	94.4	91.4	94.6
$\tau = 0$									
Bias	0.006	-0.042	0.005	0.005	-0.022	0.005	0.004	-0.004	0.004
Var	0.006	0.006	0.006	0.006	0.007	0.006	0.006	0.006	0.006
Var*	0.006	0.006	0.006	0.006	0.005	0.006	0.006	0.005	0.006
CP	94.3	90.0	94.5	92.8	89.7	92.9	94.4	92.0	94.3

Table 3.2: Bias, empirical variance (Var), average of estimated variance (Var*), coverage percentage (CP) of 95% confidence intervals of the estimate of β with censoring rate equals 0.3.

n_i/K	40/10			80/5			200/2		
	LWA	SN	New	LWA	SN	New	LWA	SN	New
				binary covariate					
$\tau = 0.8$									
Bias	0.033	-0.121	0.031	0.015	-0.117	0.014	0.007	-0.085	0.005
Var	0.023	0.016	0.014	0.019	0.011	0.011	0.016	0.010	0.012
Var*	0.022	0.007	0.014	0.018	0.005	0.011	0.016	0.005	0.012
CP	95.0	57.4	94.2	94.7	56.9	95.2	94.8	69.3	95.0
$\tau = 0.5$									
Bias	0.023	-0.475	0.023	0.000	-0.394	0.000	0.012	-0.214	0.011
Var	0.020	0.015	0.014	0.016	0.026	0.016	0.015	0.014	0.014
Var*	0.019	0.011	0.013	0.015	0.008	0.015	0.016	0.007	0.014
CP	92.8	0.9	93.5	94.5	7.3	94.4	95.6	32.7	95.9
$\tau = 0$									
Bias	0.001	-0.642	0.002	0.004	-0.387	0.004	0.008	-0.124	0.008
Var	0.014	0.025	0.015	0.015	0.026	0.015	0.016	0.023	0.016
Var*	0.015	0.011	0.015	0.015	0.008	0.015	0.015	0.007	0.015
CP	94.5	0.3	94.6	95.4	8.2	95.1	94.0	56.2	94.0
				continuous covariate					
$\tau = 0.8$									
Bias	0.022	0.092	0.024	0.015	0.072	0.013	0.002	0.046	0.002
Var	0.012	0.009	0.010	0.008	0.005	0.006	0.006	0.004	0.005
Var*	0.011	0.002	0.009	0.007	0.002	0.005	0.005	0.002	0.004
CP	93.7	45.3	94.9	93.4	52.7	95.2	91.9	75.1	92.5
$\tau = 0.5$									
Bias	0.018	0.046	0.017	0.010	0.032	0.009	0.002	0.017	0.002
Var	0.009	0.006	0.007	0.007	0.005	0.005	0.005	0.004	0.004
Var*	0.008	0.003	0.006	0.006	0.003	0.005	0.005	0.003	0.004
CP	91.4	76.9	92.7	93.2	82.7	92.7	94.6	89.9	94.7
$\tau = 0$									
Bias	0.003	-0.034	0.003	0.004	-0.019	0.004	0.001	-0.006	0.001
Var	0.004	0.004	0.004	0.005	0.005	0.005	0.004	0.004	0.004
Var*	0.004	0.005	0.005	0.005	0.005	0.005	0.005	0.004	0.005
CP	95.7	94.0	95.3	93.8	91.8	93.5	96.0	94.0	96.0

Table 3.3: Bias, empirical variance (Var), average of estimated variance (Var*), coverage percentage (CP) of 95% confidence intervals of the estimate of β with censoring rate equals 0.1.

n_i/K	40/10			80/5			200/2		
	LWA	SN	New	LWA	SN	New	LWA	SN	New
				binary covariate					
$\tau = 0.8$									
Bias	0.041	-0.137	0.040	0.010	-0.135	0.015	0.003	-0.103	0.002
Var	0.023	0.014	0.013	0.016	0.007	0.008	0.013	0.006	0.009
Var*	0.019	0.004	0.011	0.015	0.003	0.008	0.013	0.003	0.009
CP	90.0	39.5	93.5	93.4	33.3	93.2	93.6	55.6	93.9
$\tau = 0.5$									
Bias	0.035	-0.466	0.034	0.016	-0.391	0.016	0.015	-0.218	0.014
Var	0.018	0.012	0.012	0.015	0.009	0.011	0.012	0.009	0.011
Var*	0.016	0.008	0.010	0.014	0.006	0.010	0.013	0.005	0.011
CP	91.7	0.4	91.7	92.7	0.3	93.1	94.4	21.6	94.5
$\tau = 0$									
Bias	0.002	-0.741	0.002	0.003	-0.469	0.003	0.001	-0.167	0.001
Var	0.013	0.020	0.013	0.012	0.018	0.012	0.012	0.018	0.012
Var*	0.012	0.011	0.012	0.012	0.007	0.012	0.012	0.006	0.012
CP	93.5	0.0	93.1	94.3	0.9	94.0	95.5	42.3	95.5
				continuous covariate					
$\tau = 0.8$									
Bias	0.028	0.056	0.031	0.016	0.038	0.014	0.003	0.021	0.004
Var	0.012	0.008	0.010	0.007	0.004	0.005	0.004	0.003	0.004
Var*	0.010	0.001	0.009	0.007	0.001	0.005	0.004	0.001	0.003
CP	92.6	38.5	93.4	93.3	51.3	93.5	94.2	76.0	94.2
$\tau = 0.5$									
Bias	0.025	0.034	0.026	0.011	0.021	0.009	0.007	0.012	0.006
Var	0.008	0.006	0.007	0.006	0.004	0.005	0.004	0.004	0.004
Var*	0.007	0.002	0.006	0.005	0.002	0.004	0.004	0.002	0.004
CP	91.7	67.8	92.3	92.2	79.0	93.0	94.9	85.4	94.4
$\tau = 0$									
Bias	0.004	-0.012	0.004	0.002	-0.008	0.003	0.000	-0.004	0.000
Var	0.004	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004
Var*	0.004	0.005	0.004	0.004	0.004	0.004	0.004	0.003	0.004
CP	93.6	93.2	93.7	95.4	94.0	95.5	94.2	90.8	94.1

Table 3.4: Relative efficiency of $\hat{\beta}_{New}$ vs $\hat{\beta}_{LWA}$

RE = MSE($\hat{\beta}_{New}$)/MSE($\hat{\beta}_{LWA}$)						
	Binary			Normal		
τ	40/10	80/5	200/2	40/10	80/5	200/2
	censoring=0.5					
0.8	0.658	0.593	0.818	0.791	0.702	0.856
0.5	0.784	0.792	0.950	0.800	0.857	0.998
0	1.000	1.000	1.050	0.998	1.000	1.000
	censoring=0.3					
0.8	0.621	0.582	0.749	0.847	0.750	0.833
0.5	0.708	1.000	0.932	0.782	0.716	0.800
0	1.072	1.000	1.000	1.000	1.000	1.000
	censoring=0.1					
0.8	0.592	0.511	0.692	0.857	0.716	1.002
0.5	0.684	0.738	0.916	0.890	0.830	0.997
0	1.000	1.000	1.000	1.000	1.001	1.000

the relative efficiency (RE) defined as the ratio of mean squared error of the estimate from the proposed method to that from the Lee et al.'s method, i.e. RE = MSE($\hat{\beta}_{New}$)/MSE($\hat{\beta}_{LWA}$), and report them in Table 3.4. The results indicate that the proposed method can achieve considerable efficiency gain for regression parameters when the correlation is strong and the cluster size is large, and it is still comparable with the Lee et al.'s method when the correlation is weak or cluster size is small. For example, when $\tau = 0.8$ and $K/n_i = 80/5$, the REs of β can be as low as 0.511 for binary covariate, and 0.702 for continuous covariate. The REs tend to approach 1 when the correlation decreases to zero.

Table 3.5 shows the estimate of ρ_{New}, and their empirical variances and the averages of 100 bootstrap variances. To save time, we computed bootstrap variances of $\hat{\rho}_{New}$ only for 100 randomly selected data sets from the 1000 simulated data sets based on $B = 100$. The similarity of the empirical variances and the bootstrap variances indicates that the bootstrap variance estimator works well for calculating the

Table 3.5: Mean, empirical variance (Var), average of estimated variance (Var*) of $\hat{\rho}_{New}$

Type of	10/40			5/80			2/200		
covariate	Mean	Var	Var*	Mean	Var	Var*	Mean	Var	Var*
	censoring rate = 0.5								
$\tau = 0.8$									
discrete	0.173	0.005	0.005	0.180	0.007	0.006	0.179	0.014	0.013
continuous	0.173	0.007	0.006	0.175	0.008	0.007	0.181	0.017	0.015
$\tau = 0.5$									
discrete	0.081	0.002	0.002	0.084	0.003	0.003	0.086	0.007	0.006
continuous	0.081	0.002	0.002	0.083	0.002	0.003	0.082	0.006	0.004
$\tau = 0$									
discrete	0.029	0.003	0.003	0.027	0.0006	0.0007	0.028	0.002	0.002
continuous	0.028	0.003	0.003	0.027	0.0001	0.0001	0.031	0.004	0.002
	censoring rate = 0.3								
$\tau = 0.8$									
discrete	0.178	0.006	0.005	0.185	0.006	0.006	0.188	0.015	0.013
continuous	0.180	0.007	0.007	0.184	0.008	0.006	0.188	0.016	0.014
$\tau = 0.5$									
discrete	0.087	0.001	0.002	0.032	0.0007	0.0006	0.094	0.009	0.009
continuous	0.086	0.002	0.002	0.090	0.003	0.002	0.092	0.008	0.007
$\tau = 0$									
discrete	0.033	0.0004	0.0003	0.032	0.0008	0.0006	0.032	0.002	0.002
continuous	0.033	0.0003	0.0004	0.033	0.0009	0.0007	0.032	0.003	0.001
	censoring rate = 0.1								
$\tau = 0.8$									
discrete	0.181	0.005	0.005	0.186	0.006	0.007	0.189	0.014	0.012
continuous	0.182	0.007	0.006	0.191	0.009	0.008	0.192	0.017	0.016
$\tau = 0.5$									
discrete	0.091	0.001	0.002	0.094	0.002	0.003	0.094	0.006	0.005
continuous	0.093	0.002	0.002	0.092	0.003	0.003	0.092	0.007	0.006
$\tau = 0$									
discrete	0.035	0.0003	0.0003	0.034	0.0006	0.0007	0.035	0.003	0.002
continuous	0.036	0.0004	0.0005	0.036	0.0007	0.0007	0.036	0.002	0.002

variance estimate of ρ_{New}. Although ρ_{New} does not correspond to the correlation measure τ in the data generation, Table 3.5 shows that the estimated value of ρ_{New} agrees well with the value τ used in the data generation in the sense that when the latter decreases, the former tends to decrease too. When there is no correlation in clusters, the estimate of ρ_{New} is very close to zero.

3.5 Applications

3.5.1 The Diabetic Retinopathy Study

We consider a data set from a Diabetic Retinopathy study which was conducted by the National Eye Institute (Section 1.2.1). One objective of this study is to evaluate the effectiveness of laser photocoagulation in delaying the onset of blindness in patients with diabetic retinopathy. In our analysis, we consider 197 patients coming from 50% random sample of the patients with "high-risk" diabetic retinopathy as defined by the Diabetic Retinopathy Study criteria. By the end of the study, 54 treated and 101 control eyes in this group of patients had developed blindness. This data set has been widely analyzed in the literature with respect to the marginal method (Huster et al., 1989; Lee et al., 1992; Liang et al., 1993; Lin, 1994; and Segal et al., 1997).

The Kaplan-Meier survival curves by treatment and type of diabetes (Figure 1.1) show that the treatment is more effective for adult diabetes patients than for juvenile diabetes whereas in the untreated group, juvenile patients tend to have higher survival probabilities than the adult patients. Since these two age groups have very different patterns, we include the interaction term between treatment and the type of diabetes in our analysis. Therefore, we consider three covariates, i.e., treatment (1 for treated and 0 otherwise), type of diabetes (1 for adult and 0 otherwise) and the interaction between them. Additionally, we are also interested in investigating the correlation

Table 3.6: Estimated parameters from fitting the marginal proportional hazards model to the Diabetic data using Lee et al.'s method (LWA), Segal et al.'s method (SN) and the proposed method (New). The standard error estimates are given in parentheses.

Covariate	Methods		
	LWA	SN	New
Treatment	-0.425	-0.784	-0.425
	(0.185)	(0.250)	(0.184)
Diabetic Type	0.341	0.034	0.341
	(0.196)	(0.247)	(0.195)
Interaction	-0.846	-0.646	-0.846
	(0.304)	(0.362)	(0.303)
$\hat{\rho}_{SN}$		0.215	-
		(0.072)	-
$\hat{\rho}_{New}$		-	0.033
		-	(0.031)

that may exist between two eyes of a patient.

We fit the survival data with Lee et al.'s method (1992), Segal et al.'s method (1993), and the proposed method, respectively. We use the nonparametric estimate of the baseline survival function as in (3.6). The parameter estimates are summarized in Table 3.6. Note that the standard errors of the estimated correlation parameters $\hat{\rho}_{New}$ and $\hat{\rho}_{SN}$ in the table are obtained from 500 bootstrap samples separately.

From Table 3.6, we conclude that the estimates from the proposed method are similar to the results from Lee et al.'s method. In addition, we observe a positive correlation ($\hat{\rho}_{New} = 0.033$) between two eyes for each patient by incorporating an exchangeable working correlation structure in estimating function (3.5). The variances of the estimates in the proposed method are a little bit smaller than those in Lee et al.'s method (1992) although the significance of the regression parameters do not change in these two methods. Based on Segal et al.'s method (1993) with

exchangeable working correlation matrix, we also obtain a positive correlation estimate ($\hat{\rho}_{SN} = 0.215$) which is similar to the dependence estimate based on the design effect method (Segal et al., 1997). It is worth to note that we did not observe non-convergence in the regression parameters as they mentioned when using the Poisson likelihood. In addition, due to the bias of Segal et al.'s estimating function (3.4), the estimates of the regression parameters are biased although the correlation estimate $\hat{\rho}_{SN}$ is significant.

3.5.2 The Study of Infections in Kidney Patients

We consider a data set from the kidney infection study (Section 1.2.2). Two recurrence times (T_{i1}, T_{i2}) (defined in Section 1.2.2) and the corresponding censoring indicators $(\delta_{i1}, \delta_{i2})$ were recorded for the ith patient $(i = 1, \cdots, 38)$. Other variables include age (in years), gender (1 for female and 0 for male) and type of kidney disease (0 for glomerulo nephritis (GN), 1 for acute nephritis (AN), 2 for polycystic kidney disease (PKD), and 3 otherwise). This data set has been analyzed by using a multiplicative frailty model (McGilchrist and Aisbett, 1991; McGilchrist, 1993) as well as a marginal model (Chen et al., 2010).

The primary interest of the kidney patients study is to assess the the factors such as age, gender and the type of kidney disease to the development of infections. Meanwhile, we are also interested in investigating whether the recurrence times within one patient are related. We fit the survival data with Lee et al.'s method, Segal et al.'s method and the proposed method, respectively. To compare with a frailty model, we also report the results based on the ML and REML estimation with log-normal frailty for analyzing kidney patients data (McGilchrist, 1993). The parameter estimates are summarized in Table 3.7. Note that the standard errors of the estimated correlation parameters in the table are obtained from 500 bootstrap samples.

Table 3.7: Estimated parameters from fitting the marginal proportional hazards model to the Kidney Infections data using the robust method (LWA), Segal's method (SN), proposed method (New) as well as McGilchrist's methods (ML and REML). The standard error estimates are given in parentheses.

Covariate	Methods				
	LWA	SN	New	ML	REML
Age	0.003	-0.003	0.003	0.004	0.005
	(0.007)	(0.007)	(0.006)	0.013	(0.015)
Sex	-1.483	-0.241	-1.471	-1.605	-1.740
	(0.401)	(0.307)	(0.345)	(0.407)	(0.472)
GN	0.088	0.025	0.090	0.132	0.186
	(0.287)	(0.277)	(0.285)	(0.461)	(0.552)
AN	0.351	0.042	0.353	0.357	0.392
	(0.275)	(0.383)	(0.279)	(0.458)	(0.553)
PKD	-1.431	-0.034	-1.427	-1.295	-1.143
	(0.871)	(0.563)	(0.834)	(0.724)	(0.829)
$\hat{\rho}_{SN}$	-	0.301	-		-
	-	(0.144)	-		-
$\hat{\rho}_{New}$	-	-	0.057		-
	-	-	(0.097)		-
$\hat{\theta}_{ML}$	-	-	-	0.179	-
	-	-	-	(0.120)	-
$\hat{\theta}_{REML}$	-	-	-	-	0.546
	-	-	-	-	(0.310)

The results from the five methods show some substantial differences. For example, the effect of PKD disease is marginally significant in our method (p-value=0.087) and in ML method (p-value=0.074) while it is insignificant in other three methods. That is, the patients with PKD tend to have lower infection risk than those without PKD. All methods except Segal et al.'s method (1993) show that gender is a significant factor, indicating that male patients are about four to five times more likely than female patients to experience infections. Age appears to have no association with risk of infection, after adjusting for gender and type of disease. Both the proposed method

and Segal et al.'s method obtain a positive correlation between the failure times to infection measured on each patient. Different from the strong correlation estimate ($\hat{\rho}_{SN} = 0.301$) in Segal et al.'s method, the association estimate ($\hat{\rho}_{New} = 0.057$) in our method is weak. From McGilchrist (1993), we know that the variance of log-normal frailty θ is insignificant (p-value=0.136) in ML method and marginally significant (p-value=0.078) in REML method. We also notice that the regression estimates are similar between the proposed method and Lee et al.'s method.

3.6 Conclusions

Segal and Neuhaus (1993) considered a parametric marginal proportional hazards model with Weibull baseline assumption for multivariate failure time data. However, due to the bias of their estimating function, the corresponding regression estimators are biased and the variance estimates are unstable. They also observed nonconvergence (Segal and Neuhaus,1993; Segal et al., 1997) in the regression parameters when using the exchangeable working correlation matrix in their estimating equation. In this chapter, we considered a semiparametric marginal proportional hazards model and proposed an unbiased weighted estimating function for clustered survival data to accommodate the correlation within clusters. The estimates of the regression parameters are shown to be consistent and asymptotically normal under regularity conditions, and their variances can be consistently estimated by a sandwich variance estimator. The proposed estimating equation is easily implemented. Our numerical study shows that the proposed method substantially improves the estimation efficiency of the regression parameters, especially when the correlation within clusters is strong and the cluster size is large, comparing with the existing marginal method (Lee et al., 1992). The large sample approximation is reliable for the practical sample sizes. Therefore,

the proposed marginal proportional hazards model could be considered as an alternative approach to the existing marginal models for classical clustered survival data. In kidney infection study, we further compared the proposed marginal model with a frailty model (McGlichrist, 1993), the results demonstrate that both models reveal a correlation between two consecutive infection times measured on the same patient. Note that although we considered paired survival times in applications, the proposed method can readily be applied to studies with larger and unequal cluster sizes.

Chapter 4

Parametric Marginal Proportional Hazards Mixture Cure Model

4.1 Introduction

In this chapter, we consider a parametric marginal mixture cure model for clustered survival data in which individuals may have long-term censored survival times and there may also be correlations between individuals. We propose a generalized estimating equation approach by incorporating working correlation matrices into the M-step of the EM algorithm to estimate the regression coefficients and the baseline hazard function in the marginal model. The estimators of the regression parameters and the baseline hazard function are shown to be consistent and asymptotically normal, and their variances can be consistently estimated by a sandwich estimator. We conduct a simulation study to assess finite sample properties and illustrate the proposed method with an application to the analysis of a smoking cessation study.

This chapter is organized as follows. In Section 4.2, we introduce the marginal proportional hazards mixture cure model and propose a set of estimating equations for clustered survival data with a cure fraction. The asymptotic properties of the

estimators are investigated in Section 4.3. We conduct a simulation study to evaluate the finite sample performance of the proposed estimation method in Section 4.4, and illustrate this method by analyzing the smoking cessation data in Section 4.5. Conclusions and discussions are presented in Section 4.6.

4.2 Model and Estimation Method

We assume that the marginal survival function of $\tilde{T}_{ij}^*$ is from a parametric proportional hazards mixture cure model, i.e.,

$$S(t; X_{ij}, Z_{ij}) = P(\tilde{T}_{ij}^* > t; X_{ij}, Z_{ij}) = 1 - \pi(Z_{ij}) + \pi(Z_{ij}) S_u(t; X_{ij}), \tag{4.1}$$

where $\pi(Z_{ij}) = P(Y_{ij} = 1; Z_{ij})$ is in a logistic regression form

$$\pi(Z_{ij}) = \frac{\exp(\gamma' Z_{ij})}{1 + \exp(\gamma' Z_{ij})}, \tag{4.2}$$

and $S_u(t; X_{ij}) = P(\tilde{T}_{ij}^* > t | Y_{ij} = 1; X_{ij})$ is specified by the proportional hazards model

$$S_u(t; X_{ij}) = S_{u0}(t; \alpha)^{\exp(\beta' X_{ij})}, \tag{4.3}$$

and $S_{u0}(t; \alpha)$, the baseline survival function of $\tilde{T}_{ij}^* | \{Y_{ij} = 1\}$ when $X_{ij} = 0$, is assumed to follow Weibull distribution with $S_{u0}(t; \alpha) = \exp(-t^\alpha)$. Note that we use two-parameter Weibull distribution where the scale parameter is considered as an intercept term in the proportional hazards model. Here β and γ are $p_X + 1$ and $p_Z + 1$ unknown regression parameters for X_{ij} and Z_{ij}, and α is an unknown parameter in the baseline distribution.

If we ignore the correlation within clusters, the unknown parameters in the model are often estimated using the EM algorithm based on a complete log-likelihood

function from the augmented data $O_c^* = \{(T_{ij}^*, \delta_{ij}, X_{ij}, Z_{ij}, Y_{ij}), j = 1, \cdots, n_i, i = 1, \cdots, K\}$. That is

$$
\begin{aligned}
& l_{\mathbf{c}}(\theta^*; O_c^*) = l_{\mathbf{c}}(\gamma, \beta, \alpha; O_c^*) \\
= \; & \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} \pi(Z_{ij})^{y_{ij}} \{1 - \pi(Z_{ij})\}^{1-y_{ij}} \\
& + \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} [\{\lambda_{u0}(t_{ij}^*; \alpha) \exp(\beta' X_{ij})\}^{\delta_{ij}} \exp\{-\Lambda_{u0}(t_{ij}^*; \alpha) \exp(\beta' X_{ij})\}]^{y_{ij}} \\
= \; & \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} \pi(Z_{ij})^{y_{ij}} \{1 - \pi(Z_{ij})\}^{1-y_{ij}} \\
& + \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} [\exp(\beta' X_{ij})^{\kappa_{ij}} \exp\{-\exp(\beta' X_{ij})\}]^{y_{ij} t_{ij}^{*\alpha}} \\
& + \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} (\alpha t_{ij}^{*\alpha-1})^{\delta_{ij}}, \qquad (4.4)
\end{aligned}
$$

where $\lambda_{u0}(t_{ij}^*; \alpha) = \alpha t_{ij}^{*(\alpha-1)}$ and $\Lambda_{u0}(t_{ij}^*; \alpha) = t_{ij}^{*\alpha}$ are the hazard and cumulative hazard functions corresponding to $S_{u0}(t_{ij}^*; \alpha)$, and $\kappa_{ij} = \delta_{ij}/t_{ij}^{*\alpha}$. Equation (4.4) consists of three terms. The first term corresponds to a log-likelihood function of γ based on a logistic regression for y_{ij} only. The second term can be viewed as a log-likelihood function for β. The third term only contains the information about α. By differentiating $l_{\mathbf{c}}(\theta^*; O_c^*)$ with respect to θ^*, we obtain the following three estimating equations,

$$
\begin{aligned}
U_{\gamma_{EM}} &= \sum_{i=1}^{K} (U_{\gamma_{EM}})_i = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (U_{\gamma_{EM}})_{ij} \\
&= \sum_{i=1}^{K} \{\frac{\partial \pi(Z_i)}{\partial \gamma}\}^T \{A_i^{1/2} I_i A_i^{1/2}\}^{-1} \{y_i - \pi(Z_i)\} = 0, \qquad (4.5)
\end{aligned}
$$

$$
\begin{aligned}
U_{\beta_{EM}} &= \sum_{i=1}^{K}(U_{\beta_{EM}})_i = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(U_{\beta_{EM}})_{ij} \\
&= \sum_{i=1}^{K}\{\frac{\partial\mu(X_i)}{\partial\beta}\}^T\{B_i^{1/2}I_iB_i^{1/2}\}^{-1}W_i\{\kappa_i-\mu(X_i)\}=0, \qquad (4.6)
\end{aligned}
$$

$$
\begin{aligned}
U_{\alpha_{EM}} &= \sum_{i=1}^{K}(U_{\alpha_{EM}})_i = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(U_{\alpha_{EM}})_{ij} \\
&= \sum_{i=1}^{K}\sum_{j=1}^{n_i}[y_{ij}t_{ij}^{*\alpha}\log(t_{ij})\{\kappa_{ij}-\mu(X_{ij})\}+\delta_{ij}/\alpha]=0, \qquad (4.7)
\end{aligned}
$$

where $y_i = (y_{i1},\cdots,y_{in_i})^T$, $\pi(Z_i) = \{\pi(Z_{i1}),\cdots,\pi(Z_{in_i})\}^T$, $A_i = \text{diag}[\pi(Z_{i1})\{1-\pi(Z_{i1})\},\ \cdots,\ \pi(Z_{in_i})\{1-\pi(Z_{in_i})\}]$, $\kappa_i = (\kappa_{i1},\cdots,\kappa_{in_i})^T$, $\mu(X_i) = \{\mu(X_{i1}),\cdots,\mu(X_{in_i})\}^T$ with $\mu(X_{ij}) = \exp(\beta' X_{ij})$, $B_i = \text{diag}\{\mu(X_{i1}),\cdots,\mu(X_{in_i})\}$, $W_i = \text{diag}(y_{i1}t_{i1}^{*\alpha},\ \cdots, y_{in_i}t_{in_i}^{*\alpha})$, and I_i is an $n_i \times n_i$ identity matrix. Note that diag(A) implies a diagonal matrix with diagonal elements from the vector A. The E-step computes the conditional expectation of $l_{\mathbf{c}}(\theta^*; O_c^*)$ with respect to Y_{ij} given the observed data and the current estimates of the parameters. If the current estimates are denoted by $\theta^{*(m)} = (\gamma^{(m)}, \beta^{(m)}, \alpha^{(m)})$, then the E-step is equivalent to computing

$$
\begin{aligned}
g_{ij}^{(m)} &= E(Y_{ij}|\theta^{*(m)}, O^*) \\
&= \left\{\delta_{ij} + \frac{(1-\delta_{ij})\pi(Z_{ij})S_{u0}(t_{ij}^*)^{\exp(\beta' X_{ij})}}{1-\pi(Z_{ij})+\pi(Z_{ij})S_{u0}(t_{ij}^*)^{\exp(\beta' X_{ij})}}\right\}_{\theta^*=\theta^{*(m)}}, \qquad (4.8)
\end{aligned}
$$

where $O^* = \{(T_{ij}^*, \delta_{ij}, X_{ij}, Z_{ij}), j = 1,\cdots,n_i, i = 1,\cdots,K\}$ is the observed data, and the M-step is equivalent to solving equations (4.5), (4.6) and (4.7) after substituting $g_{ij}^{(m)}$ for y_{ij} in the equations. We denote the estimator $\hat{\theta}^*$ solving equations (4.5), (4.6) and (4.7) as $\hat{\theta}_{EM}^*$.

When the correlation within clusters is present, we show in Theorem 4.1 that

the above estimating equations are unbiased if the marginals are correctly specified. However, the estimates may not be efficient (Peng et al. 2007; Yu and Peng, 2008). To increase the estimation efficiency of the method above, we follow the idea of the generalized estimating equations for the generalized linear models (Liang and Zeger, 1986; Rosen et al., 2000) and propose to replace the identity matrix I_i in equations (4.5) and (4.6) with working correlation matrices to account for the potential correlations between cure statuses and between the failure times of uncured subjects in each cluster. That is, the proposed estimating equations for γ and β are

$$\begin{aligned} U_{\gamma_{ES}} &= \sum_{i=1}^{K}(U_{\gamma_{ES}})_i = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(U_{\gamma_{ES}})_{ij} \\ &= \sum_{i=1}^{K}\{\frac{\partial \pi(Z_i)}{\partial \gamma}\}^T\{A_i^{1/2}Q_i(\rho_1)A_i^{1/2}\phi_1\}^{-1}\{y_i - \pi(Z_i)\} = 0, \end{aligned} \tag{4.9}$$

$$\begin{aligned} U_{\beta_{ES}} &= \sum_{i=1}^{K}(U_{\beta_{ES}})_i = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(U_{\beta_{ES}})_{ij} \\ &= \sum_{i=1}^{K}\{\frac{\partial \mu(X_i)}{\partial \beta}\}^T\{B_i^{1/2}Q_i(\rho_2)B_i^{1/2}\phi_2\}^{-1}W_i\{\kappa_i - \mu(X_i)\} = 0, \end{aligned} \tag{4.10}$$

where $Q_i(\rho_1) = (q_{jk}(\rho_1))_{n_i \times n_i}$ and $Q_i(\rho_2) = (q_{jk}(\rho_2))_{n_i \times n_i}$ are the working correlation matrices, and ρ_1 and ρ_2 are unknown parameters in the matrices that need to be estimated. The scale parameters ϕ_1 and ϕ_2 are incorporated in the estimating equations to accommodate potential over- or under-dispersion.

In this chapter, we consider the exchangeable correlation structure for both $Q_i(\rho_1)$ and $Q_i(\rho_2)$ as it is often used for clustered data. Such a correlation structure was also considered by Segal and Neuhaus (1993) and Chatterjee and Shih (2001). Following the moment method in Liang and Zeger (1986), ρ_1 and ρ_2 can be estimated from the

standardized Pearson residuals. That is

$$\hat{\phi}_1 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\{\hat{r}_{ij}^{(1)}\}^2/(N - p_Z - 1), \tag{4.11}$$

$$\hat{\rho}_1 = \hat{\phi}_1^{-1}\sum_{i=1}^{K}\sum_{j>j'}\hat{r}_{ij}^{(1)}\hat{r}_{ij'}^{(1)}/\{\sum_{i=1}^{K}\frac{1}{2}n_i(n_i-1) - p_Z - 1\}, \tag{4.12}$$

where $\hat{r}_{ij}^{(1)} = \{g_{ij}^{(m)} - \pi(Z_{ij})\}/[\pi(z_{ij})\{1-\pi(Z_{ij})\}]^{\frac{1}{2}}$, and

$$\hat{\phi}_2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\{\hat{r}_{ij}^{(2)}\}^2/(N - p_X - 1), \tag{4.13}$$

$$\hat{\rho}_2 = \hat{\phi}_2^{-1}\sum_{i=1}^{K}\sum_{j>j'}\hat{r}_{ij}^{(2)}\hat{r}_{ij'}^{(2)}/\{\sum_{i=1}^{K}\frac{1}{2}n_i(n_i-1) - p_X - 1\}, \tag{4.14}$$

where $\hat{r}_{ij}^{(2)} = \{\kappa_{ij} - \mu(X_{ij})\}/\{\mu(X_{ij})\}^{\frac{1}{2}}$. Note that $g_{ij}^{(m)}$ is used to estimate ρ_1, which plays the role of y_{ij}.

We summarize the steps to obtain an estimate of θ^* in this modified EM algorithm as follows.

1. The E-step stays the same;

2. The M-step starts with an initial value of θ^*. We then obtain the estimates of ϕ_1, ρ_1, ϕ_2, and ρ_2 from equations (4.11), (4.12), (4.13), and (4.14), which in turn lead to a new estimate of θ^* from equations (4.7), (4.9) and (4.10) after substituting $g_{ij}^{(m)}$ for y_{ij} in the equations. This step is iterated until convergence to complete the M-step.

The E-step and the M-step are iterated until the EM algorithm converges to obtain $\hat{\theta}^*$. Following Rosen et al. (2000), we name this modified EM algorithm the ES algorithm and denote the estimator $\hat{\theta}^*$ solving equations (4.7), (4.9) and (4.10) as $\hat{\theta}^*_{ES}$. It is obvious that $\hat{\theta}^*_{ES} = \hat{\theta}^*_{EM}$ when the working correlation matrices are the identity

matrix and the scale parameters ϕ_1 and ϕ_2 are equal to 1. Due to the modifications in the M-step, some useful properties of the EM algorithm are not available in the ES algorithm. However, in Theorem 4.1, we can show that both $\hat{\theta}^*_{ES}$ and $\hat{\theta}^*_{EM}$ are consistent and asymptotically normal estimators.

4.3 Asymptotic Properties and Variance Estimation

4.3.1 Asymptotic Properties of $\hat{\theta}^*_{ES}$ and $\hat{\theta}^*_{EM}$

Theorem 4.1. *Let θ^*_0 be the true value of θ^*. Under some regularity conditions,*

(a) both $\hat{\theta}^*_{ES}$ and $\hat{\theta}^*_{EM}$ are consistent estimators of θ^*_0,

(b) $K^{\frac{1}{2}}(\hat{\theta}^*_{ES} - \theta^*_0) \to N(0, \Sigma_{ES})$ and $K^{\frac{1}{2}}(\hat{\theta}^*_{EM} - \theta^*_0) \to N(0, \Sigma_{EM})$ (in distribution) as $K \to \infty$, where

$$\Sigma_{ES} = \mathcal{A}_1^{-1}(\theta^*_0)\mathcal{V}_1(\theta^*_0)\{\mathcal{A}_1^{-1}(\theta^*_0)\}^T, \quad \Sigma_{EM} = \mathcal{A}_2^{-1}(\theta^*_0)\mathcal{V}_2(\theta^*_0)\{\mathcal{A}_2^{-1}(\theta^*_0)\}^T,$$

and $\mathcal{A}_1(\theta^*_0) = E\{\mathcal{B}_1(\theta^*_0)\} - E\{\mathcal{S}_1(\theta^*_0)\mathcal{S}_1^T(\theta^*_0)\}$, $\mathcal{V}_1(\theta^*_0) = \sum_{i=1}^K E\{\mathcal{S}_{1i}(\theta^*_0)\}E\{\mathcal{S}_{1i}^T(\theta^*_0)\}$, $\mathcal{A}_2(\theta^*_0) = E\{\mathcal{B}_2(\theta^*_0)\} - E\{\mathcal{S}_2(\theta^*_0)\mathcal{S}_2^T(\theta^*_0)\}$, $\mathcal{V}_2(\theta^*_0) = \sum_{i=1}^K E\{\mathcal{S}_{2i}(\theta^*_0)\}E\{\mathcal{S}_{2i}^T(\theta^*_0)\}$, $\mathcal{S}_1(\theta^*) = (U_{\gamma_{ES}}, U_{\beta_{ES}}, U_{\alpha_{EM}})$, $\mathcal{B}_1(\theta^*) = -\frac{\partial \mathcal{S}_1(\theta^*)}{\partial \theta^*}$ and $\mathcal{S}_{1i}(\theta^*) = (U_{\gamma_{ES(i)}}, U_{\beta_{ES(i)}}, U_{\alpha_{EM(i)}})$, $\mathcal{S}_2(\theta^*) = (U_{\gamma_{EM}}, U_{\beta_{EM}}, U_{\alpha_{EM}})$, $\mathcal{B}_2(\theta^*) = -\frac{\partial \mathcal{S}_2(\theta^*)}{\partial \theta^*}$ and $\mathcal{S}_{2i}(\theta^*) = (U_{\gamma_{EM(i)}}, U_{\beta_{EM(i)}}, U_{\alpha_{EM(i)}})$,

(c) Σ_{ES} and Σ_{EM} can be consistently estimated by $\hat{\Sigma}_{ES} = \mathcal{A}_1^{-1}(\hat{\theta}^*_{ES})\mathcal{V}_1(\hat{\theta}^*_{ES})\{\mathcal{A}_1^{-1}(\hat{\theta}^*_{ES})\}^T$ and $\hat{\Sigma}_{EM} = \mathcal{A}_2^{-1}(\hat{\theta}^*_{EM})\mathcal{V}_2(\hat{\theta}^*_{EM})\{\mathcal{A}_2^{-1}(\hat{\theta}^*_{EM})\}^T$ respectively.

Proof. Given the regularity conditions (Huber, 1967), we adapt the proof in Rosen et al. (2000) for censored data. Let (S_t, F_t, ω), (S_c, F_c, μ) and (S_y, F_y, ν) be σ-finite measure spaces, with a product measure space $(S_t \otimes S_c \otimes S_y, F_t \otimes F_c \otimes F_y, \omega \otimes \mu \otimes \nu)$,

where $S_t \subset R^{d_t}$, $S_c \subset R^{d_c}$ and $S_y \subset R^{d_y}$. We assume a marginal probability model $p_{ij}(t, c, y|\theta^*)$ for $(t^*_{ij}, c_{ij}, y_{ij}) \in S_t \otimes S_c \otimes S_y$, which are strictly positive on $S_t \otimes S_c \otimes S_y$ and may depend on subscripts i and j via covariates, for each $j = 1, \cdots, n_i, i = 1, \cdots, K$, associated with the product measure $\omega \otimes \mu \otimes \nu$. Here θ^* is some vector-valued parameter in a subset Θ of $R^{d_{\theta^*}}$ with $d_{\theta^*} = \dim(\theta^*)$, and $p_{ij}(t, c, y|\cdot)$ is continuously differentiable on Θ for each $(t^*_{ij}, c_{ij}, y_{ij}) \in S_t \otimes S_c \otimes S_y$. Let $p_{ij}(y|t^*_{ij}, c_{ij}, \theta^*)$ be the conditional probability model for all $y_{ij} \in S_y$. That is

$$p_{ij}(y|t^*_{ij}, c_{ij}, \theta^*) = p_{ij}(t^*_{ij}, c_{ij}, y|\theta^*) / \int_{S_y} p_{ij}(t^*_{ij}, c_{ij}, u|\theta^*) d\nu(u).$$

Let $q_{ij}(\cdot, \cdot, \cdot; \cdot)$ be a $d_{\theta^*} \times 1$ vector-valued function composed by $\{(U_{\gamma_{ES}})_{ij}, (U_{\beta_{ES}})_{ij}, (U_{\alpha_{EM}})_{ij}\}^T$. It is defined on $S_t \otimes S_c \otimes S_y \otimes \Theta \mapsto R^{d_{\theta^*}}$ such that $q_{ij}(\cdot, \cdot, \cdot; \varphi) : S_t \otimes S_c \otimes S_y \mapsto R^{d_{\theta^*}}$ is measurable and integrable with respect to $p_{ij}(\cdot, \cdot, \cdot|\theta^*)$ for each $\varphi \in \Theta$, and $q_{ij}(t, c, y; \cdot) : \Theta \mapsto R^{d_{\theta^*}}$ is continuously differentiable on Θ for each $(t^*_{ij}, c_{ij}, y_{ij}) \in S_t \otimes S_c \otimes S_y$. We then define a bivariate function $H(\cdot|\cdot) : \Theta \otimes \Theta \mapsto R^{d_{\theta^*}}$ by

$$H(\varphi|\theta^{*(m)}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \int_{S_y} q_{ij}(t^*_{ij}, c_{ij}, y|\varphi) p_{ij}(y|t^*_{ij}, c_{ij}, \theta^{*(m)}).$$

The E-step of the ES algorithm computes $H(\varphi|\theta^{*(m)})$, and the S-step solves for $\varphi = \theta^{*(m+1)}$ from the equation $H(\varphi|\theta^{*(m)}) = 0$. Furthermore, we require the following proposition given by Rosen et al. (2000).

Proposition 4.1. *Assuming that the following conditions hold:*

(a) $H(\cdot|\cdot)$ *is a bivariate continuous function on* $\Theta \otimes \Theta$, *where* $\Theta \subseteq R^{d_{\theta^*}}$, *and*

(b) $q_{ij}(\cdot, \cdot, \cdot; \cdot)$ *is an unbiased estimating function satisfying* $E\{q_{ij}(t^*_{ij}, c_{ij}, y_{ij}; \theta^*)|\theta^*\} = 0$ *for all* $\theta^* \in \Theta$ *and all* $j = 1, \cdots, n_i$ *and* $i = 1, \cdots, K$.

If there exists a point $\hat{\theta}^* \in \Theta$ *such that* $\lim_{m\to\infty} \theta^{*(m)} = \hat{\theta}^*$*, where* $\theta^{*(m)}$ *is a sequence generated by the Expectation-Solution algorithm for* $m = 0, 1, 2, \cdots$*, then*

(i) $\hat{\theta}^*$ *satisfies the estimating equation* $\Psi(\hat{\theta}^*) = H(\hat{\theta}^*|\hat{\theta}^*) = 0$*;*

(ii) $\Psi(\theta^*) = \sum_{i=1}^{K}\sum_{j=1}^{n_i} q_{ij} = 0$ *is an unbiased estimating equation, satisfying* $E\{\Psi(\theta^*)|\theta^*\} = 0$ *for each* $\theta^* \in \Theta$*.*

We now show that $H(\cdot|\cdot)$ and $q_{ij}(\cdot,\cdot,\cdot;\cdot)$ satisfy conditions (a) and (b). Condition (a) holds since

$$\begin{aligned} H(\varphi|\theta^{*(m)}) &= \sum_{i=1}^{K}\sum_{j=1}^{n_i} E\{q_{ij}(t^*_{ij}, c_{ij}, y_{ij};\varphi)|t^*_{ij}, c_{ij}, \theta^{*(m)}\} \\ &= \sum_{i=1}^{K}\sum_{j=1}^{n_i}\sum_{y_{ij}} q_{ij}(t^*_{ij}, c_{ij}, y_{ij};\varphi) p_{ij}(y_{ij}|t^*_{ij}, c_{ij}, \theta^{*(m)}), \end{aligned}$$

where

$$p_{ij}(y_{ij}=1|t^*_{ij}, c_{ij}, \theta^{*(m)}) = \left\{\delta_{ij} + (1-\delta_{ij})\frac{\pi(Z_{ij})S_{u0}(t^*_{ij})^{\exp(\beta' X_{ij})}}{1-\pi(Z_{ij})+\pi(Z_{ij})S_{u0}(t^*_{ij})^{\exp(\beta' X_{ij})}}\right\}_{\theta=\theta^{*(m)}},$$

$$p_{ij}(y_{ij}=0|t^*_{ij}, c_{ij}, \theta^{*(m)}) = \left\{(1-\delta_{ij})\frac{1-\pi(Z_{ij})}{1-\pi(Z_{ij})+\pi(Z_{ij})S_{u0}(t^*_{ij})^{\exp(\beta' X_{ij})}}\right\}_{\theta=\theta^{*(m)}},$$

and the $q_{ij}(t,c,y;\cdot)$'s and $p_{ij}(t,c,y|\cdot)$'s are continuous functions for each $(t^*_{ij}, c_{ij}, y_{ij}) \in S_t \otimes S_c \otimes S_y$.

To prove condition (b) holds, i.e.,

$$E\{q_{ij}(t^*_{ij}, c_{ij}, y_{ij};\theta^*)|\theta^*\} = \int\int \sum_y p_{ij}(y|t,c,\theta^*) q_{ij}(t,c,y;\theta^*) dF_{ij}(t|c,\theta^*) dF_{ij}(c|\theta^*) = 0,$$

we first investigate the unbiasedness of the components of $q_{ij}(t^*_{ij}, c_{ij}, y_{ij};\theta^*)$ corresponding to γ.

Let $\pi_{ij} = \pi(Z_{ij})$. Then

$$
\begin{aligned}
& E\{(U_{\gamma_{ES}})_{ij}|\theta^*\} = E[E\{(U_{\gamma_{ES}})_{ij}|c,\theta^*\}|\theta^*] \\
= & \int\int \sum_y p_{ij}(y|t,c,\theta^*) \sum_{l=1}^{n_i} \phi_1^{-1} Z_{il}\{\pi_{il}(1-\pi_{il})\}^{1/2} \tilde{Q}_i(\rho_1)_{lj} \\
& \times\{\pi_{ij}(1-\pi_{ij})\}^{-1/2}(y_{ij}-\pi_{ij}) dF_{ij}(t|c,\theta^*) dF_{ij}(c|\theta^*) \\
= & \phi_1^{-1} \sum_{l=1}^{n_i} Z_{il} \int\int \sum_y p_{ij}(y|t,c,\theta^*)\{\pi_{il}(1-\pi_{il})\}^{1/2} \tilde{Q}_i(\rho_1)_{lj} \\
& \times\{\pi_{ij}(1-\pi_{ij})\}^{-1/2}(y_{ij}-\pi_{ij}) dF_{ij}(t|c,\theta^*) dF_{ij}(c|\theta^*) \\
= & \phi_1^{-1} \sum_{l=1}^{n_i} Z_{il}\{\pi_{il}(1-\pi_{il})\}^{1/2} \tilde{Q}_i(\rho_1)_{lj}\{\pi_{ij}(1-\pi_{ij})\}^{-1/2} \\
& \times\{\int\int \sum_y y p_{ij}(y|t,c,\theta^*) dF_{ij}(t|c,\theta^*) dF_{ij}(c|\theta^*) - \pi_{ij}\} \\
= & \phi_1^{-1} \sum_{l=1}^{n_i} Z_{il}\{\pi_{il}(1-\pi_{il})\}^{1/2} \tilde{Q}_i(\rho_1)_{lj}\{\pi_{ij}(1-\pi_{ij})\}^{-1/2}\{\pi_{ij}-\pi_{ij}\} = \mathbf{0},
\end{aligned}
$$

where $\tilde{Q}_i(\rho_1) = Q_i^{-1}(\rho_1)$ and $Q_i(\rho_1)$ is an $n_i \times n_i$ working correlation matrix for $i = 1, \cdots, K$.

To prove the unbiasedness of the components of $q_{ij}(t_{ij}^*, c_{ij}, y_{ij}; \theta^*)$ corresponding to β, let $\mu_{ij} = \mu(X_{ij})$. We first look at the unbiasedness of $\{(U_{\beta_{ES}})_{ij}|y_{ij}, c_{ij}, \theta^*\}$

$$
\begin{aligned}
& E\{(U_{\beta_{ES}})_{ij}|y_{ij}, c_{ij}, \theta^*\} \\
= & \int \sum_{l=1}^{n_i} \phi_2^{-1} X_{il} \mu_{il}^{1/2} \tilde{Q}_i(\rho_2)_{lj} \mu_{ij}^{-1/2} y_{ij} t^{\alpha}(\kappa_{ij} - \mu_{ij}) dF_{ij}(t|y_{ij}, c_{ij}, \theta^*) \\
= & \{\phi_2^{-1} \sum_{l=1}^{n_i} X_{il} \mu_{il}^{1/2} \tilde{Q}_i(\rho_2)_{lj} \mu_{ij}^{-1/2}\} \int y_{ij} t^{\alpha}(\kappa_{ij} - \mu_{ij}) dF_{ij}(t|y_{ij}, c_{ij}, \theta^*) \\
= & M_{ij} \int X_{ij} y_{ij} t^{\alpha}(\kappa_{ij} - \mu_{ij}) dF_{ij}(t|y_{ij}, c_{ij}, \theta^*) \\
= & M_{ij} \int \frac{\partial \mu_{ij}}{\partial \beta} \frac{1}{\mu_{ij}} y_{ij}(\delta_{ij} - t^{\alpha}\mu_{ij}) dF_{ij}(t|y_{ij}, c_{ij}, \theta^*) \\
= & M_{ij} \int \frac{\partial}{\partial \beta} \log[\{\mu_{ij}^{\delta_{ij}} \exp(-t^{\alpha}\mu_{ij})\}^{y_{ij}} (t^{\alpha-1}\alpha)^{\delta_{ij}}] dF_{ij}(t|y_{ij}, c_{ij}, \theta^*)
\end{aligned}
$$

$$
\begin{aligned}
&= M_{ij}\int \frac{\partial}{\partial\beta}\log[\{(t^{\alpha}\mu_{ij})^{\delta_{ij}}\exp(-t^{\alpha}\mu_{ij})\}^{y_{ij}}(\frac{\alpha t^{\alpha-1}\mu_{ij}}{t^{\alpha}\mu_{ij}})^{\delta_{ij}}]dF_{ij}(t|y_{ij},c_{ij},\theta^*)\\
&= M_{ij}\int \frac{\partial}{\partial\beta}\log([\lambda_u^{\delta_{ij}}(t;X_{ij})\exp\{-\Lambda_u(t;X_{ij})\}]^{y_{ij}})dF_{ij}(t|y_{ij},c_{ij},\theta^*)\\
&= M_{ij}y_{ij}\int \frac{\partial}{\partial\beta}\log[\lambda_u^{\delta_{ij}}(t;X_{ij})\exp\{-\Lambda_u(t;X_{ij})\}]dF_{ij}(t|y_{ij},c_{ij},\theta^*),
\end{aligned}
$$

where $M_{ij}=\{\phi_2^{-1}\sum_{l=1}^{n_i}X_{il}\mu_{il}^{1/2}\tilde{Q}_i(\rho_2)_{lj}\mu_{ij}^{-1/2}/X_{ij}\}$ and $\tilde{Q}_i(\rho_2)=Q_i^{-1}(\rho_2)$ and $Q_i(\rho_2)$ is an $n_i\times n_i$ working correlation matrix for $i=1,\cdots,K$. Therefore

$$
\begin{aligned}
&E\{(U_{\beta_{ES}})_{ij}|c_{ij},\theta^*\}\\
=\;&\sum_y p_{ij}(y|c_{ij},\theta^*)E\{(U_{\beta_{ES}})_{ij}|y,c_{ij},\theta^*\}\\
=\;&\sum_y p_{ij}(y|c_{ij},\theta^*)M_{ij}y\int \frac{\partial}{\partial\beta}\log[\lambda_u^{\delta_{ij}}(t|c_{ij},\theta^*)\exp\{-\Lambda_u(t|c_{ij},\theta^*)\}]dF_{ij}(t|y,c_{ij},\theta^*)\\
=\;&\pi_{ij}M_{ij}\int \frac{\partial}{\partial\beta}[\log\{f_u^{\delta_{ij}}(t|c_{ij},\theta^*)S_u^{1-\delta_{ij}}(t|c_{ij},\theta^*)\}]dF_u(t|c_{ij},\theta^*)\\
=\;&\pi_{ij}M_{ij}\{\int \delta_{ij}\frac{\partial}{\partial\beta}\log f_u(t|c_{ij},\theta^*)dF_u(t|c_{ij},\theta^*)\\
&+\int(1-\delta_{ij})\frac{\partial}{\partial\beta}\log S_u(t|c_{ij},\theta^*)dF_u(t|c_{ij},\theta^*)\}\\
=\;&\pi_{ij}M_{ij}\{\int_0^{c_{ij}}\frac{1}{f_u(t|c_{ij},\theta^*)}\frac{\partial f_u(t|c_{ij},\theta^*)}{\partial\beta}dF_u(t|c_{ij},\theta^*)\\
&+\int_{c_{ij}}^{\infty}\frac{1}{S_u(t|c_{ij},\theta^*)}\frac{\partial S_u(t|c_{ij},\theta^*)}{\partial\beta}dF_u(t|c_{ij},\theta^*)\}\\
=\;&\pi_{ij}M_{ij}\{\int_0^{c_{ij}}\frac{\partial f_u(t|c_{ij},\theta^*)}{\partial\beta}dt+\int_{c_{ij}}^{\infty}\frac{\partial S_u(t|c_{ij},\theta^*)}{\partial\beta}d\Lambda_u(t|c_{ij},\theta^*)\}\\
=\;&\pi_{ij}M_{ij}\frac{\partial}{\partial\beta}\{\int_0^{c_{ij}}f_u(t|c_{ij},\theta^*)dt+\int_{c_{ij}}^{\infty}\lambda_u(t|c_{ij},\theta^*)S_u(t|c_{ij},\theta^*)dt\}\\
=\;&\pi_{ij}M_{ij}\frac{\partial}{\partial\beta}\{F_u(c_{ij}|c_{ij},\theta^*)+S_u(c_{ij}|c_{ij},\theta^*)\}=\pi_{ij}M_{ij}\frac{\partial}{\partial\beta}(1)=\mathbf{0}.
\end{aligned}
$$

Since the component of $q_{ij}(t_{ij}^*,c_{ij},y_{ij};\theta^*)$ corresponding to α is a score function based on the complete log likelihood function $l_{\mathbf{c}}$, following the same idea used in the

unbiasedness of $\{(U_{\beta_{ES}})_{ij}|c_{ij},\theta^*\}$, we can show that

$$E\{(U_{\alpha_{EM}})_{ij}|c_{ij},\theta^*\}=\sum_{y}p_{ij}(y|c_{ij},\theta)E\{(U_{\alpha_{EM}})_{ij}|y,c_{ij},\theta^*\}=0,$$

where $U_{\alpha_{EM}}=\sum_{i=1}^{K}\sum_{j=1}^{n_i}(U_{\alpha_{EM}})_{ij}$. Therefore, we have

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}E\{q_{ij}(t_{ij}^*,c_{ij},y_{ij};\theta^*)|\theta^*\}=E[E\{q_{ij}(t_{ij}^*,c_{ij},y_{ij};\theta^*)|c_{ij},\theta^*\}|\theta^*]=\mathbf{0},$$

and this completes the verification of condition (b) of Proposition 4.1. That is, we prove that if the ES algorithm converges, $\hat{\theta}_{ES}^*$ is a solution of unbiased estimating equations $\Psi(\theta^*)$ of θ^*. Therefore, based on the regularity conditions, the asymptotic normality of $\hat{\theta}_{ES}^*$ and its covariance estimation follow from the result of Huber (1967) as the number of clusters $K\to\infty$.

The consistency and asymptotic normality of $\hat{\theta}_{EM}^*$ can be established similarly after replacing the working correlation structures with the identity matrix and replacing ϕ_1 and ϕ_2 with 1 in the estimating functions (4.9) and (4.10), which reduce to (4.5) and (4.6).

4.3.2 Variance Estimation for $\hat{\theta}_{ES}^*$, $\hat{\theta}_{EM}^*$, $\hat{\rho}_1$ and $\hat{\rho}_2$

The variance and standard error of $\hat{\theta}_{ES}^*$ can be obtained based on Theorem 4.1 (c). Obtaining $\mathcal{V}_1(\hat{\theta}_{ES}^*)$ is straightforward. We need to find $E\{\mathcal{S}_{1i}(\hat{\theta}_{ES}^*)\}=\{E(U_{\gamma_{ES}(i)}),$ $E(U_{\beta_{ES}(i)}),$ $E(U_{\alpha_{EM}(i)})\}$ for $i=1,\cdots,K$ and calculate $\mathcal{V}_1(\theta^*)=\sum_{i=1}^{K}E\{\mathcal{S}_{1i}(\hat{\theta}_{ES}^*)\}$ $E\{\mathcal{S}_{1i}^T(\hat{\theta}_{ES}^*)\}$. Here we provide some details for calculating $\mathcal{A}_1(\theta^*)$. To simplify the notations, we let $(U_\gamma,U_\beta,U_\alpha)=(U_{\gamma_{ES}},U_{\beta_{ES}},U_{\alpha_{EM}})$. The first term in $\mathcal{A}_1(\hat{\theta}_{ES}^*)$,

$E(\mathcal{B}_1(\hat{\theta}^*_{ES}))$, can be written as

$$E(\mathcal{B}_1(\hat{\theta}^*_{ES})) = -E\begin{pmatrix} U_{\gamma\gamma} & U_{\gamma\beta} & U_{\gamma\alpha} \\ U_{\beta\gamma} & U_{\beta\beta} & U_{\beta\alpha} \\ U_{\alpha\gamma} & U_{\alpha\beta} & U_{\alpha\alpha} \end{pmatrix} = -E\begin{pmatrix} U_{\gamma\gamma} & 0 & 0 \\ 0 & U_{\beta\beta} & U_{\beta\alpha} \\ 0 & U_{\alpha\beta} & U_{\alpha\alpha} \end{pmatrix}.$$

The (ν, ω) $\{\omega, \nu = 1, 2, \cdots, \dim(\gamma)\}$ element in the first $\dim(\gamma)\times\dim(\gamma)$ matrix $U_{\gamma\gamma}$ is

$$\sum_{i=1}^{K} A_i^{(\gamma)}(B_i^{(\gamma)}C_i^{(\gamma)} - D_i^{(\gamma)}E_i^{(\gamma)}),$$

where $A_i^{(\gamma)} = (Z_{i1\nu}, Z_{i2\nu}, \cdots, Z_{in_i\nu})_{1\times n_i}$,

$$B_i^{(\gamma)} = \begin{pmatrix} B_{i11}^{(\gamma)} & \cdots & B_{i1n_i}^{(\gamma)} \\ \vdots & \ddots & \vdots \\ B_{in_i1}^{(\gamma)} & \cdots & B_{in_in_i}^{(\gamma)} \end{pmatrix}_{n_i\times n_i}$$

with $B_{imn}^{(\gamma)} = \frac{1}{2\phi_1}\{Z_{im\omega}(1-2\pi_{im}) - Z_{in\omega}(1-2\pi_{in})\}\{\pi_{im}(1-\pi_{im})\}^{1/2}\{\pi_{in}(1-\pi_{in})\}^{-1/2}$ $\tilde{Q}_i(\rho_1)_{mn}$ for $m \neq n$, otherwise $B_{imm}^{(\gamma)} = 0$, $m, n = 1, 2, \cdots, n_i$. $C_i^{(\gamma)} = (y_{i1} - \pi_{i1}, y_{i2} - \pi_{i2}, \cdots, y_{in_i} - \pi_{in_i})^T$,

$$D_i^{(\gamma)} = \begin{pmatrix} D_{i11}^{(\gamma)} & \cdots & D_{i1n_i}^{(\gamma)} \\ \vdots & \ddots & \vdots \\ D_{in_i1}^{(\gamma)} & \cdots & D_{in_in_i}^{(\gamma)} \end{pmatrix}_{n_i\times n_i}$$

with $D_{imn}^{(\gamma)} = \phi_1^{-1}\{\pi_{im}(1-\pi_{im})\}^{1/2}\{\pi_{in}(1-\pi_{in})\}^{-1/2}\tilde{Q}_i(\rho_1)_{mn}$ for $m \neq n$, otherwise $D_{imm}^{(\gamma)} = \tilde{Q}_i(\rho_1)_{mm}$. $E_i^{(\gamma)} = \{Z_{i1\omega}\pi_{i1}(1-\pi_{i1}), Z_{i2\omega}\pi_{i2}(1-\pi_{i2}), \cdots, Z_{in_i\omega}\pi_{in_i}(1-\pi_{in_i})\}_{1\times n_i}^T$.

The (ν,ω) $\{\omega,\nu = 1,2,\cdots,\dim(\beta)\}$ element in the second block diagonal $\dim(\beta) \times \dim(\beta)$ matrix $U_{\beta\beta}$ is

$$\sum_{i=1}^{K} A_i^{(\beta)}(B_i^{(\beta)}W_iC_i^{(\beta)} - D_i^{(\beta)}W_iE_i^{(\beta)}),$$

where $A_i^{(\beta)} = (X_{i1\nu}, X_{i2\nu}, \cdots, X_{in_i\nu})_{1\times n_i}$,

$$B_i^{(\beta)} = \begin{pmatrix} B_{i11}^{(\beta)} & \cdots & B_{i1n_i}^{(\beta)} \\ \vdots & \ddots & \vdots \\ B_{in_i1}^{(\beta)} & \cdots & B_{in_in_i}^{(\beta)} \end{pmatrix}_{n_i\times n_i}$$

with $B_{imn}^{(\beta)} = \frac{1}{2\phi_2}(X_{im\omega} - X_{in\omega})(\mu_{im})^{1/2}(\mu_{in})^{-1/2}\tilde{Q}_i(\rho_2)_{mn}$ for $m \neq n$, otherwise $B_{imm}^{(\beta)} = 0$, $m,n = 1,2,\cdots,n_i$. $C_i^{(\beta)} = (\kappa_{i1}-\mu_{i1}, \kappa_{i2}-\mu_{i2}, \cdots, \kappa_{in_i}-\mu_{in_i})^T$,

$$D_i^{(\beta)} = \begin{pmatrix} D_{i11}^{(\beta)} & \cdots & D_{i1n_i}^{(\beta)} \\ \vdots & \ddots & \vdots \\ D_{in_i1}^{(\beta)} & \cdots & D_{in_in_i}^{(\beta)} \end{pmatrix}_{n_i\times n_i}$$

with $D_{imn}^{(\beta)} = \phi_2^{-1}(\mu_{im})^{1/2}(\mu_{in})^{-1/2}\tilde{Q}_i(\rho_2)_{mn}$ for $m \neq n$, otherwise $D_{imm}^{(\beta)} = \tilde{Q}_i(\rho_2)_{mm}$. $E_i^{(\beta)} = (X_{i1\omega}\mu_{i1}, X_{i2\omega}\mu_{i2}, \cdots, X_{in_i\omega}\mu_{in_i})^T$, and $W_i = \mathrm{diag}(y_{i1}t_{i1}^{*\alpha}, y_{i2}t_{i2}^{*\alpha}\cdots, y_{in_i}t_{in_i}^{*\alpha})$.

The last diagonal element $U_{\alpha\alpha}$ is

$$-\sum_{i=1}^{K}\sum_{j=1}^{n_i}[y_{ij}\{\log(t_{ij})\}^2 t_{ij}^{\alpha}\mu_{ij} + \delta_{ij}/\alpha^2].$$

The off-diagonal element $U_{\beta\alpha}$, a matrix of order $\dim(\beta) \times 1$, is

$$-\sum_{i=1}^{K} A_i^{(\beta)} D_i^{(\beta)} W_i F_i^{(\beta)},$$

where $F_i^{(\beta)} = \{\mu_{i1}\log(t_{i1}^*), \mu_{i2}\log(t_{i2}^*), \cdots, \mu_{in_i}\log(t_{in_i}^*)\}^T$.

The off-diagonal $1\times\dim(\beta)$ matrix $U_{\alpha\beta}$ has elements

$$-\sum_{i=1}^{K} A_i^{(\beta)} W_i F_i^{(\beta)}.$$

Next we compute the second term of $\mathcal{A}_1(\hat{\theta}_{ES}^*)$, i.e.,

$$E\{\mathcal{S}_1(\hat{\theta}_{ES}^*)\mathcal{S}_1^T(\hat{\theta}_{ES}^*)\} = E\begin{pmatrix} U_\gamma U_\gamma & U_\gamma U_\beta & U_\gamma U_\alpha \\ U_\beta U_\gamma & U_\beta U_\beta & U_\beta U_\alpha \\ U_\alpha U_\gamma & U_\alpha U_\beta & U_\alpha U_\alpha \end{pmatrix},$$

where

$$\begin{aligned} E(U_\gamma U_\gamma) &= E\{\sum_{i=1}^{K} R_{1i}^T V_{1i}^{-1}(y_i - g_i)\}^2 \\ &= E\{\sum_{i=1}^{K} R_{1i}^T V_{1i}^{-1}(y_i - g_i)(y_i - g_i)^T V_{1i}^{-T} R_{1i}\} \\ &= \sum_{i=1}^{K} R_{1i}^T V_{1i}^{-1} E\{(y_i - g_i)(y_i - g_i)^T\} V_{1i}^{-T} R_{1i}, \end{aligned}$$

$$
\begin{aligned}
E(U_\beta U_\beta) &= E\{\sum_{i=1}^{K} R_{2i}^T V_{2i}^{-1}(y_i - g_i)(\delta_i - t_i^{*\alpha}\mu_i)\}^2 \\
&= E\{\sum_{i=1}^{K} R_{2i}^T V_{2i}^{-1}(y_i - g_i)(\delta_i - t_i^{*\alpha}\mu_i)(\delta_i - t_i^{*\alpha}\mu_i)^T(y_i - g_i)^T V_{2i}^{-T} R_{2i}\} \\
&= \sum_{i=1}^{K} R_{2i}^T V_{2i}^{-1} E\{(y_i - g_i)(\delta_i - t_i^{*\alpha}\mu_i)(\delta_i - t_i^{*\alpha}\mu_i)^T(y_i - g_i)^T\} V_{2i}^{-T} R_{2i},
\end{aligned}
$$

$$
\begin{aligned}
E(U_\alpha U_\alpha) &= E[\sum_{i=1}^{K} \{\log(t_i^*)(\delta_i - t_i^{*\alpha}\mu_i)\}^T (y_i - g_i)]^2 \\
&= E[\sum_{i=1}^{K} \{\log(t_i^*)(\delta_i - t_i^{*\alpha}\mu_i)\}^T (y_i - g_i)(y_i - g_i)^T \log(t_i^*)(\delta_i - t_i^{*\alpha}\mu_i)] \\
&= \sum_{i=1}^{K} \{\log(t_i^*)(\delta_i - t_i^{*\alpha}\mu_i)\}^T E\{(y_i - g_i)(y_i - g_i)^T\} \log(t_i^*)(\delta_i - t_i^{*\alpha}\mu_i),
\end{aligned}
$$

$$
\begin{aligned}
E(U_\gamma U_\beta) &= E\{\sum_{i=1}^{K} R_{1i}^T V_{1i}^{-1}(y_i - g_i) \sum_{i=1}^{K} R_{2i}^T V_{2i}^{-1}(y_i - g_i)(\delta_i - t_i^{*\alpha}\mu_i)\} \\
&= E\{\sum_{i=1}^{K} R_{1i}^T V_{1i}^{-1}(y_i - g_i)(\delta_i - t_i^{*\alpha}\mu_i)^T(y_i - g_i)^T V_{2i}^{-T} R_{2i}\} \\
&= \sum_{i=1}^{K} R_{1i}^T V_{1i}^{-1} E\{(y_i - g_i)(\delta_i - t_i^{*\alpha}\mu_i)^T(y_i - g_i)^T\} V_{2i}^{-T} R_{2i},
\end{aligned}
$$

$$
\begin{aligned}
E(U_\gamma U_\alpha) &= E[\sum_{i=1}^{K} R_{1i}^T V_{1i}^{-1}(y_i - g_i) \sum_{i=1}^{K} \{\log(t_i^*)(\delta_i - t_i^{*\alpha}\mu_i)\}^T (y_i - g_i)] \\
&= E\{\sum_{i=1}^{K} R_{1i}^T V_{1i}^{-1}(y_i - g_i)(y_i - g_i)^T \log(t_i^*)(\delta_i - t_i^{*\alpha}\mu_i)\} \\
&= \sum_{i=1}^{K} R_{1i}^T V_{1i}^{-1} E\{(y_i - g_i)(y_i - g_i)^T\} \log(t_i^*)(\delta_i - t_i^{*\alpha}\mu_i),
\end{aligned}
$$

$$
\begin{aligned}
E(U_\beta U_\alpha) &= E[\sum_{i=1}^{K} R_{2i}^T V_{2i}^{-1}(y_i - g_i)(\delta_i - t_i^{*\alpha}\mu_i) \sum_{i=1}^{K} \{\log(t_i^*)(\delta_i - t_i^{*\alpha}\mu_i)\}^T (y_i - g_i)] \\
&= E[\sum_{i=1}^{K} R_{2i}^T V_{2i}^{-1}(y_i - g_i)(\delta_i - t_i^{*\alpha}\mu_i)(y_i - g_i)^T \{\log(t_i^*)(\delta_i - t_i^{*\alpha}\mu_i)\}] \\
&= \sum_{i=1}^{K} R_{2i}^T V_{2i}^{-1} E\{(y_i - g_i)(\delta_i - t_i^{*\alpha}\mu_i)(y_i - g_i)^T\}\{\log(t_i^*)(\delta_i - t_i^{*\alpha}\mu_i)\}.
\end{aligned}
$$

Here $R_{1i} = \frac{\partial \pi(Z_i)}{\partial \gamma}, V_{1i} = A_i^{1/2} Q_i(\rho_1) A_i^{1/2} \phi_1, R_{2i} = \frac{\partial \mu(X_i)}{\partial \beta}$ and $V_{2i} = B_i^{1/2} Q_i(\rho_2) B_i^{1/2} \phi_2$, $g_i = (g_{i1}, \cdots, g_{in_i})^T$, $\delta_i = (\delta_{i1}, \cdots, \delta_{in_i})^T$, $t_i^{*\alpha} = (t_{i1}^{*\alpha}, \cdots, t_{in_i}^{*\alpha})^T$ and $\mu_i = (\mu_{i1}, \cdots, \mu_{in_i})^T$.

For $\hat{\theta}_{EM}^*$, the formulas for variance estimation are similar. Note that Yu and Peng (2008) also provided a jackknife variance estimate for $\hat{\theta}_{EM}^*$.

To obtain the variance estimates of the correlation coefficients $\hat{\rho}_1$ and $\hat{\rho}_2$, we consider a bootstrap method which is similar to the method we used for $\hat{\rho}_{New}$ in Section 3.3.2. That is, a bootstrap sample is obtained from sampling clusters with replacement and is fitted with the proposed method to obtain the parameter estimates $\hat{\rho}_{1b}$ and $\hat{\rho}_{2b}$, separately. Then the variance estimates of $\hat{\rho}_1$ and $\hat{\rho}_2$ can be estimated by $\sum_{i=1}^{B}(\hat{\rho}_{1b}^{(i)} - \sum_{j=1}^{B} \hat{\rho}_{1b}^{(j)}/B)^2/(B-1)$ and $\sum_{i=1}^{B}(\hat{\rho}_{2b}^{(i)} - \sum_{j=1}^{B} \hat{\rho}_{2b}^{(j)}/B)^2/(B-1)$, where B is the number of bootstrap samples for each simulated data set.

4.4 A Simulation Study

We conduct an extensive simulation study to investigate the performance of the proposed method and to compare its estimates with those from existing methods. The study considers various cluster sizes with different correlation settings for both discrete and continuous covariates.

The data in the simulation are generated from the marginal proportional hazards mixture cure model (4.1), (4.2) and (4.3) with a single covariate in both X and Z. The parameters in the model are set to $\theta^* = (\gamma_0, \gamma_1, \beta_0, \beta_1, \alpha) = (0.4, -1, \log(2), -1, 1)$.

The single covariate is either a binary covariate generated from the Bernoulli distribution with mean 0.5 or a continuous covariate generated from the standard normal distribution.

For each data set, we consider the following pairs of the number of clusters and the cluster size: (40,10), (80,5), and (200,2). For each cluster in a data set, correlated Y_{ij}'s and $\tilde{T}^*_{ij}$'s given $Y_{ij} = 1$ are generated. Specifically, to generate the Y_{ij} and $Y_{ij'}$ that satisfy $P(Y_{ij} = 1; Z_{ij}) = \pi(Z_{ij})$, $P(Y_{ij'} = 1; Z_{ij'}) = \pi(Z_{ij'})$ and $\text{corr}(Y_{ij}, Y_{ij'}) = \zeta$, we adopt the method proposed by Emrich and Piedmonte (1991). That is, given $\pi(Z_{ij})$, $\pi(Z_{ij'})$ and ζ, we solve for $\tilde{\rho}_{ijj'}$ through

$$\frac{\Phi[\{z_{\pi(Z_{ij})}, z_{\pi(Z_{ij'})}\}, \tilde{\rho}_{ijj'}] - \pi(Z_{ij})\pi(Z_{ij'})}{\pi(Z_{ij})\pi(Z_{ij'})\{1 - \pi(Z_{ij})\}\{1 - \pi(Z_{ij'})\}} = \zeta,$$

where $\Phi(\cdot, \tilde{\rho}_{ijj'})$ is the standard bivariate normal distribution function with the correlation coefficient equal to $\tilde{\rho}_{ijj'}$. After obtaining $\tilde{\rho}_{ijj'}$ for the ith cluster, we use them to form a correlation matrix $\mathbf{\Sigma_i}$ and then generate $(z_{i1}, \cdots, z_{in_i})$ from the multivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma_i})$. The correlated Y_{ij}'s in the cluster with specified correlation ζ can be obtained from $(z_{i1}, \cdots, z_{in_i})$ via $Y_{ij} = I\{z_{ij} < z_{\pi(Z_{ij})}\}$, where $z_{\pi(Z_{ij})}$ is the $\pi(Z_{ij})$th quantile of the standard normal distribution. An R software package "mvtBinaryEP" (By and Qaqishi, 2011) is available to produce binary data following the procedure above.

To generate the correlated failure times for uncured patients with the given marginal survival function in (4.3), we use the Clayton copula model (Clayton, 1978),

$$\begin{aligned} P(\tilde{T}^*_{i1} > t_{i1}, \cdots, \tilde{T}^*_{in_i} > t_{in_i} | Y_{ij} = 1, X_{ij}, j = 1, \cdots, n_i) \\ = \{\sum_{j=1}^{n_i} S_u(t^*_{ij}; X_{ij})^{-\xi} - n_i + 1\}^{-1/\xi}, \end{aligned}$$

where ξ measures the degree of dependence among the failure times of uncured patients within cluster i and it relates to Kendall's tau by $\xi = 2\tau/(1-\tau)$. When τ takes 0.8, 0.5 and 0, the corresponding values of ξ are 8, 2 and 0, respectively, and they correspond to strong, weak and zero correlation.

We consider three configurations of (ζ, τ): $(0.4, 0.8)$, $(0.2, 0.5)$ and $(0, 0)$. A pair of larger values of (ζ, τ) imply a stronger correlation of the cure status and the failure times of uncured patients in a cluster. When $(\zeta, \tau) = (0, 0)$, clustering does not produce correlations. Finally, the censoring times are generated independently from the uniform distribution in (0,12).

For each setting above, we generate 1000 data sets and fit each data set with the marginal model using the proposed estimation equations. As a comparison, we also estimate the parameters in the marginal model using the EM algorithm (Yu and Peng, 2008). The initial value of θ^* in our simulation study is set to $\mathbf{0}$. The biases, empirical variances (Var), the averages of estimated variances (Var*), and the coverage probabilities of 95% confidence intervals of the parameter estimates, under each method, are reported in Table 4.1 and Table 4.2 for the binary and continuous covariate cases, respectively.

The results indicate that the proposed ES estimation method outperforms the existing EM estimation method. The average estimated variances of the regression parameters from the ES method are close to their empirical variances in all cases, and the 95% confidence interval coverage rates are satisfactory and close to the nominal level. When the correlation within a cluster is strong and the cluster size is large, the biases and variances, particularly the variances of γ_1 and β_1, from the ES method are smaller than those from the EM method. Given the same sample size, the variances of the parameters except γ_1 tend to decrease as the cluster size decreases or as the correlation decreases. When the correlation reduces to zero, the biases and empirical

Table 4.1: Bias, empirical variance (Var), average of estimated variance (Var*), coverage percentage (CP) of 95% confidence intervals of the estimate of $\theta^* = (\gamma_0, \gamma_1, \beta_0, \beta_1, \alpha)$ with binary covariate.

n_i/K		EM					ES				
		γ_0	γ_1	β_0	β_1	α	γ_0	γ_1	β_0	β_1	α
$(\zeta, \tau) = (0.4, 0.8)$											
10/40	Bias	0.005	-0.019	0.031	-0.047	0.045	0.004	-0.017	0.030	-0.046	0.044
	Var	0.061	0.048	0.043	0.057	0.022	0.055	0.034	0.041	0.047	0.022
	Var*	0.061	0.055	0.040	0.057	0.020	0.052	0.038	0.038	0.042	0.022
	CP	95.2	95.1	92.2	93.2	93.0	94.2	95.6	92.5	93.1	94.4
5/80	Bias	0.012	-0.013	0.021	-0.028	0.026	0.011	-0.012	0.020	-0.027	0.025
	Var	0.040	0.052	0.024	0.044	0.012	0.038	0.040	0.023	0.038	0.012
	Var*	0.040	0.051	0.022	0.043	0.010	0.035	0.047	0.022	0.039	0.012
	CP	94.9	95.1	92.3	92.6	92.7	94.1	96.0	93.0	93.7	93.5
2/200	Bias	0.008	0.002	0.011	-0.013	0.018	0.008	0.003	0.010	-0.013	0.017
	Var	0.025	0.046	0.014	0.038	0.006	0.024	0.040	0.014	0.036	0.006
	Var*	0.027	0.051	0.013	0.037	0.006	0.026	0.048	0.013	0.035	0.006
	CP	96.0	96.0	92.9	93.8	93.8	96.0	96.6	92.9	94.3	94.6
$(\zeta, \tau) = (0.2, 0.5)$											
10/40	Bias	-0.005	0.005	0.026	-0.038	0.035	-0.006	0.006	0.025	-0.036	0.035
	Var	0.040	0.050	0.034	0.043	0.011	0.038	0.044	0.033	0.039	0.011
	Var*	0.041	0.051	0.031	0.041	0.010	0.038	0.052	0.029	0.039	0.011
	CP	95.9	95.5	92.9	93.6	93.8	95.2	96.7	91.3	92.3	93.4
5/80	Bias	-0.003	-0.008	0.015	-0.012	0.017	-0.004	-0.006	0.015	-0.012	0.016
	Var	0.032	0.048	0.021	0.038	0.007	0.031	0.044	0.020	0.035	0.007
	Var*	0.031	0.051	0.019	0.039	0.006	0.029	0.049	0.019	0.036	0.007
	CP	94.9	95.0	92.0	93.6	93.8	93.8	95.5	93.1	94.6	94.4
2/200	Bias	0.001	0.002	0.013	-0.007	0.014	0.001	0.002	0.013	-0.006	0.014
	Var	0.025	0.051	0.013	0.033	0.005	0.025	0.049	0.013	0.032	0.005
	Var*	0.025	0.051	0.012	0.036	0.004	0.024	0.050	0.012	0.034	0.004
	CP	95.2	94.9	93.9	95.3	94.2	95.7	95.1	93.9	96.1	94.2
$(\zeta, \tau) = (0, 0)$											
10/40	Bias	-0.002	0.002	0.007	-0.003	0.011	-0.003	0.002	0.006	-0.002	0.011
	Var	0.021	0.047	0.011	0.034	0.004	0.021	0.047	0.011	0.034	0.004
	Var*	0.022	0.050	0.010	0.034	0.004	0.022	0.049	0.010	0.033	0.004
	CP	95.0	95.3	93.2	93.4	93.2	95.1	94.9	93.6	93.2	93.2
5/80	Bias	0.005	-0.006	0.005	-0.004	0.008	0.005	-0.006	0.005	-0.003	0.008
	Var	0.023	0.048	0.011	0.034	0.004	0.023	0.048	0.011	0.034	0.004
	Var*	0.023	0.051	0.010	0.035	0.004	0.023	0.050	0.010	0.033	0.004
	CP	95.3	96.2	93.2	93.9	94.9	95.4	95.7	93.4	93.1	94.9
2/200	Bias	0.003	-0.006	0.013	-0.015	0.010	0.003	-0.007	0.013	-0.015	0.010
	Var	0.023	0.047	0.011	0.035	0.004	0.023	0.047	0.011	0.035	0.004
	Var*	0.023	0.051	0.010	0.035	0.004	0.023	0.050	0.010	0.033	0.004
	CP	95.2	95.3	93.6	93.9	94.5	95.1	95.4	93.6	94.6	94.3

Table 4.2: Bias, empirical variance (Var), average of estimated variance (Var*), coverage percentage (CP) of 95% confidence intervals of the estimate of $\theta^* = (\gamma_0, \gamma_1, \beta_0, \beta_1, \alpha)$ with continuous covariate.

n_i/K		EM					ES				
		γ_0	γ_1	β_0	β_1	α	γ_0	γ_1	β_0	β_1	α
$(\zeta, \tau) = (0.4, 0.8)$											
10/40	Bias	0.003	-0.027	0.037	-0.047	0.040	0.003	-0.024	0.035	-0.036	0.037
	Var	0.052	0.074	0.044	0.048	0.019	0.052	0.034	0.043	0.032	0.019
	Var*	0.053	0.065	0.039	0.046	0.017	0.048	0.043	0.037	0.033	0.018
	CP	95.0	91.4	92.0	92.7	92.1	93.2	96.2	90.8	93.8	93.3
5/80	Bias	0.000	-0.007	0.019	-0.027	0.025	0.000	-0.008	0.018	-0.020	0.023
	Var	0.033	0.044	0.021	0.028	0.010	0.032	0.030	0.021	0.020	0.010
	Var*	0.032	0.045	0.019	0.027	0.009	0.029	0.034	0.018	0.020	0.009
	CP	94.4	95.4	92.1	93.2	93.3	93.5	96.0	92.4	93.8	94.2
2/200	Bias	-0.001	-0.006	0.014	-0.019	0.016	-0.001	-0.003	0.013	-0.016	0.014
	Var	0.018	0.027	0.011	0.014	0.004	0.018	0.025	0.011	0.012	0.004
	Var*	0.018	0.029	0.010	0.014	0.004	0.018	0.027	0.010	0.014	0.005
	CP	95.5	96.9	93.3	94.0	94.5	94.8	95.8	93.0	95.0	95.0
$(\zeta, \tau) = (0.2, 0.5)$											
10/40	Bias	0.001	-0.021	0.028	-0.031	0.031	-0.001	-0.021	0.028	-0.029	0.030
	Var	0.036	0.025	0.029	0.022	0.011	0.036	0.022	0.028	0.019	0.011
	Var*	0.035	0.025	0.026	0.019	0.009	0.032	0.025	0.026	0.017	0.009
	CP	94.9	94.0	92.2	92.9	92.2	93.8	96.2	91.3	91.8	91.9
5/80	Bias	0.004	-0.014	0.021	-0.021	0.018	0.003	-0.013	0.021	-0.017	0.017
	Var	0.023	0.025	0.017	0.014	0.006	0.023	0.023	0.017	0.013	0.006
	Var*	0.023	0.024	0.016	0.013	0.005	0.022	0.023	0.015	0.012	0.006
	CP	94.9	94.2	91.5	92.3	93.9	94.8	95.3	92.1	94.0	94.2
2/200	Bias	0.002	-0.013	0.009	-0.012	0.011	0.002	-0.013	0.009	-0.011	0.011
	Var	0.016	0.023	0.010	0.011	0.004	0.016	0.022	0.010	0.010	0.004
	Var*	0.017	0.022	0.009	0.010	0.004	0.016	0.022	0.009	0.010	0.004
	CP	95.7	95.4	94.0	93.9	94.9	95.5	95.5	93.5	94.5	94.8
$(\zeta, \tau) = (0, 0)$											
10/40	Bias	0.007	-0.004	0.007	-0.008	0.007	0.007	-0.003	0.007	-0.008	0.007
	Var	0.014	0.020	0.007	0.009	0.003	0.014	0.020	0.007	0.009	0.003
	Var*	0.014	0.021	0.007	0.009	0.003	0.014	0.021	0.006	0.009	0.003
	CP	93.3	95.4	92.8	93.9	95.0	93.7	95.1	92.4	94.1	94.7
5/80	Bias	0.003	-0.013	0.009	-0.014	0.010	0.003	-0.013	0.009	-0.013	0.009
	Var	0.013	0.022	0.007	0.009	0.003	0.013	0.022	0.007	0.009	0.003
	Var*	0.014	0.022	0.007	0.009	0.003	0.014	0.022	0.007	0.009	0.003
	CP	95.4	94.5	94.0	94.3	94.0	95.6	94.2	93.6	94.8	94.1
2/200	Bias	0.007	-0.010	0.013	-0.008	0.012	0.007	-0.010	0.013	-0.008	0.012
	Var	0.015	0.020	0.007	0.010	0.003	0.015	0.020	0.007	0.010	0.003
	Var*	0.014	0.022	0.007	0.009	0.003	0.014	0.022	0.007	0.009	0.003
	CP	94.6	95.7	94.0	93.0	93.8	94.6	95.6	94.1	93.7	94.0

variances based on the ES method and the EM method are comparable.

To further evaluate the efficiency gains from the ES method, we calculate the relative efficiency (RE) defined as the ratio of mean squared errors of the estimates from the ES method to that from the EM method, i.e., $\text{RE} = \text{MSE}_{ES}/\text{MSE}_{EM}$, and report them in Table 4.3. The results indicate that the proposed ES method can

Table 4.3: Relative efficiency of $\hat{\theta}^*_{ES}$ vs $\hat{\theta}^*_{EM}$.

n_i/K	Type of covariate	RE = MSE($\hat{\theta}^*_{ES}$)/MSE($\hat{\theta}^*_{EM}$)				
		γ_0	γ_1	β_0	β_1	α
$\zeta = 0.4$, $\tau = 0.8$						
10/40	discrete	0.902	0.709	0.953	0.830	0.996
	continuous	1.000	0.463	0.975	0.663	0.989
5/80	discrete	0.950	0.769	0.957	0.865	0.996
	continuous	0.970	0.683	0.998	0.710	0.991
2/200	discrete	0.960	0.870	0.999	0.948	0.994
	continuous	1.000	0.925	0.998	0.853	0.986
$\zeta = 0.2$, $\tau = 0.5$						
10/40	discrete	0.950	0.880	0.970	0.907	1.000
	continuous	1.000	0.882	0.966	0.864	0.995
5/80	discrete	0.969	0.916	0.953	0.921	0.995
	continuous	1.000	0.920	1.000	0.920	0.994
2/200	discrete	1.000	0.961	1.000	0.969	1.000
	continuous	1.000	0.957	1.000	0.908	1.000
$\zeta = 0$, $\tau = 0$						
10/40	discrete	1.000	1.000	0.999	1.000	1.000
	continuous	1.000	1.000	1.000	1.000	1.000
5/80	discrete	1.000	1.000	1.000	1.000	1.000
	continuous	1.000	1.000	1.000	0.997	0.994
2/200	discrete	1.000	1.000	1.000	1.000	1.000
	continuous	1.000	1.000	1.000	1.000	1.000

achieve considerable efficiency gain for regression parameters, particularly for γ_1 and β_1, when the correlation is strong, and it is still comparable with the EM method when the correlation is weak. For example, when the correlation is strong, that is $(\zeta, \tau) = (0.4, 0.8)$, the REs of γ_1 can be as low as 0.709 for binary covariate and 0.463

for continuous covariate and the REs of β_1 can be as low as 0.830 for binary covariate and 0.663 for continuous covariate. The REs tend to approach 1 when the correlation becomes weak.

The proposed estimation method also produces estimates of ρ_1 and ρ_2, the correlation coefficients in the two working correlation matrices. Even though they do not correspond to the correlation measures ζ and τ in the data generation, Table 4.4 shows that the estimated values of ρ_1 and ρ_2 agree well with the values of ζ and τ used in the data generation in the sense that when the latter decrease, the former tend to decrease too. When there is no correlation in clusters, the estimates of ρ_1 and ρ_2 are very close to zero.

Table 4.4: Mean, empirical variance (Var), average of estimated variance (Var*) of $(\hat{\rho}_1, \hat{\rho}_2)$.

Type of		10/40			5/80			2/200		
covariate		Mean	Var	Var*	Mean	Var	Var*	Mean	Var	Var*
$(\zeta, \tau) = (0.4, 0.8)$										
discrete	$\hat{\rho}_1$	0.354	0.003	0.004	0.357	0.003	0.003	0.367	0.004	0.004
	$\hat{\rho}_2$	0.133	0.013	0.011	0.136	0.015	0.013	0.139	0.023	0.022
continuous	$\hat{\rho}_1$	0.363	0.004	0.004	0.371	0.003	0.003	0.373	0.005	0.004
	$\hat{\rho}_2$	0.138	0.012	0.010	0.148	0.014	0.011	0.146	0.024	0.023
$(\zeta, \tau) = (0.2, 0.5)$										
discrete	$\hat{\rho}_1$	0.178	0.002	0.002	0.180	0.003	0.002	0.186	0.005	0.005
	$\hat{\rho}_2$	0.055	0.003	0.002	0.057	0.004	0.002	0.056	0.007	0.005
continuous	$\hat{\rho}_1$	0.179	0.002	0.002	0.179	0.003	0.002	0.184	0.005	0.005
	$\hat{\rho}_2$	0.056	0.003	0.002	0.061	0.004	0.004	0.061	0.008	0.007
$(\zeta, \tau) = (0, 0)$										
discrete	$\hat{\rho}_1$	-0.002	0.0005	0.0006	-0.003	0.0013	0.0012	-0.004	0.0053	0.0051
	$\hat{\rho}_2$	0.013	0.0003	0.0002	0.011	0.0005	0.0003	0.012	0.0018	0.0020
continuous	$\hat{\rho}_1$	-0.005	0.0006	0.0006	-0.002	0.0013	0.0012	-0.002	0.0049	0.0047
	$\hat{\rho}_2$	0.018	0.0004	0.0004	0.017	0.0008	0.0007	0.049	0.0019	0.0017

For the bootstrap method to estimate the variances of $\hat{\rho}_1$ and $\hat{\rho}_2$, we select 25 data sets randomly from the 1000 simulated data sets and choose $B = 100$ in our simulation studies. Table 4.4 indicates that the empirical variance estimates and the

average of 25 bootstrap variances are quite close, which indicates that the bootstrap variance estimators work well for calculating the variance estimates of $\hat{\rho}_1$ and $\hat{\rho}_2$.

4.5 Analysis of the Smoking Cessation Data

We consider data from a smoking cessation study (Section 1.2.3). Observed covariates include sex, duration as smokers in years, intervention type and the average number of cigarettes smoked per day just prior to quitting. The survival time is defined as the time required for a failed quitter to resume smoking. Banerjee and Carlin (2004) considered the data as interval-censored survival data and analyzed them based on a parametric mixture cure model with a Bayesian method. They assumed the same cure rate for different smokers. Due to the potential spatial correlation among subjects residing the area with same zip code, Yu and Peng (2008) considered the survival times as right-censored by defining the midpoint of the intervals of the relapse time as the survival time and applied a marginal mixture cure model with a Weibull baseline distribution to the data. Chen and Lu (2012) considered a marginal semiparametric transformation cure model for the right-censored times. Neither of the two works considered the correlation structures within clusters.

As an illustration of the right-censored survival times, we plot the Kaplan-Meier survival curves by sex and intervention type in Figure 1.3. We observe that male smokers tend to have a higher cure rate than female smokers in the SI group whereas in the UC group, female smokers tend to have a higher cure rate than male smokers. This indicates that the interaction between sex and intervention type should be considered in a cure model.

To examine the impact of using correlation structures on the marginal parameter estimation, we propose to fit the survival data with the proposed marginal mixture

cure model with an exchangeable correlation structure for both the cure statues and the failure times of uncured patients from the same zip code area. The model includes sex, duration as a smoker, intervention type, average number of cigarettes smoked per day as well as the interaction between sex and intervention type in both logistic regression (4.2) and proportional hazards model (4.3). The parameter estimates from the proposed estimating equations are reported in Table 4.5. Note that the standard errors of the estimated correlation parameters in the table are obtained from 200 bootstrap samples. As a comparison, we also include estimates from the marginal mixture cure model proposed by Yu and Peng (2008) where the correlation structures within clusters are ignored.

Table 4.5: Estimated parameters from fitting the marginal mixture cure model to the smoking cessation data using the ES method and EM method.

	ES method			EM method		
Covariate	$\hat{\theta}^*$	$s.e.(\hat{\theta}^*)$	p-value	$\hat{\theta}^*$	$s.e.(\hat{\theta}^*)$	p-value
PH Survival Model						
Intercept	-2.966	1.011	0.003	-2.833	1.072	0.008
Sex (male=0)	1.048	0.549	0.056	0.954	0.654	0.145
Duration as smoker	0.014	0.039	0.712	0.016	0.038	0.675
SI/UC (usual care=0)	0.713	0.692	0.302	0.707	0.757	0.350
Cigarettes/day	-0.043	0.036	0.234	-0.042	0.024	0.071
Sex*SI/UC	-0.843	0.860	0.327	-0.752	0.899	0.403
α	2.931	0.265	0.000	2.782	0.139	0.000
ρ_2	-0.020	0.067	0.767	-	-	-
Logistic Model						
Intercept	0.265	0.653	0.685	0.183	0.650	0.778
Sex (male=0)	-0.214	0.542	0.692	-0.248	0.613	0.686
Duration as smoker	-0.041	0.020	0.036	-0.039	0.020	0.046
SI/UC (usual care=0)	-0.969	0.339	0.004	-0.982	0.360	0.006
Cigarettes/day	0.024	0.021	0.257	0.025	0.016	0.116
Sex*SI/UC	0.811	0.574	0.158	0.859	0.626	0.170
ρ_1	-0.023	0.016	0.143	-	-	-

The estimates from the two estimation methods are generally similar. The noticeable difference is in the effect of sex on the relapse (resume smoking) time of subject, which is marginally significant in the ES method but insignificant in the EM method. That is, women tend to resume smoking sooner than men, which may be attributed to the risk of weight gain following smoking cessation (Banerjee and Carlin, 2004). The number of cigarettes smoked per day, on the other hand, becomes insignificant in the ES method instead of marginally significant in the EM method. It indicates that the daily consumption of cigarettes may have little impact on the relapse time or on the probability of being cured. The effects of the remaining covariates are similar in the two models. The similarity of the estimates from the two methods may indicate that the correlation within clusters may not be strong enough to make differences in parameter estimates. This is evident from the estimates of ρ_1 and ρ_2. Both values are close to zero. Their large variances make the correlations insignificant.

4.6 Conclusions

Existing marginal cure models and estimation methods for analyzing clustered survival data with a cure fraction do not impose specific dependence structures on the correlated failure times or cure statuses. They are useful when there is little information about the correlation structures. However, when the correlation is of interest and there is partial information available for the correlation structures, the methods may not be efficient.

Rosen et al. (2000) extended the estimating equations of Liang and Zeger (1986) to mixtures of the generalized linear models. This idea was also explored by Hall and Zhang (2004) for zero-inflated count data. In this chapter, we propose an approach to extend the generalized estimating equation approach from generalized linear models to

the marginal mixture cure model for censored survival data. The estimating equations incorporate two working correlation structures, one for the failure times of uncured subjects and the other for the cure statuses within a cluster. We show that the estimates of the regression parameters and the baseline distribution are consistent and asymptotically normal, and their variances can be consistently estimated by a sandwich variance estimator. Our numerical study demonstrates that the proposed method substantially improves the estimation efficiency of the regression parameters, especially when the correlation within clusters is strong and the cluster size is large. Therefore, the proposed marginal proportional hazards mixture cure model is a useful alternative to the existing marginal models for clustered survival data with a possible cure fraction.

Our method generalizes the marginal proportional hazards model proposed for the correlated failure time data without cure fraction (Segal and Neuhaus, 1993) to the marginal proportional hazards mixture cure model for clustered survival data with a cure fraction. The proposed method also extends the marginal proportional hazards mixture cure model (Yu and Peng, 2008) by explicitly including correlation structures such as the exchangeable working matrix in the model estimation. Future work for this model includes a method to consider a correlation structure when estimating the parameters in the baseline distribution.

Chapter 5

Semiparametric Marginal Proportional Hazards Mixture Cure Model

5.1 Introduction

As we discussed in Chapter 4, the proposed estimating equations approach can improve the estimation efficiency in a parametric proportional hazards mixture cure model for clustered survival data with a cure fraction. In this chapter, we further consider a semiparametric proportional hazards mixture cure model where the survival function for the uncured patients is modeled by a semiparametric proportional hazards model. Peng et al. (2007) considered the same model for clustered failure time data and proposed a robust variance estimation method. However, their method may lose efficiency when potential correlation exists within clusters. To improve the estimation efficiency, we follow the idea in Chapter 4 and apply the ES method in a semiparametric proportional hazards mixture cure model. Similarly, the dependence among the cure statuses and among the survival times of uncured patients within clusters are modeled by working correlation matrices in the proposed estimating equations. A bootstrap method is used to obtain the variances of the estimates.

We report a simulation study to demonstrate a substantial efficiency gain of the proposed method over the existing marginal method. Finally, we apply the model and the proposed method to two sets of data including a multi-institutional study of tonsil cancer patients treated with radiation therapy and a multi-center study of leukemia patients treated with bone marrow transplantation.

The rest of this chapter is organized as follows. In Section 5.2, we introduce the marginal semiparametric proportional hazards mixture cure model with clustered observations. Then we present the proposed estimating equations and the corresponding estimation steps. The variance estimation is discussed in Section 5.3. We conduct a simulation study to evaluate the performance of the proposed marginal method in Section 5.4. The proposed model and estimation method are applied to a tonsil cancer data in Section 5.5. Finally, we provide conclusions in Section 5.6.

5.2 Model and Estimation

We assume that the marginal survival function of $\tilde{T}_{ij}^*$ is from a semiparametric proportional hazards mixture cure model, i.e.,

$$S(t; X_{ij}, Z_{ij}) = 1 - \pi(Z_{ij}) + \pi(Z_{ij}) S_u(t; X_{ij}), \tag{5.1}$$

where $\pi(Z_{ij}) = P(Y_{ij} = 1; Z_{ij})$ is in a logistic regression form

$$\pi(Z_{ij}) = \frac{\exp(\gamma' Z_{ij})}{1 + \exp(\gamma' Z_{ij})}, \tag{5.2}$$

and $S_u(t; X_{ij})$ is specified by the proportional hazards model

$$S_u(t; X_{ij}) = S_{u0}(t)^{\exp(\beta' X_{ij})}, \tag{5.3}$$

where $S_{u0}(t)$ is the baseline survival function of $\tilde{T}^*_{ij}|\{Y_{ij}=1\}$ when $X_{ij}=0$ and is usually unspecified, and β and γ are $p_X \times 1$ and $p_Z \times 1$ parameter vectors for X_{ij} and Z_{ij}.

Let t^*_{ij} be the observed value of T^*_{ij}, and $O^* = \{(T^*_{ij}, \delta_{ij}, X_{ij}, Z_{ij}), j = 1, \cdots, n_i, i = 1, \cdots, K\}$ be the observed data. We also augment the observed data to include the latent values of Y_{ij} and denote the augmented data as $O^*_c = \{(T^*_{ij}, \delta_{ij}, X_{ij}, Z_{ij}, Y_{ij}), j = 1, \cdots, n_i, i = 1, \cdots, K\}$. If we ignore the potential correlation between $\tilde{T}^*_{ij}|\{Y_{ij}=1\}$ and $\tilde{T}^*_{ij'}|\{Y_{ij'}=1\}$, and between Y_{ij} and $Y_{ij'}$, the unknown parameters in the marginal model specified in (5.1), (5.2), and (5.3) can be estimated using the EM algorithm (Peng and Dear, 2000; Sy and Taylor, 2000). The E-step in the EM algorithm computes the expectation of a log likelihood function based on data O^*_c,

$$\begin{aligned}
l_{\mathbf{c}}(\gamma, \beta, \alpha; O^*_c) &= \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} \pi(z_{ij})^{y_{ij}} \left(1 - \pi(z_{ij})\right)^{1-y_{ij}} \left[f_u(t^*_{ij}; x_{ij})^{\delta_{ij}} S_u(t^*_{ij}; x_{ij})^{1-\delta_{ij}} \right]^{y_{ij}} \\
&= \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} \pi(z_{ij})^{y_{ij}} \left(1 - \pi(z_{ij})\right)^{1-y_{ij}} \\
&+ \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} \left(\left(\Lambda_{u0}(t^*_{ij}; \alpha) \exp(\beta' x_{ij})\right)^{\delta_{ij}} \exp(-\Lambda_{u0}(t^*_{ij}; \alpha) \exp(\beta' x_{ij})) \right)^{y_{ij}} \\
&+ \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} \left(\frac{\lambda_{u0}(t^*_{ij}; \alpha)}{\Lambda_{u0}(t^*_{ij}; \alpha)} \right)^{\delta_{ij}}, \qquad (5.4)
\end{aligned}$$

where $\lambda_{u0}(t; \alpha)$ and $\Lambda_{u0}(t; \alpha)$ are the corresponding baseline hazard and cumulative baseline hazard functions for $S_{u0}(t)$, and α is a set of unknown parameters in the baseline distribution. For given y_{ij}, the first term corresponds to a log-likelihood function of the logistic regression for y_{ij}. The second term can be viewed as a log-likelihood function if δ_{ij} is assumed to follow a Poisson distribution with mean equal to $y_{ij}\Lambda_{u0}(t^*_{ij}; \alpha) \exp(\beta' x_{ij})$, and the last term does not depend on β and γ. A similar approach of treating a likelihood function of the proportional hazards model as a

likelihood function of a Poisson model was discussed in Chapter 3. The expectation in E-step is taken with respect to the conditional distribution of the latent variable Y_{ij} given the observed data and the current estimates of the parameters. If the current estimate is denoted by $\theta^{*(m)} = (\gamma^{(m)}, \beta^{(m)}, \alpha^{(m)})$, then

$$\begin{aligned} g_{ij}^{(m)} &= E(Y_{ij}|\theta^{*(m)}, O^*) \\ &= \left[\delta_{ij} + \frac{(1-\delta_{ij})\pi(Z_{ij})S_{u0}(t_{ij}^*)^{\exp(\beta' X_{ij})}}{1-\pi(Z_{ij})+\pi(Z_{ij})S_{u0}(t_{ij}^*)^{\exp(\beta' X_{ij})}}\right]_{\theta^*=\theta^{*(m)}}, \end{aligned} \tag{5.5}$$

which is the same as (4.8), and the E-step is equivalent to substituting $g_{ij}^{(m)}$ for y_{ij} in (5.1). The M-step maximizes $E(l_{\mathbf{c}})$ with respect to γ, β and α. It results in the following estimating equations for γ and β respectively

$$\sum_{i=1}^{K} (\frac{\partial \pi(Z_i)}{\partial \gamma})^T (A_i^{1/2} I_i A_i^{1/2})^{-1} (g_i^{(m)} - \pi(Z_i)) = 0, \tag{5.6}$$

$$\sum_{i=1}^{K} (\frac{\partial \mu(X_i)}{\partial \beta})^T (B_i^{1/2} I_i B_i^{1/2})^{-1} (\delta_i - \mu(X_i)) = 0, \tag{5.7}$$

where $g_i^{(m)} = (g_{i1}^{(m)}, \cdots, g_{in_i}^{(m)})^T$, $\pi(Z_i) = (\pi(Z_{i1}), \cdots, \pi(Z_{in_i}))^T$, $A_i = \text{diag}(\pi(Z_{i1})(1-\pi(Z_{i1})), \cdots, \pi(Z_{in_i})(1-\pi(Z_{in_i})))$, $\delta_i = (\delta_{i1}, \cdots, \delta_{in_i})^T$, $\mu(X_i) = (\mu(X_{i1}), \cdots, \mu(X_{in_i}))^T$ with $\mu(X_{ij}) = g_{ij}^{(m)} \Lambda_{u0}(t_{ij}^*; \alpha) \exp(\beta' X_{ij})$, $B_i = \text{diag}(\mu(X_{i1}), \cdots, \mu(X_{in_i}))$, and I_i is an $n_i \times n_i$ identity matrix.

The baseline survival function $S_{u0}(t)$ in the M-step can be estimated using the nonparametric maximum likelihood estimation method as we discussed in Section 2.1.3. That is, the estimating equations for α_s are

$$\sum_{(i,j)\in D_s} \frac{e^{\beta' X_{ij}}}{1-\alpha_s^{\exp(\beta' X_{ij})}} - \sum_{(i,j)\in R_s} g_{ij}^{(m)} \exp(\beta' X_{ij}) = 0\,, \quad s = 1, \cdots, k, \tag{5.8}$$

The equation (5.8) is the same as equation (2.9), and it does not have a closed solution of α_s when there is a $d_s > 1$. An approximate estimator for the baseline survival function $S_{u0}^{(m)}(t)$ can be obtained (Peng and Dear, 2000)

$$\hat{S}_{u0}^{(m)}(t) = \exp(-\sum_{s:\tau_s \le t} \frac{d_s}{\sum_{(i,j)\in R_s} g_{ij}^{(m)} \exp(\beta' X_{ij})}). \tag{5.9}$$

To enhance the identifiability of the parameter estimation, it is often assumed that $\hat{S}_{u0}^{(m)}(t) = 0$ if $t_{ij}^* > \tau_k$ (Taylor, 1995; Peng and Dear, 2000; and Sy and Taylor, 2000).

Due to the potential correlation between $\tilde{T}_{ij}^*|\{Y_{ij} = 1\}$ and $\tilde{T}_{ij'}^*|\{Y_{ij'} = 1\}$, and between Y_{ij} and $Y_{ij'}$ for $j \neq j'$, the aforementioned method may not be efficient, even though the marginal model is correctly specified. Peng et al. (2007) considered this method for clustered data and proposed a sandwich variance estimate for the estimated parameters in the marginal model. Their method may still lack efficiency due to the absence of the correlation modeling in the model. To increase the estimation efficiency of the method above, we use the ES algorithm proposed in Chapter 4. That is, the M-step in the EM algorithm is replaced by the S-step where the identity matrix I_i in (5.6) and (5.7) are replaced by working correlation matrices to account for the potential correlation in each cluster. Therefore, the proposed estimating equations for γ and β are

$$\sum_{i=1}^{K} (\frac{\partial \pi(Z_i)}{\partial \gamma})^T (A_i^{1/2} Q_i(\rho_1) A_i^{1/2} \phi_1)^{-1} (g_i^{(m)} - \pi(Z_i)) = 0, \tag{5.10}$$

$$\sum_{i=1}^{K} (\frac{\partial \mu(X_i)}{\partial \beta})^T (B_i^{1/2} Q_i(\rho_2) B_i^{1/2} \phi_2)^{-1} (\delta_i - \mu(X_i)) = 0, \tag{5.11}$$

where $Q_i(\rho_1) = (q_{jj'}(\rho_1))_{n_i \times n_i}$ and $Q_i(\rho_2) = (q_{jj'}(\rho_2))_{n_i \times n_i}$ are the working correlation matrices, and ρ_1 and ρ_2 are unknown parameters in the matrices that need to be estimated. The scale parameters ϕ_1 and ϕ_2 are incorporated in the estimating

equations to accommodate the over- or under-dispersion. Note that the proposed estimating equation (5.11) for β is different from the estimating equation (4.10) which has a weighted form. Furthermore, through the simulation study, we observe that equation (5.11) produces biased estimate of β. To reduce the bias of the estimate of β, motivated by Ritov (1990), we standardize X_{ij} in (5.11), i.e., we use $(X_{ij} - \bar{X}_i)/(var(X_i)^{1/2})$ instead of X_{ij} as the covariate, where $\bar{X}_i$ and $var(X_i)$ are the mean and variance of X_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, K$.

Similar to Chapter 4, we apply the exchangeable correlation structure with $q_{jj'}(\rho_1) = \rho_1$ and $q_{jj'}(\rho_2) = \rho_2$ for $j \neq j'$ to estimating equations (5.10) and (5.11). Following the moment method, ρ_1 and ρ_2 in the two exchangeable correlation structures can be estimated from the standardized Pearson residuals $\hat{r}_{ij}$ by $\hat{\phi}_1^{-1} \sum_{i=1}^{K} \sum_{j>j'} \hat{r}_{ij}\hat{r}_{ij'} / \{\sum_{i=1}^{K} \frac{1}{2} n_i(n_i-1) - p_z\}$ and $\hat{\phi}_2^{-1} \sum_{i=1}^{K} \sum_{j>j'} \hat{r}_{ij}\hat{r}_{ij'} / \{\sum_{i=1}^{K} \frac{1}{2} n_i(n_i-1) - p_x\}$ respectively, where $\hat{\phi}_1 = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \hat{r}_{ij}^2/(N - p_z)$ and $\hat{r}_{ij} = (g_{ij}^{(m)} - \pi(z_{ij}))/(\pi(z_{ij})(1 - \pi(z_{ij})))^{\frac{1}{2}}$ for ρ_1 and $\hat{\phi}_2 = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \hat{r}_{ij}^2/(N - p_x)$ and $\hat{r}_{ij} = (\delta_{ij} - \mu(x_{ij}))/\mu(x_{ij})^{\frac{1}{2}}$ for ρ_2.

Due to the substitution of (5.10) and (5.11) for (5.6) and (5.7) respectively, the solution of γ, β, and $S_{u0}(t)$ from the S-step in the ES algorithm have to be found iteratively. We summarize this algorithm as follows:

1. Set initial values for γ, β and $S_{u0}(t)$.

2. E-step: calculate the conditional expectation of Y_{ij} via (5.5).

3. S-step:

   (a) Given current estimates of ρ_1, ρ_2, ϕ_1 and ϕ_2, calculate the updated estimates of γ and β from (5.10) and (5.11) using the Newton-Raphson method and an updated estimate of $S_{u0}(t)$ from (5.9).

   (b) Given the estimates of γ, β and $S_{u0}(t)$, calculate the standardized Pearson residuals $\hat{r}_{ij}$.

(c) Use the residuals $\hat{r}_{ij}$ to estimate ρ_1, ρ_2, ϕ_1, and ϕ_2.

(d) Repeat steps (a), (b), and (c) until convergence.

4. Iterate steps 2 and 3 until the algorithm converges to obtain $\hat{\theta}^*$.

5.3 Variance Estimation for $\hat{\theta}^*$, $\hat{\rho}_1$ and $\hat{\rho}_2$

Obtaining the variances of the estimated parameters $\hat{\theta}^*$ in the proposed estimating equations for the semiparametric proportional hazards mixture cure model is not straightforward. Rosen et al. (2000) proposed a sandwich variance estimator for the mixtures-of-experts model. However, it is difficult to use the estimator due to the nonparametric baseline estimation in the proposed ES algorithm. Therefore, we consider a bootstrap method as we did in Section 4.3.2 to obtain the variance estimates of $\hat{\theta}^*$. A bootstrap sample from this approach is obtained from sampling clusters with replacement. That is, all observations from one cluster will be either selected or excluded in a bootstrap sample. Let $\hat{\theta}_b^*$ be the estimate of $\theta^* = (\gamma, \beta, \alpha)$ from the bth bootstrap sample, $b = 1, \cdots, B$, and B is the number of bootstrap samples. The variance of $\hat{\theta}^*$ can be estimated by

$$\hat{\text{Var}}(\hat{\theta}^*) = \sum_{b=1}^{B} (\hat{\theta}_b^* - \sum_{a=1}^{B} \hat{\theta}_a^*/B)^2/(B-1).$$

The same bootstrap approach can be applied to the variance estimates of $\hat{\rho}_1$ and $\hat{\rho}_2$ in (5.10) and (5.11).

5.4 A Simulation Study

The design of our simulation study is similar to that in Chapter 4. Our objective is to investigate the performances of the proposed method and to compare the results with

those from Peng et al. (2007). The data in the simulation study are generated from a cure model for clustered survival data with the exchangeable correlation structure and the marginal equal to (5.1), (5.2), and (5.3). In particular, we consider a single covariate in the model and assume that the covariate has effects on both $\pi(Z_{ij})$ and $S_u(t; X_{ij})$. The effect on $\pi(Z_{ij})$ are specified by $(\gamma_0, \gamma_1) = (0.4, -1)$, the effect of the covariate on $S_u(t; X_{ij})$ is specified by $\beta = -1$, and the baseline distribution in (5.3) is the Weibull distribution with $S_{u0}(t; \alpha) = e^{-(\alpha_2 t)^{\alpha_1}}$, where $\alpha = (\alpha_1, \alpha_2) = (2, 2)$. The covariate is either a binary covariate with value 0 for a control group and 1 for a treatment group, or a continuous covariate with values generated from the standard normal distribution. Under the binary covariate case, the marginal cure rates are 40% and 64% in the control and the treatment groups respectively.

The correlation coefficient between Y_{ij} and $Y_{ij'}$, denoted as ζ, is set to 0.4, 0.2, and 0. To generate data Y_{ij} and $Y_{ij'}$ so that the correlation of Y_{ij} and $Y_{ij'}$ is ζ with $P(Y_{ij} = 1) = \pi_{ij}$ and $P(Y_{ij'} = 1) = \pi_{ij'}$ given in (5.2), we adopt the method as described in Section 4.4. That is, given ζ, π_{ij} ,$\pi_{ij'}$, we solve for $\tilde{\rho}_{ijj'}$ through

$$\frac{\Phi((z_{\pi_{ij}}, z_{\pi_{ij'}}), \tilde{\rho}_{ijj'}) - \pi_{ij}\pi_{ij'}}{\pi_{ij}\pi_{ij'}(1 - \pi_{ij})(1 - \pi_{ij'})} = \zeta,$$

where $\Phi(\cdot, \tilde{\rho}_{ijj'})$ is the standard bivariate normal distribution function and the correlation coefficient equals to $\tilde{\rho}_{ijj'}$. We use $z_{\pi_{ij}}$ and $z_{\pi_{ij'}}$ to denote the π_{ij}th and $\pi_{ij'}$th quantiles of the standard normal distribution. After obtaining $\tilde{\rho}_{ijj'}$ for the ith cluster, we generate $(z_{i1}, \cdots, z_{in_i})$ from the multivariate normal distribution $N(\mathbf{0}, \Sigma_i)$ and obtain $(y_{i1}, \cdots, y_{in_i})$ with $y_{ij} = 1$ if $z_{ij} \leq z_{\pi_{ij}}$ and 0 otherwise, where the diagonal elements of the covariance matrix Σ_i are 1 and the rest are $\tilde{\rho}_{ijj'}$.

To generate the correlated failure times for uncured patients with the given marginal

survival function in (5.3), we use the Clayton copula model (Clayton, 1978),

$$P(\tilde{T}^*_{i1} > t^*_{i1}, \cdots, \tilde{T}^*_{in_i} > t^*_{in_i} | Y_{ij} = 1, X_{ij}, j = 1, \cdots, n_i)$$
$$= \{\sum_{j=1}^{n_i} S_u(t^*_{ij}; X_{ij})^{-\xi} - n_i + 1\}^{-1/\xi},$$

where ξ measures the degree of dependence among the failure times of uncured patients within cluster i and it relates to Kendall's tau by $\xi = 2\tau/(1-\tau)$. We set $\xi = 8, 2$, and 0. The corresponding values of τ are 0.8, 0.5 and 0 respectively, and the larger value implies the stronger correlation of the failure times. When $\xi = 0$ or $\tau = 0$, it implies the independence among the failure times.

To save computational time, we only consider three configurations of (ζ, τ): (0.4, 0.8), (0.2, 0.5) and (0, 0), and equal cluster sizes ($n_1 = \cdots = n_K$). For each configuration of (ζ, τ) above, we generate clustered failure time data with the following pairs of the number of clusters and cluster sizes: (40, 10), (80, 5) and (200, 2). The censoring times are non-informative and generated from the uniform distribution in $(0, 3)$. For each setting above, we generate 1000 data sets and estimate the parameters in the marginal model using the proposed ES algorithm for each data set. The bias, empirical variance (Var), and the average of bootstrap variance (Var*) of the parameter estimates are computed. The bootstrap variance and the coverage probability of 95% confidence intervals are based on 200 randomly selected data sets from the 1000 data sets to save some computational time. As a comparison, we also estimate the parameters in the marginal model using the method by Peng et al. (2007) (denoted as PTY in the following tables) and calculate the relative efficiency (RE), defined as the ratio of the mean squared error of the estimates from the ES method to that from the PTY method, to measure the efficiency gains from using the ES method relative to the PTY method.

From the simulation study, we observe the similar features as the results based on the parametric proportional hazards mixture cure model in Chapter 4. Tables 5.1 and 5.2 present results from the data generated with the binary covariate and continuous covariate, respectively. They show that when the cure statuses and the failure times of uncured patients within a cluster are correlated, the empirical variance estimates of the regression parameters, particularly γ_1 and β, from the ES method are less than those from the PTY method, and the REs are generally less than 1 and can be as low as 0.56 for γ_1 and 0.61 for β when the correlation is strong. Given the same total number of observations, the most empirical variance estimates of γ_0 and β tend to decrease as the number of clusters increases. However, this trend does not apply to γ_1 for binary covariate. For example, when the correlation is strong, the empirical variance of γ_1 firstly increases then decreases in the PTY method and consistently increases in the ES method. When the correlation is moderate, the empirical variance of γ_1 in both methods firstly increases then decreases as the number of clusters increases. The REs tend to approach 1 when the correlation decreases. When the correlation reduces to zero, the empirical variances based on the ES method and the PTY method are almost the same, and the REs are close to 1. It indicates that the proposed ES method can achieve a considerable efficiency gain when the correlation is strong and is still comparable with the existing method when the correlation is weak.

Table 5.1: Bias, empirical variance, average of estimated variance, coverage percentage and relative efficiency of the estimates of $(\gamma_0, \gamma_1, \beta)$ with a binary covariate and high cure rate.

n_i/K				10/40			5/80			2/200	
(ζ,τ)	θ		PTY	ES	RE	PTY	ES	RE	PTY	ES	RE
(0.4,0.8)	γ_0	Bias	0.008	0.008	0.92	-0.008	-0.007	0.93	0.011	0.010	0.96
		Var	0.065	0.060		0.043	0.040		0.031	0.030	
		Var*	0.063	0.062		0.043	0.041		0.031	0.030	
		CP	94.6	96.0		94.1	96.0		94.4	96.5	
	γ_1	Bias	0.009	0.013	0.76	-0.001	0.002	0.80	-0.002	0.003	0.89
		Var	0.055	0.042		0.063	0.050		0.060	0.053	
		Var*	0.055	0.045		0.058	0.050		0.058	0.055	
		CP	94.0	94.0		94.3	95.0		94.9	93.0	
	β	Bias	-0.079	-0.119	0.69	-0.022	-0.073	0.67	-0.001	-0.034	0.79
		Var	0.077	0.043		0.061	0.036		0.044	0.034	
		Var*	0.061	0.055		0.046	0.038		0.038	0.038	
		CP	90.6	97.5		90.9	96.5		93.0	93.5	
(0.2,0.5)	γ_0	Bias	0.013	0.014	0.99	-0.001	-0.001	0.96	0.016	0.016	0.99
		Var	0.046	0.045		0.036	0.035		0.027	0.027	
		Var*	0.044	0.043		0.034	0.035		0.028	0.029	
		CP	94.8	94.0		94.2	92.5		95.2	95.5	
	γ_1	Bias	-0.012	-0.014	0.95	0.009	0.009	0.95	-0.010	-0.011	0.98
		Var	0.057	0.054		0.060	0.058		0.054	0.053	
		Var*	0.056	0.055		0.057	0.056		0.057	0.057	
		CP	94.5	93.5		93.9	90.0		95.6	94.5	
	β	Bias	-0.030	-0.013	0.68	-0.028	-0.015	0.74	-0.006	-0.001	0.94
		Var	0.058	0.040		0.051	0.038		0.044	0.041	
		Var*	0.045	0.043		0.040	0.039		0.037	0.041	
		CP	90.6	94.5		91.6	96.0		93.2	95.5	
(0,0)	γ_0	Bias	0.010	0.012	1.02	0.016	0.016	1.00	0.016	0.016	1.00
		Var	0.025	0.025		0.026	0.026		0.028	0.028	
		Var*	0.026	0.027		0.026	0.026		0.026	0.027	
		CP	95.1	95.5		94.4	94.5		93.8	95.5	
	γ_1	Bias	-0.007	-0.014	1.03	-0.015	-0.018	1.01	-0.019	-0.021	1.00
		Var	0.058	0.060		0.060	0.060		0.061	0.061	
		Var*	0.057	0.061		0.057	0.061		0.058	0.062	
		CP	94.4	97.0		93.8	93.0		94.6	92.5	
	β	Bias	-0.011	0.032	1.05	-0.008	0.013	1.02	-0.015	-0.008	1.02
		Var	0.042	0.044		0.043	0.044		0.045	0.046	
		Var*	0.035	0.045		0.036	0.046		0.036	0.047	
		CP	91.8	92.5		92.2	94.5		92.4	94.5	

Table 5.2: Bias, empirical variance, average of estimated variance, coverage percentage and relative efficiency of the estimates of $(\gamma_0, \gamma_1, \beta)$ with a continuous covariate and high cure rate.

n_i/K				10/40			5/80			2/200	
(ζ, τ)	θ		PTY	ES	RE	PTY	ES	RE	PTY	ES	RE
(0.4,0.8)	γ_0	Bias	0.015	0.016	0.99	0.009	0.010	1.00	0.005	0.007	1.00
		Var	0.053	0.053		0.034	0.034		0.020	0.020	
		Var*	0.054	0.058		0.033	0.034		0.020	0.021	
		CP	95.5	96.0		95.4	98.0		94.4	93.0	
	γ_1	Bias	-0.033	-0.017	0.56	-0.031	-0.018	0.72	-0.016	-0.015	0.92
		Var	0.073	0.041		0.054	0.039		0.032	0.029	
		Var*	0.069	0.046		0.049	0.039		0.031	0.032	
		CP	92.6	97.0		92.3	94.0		95.5	94.0	
	β	Bias	-0.071	-0.090	0.61	-0.024	-0.050	0.65	-0.008	-0.017	0.80
		Var	0.071	0.038		0.033	0.019		0.016	0.013	
		Var*	0.052	0.035		0.030	0.021		0.015	0.013	
		CP	89.7	94.5		92.4	95.5		94.1	93.5	
(0.2,0.5)	γ_0	Bias	0.015	0.014	0.98	0.005	0.005	0.99	0.010	0.010	1.00
		Var	0.039	0.038		0.025	0.025		0.020	0.020	
		Var*	0.037	0.038		0.025	0.027		0.019	0.019	
		CP	94.0	95.0		94.8	96.0		94.5	93.0	
	γ_1	Bias	-0.022	-0.025	0.91	-0.015	-0.019	0.94	-0.018	-0.017	1.00
		Var	0.031	0.028		0.029	0.027		0.024	0.024	
		Var*	0.027	0.028		0.027	0.028		0.025	0.026	
		CP	92.6	93.5		94.5	94.5		95.2	94.5	
	β	Bias	-0.028	-0.003	0.67	-0.018	-0.001	0.70	-0.006	0.005	0.91
		Var	0.026	0.018		0.017	0.012		0.012	0.011	
		Var*	0.020	0.018		0.015	0.013		0.011	0.011	
		CP	90.5	93.5		93.2	96.0		94.5	93.0	
(0,0)	γ_0	Bias	0.009	0.004	0.99	0.011	0.009	1.00	0.004	0.004	1.00
		Var	0.016	0.016		0.016	0.016		0.018	0.018	
		Var*	0.016	0.016		0.016	0.017		0.016	0.017	
		CP	94.2	96.5		95.6	97.0		94.3	94.5	
	γ_1	Bias	-0.016	-0.025	1.03	-0.014	-0.019	1.02	-0.015	-0.017	1.01
		Var	0.027	0.027		0.026	0.026		0.027	0.027	
		Var*	0.024	0.027		0.023	0.025		0.024	0.026	
		CP	93.1	94.0		93.8	92.5		94.3	94.5	
	β	Bias	-0.014	0.054	1.24	-0.008	0.030	1.08	-0.003	0.008	1.03
		Var	0.011	0.011		0.011	0.011		0.011	0.011	
		Var*	0.010	0.011		0.010	0.011		0.010	0.011	
		CP	92.8	89.5		93.5	94.5		92.9	95.0	

Table 5.3: Bias, empirical variance, average of estimated variance, coverage percentage and relative efficiency of the estimates of $(\gamma_0, \gamma_1, \beta)$ with a binary covariate and low cure rate.

n_i/K				10/40			5/80			2/200	
(ζ,τ)	θ		PTY	ES	RE	PTY	ES	RE	PTY	ES	RE
(0.4,0.8)	γ_0	Bias	0.095	0.080	0.93	0.064	0.051	0.93	0.042	0.035	0.98
		Var	0.220	0.207		0.136	0.128		0.098	0.096	
		Var*	0.188	0.229		0.129	0.139		0.094	0.105	
		CP	93.5	95.5		95.4	95.5		95.6	98.0	
	γ_1	Bias	-0.064	-0.032	0.81	-0.034	-0.005	0.83	-0.016	0.001	0.97
		Var	0.163	0.134		0.144	0.121		0.132	0.127	
		Var*	0.143	0.152		0.134	0.130		0.131	0.144	
		CP	93.5	94.0		94.4	93.0		95.2	96.0	
	β	Bias	-0.051	-0.111	0.83	-0.022	-0.082	0.78	-0.011	-0.053	0.80
		Var	0.040	0.023		0.029	0.016		0.023	0.015	
		Var*	0.034	0.026		0.025	0.016		0.020	0.014	
		CP	92.9	89.0		93.0	90.5		92.2	92.5	
(0.2,0.5)	γ_0	Bias	0.069	0.064	0.98	0.043	0.039	0.96	0.038	0.035	0.99
		Var	0.151	0.148		0.108	0.104		0.088	0.087	
		Var*	0.133	0.161		0.104	0.127		0.088	0.100	
		CP	93.7	94.0		95.2	98.5		96.1	95.5	
	γ_1	Bias	-0.044	-0.033	0.96	-0.016	-0.008	0.94	-0.025	-0.018	0.99
		Var	0.143	0.138		0.132	0.125		0.127	0.126	
		Var*	0.134	0.163		0.131	0.153		0.130	0.148	
		CP	94.7	93.0		95.4	97.0		96.4	97.5	
	β	Bias	-0.044	-0.050	0.66	-0.028	-0.040	0.74	-0.007	-0.016	0.88
		Var	0.034	0.022		0.024	0.017		0.021	0.019	
		Var*	0.026	0.020		0.022	0.017		0.020	0.019	
		CP	89.7	92.0		93.2	91.5		93.4	93.5	
(0,0)	γ_0	Bias	0.041	0.047	1.02	0.027	0.031	1.02	0.026	0.028	1.01
		Var	0.083	0.084		0.085	0.086		0.081	0.081	
		Var*	0.081	0.091		0.080	0.097		0.081	0.100	
		CP	94.8	95.5		95.2	96.0		95.0	96.5	
	γ_1	Bias	-0.020	-0.037	1.03	-0.002	-0.012	1.04	0.014	0.010	1.02
		Var	0.133	0.137		0.133	0.138		0.137	0.140	
		Var*	0.127	0.144		0.128	0.151		0.129	0.157	
		CP	94.0	95.0		94.7	97.0		94.9	96.5	
	β	Bias	-0.009	0.041	1.12	-0.014	0.013	1.05	-0.008	0.001	1.04
		Var	0.021	0.022		0.021	0.022		0.021	0.022	
		Var*	0.018	0.022		0.019	0.022		0.019	0.022	
		CP	93.0	91.5		93.5	95.0		93.6	96.0	

Table 5.4: Bias, empirical variance, average of estimated variance, coverage percentage and relative efficiency of the estimates of $(\gamma_0, \gamma_1, \beta)$ with a continuous covariate and low cure rate.

n_i/K				10/40			5/80			2/200	
(ζ, τ)	θ		PTY	ES	RE	PTY	ES	RE	PTY	ES	RE
(0.4,0.8)	γ_0	Bias	0.103	0.100	0.98	0.069	0.071	0.99	0.054	0.057	1.01
		Var	0.181	0.177		0.115	0.114		0.066	0.066	
		Var*	0.153	0.216		0.097	0.127		0.063	0.077	
		CP	92.6	95.5		94.0	97.5		95.6	97.5	
	γ_1	Bias	-0.064	-0.053	0.69	-0.028	-0.018	0.82	-0.030	-0.021	1.00
		Var	0.222	0.152		0.138	0.114		0.085	0.085	
		Var*	0.183	0.189		0.121	0.131		0.071	0.081	
		CP	90.4	95.0		91.9	98.0		93.5	95.0	
	β	Bias	-0.061	-0.091	0.58	-0.023	-0.063	0.69	-0.004	-0.032	0.81
		Var	0.053	0.025		0.023	0.012		0.012	0.008	
		Var*	0.041	0.024		0.021	0.014		0.010	0.008	
		CP	90.8	89.0		94.2	95.5		93.5	93.5	
(0.2,0.5)	γ_0	Bias	0.071	0.074	1.01	0.059	0.061	0.99	0.051	0.052	1.00
		Var	0.122	0.122		0.081	0.080		0.059	0.059	
		Var*	0.104	0.132		0.076	0.090		0.058	0.075	
		CP	92.4	96.5		94.3	95.0		94.9	96.0	
	γ_1	Bias	-0.053	-0.049	0.94	-0.037	-0.034	0.96	-0.033	-0.032	0.99
		Var	0.082	0.077		0.073	0.070		0.060	0.060	
		Var*	0.068	0.084		0.061	0.071		0.055	0.068	
		CP	90.0	92.5		91.7	95.0		94.7	96.0	
	β	Bias	-0.039	-0.043	0.68	-0.012	-0.018	0.76	-0.005	-0.011	0.90
		Var	0.019	0.012		0.012	0.009		0.008	0.007	
		Var*	0.015	0.013		0.010	0.009		0.007	0.007	
		CP	88.9	90.5		91.6	96.0		92.5	96.0	
(0,0)	γ_0	Bias	0.042	0.039	1.00	0.043	0.041	1.00	0.060	0.059	1.00
		Var	0.054	0.054		0.056	0.056		0.050	0.050	
		Var*	0.051	0.061		0.051	0.059		0.053	0.062	
		CP	94.5	95.5		93.9	93.5		95.8	96.0	
	γ_1	Bias	-0.028	-0.045	1.03	-0.029	-0.040	1.02	-0.044	-0.047	1.01
		Var	0.055	0.055		0.058	0.059		0.053	0.053	
		Var*	0.050	0.062		0.051	0.059		0.051	0.060	
		CP	93.5	95.0		92.9	94.0		94.3	95.0	
	β	Bias	-0.006	0.053	1.22	-0.003	0.034	1.16	0.000	0.011	1.04
		Var	0.008	0.007		0.007	0.007		0.007	0.007	
		Var*	0.007	0.007		0.006	0.007		0.006	0.007	
		CP	93.7	89.0		93.3	92.0		92.5	91.5	

Table 5.5: Mean, empirical variance (Var), average of estimated variance (Var*) of the estimates of $(\hat{\rho}_1, \hat{\rho}_2)$. Var* are from 200 bootstrap samples.

covariates	cure rate		10/40			5/80			2/200		
			Mean	Var	Var*	Mean	Var	Var*	Mean	Var	Var*
	$(\zeta, \tau) = (0.4, 0.8)$										
discrete	high	$\hat{\rho}_1$	0.309	0.003	0.003	0.311	0.003	0.003	0.317	0.004	0.004
		$\hat{\rho}_2$	0.380	0.009	0.006	0.387	0.008	0.006	0.398	0.013	0.009
	low	$\hat{\rho}_1$	0.277	0.007	0.006	0.285	0.006	0.005	0.287	0.008	0.008
		$\hat{\rho}_2$	0.502	0.006	0.005	0.513	0.006	0.005	0.517	0.009	0.006
continuous	high	$\hat{\rho}_1$	0.320	0.003	0.003	0.324	0.003	0.003	0.323	0.005	0.005
		$\hat{\rho}_2$	0.432	0.008	0.006	0.439	0.008	0.006	0.445	0.013	0.009
	low	$\hat{\rho}_1$	0.281	0.011	0.010	0.288	0.009	0.008	0.292	0.014	0.012
		$\hat{\rho}_2$	0.538	0.007	0.004	0.539	0.007	0.005	0.538	0.010	0.007
	$(\zeta, \tau) = (0.2, 0.5)$										
discrete	high	$\hat{\rho}_1$	0.153	0.002	0.002	0.157	0.003	0.002	0.156	0.005	0.005
		$\hat{\rho}_2$	0.203	0.003	0.003	0.206	0.004	0.003	0.214	0.008	0.006
	low	$\hat{\rho}_1$	0.144	0.004	0.003	0.145	0.004	0.005	0.149	0.008	0.007
		$\hat{\rho}_2$	0.314	0.003	0.003	0.315	0.003	0.004	0.318	0.006	0.005
continuous	high	$\hat{\rho}_1$	0.158	0.002	0.002	0.162	0.002	0.002	0.165	0.005	0.005
		$\hat{\rho}_2$	0.244	0.004	0.003	0.245	0.004	0.003	0.249	0.008	0.006
	low	$\hat{\rho}_1$	0.139	0.006	0.004	0.141	0.006	0.004	0.146	0.011	0.009
		$\hat{\rho}_2$	0.326	0.005	0.003	0.335	0.005	0.003	0.333	0.008	0.005
	$(\zeta, \tau) = (0, 0)$										
discrete	high	$\hat{\rho}_1$	-0.003	0.0006	0.0005	-0.004	0.0013	0.0012	-0.003	0.005	0.005
		$\hat{\rho}_2$	0.034	0.0005	0.0004	0.032	0.0010	0.0008	0.034	0.004	0.004
	low	$\hat{\rho}_1$	-0.002	0.0006	0.0005	-0.004	0.0013	0.0011	-0.002	0.005	0.005
		$\hat{\rho}_2$	0.075	0.0005	0.0004	0.075	0.0010	0.0008	0.080	0.005	0.003
continuous	high	$\hat{\rho}_1$	-0.004	0.0005	0.0005	-0.003	0.0012	0.0012	0.001	0.005	0.005
		$\hat{\rho}_2$	0.048	0.0005	0.0004	0.049	0.0011	0.0008	0.049	0.005	0.003
	low	$\hat{\rho}_1$	-0.003	0.0006	0.0005	-0.004	0.0013	0.0010	-0.007	0.005	0.004
		$\hat{\rho}_2$	0.083	0.0006	0.0004	0.084	0.0012	0.0008	0.083	0.004	0.003

Besides the simulation with high cure rate ($(\gamma_0, \gamma_1) = (0.4, -1)$) we consider above, the cases with low cure rate, i.e., $(\gamma_0, \gamma_1) = (2.2, -1)$, are also investigated and summarized in Tables 5.3 and 5.4 which are corresponding to the binary covariate and continuous covariate, separately. The observation from these tables are similar to that in Tables 5.1 and 5.2. For example, the REs can be as low as 0.69 for γ_1 and 0.58 for β when the correlation in a cluster is strong. Therefore, the efficiency gain is maintained in the low cure rate cases too.

As in other estimating equations with working correlation matrices, ρ_1 and ρ_2 in the two working correlation matrices in the ES algorithm do not necessarily correspond to the correlation measures ζ and τ in the data generation. However, the estimated values of ρ_1 and ρ_2 provide good measures of the strength of the correlations between the cure statuses and between the failure times of uncured subjects in a cluster. That is, the stronger associations specified by ζ and τ in the data set indicate the larger correlation coefficients estimated by the proposed method in the working correlation matrices. Table 5.5 clearly shows that when the strengths of the correlation measures between the cure statuses and between the failure times of uncured patients in a cluster become strong, the estimated working correlation coefficients become large correspondingly. When (ζ, τ) reduce to $(0, 0)$, which implies that there is no correlation within a cluster, the estimates of ρ_1 and ρ_2 are also close to zero, which indicates that the working correlation matrices could be considered as the identity matrices.

5.5 Applications

5.5.1 Multi-Center Clinical Trial of Tonsil Carcinoma

We consider a data set from a tonsil cancer clinical trial study introduced in Section 1.2.4. A part of the data from the study is available in Kalbfleisch and Prentice (2003), which includes times (in days) from diagnosis to death of 195 patients with squamous cell carcinoma of three sites in the oropharynx between 1968 and 1972 in six participating institutions. Other variables include censoring indicator, treatment, sex, tumor stage (a binary variable with 1 for T_4 stage corresponding to a massive invasive tumor and 0 for T_1, T_2 and T_3 stages corresponding to a primary tumor measuring 2cm or less in the largest diameter, a primary tumor measuring 2 to 4cm in the largest diameter, or a primary tumor measuring more than 4cm), node stage, age, general condition (0 for no disability and 1 for the cases including restricted work capability, requiring assistance with self-care or bed confined), grade (1, 2, and 3 for well, moderate, and poorly differentiated respectively), and the institution code. We delete observations from 3 patients because of the presence of missing values and the actual number of patients analyzed is 192. As we discussed in Section 1.2.4, the Kaplan-Meier survival curve (Figure 1.4) suggests that the cure fraction should be considered in the model for the data.

Yau and Ng (2001) and Lai and Yau (2008) considered a mixture cure model with random effects for the data. However, they only analyzed the effect of the dichotomized T-stage on the cure probability and on the failure time distribution of uncured patients based on a subset (carcinoma of the pharyngeal tongue) of the data.

We apply the proposed marginal mixture cure model in the previous sections to the data. The covariates in the model include treatment, sex, grade, age, condition, and tumor stage, and they are considered in both (5.2) and (5.3). We assume the

exchangeable correlation structure for both the cure statuses and the failure times of uncured patients from one institution. The standard errors of the estimated parameters are obtained from 500 bootstrap samples. As a comparison, we also fit the data with the marginal mixture cure model using the PTY method. All results are summarized in Table 5.6.

Table 5.6: Estimated parameters from fitting the marginal mixture cure model to the tonsil cancer data using the ES algorithm and the PTY method.

Covariate	ES		PTY	
	$\hat{\theta}^*$	$\frac{\hat{\theta}^*}{s.e.(\hat{\theta}^*)}$	$\hat{\theta}^*$	$\frac{\hat{\theta}^*}{s.e.(\hat{\theta}^*)}$
PH Survival Model				
Treatment: (test vs. standard)	0.157	0.174	0.107	0.081
Sex: (female vs. male)	-0.439	-1.115	-0.385	-0.157
Grade 2 (vs. Grade 1)	-0.295	-0.802	-0.217	-0.161
Grade 3 (vs. Grade 1)	0.245	0.632	0.148	0.583
Age	-0.011	-1.154	-0.009	-0.351
Condition	1.660	7.048	1.724	0.619
Tumor	0.640	2.558	0.924	0.468
ρ_2	0.095	4.476	-	-
Logistic Model				
Intercept	-0.388	-0.156	-0.487	-0.047
Treatment: (test vs. standard)	-0.141	-0.354	-0.105	-0.193
Sex: (female vs. male)	-0.388	-0.429	-0.436	-0.173
Grade 2 (vs. Grade 1)	1.192	0.788	1.163	0.126
Grade 3 (vs. Grade 1)	-0.817	-0.963	-0.750	-0.079
Age	0.030	0.773	0.035	0.918
Condition	0.609	0.435	0.454	0.280
Tumor	0.102	0.086	-0.108	-0.063
ρ_1	0.007	0.443	-	-

The results from the two methods show some substantial differences. For example, condition effect (p-value$<$0.001) and tumor stage effect (p-value=0.011) on the failure time of uncured patients become highly significant in the ES method instead of insignificant in the PTY method. That is, the proposed model suggests that, if not cured, patients with disability or with massive invasive tumors tend to have shorter

failure times than those without massive invasive tumors.

Yau and Ng (2001) concluded based on their single-covariate model and a subset analysis of data that there is no significant correlation induced by the institution among the cure statuses and the failure times of uncured patients from the same institution. Table 5.6 shows that our model, which is based on the whole data and multiple covariates, suggests that the correlation induced by the institution among the failure times of uncured patients cannot be ignored.

5.5.2 Multi-Center Leukemia Data

We also apply the proposed method to the bone marrow transplantation (BMT) data (Section 1.2.5) which has been studied by Lai and Yau (2008). Several potential risk factors were collected at the time of transplantation. They are AML high-risk, AML low-risk, ALL, recipient and donor gender, recipient and donor age, recipient and donor cytomegalovirus immune status (CMV), waiting time from diagnosis to transplantation, and, for AML patients, their French-American-British (FAB) classification based on standard morphological criteria. Specifically, as pointed by Copelan et al. (1991), the risk of relapse or treatment-related death for patients with FAB classification of M4 or M5 was higher than that for patients in other FAB groups. Details of the study can be found in Copelan et al. (1991).

Based on the Kaplan-Meier survival curve (Figure 1.5) in Section 1.2.4, both the cure fraction and the cluster effect should be considered in the model introduced in Section 5.2. We consider the covariates AML low-risk, AML high-risk and FAB in both the logistic and the proportional hazards regression components. The standard errors of the estimated parameters are obtained from 500 bootstrap samples as we did in Section 5.5.1. All results are summarized in Table 5.7.

The results based on the two estimation methods are generally similar. The

Table 5.7: Estimated parameters from fitting the marginal mixture cure model to the leukemia data using the ES method and the PTY method.

	ES method		PTY method	
Covariate	$\hat{\theta}^*$	$\frac{\hat{\theta}^*}{s.e.(\hat{\theta}^*)}$	$\hat{\theta}^*$	$\frac{\hat{\theta}^*}{s.e.(\hat{\theta}^*)}$
PH Survival Model				
AML low-risk (vs. ALL)	0.338	0.636	-0.670	-0.959
AML high-risk (vs. ALL)	1.121	2.905	0.438	1.646
FAB	-0.133	-0.411	-0.044	-0.071
ρ_2	0.060	1.293	-	-
Logistic Model Intercept	1.768	2.458	0.709	4.838
AML low-risk (vs. ALL)	-2.109	-2.589	-0.989	-2.070
AML high-risk (vs. ALL)	-1.348	-1.000	-0.295	-1.212
FAB	1.497	1.038	1.440	1.596
ρ_1	0.005	0.154	-	-

noticeable difference is in the effect of AML high-risk in the proportional hazards component. That is, the effect of AML high-risk is highly significant (p-value=0.004) in the ES method instead of marginally significant (p-value=0.01) in the PTY method. In other words, the uncured patients of AML high-risk are at a higher risk of death or relapse comparing with the patients in other two groups (AML low-risk or ALL). The effect of AML low-risk on the logistic component is marginally significant in both methods. It indicates that the patients of AML low-risk may have higher chance of being cured. The effects of the remaining covariates are similar in the two models. The similarity of the estimates from the two methods may indicate that the correlation within clusters may not be strong enough to make differences in parameter estimates. This is evident from the estimates of ρ_1 and ρ_2. Both values are close to zero and their large variances make the correlation insignificant. Lai and Yau (2008) also concluded that there are no significant differences in the cured proportion and in the survival for the uncured patients between the participating clinics though a high-positive correlation between the random effects was obtained.

5.6 Conclusions

We considered a semiparametric marginal proportional hazards mixture model for clustered failure time data with a possible cure fraction and proposed a novel approach based on the generalized estimating equations to incorporate a correlation structure in the marginal model. Our method generalizes the parametric marginal proportional hazards mixture cure model investigated in Chapter 4 to the semiparametric one for clustered survival data with a cure fraction. The proposed method also extends the existing marginal proportional hazards mixture cure model (Peng et al., 2007) by explicitly including the correlation structures in the model estimation. A simulation study demonstrates that the proposed method can substantially improve the estimation efficiency compared to the method in Peng et al. (2007) when the cluster size is large and the correlation within a cluster is strong. These two methods are comparable when the cluster size or the strength of the correlation decreases. Therefore, the proposed semiparametric marginal PH mixture model is a useful alternative to the existing marginal models for clustered data with a possible cure fraction, particularly when the correlation structures among the cure statuses and among the failure times of uncured patients can be specified up to a few unknown parameters.

We employ the bootstrap method to estimate the variances of the estimated parameters in the model. This method is straightforward but computationally intensive. Future work for this model includes the asymptotic properties of the estimates (particularly their asymptotic variance estimation). As we discussed in Section 4.6, we will also consider methods to include a correlation structure in estimating the parameters in the baseline survival distribution and hope to further improve the estimation efficiency. Since the random effects approach is widely used in the proportional hazards mixture cure model, the performance of the proposed approach comparing with

the existing random effects approach deserves a further study.

A paper (Niu and Peng, 2012) based on the main results of this chapter has been accepted by Statistics in Medicine.

Chapter 6

Summary and Future Work

In this thesis, we proposed novel marginal methods, based on the proportional hazards model and the proportional hazards mixture cure model, for modeling clustered survival data with/without a cure fraction. We developed a set of estimating equations to accommodate the potential correlation within clusters through flexible working correlation structures such as the exchangeable working matrix.

Motivated by Segal and Neuhaus (1993), in Chapter 3, we investigated the classical clustered failure time data without a cure fraction by a semiparametric marginal proportional hazards model. The dependence among failure times within a cluster are modeled explicitly by an exchangeable working correlation matrix through a new unbiased weighted estimating equation. We showed that the regression estimators from the proposed estimating equation are consistent and asymptotically normal under some regularity conditions. The variance estimates have a closed form and can be consistently estimated by a sandwich method. The finite sample properties were investigated by a simulation study which shows that the estimators of the regression parameters based on the proposed estimating equation are more efficient than those with the existing method (Lee et al., 1992).

In Chapter 4, we extended the marginal proportional hazards model for clustered

failure time data without a cure fraction to a marginal proportional hazards mixture cure model for clustered survival data with a cure fraction. The baseline survival function was assumed to follow the Weibull distribution. Yu and Peng (2008) considered the same marginal model for a smoking cessation study. They applied the EM algorithm to handle the missing information in the estimation procedure and provided jackknife variance estimates of the parameters in the model. Their estimation method is robust to misspecification of the correlation structure but may incur a substantial efficiency loss of the parameters when there is information available for the correlation structure. In this chapter, we extended the EM algorithm to the ES algorithm to handle the substantial correlation among the cure statuses and among the failure times of uncured patients in one cluster. Specifically, the S-step in the ES algorithm for the regression parameters in the survival function of the susceptible group inherits the generalized estimating equation approach as we did in Chapter 3. Following the same idea, to accommodate the correlation among cure statues, we also proposed an estimating equation for the regression parameters in the logistic model of the incidence in the S-step. We proved that the proposed estimating functions are unbiased and the corresponding estimators are consistent and asymptotically normal under some regularity conditions.

In Chapter 5, we generalized the parametric marginal proportional hazards mixture cure model to the semiparametric one where the baseline survival function is nonparametrically specified for modeling clustered survival data with a cure fraction. Peng et al. (2007) considered the same model for a multi-institutional tonsil cancer data. They provided the robust variance estimates of parameters but did not explicitly model the correlation within an institution in the study. Alternatively, we proposed two sets of estimating equations for the regression parameters in the ES algorithm as we did in Chapter 4. However, the unweighted estimating equation for

the regression parameters in the survival function of uncured patients produces biased estimates. We standardized the corresponding covariates to reduce such biases. A bootstrap method was applied to obtain the variance estimates because of the bias of the estimating function of β and the complexity of the nonparametric baseline estimation in the proposed ES algorithm.

Based on the extensive simulation studies in Chapters 3, 4, and 5, we considered the substantial improvement of estimation efficiency as the contribution of the proposed methods comparing with the existing marginal methods for modeling clustered failure time data, especially when the correlation within cluster is strong and the cluster size is large given the total number of observations. The applications to the real data sets from the biomedical research demonstrate that the proposed methods are feasible in the practical applications. Therefore, the proposed marginal methods are useful alternatives to the existing marginal methods for analyzing clustered survival data with/without a cure fraction.

We would like to point out that, unlike the weighted estimating functions of β in Chapters 3 and 4, the estimating function we proposed for β in Chapter 5 is biased. We suggested to standardize the corresponding covariates to reduce the biases in estimates. Future work include applying the unbiased weighted estimating equation approach for β in the semiparametric marginal proportional hazards mixture cure model and developing the asymptotic properties of the estimates.

As we mentioned in the conclusions of Chapters 4 and 5, the estimates of baseline parameters in the marginal proportional hazards mixture cure models are based on the independent estimating equations. Including a correlation structure in estimating the parameters in the baseline survival distribution to further improve the estimation efficiency is under consideration. The comparison between the marginal models and the random effects models is an interesting topic in simulation studies.

Bibliography

[1] S. Banerjee and B. P. Carlin. Parametric spatial cure rate models for interval-censored time-to-relapse data. *Biometrics*, 60:268–275, 2004.

[2] J. Berkson and R. P. Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47:501–515, 1952.

[3] J. W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B*, 11:15–44, 1949.

[4] K. By and B. Qaqishi. 'mvtBinaryEP': An R package to generate correlated binary data, version 1.0.1. 2011.

[5] J. Cai and R. L. Prentice. Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika*, 82:151–164, 1995.

[6] J. Cai and R. L. Prentice. Regression analysis for correlated failure time data. *Lifetime Data Analysis*, 3:197–213, 1997.

[7] R. J. Carroll, D. Ruppert, and L. A. Stefanski. *Measurement Error in Nonlinear Models*. Chapman & Hall, New York, 1995.

[8] N. Chatterjee and J. Shih. A bivariate cure-mixture approach for modeling familial association in diseases. *Biometrics*, 57:779–786, 2001.

[9] C. M. Chen and T. F. C. Lu. Marginal analysis of multivariate failure time data with a surviving fraction based on semiparametric transformation cure models. *Computational Statistics & Data Analysis*, 56:645–655, 2012.

[10] Y. Chen, K. Chen, and Z. Ying. Analysis of multivariate failure time data using marginal proportional hazards model. *Statistica Sinica*, 20:1025–1041, 2010.

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

[12] D. G. Clayton and J. Cuzick. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A*, 148:82–117, 1985.

[13] L. X. Clegg, J. Cai, and P. K. Sen. A marginal mixed baseline hazards model for multivariate failure time data. *Biometrics*, 55:805–812, 1999.

[14] M. Cleves, W. Gould, R. Gutierrez, and Y. Marchenko. *An Introduction to Survival Analysis Using Stata*. Stata Press, Texas, 2nd edition, 2008.

[15] E. A. Copelan, J. C. Biggs, J. M. Thompson, P. Crilley, J. Szer, J. P. Klein, N. Kapoor, B. R. Avalos, I. Cunningham, K. Atkinson, K. Downs, G. S. Harmon, M. B. Daly, I. Brodsky, S. I. Bulova, and P. J. Tutschka. Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with Bu/Cy2. *Blood*, 78:838–843, 1991.

[16] D. R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.

[17] D. R. Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975.

[18] M. Crowder. On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, 82:407–410, 1995.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[20] A. J. Dobson. *An Introduction to Generalized Linear Models*. Chapman & Hall, Boca Raton, 2nd edition, 2002.

[21] B. Efron and T. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[22] L. J. Emrich and M. R. Piedmonte. A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45:302–304, 1991.

[23] V. T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:1041–1046, 1982.

[24] V. T. Farewell. Mixture models in survival analysis: Are they worth the risk? *The Canadian Journal of Statistics*, 14:257–262, 1986.

[25] R. J. Gray and Y. Li. Optimal weight functions for marginal proportional hazards analysis of clustered failure time data. *Lifetime Data Analysis*, 8:5–19, 2002.

[26] Diabetic Retinopathy Study Research Group. Diabetic retinopathy study. *Investigative Ophthalmology and Visual Science*, 21, Part 2:149–226, 1981.

[27] D. B. Hall and Z. Zhang. Marginal models for zero inflated clustered data. *Statistical Modelling*, 4:161–180, 2004.

[28] P. Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73:387–396, 1986a.

[29] P. Hougaard. A class of multivariate failure time distribution. *Biometrika*, 73:671–678, 1986b.

[30] P. Hougaard. Frailty models for survival data. *Lifetime Data Analysis*, 1:255–273, 1995.

[31] P. Hougaard. *Analysis of Multivariate Survival Data*. Springer-Verlag, New York, 2000.

[32] P. J. Huber. The behaviou of maximum likelihood estmates under nonstandard conditions. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–233, 1967.

[33] W. J. Huster, R. Brookmeyer, and S. G. Self. Modelling paired survival data with covariates. *Biometrics*, 84:1065–1073, 1989.

[34] M. Jamshidian and R. I. Jennrich. Standard errors for EM estimation. *Journal of the Royal Statistical Society, Series B*, 62:257–270, 2000.

[35] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Hoboken, 2nd edition, 2002.

[36] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2nd edition, 2003.

[37] A. Y. C. Kuk and C. H. Chen. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79:531–541, 1992.

[38] X. Lai and K. K. W. Yau. Long-term survivor model with bivariate random effects: Applications to bone marrow transplant and carcinoma study data. *Statistics in Medicine*, 27:5692–5708, 2008.

[39] E. W. Lee, L. J. Wei, and D. A. Amato. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. *Survival Analysis: State of the Art*, eds. J. P. Klein and P. K. Goel:237–247, 1992.

[40] C. S. Li and J. M. G. Taylor. A semi-parametric accelerated failure time cure model. *Statistics in Medicine*, 21:3235–3247, 2002.

[41] K.-Y. Liang, S. G. Self, and Y. C. Chang. Modeling for marginal hazards in multivariate failure-time data. *Journal of the Royal Statistical Society, Series B*, 55:441–453, 1993.

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

[43] D. Y. Lin. Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine*, 13:2233–2247, 1994.

[44] T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, B*, 44:226–233, 1982.

[45] W. Lu and Z. Ying. On semiparametric transformation cure models. *Biometrika*, 91:331–343, 2004.

[46] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, London, 2nd edition, 1989.

[47] C. A. McGilchrist. REML estimation for survival models with frailty. *Biometrics*, 49:221–225, 1993.

[48] C. A. McGilchrist and C. Aisbett. Regression with frailty in survival analysis. *Biometrics*, 47:461–466, 1991.

[49] X.-L. Meng and D. B. Rubin. Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, 86:899–909, 1991.

[50] J. Monaco, J. Cai, and J. Grizzle. Bootstrap analysis of multivariate failure time data. *Statistics in Medicine*, 24:3387–3400, 2005.

[51] R. P. Murray, N. R. Anthonisen, J. E. Connett, R. A. Wise, P. G. Lindgren, P. G. Greene, and M. A. Nides (for the Lung Health Study Research Group). Effects of multiple attempts to quit smoking and relapses to smoking on pulmonary function. *Journal of Clinical Epidemiology*, 51:1317–1326, 1998.

[52] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical, Society A*, 135:370–384, 1972.

[53] Y. Niu and Y. Peng. A semiparametric marginal mixture cure model for clustered survival data. *Statistics in Medicine*, published online DOI: 10.1002/sim5687, 2012.

[54] Y. Peng. Fitting semiparametric cure models. *Computational Statistics & Data Analysis*, 41:481–490, 2003.

[55] Y. Peng and K. B. G. Dear. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56:237–243, 2000.

[56] Y. Peng, K. B. G. Dear, and J. W. Denham. A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, 17:813–830, 1998.

[57] Y. Peng and J. M. G. Taylor. Mixture cure model with random effects for the analysis of a multi-center tonsil cancer study. *Statistics in Medicine*, 30:211–223, 2011.

[58] Y. Peng, J. M. G. Taylor, and B. Yu. A marginal regression model for multivariate failure time data with a surviving fraction. *Lifetime Data Analysis*, 13:351–369, 2007.

[59] R. L. Prentice and L. Hsu. Regression on hazard ratios and cross ratios in multivariate failure time analysis. *Biometrika*, 84:349–363, 1997.

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

[61] Y. Ritov. Estimation in a linear regression model with censored data. *The Annals of Statistics*, 18:303–328, 1990.

[62] O. Rosen, W. Jiang, and M. A. Tanner. Mixtures of marginal models. *Biometrika*, 87:391–404, 2000.

[63] M. R. Segal and J. M. Neuhaus. Robust inference for multivariate survival data. *Statistics in Medicine*, 12:1019–1031, 1993.

[64] M. R. Segal, J. M. Neuhaus, and I. R. James. Dependence estimation for marginal models of multivariate survival data. *Lifetiem Data Analysis*, 3:251–268, 1997.

[65] C. F. Spiekerman and D. Y. Lin. Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association*, 93:1164–1175, 1998.

[66] J. A. Stoner and B. G. Leroux. Analysis of clustered data: a combined estimating equations approach. *Biometrika*, 89:567–578, 2002.

[67] J. P. Sy and J. M. G. Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, 56:227–236, 2000.

[68] J. M. G. Taylor. Semi-parametric estimation in failure time mixture models. *Biometrics*, 51:899–907, 1995.

[69] T. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, 2000.

[70] L. J. Wei, D. Y. Lin, and L. Weissfeld. Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association*, 84:1065–1073, 1989.

[71] A. Wienke, P. Lichtenstein, and A. I. Yashin. A bivariate frailty model with a cure fraction for modeling familial correlations in diseases. *Biometrics*, 59:1178–1183, 2003.

[72] K. Yamaguchi. Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of 'permanent employment' in Japan. *Journal of the American Statistical Association*, 87:284–292, 1992.

[73] Y. Yang and Z. Ying. Marginal proportional hazards models for multiple event-time data. *Biometrika*, 88:581–586, 2001.

[74] K. K. W. Yau and A. S. K. Ng. Long-term survivor mixture model with random effects: application to a multi-centre clinical trial of carcinoma. *Statistics in Medicine*, 20:1591–1607, 2001.

[75] B. Yu and Y. Peng. Mixture cure models for multivariate survival data. *Computational Statistics & Data Analysis*, 52:1524–1532, 2008.

[76] K.-H. Yuan and R. J. Jennrich. Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, 65:245–260, 1998.

[77] J. Zhang and Y. Peng. A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine*, 26:3157–3171, 2007.