ABSTRACT

SONG, XIAO. Semiparametric approaches to inference in joint models for longitudinal and time-to-event data. (Under the direction of Marie Davidian and Anastasios A. Tsiatis)

In many longitudinal studies, it is of interest to characterize the relationship between a time-to-event (e.g. survival) and time-dependent and time-independent covariates. Time-dependent covariates are generally observed intermittently and with error.

For a single time-dependent covariate, a popular approach is to assume a joint longitudinal data-survival model, where the time-dependent covariate follows a linear mixed effects model and the hazard of failure depends on random effects and time-independent covariates via a proportional hazards relationship. Interest may focus on inference on the longitudinal data process, which is informatively censored by death or withdrawal, or on the hazard relationship. Several methods for fitting such models have been proposed, including regression calibration and likelihood or Bayesian methods. However, most approaches require a parametric distributional assumption (normality) on the random effects. In addition, generalization to more than one time-dependent covariate may become prohibitive. For a single time-dependent covariate, Tsiatis and Davidian (2001) have proposed an approach that is easily implemented and does not require an assumption on the distribution of the random effects. We extend this technique to multiple, possibly correlated, time-dependent covariates. This approach is easy to compute.

However, the conditional score approach might be less efficient relative to the

likelihood approaches. In addition, inference on the longitudinal data process is not available. To improve the efficiency and meanwhile obtain an estimator for the random effects distribution, we propose to approximate the random effects distribution by the seminonparametric (SNP) densities of Gallant and Nychka (1987), which requires only the assumption that the random effects have a "smooth" density, and take a semiparametric likelihood approach. The EM algorithm is used for implementation. We demonstrate the approaches via simulations and apply them to data from an HIV clinical trial.

SEMIPARAMETRIC APPROACHES TO INFERENCE IN JOINT MODELS FOR LONGITUDINAL AND TIME-TO-EVENT DATA

by

XIAO SONG

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

STATISTICS

Raleigh

2002

APPROVED BY:

Professor Marie Davidian Co-Chair of Advisory Comittee	Professor Anastasios A. Tsiatis Co-Chair of Advisory Comittee
Professor Daowen Zhang	Professor Sujit Ghosh
Professor Charles E. Smith	

To my family

Biography

SONG, XIAO was born in Pingxiang, China on March 25, 1971. She entered the Department of Mathematics at Peking University in 1988 and obtained her B.S. and M.S. in Mathematics in 1992 and 1995, respectively. Then she worked as a software engineer in the Institute of Computing Science and Technology, Peking University. Three years later, in August 1998, she joined the Statistics Graduate Program at North Carolina State University. She obtained her M.S. in statistics in May 2000 and continued to pursue her Ph.D. in statistics at North Carolina State University.

Acknowledgements

I would like to express my deepest gratitude and appreciation to my advisors Dr. Marie Davidian and Dr. Anastasios A. Tsiatis for their guidance, encouragement and support in my dissertation research. I would also like to thank the other committee members, Dr. Daowen Zhang, Dr. Sujit Ghosh and Dr. Charles E. Smith for their careful reading of this manuscript and their helpful comments.

It has been a very pleasant experience to study in this department. I would like to convey my gratitude to all faculty and staff for their help and assistance. Thanks also go to my fellow students here.

Finally, I would like to thank my family, especially my husband Hao, for their love, understanding, encouragement and support.

Contents

List of Tables **vii**

List of Figures **viii**

1 Introduction **1**
1.1 Introduction 1
1.2 Proportional Hazards Model 2
1.3 Naive Approaches 4
1.4 Joint Model Approach 4
1.4.1 Model Definition 5
1.4.2 Past Approaches 9
1.4.3 This Research 13

2 Conditional Score Approach **16**
2.1 Introduction 16
2.2 Conditional Score Estimator 17
2.3 Large Sample Properties 21
2.4 Technical Details 22
2.4.1 Derivation of (2.2) 22
2.4.2 Derivation of (2.5) 24
2.4.3 Heuristic proof of asymptotic properties of the conditional score estimator 25

3 Application and Performance of the Conditional Score Approach **28**
3.1 Application to AIDS Clinical Trials Group 175 28
3.2 Simulation Studies 37
3.3 Discussion 40

4 Semiparametric Likelihood Approach **51**
4.1 Introduction 51
4.2 Model and Density Approximation 52
4.3 Estimation Procedure 56

4.3.1 Likelihood Function and Existence of Estimators 56
4.3.2 EM Algorithm 57
4.3.3 Choosing K 61
4.4 Technical details 62
4.4.1 Proof of Theorem 1 62
4.4.2 Proof of Theorem 2 63
4.4.3 Computation of $E_{ik}\{g(\alpha_i)\}$ 64

5 Application and Performance of the Semiparametric Likelihood Approach 66
5.1 Simulation Studies 66
5.2 Application to ACTG 175 71
5.3 Discussion 74

Biography 83

List of Tables

3.1 Results for several models fitted to the ACTG 175 data 42
3.2 Diagnostic checks of the proportional hazards assumption in Model 2.5 42
3.3 Fits of Models 1, 2.5 and 2.5*, and 3.5 and 3.5* for the ACTG 175 data 43
3.4 Simulation results for sample size $n = 250$ 44
3.5 Simulation results for sample size $n = 500$ 45
3.6 Simulation results for short-tailed (truncated normal) error 46
3.7 Simulation results for nonlinear $G(u, \alpha_i)$ 47
3.8 Simulation results for the estimation of ω^T 48

5.1 Simulation results, hazard parameters 76
5.2 Simulation results, means and covariance matrix of random effects distribution and intra-subject variance using the SNP method 77
5.3 Fit of the model in (4.1) and (4.4) to the ACTG 175 data 78

List of Figures

1.1 Plots for ACTG data 15

3.1 Trajectories of log CD4 and log CD8 49
3.2 Q-Q plots for studentized residuals for log CD4 and log CD8 50

5.1 Simulation results for the bimodal case 79
5.2 Monte Carlo average of estimated baseline cumulative hazard functions 80
5.3 Fit of Model (4.1),(4.4) to the ACTG data using the SNP estimator . 81
5.4 Plots of $\widehat{\alpha}_{i1}$ versus various covariates for the ACTG 175 data 82

Chapter 1

Introduction

1.1 Introduction

In longitudinal studies, information is often collected on a time-to-event (e.g. survival, failure, time to progression of disease) and time-dependent and time-independent covariates. A routine objective is to model the association between survival and the time-independent and time-dependent covariates, which is usually placed in the framework of the proportional hazards model (Cox, 1972). Knowledge of true values of the time-dependent covariates at each unique failure time is a requirement for implementation; however, longitudinal covariates are collected intermittently and are subject to error.

Such modeling is often carried out to identify covariates, both time-independent and -dependent, that are associated with prognosis. An additional objective might be to assess the value of time-dependent covariates as potential surrogate markers (Pren-

tice, 1989), which is often done within the context of these models by examining the interrelationship between time-dependent covariates and treatment effect. A further goal may be to characterize the longitudinal covariate process, which is complicated by the fact that subjects may experience the event or withdraw from the study.

An application where this approach has seen widespread use is in the analysis of HIV clinical trials. An example is AIDS Clinical Trials Group (ACTG) 175, a randomized clinical trial to compare four therapies, zidovudine alone, zidovudine plus didanosine, zidovudine plus zalcitabine, or didanosine alone, in HIV-infected subjects (Hammer *et al.*, 1996). Between December 1991 and October 1992, 2467 subjects were recruited and followed until November 1994. CD4 and CD8 counts, both changed over time (time-dependent) and were reflections of immune status, were measured for each participant about every 12 weeks after randomization, and the time-to-event endpoint was time to progression to AIDS or death. We will use this example in the sequel to illustrate these data-analytic objectives and to demonstrate the methods we will develop.

1.2 Proportional Hazards Model

A standard framework to model the relationship between a time-to-event process and covariates is the proportional hazards model (Cox, 1972). For each subject i, $i = 1, \ldots, n$, let T_i denote failure time and C_i denote censoring time. The observed survival data are $V_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function; these and all other variables are independent across i. The hazard of failure

is assumed to depend on possibly time-dependent covariates $X_i(u)$ through the form

$$\begin{aligned}\lambda_i(u) &= \lim_{du\to 0} du^{-1}\mathrm{pr}\{u \le T_i < u+du | T_i \ge u, X_i^H(u)\} \\ &= \lambda_0(u)\exp\left[\beta^T g\{u, X_i(u)\}\right] \qquad (1.1)\end{aligned}$$

where $X_i^H(u) = \{X_i(t), t \le u\}$ is the history of $X_i(u)$ up to time u for subject i, $\lambda_0(u)$ is an unspecified baseline hazard function, β is a vector of regression parameters, and $g\{u, X_i(u)\}$ is a function of the same dimension. This is a semiparametric model, as the baseline hazard function is not taken to have a parametric form. Censoring is assumed to be noninformative, which usually means that C_i and T_i are independent conditional on the history of the covariates. Cox (1972) suggested to estimate β based on

$$\prod_{i=1}^{n}\left(\frac{\exp[\beta^T g\{V_i, X_i(V_i)\}]}{\sum_{j=1}^{n}\exp[\beta^T g\{V_i, X_j(V_i)\}]I(V_j \ge V_i)}\right)^{\Delta_i}, \qquad (1.2)$$

which he called partial likelihood (Cox 1976). Let the counting process increment be $dN_i(u) = I(u \le V_i < u+du, \Delta_i = 1)$ and let $Y_i(u) = I(V_i \ge u)$ denote the "at risk" process. The partial likelihood score equations are

$$\sum_{i=1}^{n}\int\left(g\{u, X_i(u)\} - \frac{\sum_{j=1}^{n} g\{u, X_i(u)\}\exp[\beta^T g\{u, X_i(u)\}]Y_j(u)}{\sum_{j=1}^{n}\exp[\beta^T g\{u, X_j(u)\}]Y_j(u)}\right)dN_i(u) = 0. \qquad (1.3)$$

Under some regularity conditions, there exists a consistent and asymptotically normal solution $\hat{\beta}$ for (1.3). Johansen (1983) demonstrated that (1.2) is the likelihood profile in the sense that (1.2) $= \max_{\Lambda_0} L(\beta, \Lambda_0)$, where $L(\beta, \Lambda_0)$ is the joint likelihood for the regression coefficient β and the baseline cumulative hazard $\Lambda_0(u) = \int_0^u \lambda_0(u)du$. The domain of $\Lambda_0(u)$ was extended to allowing for multiple jumps in $L(\beta, \Lambda_0)$. The model (1.1) may be extended to include time-independent covariates Z_i, say. The above results are still applicable.

As noted above, knowledge of the true history of time-dependent covariates is required to implement the Cox model; however, time-dependent covariates, such as the CD4 and CD8 counts in the ACTG 175, are usually measured intermittently and with possible measurement error. The usual analysis is complicated by the intermittent, error-prone nature of the time-dependent covariates.

1.3 Naive Approaches

Early approaches attempted to imputed the values of time-dependent covariates at each failure time by what are now regarded as naive methods. These include "last value carried forward," which substitutes the last covariate value prior to the failure time, and "naive regression," where least squares estimates based on each subject's entire longitudinal profile are substituted. Such methods may perform adequately in some situations (e.g. Raboud *et al.*, 1993), but can lead to biased and misleading inference on the Cox model parameters (Prentice, 1982; Tsiatis and Davidian, 2001) and biased estimates of treatment effects and thus erroneous conclusions on surrogacy (Dafni and Tsiatis, 1998).

1.4 Joint Model Approach

A popular approach is base on a so-called "joint model" for the longitudinal covariate and time-to-event. Specifically, the model assumes that the longitudinal covariate data follow a linear mixed effects model and that the hazard of failure depends both on the random effects and other time-independent covariates through a proportional

hazards relationship. A general model definition on which this research is based is given in the following.

As noted by Hogan and Laird (1997), two issues may be of interest. Characterizing changes over time in the longitudinal data process (e.g., change in CD4 in different treatment groups) may be the primary objective, inference on which is complicated by the informative censoring of longitudinal profiles by the event or study withdrawal. (e.g., Wu and Carroll, 1988). Failure to take this into account may result in biased inferences on the longitudinal data process. Alternatively, the survival endpoint (e.g., time to AIDS or death) may be the main focus, and interest is in elucidating the relationship of prognosis to the longitudinal process and other covariates, such as treatment group.

1.4.1 Model Definition

For each subject i, $i = 1, \ldots, n$, let T_i denote failure time and C_i denote censoring time. The observed survival data are $V_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function; these and all other variables are independent across i. Let $X_{ik}(u)$, $k = 1, \ldots, M$, denote M time-dependent covariates at time u, and let the r-dimensional vector Z_i denote r time-independent covariates for subject i. We assume each covariate process $X_{ik}(u)$ satisfies

$$X_{ik}(u) = \alpha_{ik}^T f_k(u), \tag{1.4}$$

where $f_k(u)$ is a $(q_k \times 1)$ vector of functions of u, α_{ik} is a $(q_k \times 1)$ random effect, and f_k and α_{ik} may be different for each k. This allows great flexibility in modeling the time

trajectory of each covariate. E.g. $\alpha_{ik}^T f_k(u) = \alpha_{ik0} + \alpha_{ik1}u + \cdots + \alpha_{ik(q_k-1)}u^{q_k-1}$; spline models are also possible. The covariate processes $X_{ik}(u)$ are not observed directly; rather, longitudinal measurements W_{ikj} on the kth covariate are taken at the set of times t_{ikj}, $j = 1, 2, \ldots, m_{ik}$, for each i, where

$$W_{ikj} = X_{ik}(t_{ikj}) + e_{ikj}, \tag{1.5}$$

and e_{ikj} are normally-distributed, mean-zero "errors" with variance σ_{kk} that may reflect both biological variation and measurement error. Thus, (1.5) with (1.4) is a linear mixed effects model, and $X_{ik}(u)$ may be regarded as the "inherent" trajectory for subject i and covariate k.

We assume that the available measurements are sufficiently separated in time that serial correlation associated with within-subject biological variation is negligible; however, for the hazard formulation given below, this could be relaxed. We assume that measurements on different covariates taken at the same time might be correlated. More formally, noting that j indexes measurement times separately for each k, we may write for $k, k' = 1, \ldots, M$, $j = 1, \ldots, m_{ik}$, and $j' = 1, \ldots, m_{ik'}$,

$$\mathrm{cov}(e_{ikj}, e_{ik'j'}) = \sigma_{kk'} I(t_{ikj} = t_{ik'j'}),$$

where $\sigma_{kk'}$ is the covariance between errors from covariates k and k' measured at the same time point, reflecting correlation of components of within-subject biological variation, the measurement error, or both; of course, this formulation subsumes the case $\sigma_{kk'} = 0$ for all $k \neq k'$. Ordinarily, measurement errors across k might be thought of as uncorrelated; however, we allow the possibility that some covariate measurements may be derived from a common blood sample or measuring technique,

e.g. CD4 and CD8.

Let $e_i = (e_{i1}^T, \ldots, e_{iM}^T)^T$, where $e_{ik} = (e_{ik1}, \ldots, e_{ikm_{ik}})^T$; $t_{ik} = (t_{ik1}, \ldots, t_{ikm_{ik}})^T$ be the ordered times for subject i, covariate k, and $t_i = (t_{i1}^T, \ldots, t_{iM}^T)^T$ be the set of time points where observations on all M covariates are available; $m_i = (m_{i1}, \ldots, m_{iM})^T$; and $\alpha_i = (\alpha_{i1}^T, \ldots, \alpha_{iM}^T)^T$ $(q \times 1)$, $q = \sum_k q_k$. Assume that the conditional distribution of e_i given $(T_i, C_i, \alpha_i, Z_i, t_i, m_i)$ is normal with covariance matrix depending only on m_i and the parameters $\sigma_{kk'}$. The α_{ik} may be correlated across k; however, this will be of no consequence for the methods we propose in Chapter 2.

A proportional hazards regression model is assumed for the relationship between the hazard of failure and the covariates; that is, the hazard for subject i is

$$\begin{aligned} \lambda_i(u) &= \lim_{du \to 0} du^{-1} \mathrm{pr}\{u \le T_i < u + du | T_i \ge u, \alpha_i, Z_i, C_i, e_i(u), t_i(u)\} \\ &= \lambda_0(u) \exp[\gamma^T \{G(u)\alpha_i\} + \eta^T Z_i], \end{aligned} \tag{1.6}$$

where $\lambda_0(u)$ is an unspecified baseline hazard function; $G(u)$ is a $(s \times q)$ matrix whose elements are functions of u; γ and η are $(s \times 1)$ and $(r \times 1)$, respectively; $t_i(u) = (t_{ikj} \le u;\ k = 1, \ldots, M)$ denotes the observation times up to and including u; and $e_i(u) = (e_{ikj} : t_{ikj} \le u;\ k = 1, \ldots, M)$. As with modeling of $X_{ik}(u)$, the matrix $G(u)$ may contain complicated functions of time, so that flexibility in modeling the hazard relationship is possible. Equation (1.6) makes explicit the assumption that censoring, timing of measurements, and covariate errors are noninformative.

The vector $G(u, \alpha_i)$ allows flexibility in modeling the hazard relationship. For example, for

$$X_{i1}(u) = \alpha_{i10} + \alpha_{i11}u + \alpha_{i12}u^2, \quad X_{i2}(u) = \alpha_{i20} + \alpha_{i21}u + \alpha_{i22}u^2, \quad M = 2, \tag{1.7}$$

if $G(u, \alpha_i) = G(u)\alpha_i$,

$$G(u) = \begin{bmatrix} 1 & u & u^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & u & u^2 \end{bmatrix}, \ \alpha_i = (\alpha_{i10}, \alpha_{i11}, \alpha_{i12}, \alpha_{i20}, \alpha_{i21}, \alpha_{i22})^T,$$

$$\gamma = (\gamma_1, \gamma_2)^T \ (s = 2),$$

then $\gamma^T G(u, \alpha_i) = \gamma_1 X_{i1}(u) + \gamma_2 X_{i2}(u)$. Here, dependence of the hazard on α_i is linear. Alternatively, the dependence could be nonlinear; for $X_{i1}(u)$ and $X_{i2}(u)$ in (1.7) and $\gamma = (\gamma_1, \gamma_2)^T$, $G(u, \alpha_i) = (\alpha_{i10} + \alpha_{i11}u + \alpha_{i12}u^2, \alpha_{i20}^2)^T$ yields $\gamma^T G(u, \alpha_i) = \gamma_1 X_{i1}(u) + \gamma_2 \alpha_{i20}^2$. The representation in (1.6) also accommodates models involving interactions between covariates and time, which are useful for assessing the relevance of the proportional hazards assumption as suggested by Fleming and Harrington (1991, p. 173) and demonstrated in Section 3.1. To illustrate, for the representations of covariates in (1.7), the hazard relationship linear in the α_i given by

$$\gamma_1 X_{i1}(u) + \gamma_2 X_{i2}(u) + \gamma_3 X_{i2}(u)u \tag{1.8}$$

may be expressed as $\gamma^T G(u)\alpha_i$, with $\gamma = (\gamma_1, \gamma_2, \gamma_3)^T$, $s = 3$ and

$$G(u) = \begin{bmatrix} 1 & u & u^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & u & u^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & u & u^2 & u^3 \end{bmatrix}.$$

The hazard formulation in (1.6) assumes that the survival distribution depends on the time-dependent covariates only through random effects that characterize the individual "inherent trajectories." An alternative modeling strategy would be to decompose the "error" in (1.5) into components associated with serially correlated biological variation and measurement error and to assume the hazard depends both on random effects and biological variation (e.g. Henderson *et al.* 2000). Specification of the hazard would be guided by subject-matter considerations.

1.4.2 Past Approaches

Many approaches were proposed for implementation of these joint models in the case of a single time-dependent covariate, that is, $M = 1$.

One strategy is a two-stage, "regression calibration" technique (Carroll *et al.*, 1995), where the mixed effects model is fitted to data at each risk set under the usual normality assumptions and best linear unbiased predictors from these fits are used to impute covariate values at each failure time(e.g. Pawitan and Self, 1993; Tsiatis *et al.*}, 1995; Dafni and Tsiatis, 1998). For example, consider the simple case of a single time-dependent covariate $X_i(u)$ and no time-independent covariate with the model

$$\begin{aligned} \lambda_i(u) &= \lambda_0(u)\exp\{\gamma X_i(u)\}; \\ W_{ij} &= X_i(t_{ij}) + e_{ij}; \\ X_i(u) &= \alpha_i^T f(u). \end{aligned} \tag{1.9}$$

The hazard based on the observed covariate can be expressed as the conditional expectation of $\lambda_i(u)$ given $(W_{ij}, t_{ij} \leq u)$, that is,

$$\lambda_i\{u|W_{ij}, t_{ij} \leq u\} = \lambda_0(u)E[\exp\{\gamma X_i(V_i)\}|W_{ij}, t_{ij} \leq V_i].$$

Consequently, the partial likelihood can be written as

$$\prod_{i=1}^{n}\left(\frac{E[\exp\{\gamma X_i(V_i)\}|W_{ij}, t_{ij} \leq V_i]}{\prod_{i'=1}^{n} E[\exp\{\gamma X_{i'}(V_i)\}|W_{i'j}, t_{i'j} \leq V_i]Y_j(V_i)}\right)^{\Delta_i}. \tag{1.10}$$

Assuming α_i or $X_i(u)$ is normally distributed, the conditional expectation in (1.10) can be computed by using the moment generating function of the normal distribution. An alternative strategy is using simple first-order approximation to the conditional

expectation, which is called regression calibration (Carroll, Ruppert, and Stefanski, 1995). This leads to

$$E[\exp\{\gamma X_i(V_i)\}|W_{ij}, t_{ij} \leq V_i] \approx \exp\left[\gamma E\left\{X_i(V_i)|W_{ij}, t_{ij} \leq V_i\right\}\right]. \tag{1.11}$$

Estimate (1.11) by substituting the estimated best linear unbiased predictor from the mixed model for $E\left\{X_i(V_i)|W_{ij}, t_{ij} \leq V_i\right\}$. The result is then substituted in (1.10), and the estimator of γ is obtained by solving the estimating equation from this "modified" partial likelihood. The advantage of the approximation (1.11) is that standard software for Cox model can be used. Although this method reduces bias relative to the naive approaches, it may still yield erroneous results (Tsiatis and Davidian, 2001).

An alternative approach that yields sound inferences is to base estimation on the joint likelihood of the survival and longitudinal data under parametric (normal) assumptions on the mixed model random effects. DeGruttola and Tu (1994) took a parametric likelihood approach, assuming that the covariate process and logarithm of survival times are jointly normal. Wulfsohn and Tsiatis (1997) and Henderson *et al.* (2000) took a nonparametric likelihood approach, adopting a less rigid proportional hazards relationship between the hazard of failure and the covariate process, but still assuming normality of the covariate process. Faucett and Thomas (1996) and Xu and Zeger (2001a) took a Bayesian approach, also relying on normality assumption.

The nonparametric approach by Wulfsohn and Tsiatis (1997) can be demonstrated using model (1.9). The random effect α_i is assumed to be normally distributed with mean μ and variance $V = RR^T$. Denote the variance of e_i by σ^2, and the distinct parameters in R by θ. Thus the joint likelihood of the observed longitudinal and

survival data is given by

$$\prod_{i=1}^{n} \int \left\{ \prod_{j=1}^{m_i} p(W_{ij}|\alpha_i, \sigma^2) \right\} p(V_i, \Delta_i|\alpha_i, \lambda_0) p(\alpha_i|\mu, \theta) d\alpha_i \ , \tag{1.12}$$

where

$$\begin{aligned} p(W_{ij}|\alpha_i, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\{W_{ij} - \alpha_i^T f(t_{ij})\}^2}{2\sigma^2} \right], \\ p(V_i, \Delta_i|\alpha_i, \lambda_0) &= \left(\lambda_0(V_i) \exp[\gamma\{\alpha_i^T f(V_i)\}]\right)^{\Delta_i} \\ &\quad \times \exp\left(-\int_0^{V_i} \lambda_0(u) \exp[\gamma\{\alpha_i^T f(u)\}] du \right), \\ p(\alpha_i|\mu, \theta) &= \varphi\{R^{-1}(\alpha_i - \mu)\}|R|^{-1}, \end{aligned}$$

and $\varphi(\cdot)$ is the density of the standard normal distribution. Parameters are estimated using the EM algorithm. Guass-Hermite quadrature is used to approximate the integration in the E-step.

Extension of any of these approaches to multiple time-dependent covariates in the hazard relationship is complicated by the need to model the joint relationship among all covariates. Moreover, these methods rely on normality or other parametric assumptions for the random effects. Hu *et al.* (1998) proposed likelihood methods under weakening of this assumption for a single error-prone time-independent covariate, but adaptation to multiple time-dependent covariates would suffer the same drawbacks as above.

The ACTG 175 data introduced in Section 1.1 exemplifying the need for methods to address these issues. The relationship between $M = 2$ time-dependent covariates and time to death or AIDS is of interest. Moreover, the evidence in the data suggests that the normality assumption for random effects characterizing the CD4 or CD8 longitudinal processes may be unrealistic. Figure 1.1(a) shows $\log_{10}$-transformed CD4

profiles for ten randomly-chosen subjects from ACTG 175 and suggests that the underlying trend is well-approximated by a straight line after week twelve; as only nine events occurred prior to this time, we consider the post-week-twelve data henceforth for simplicity. Residuals from individual simple linear regression fits suggest that intra-subject errors are approximately normally distributed with constant variance (not shown). However, Figure 1.1(b), which shows histograms of the least squares estimates for the zidovudine-alone group, suggests that, although underlying random intercepts may be approximately marginally normal, the distribution of random slopes is possibly skewed; the plot for the combined treatment group is virtually identical. This informal evidence indicates that the normality assumption for log CD4 random effects may be inappropriate, raising concern that inference using the foregoing methods may be compromised.

A procedure that allows relaxation of the normality assumption is available. Tsiatis and Davidian (2001) proposed a conditional score approach to estimation of the Cox parameters that requires no distributional assumption on the random effects. Because the conditional score approach "conditions away" random effects associated with time-dependent covariates, no distributional assumption is required, and an extension of it to handle multiple such covariates is an appealing alternative to other methods. However, this approach is limited to the case of a single, time-dependent covariate, $M = 1$.

When interest focuses on the longitudinal data process, intuition and experience in mixed models with no informative censoring (e.g., Verbeke and Lesaffre, 1997; Tao et al., 1999; Heagerty and Kurland, 2001; Zhang and Davidian, 2001) suggest that

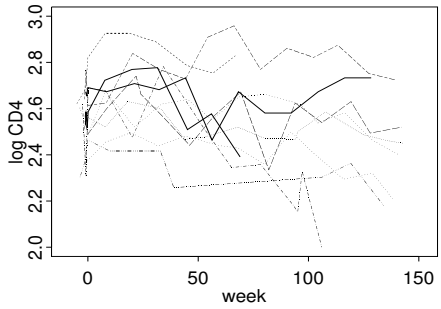
misspecification of the random effects distribution can lead to misleading inferences on certain model parameters. None of the previous approaches to joint modeling allow for straightforward relaxation of a parametric random effects assumption however. Consequently, little is known about the effects of misspecification of the random effects distribution, nor have the advantages of moving away from a parametric assumption been studied systematically.

1.4.3 This Research

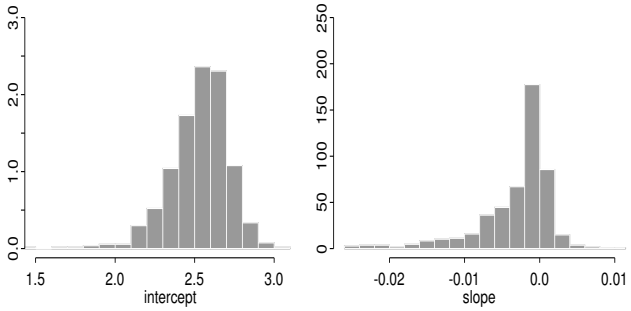
In this research, we address these issues. In the first part, we extend the conditional score approach to multiple time-dependent covariates. In Chapter 2 , we carry out this extension, allowing several time-dependent covariates, as well as time-independent covariates, where the time-dependent covariates may be correlated. In Chapter 3, we assess the finite sample properties of this approach by simulation and present an analysis of the ACTG 175 data, showing the utility of this method in the simultaneous evaluation of several potential surrogate markers.

The conditional score approach is fast and easy to compute. However, a drawback is that it may make inefficient use of the information in the longitudinal data. Moreover, as it is unlikely that random effects associated with continuous covariates such as CD4 have a discrete distribution, the method accommodates too wide a class, including implausible candidates. As the conditioning argument removes the random effects from consideration, no insight into the nature of the true distribution is possible. In addition, inference on the longitudinal data process is not available. In the second part of this dissertation we take another approach, focusing on the issue of the

random effects distribution. We propose to approximate the random effects distribution by the seminonparametric (SNP) densities of Gallant and Nychka (1987), which requires only the assumption that the random effects have a "smooth" density, and take a semiparametric likelihood approach to obtain estimators for both the model parameters and the random effects density. As the approach is likelihood-based, it offers the potential for improved efficiency. We discuss this approach in Chapter 4. We illustrate the method by simulations and application to ACTG 175 data in Chapter 5.



(a)



(b)

Figure 1.1: **a.** Trajectories of $\log_{10}$ CD4 for 10 randomly selected subjects. **b.** Histograms of subject-specific intercept and slope estimates from simple least square fits.

Chapter 2

Conditional Score Approach

2.1 Introduction

Recently, Tsiatis and Davidian (2001) proposed for such joint models a semiparametric conditional score estimator for the parameters in the hazard relationship in the case of a single time-dependent covariate and time-independent covariates. In this approach, the random effects are treated as nuisance parameters for which a sufficient statistic may be derived, and a set of estimating equations based on conditioning on the sufficient statistic may be deduced that remove the dependence on the random effects. The estimators are shown to be consistent and asymptotically normal and to yield unbiased and reliable inferences on the parameters of the Cox model in finite samples. An attractive feature of this approach is that no assumption need be made on the distribution of the random effects. Moreover, the estimator is simple to compute.

Because this approach "conditions away" random effects associated with time-

dependent covariates, an extension of it to handle multiple such covariates is an appealing alternative to other methods. The extension is not trivial, however, as different covariates may be recorded on different schedules by happenstance or design and may be subject to errors that are correlated. In this chapter, we carry out this extension, allowing several time-dependent covariates, as well as time-independent covariates, where the time-dependent covariates may be correlated. We consider the joint longitudinal-survival model with multiple time-dependent and time-independent covariates given in Section 1.4.1. The extension of the conditional score approach is derived in Section 2.2, and in Section 2.3, we remark on large-sample properties. Technical details are given in Section 2.4.

2.2 Conditional Score Estimator

The derivation of the proposed estimator follows the same reasoning as that in Tsiatis and Davidian (2001), motivated by the conditional score method of Stefanski and Carroll (1987). No distributional assumption is placed on the α_i, nor is one needed. Let $\hat{\alpha}_{ik}(u)$ be the ordinary least squares estimator of α_{ik} based on all the longitudinal data measured before time u for the kth covariate for subject i, that is, based on $t_{ik}(u) = (t_{ikj} \leq u)$. Note that in order to obtain $\hat{\alpha}_{ik}(u)$ for each k, i must have at least q_k measurements on the kth covariate; thus, letting $m_{ik}(u)$ denote the number of time points in $t_{ik}(u)$, to ensure estimation of α_{ik} at u is possible for all $k = 1, \ldots, M$, we require $m_{ik}(u) \geq q_k$ for each k. Let the counting process increment be $dN_i(u) = I\{u \leq V_i < u + du, \Delta_i = 1, m_{ik}(u) \geq q_k, k = 1, \ldots, M\}$; let $Y_i(u) = I\{V_i \geq u, m_{ik}(u) \geq q_k, k = 1, \ldots, M\}$ denote the "at risk" process; and

define $\omega = \{\sigma_{kk'} : k \geq k'\}$, the distinct parameters characterizing the covariances of the errors.

For $k = 1, \ldots, M$, define

$$F_{ik}(u) = \begin{bmatrix} f_k^T(t_{ik1}) \\ \vdots \\ f_k^T(t_{ikm_{ik}(u)}) \end{bmatrix} \quad \{m_{ik}(u) \times q_k\},$$

and let $I_{ikk'}(u)$ be the $\{m_{ik}(u) \times m_{ik'}(u)\}$ matrix whose (j, j') entry is $I(t_{ikj} = t_{ik'j'})$, for $j = 1, \ldots, m_{ik}(u)$, $j' = 1, \ldots, m_{ik'}(u)$. It is straightforward to deduce that, conditional on $\{Y_i(u) = 1, \alpha_i, Z_i, t_i(u)\}$, $G(u)\hat{\alpha}_i(u)$, $i = 1, \ldots, n$, are independently distributed as $\mathcal{N}\{G(u)\alpha_i, \Sigma_i(u, \omega)\}$, where $\Sigma_i(u, \omega) = G(u)\Gamma_i(u, \omega)G^T(u)$ $(s \times s)$. The symmetric $(q \times q)$ matrix $\Gamma_i(u, \omega)$ may be written as

$$\Gamma_i(u, \omega) = \begin{bmatrix} \Gamma_{i11}(u, \omega) & \Gamma_{i12}(u, \omega) & \cdots & \Gamma_{i1M}(u, \omega) \\ \Gamma_{i21}(u, \omega) & \Gamma_{i22}(u, \omega) & \cdots & \Gamma_{i1M}(u, \omega) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{iM1}(u, \omega) & \Gamma_{iM2}(u, \omega) & \cdots & \Gamma_{iMM}(u, \omega) \end{bmatrix},$$

where for $k, k' = 1, \ldots, M$,

$$\Gamma_{ikk'}(u, \omega) = \sigma_{kk'}\{F_{ik}^T(u)F_{ik}(u)\}^{-1}F_{ik}^T(u)I_{ikk'}(u)F_{ik'}(u)\{F_{ik'}^T(u)F_{ik'}(u)\}^{-1} \quad (q_k \times q_{k'}).$$

For now we assume that ω is known. Conditional on $\{Y_i(u) = 1, \alpha_i, Z_i, t_i(u)\}$ and by the proportional hazards assumption, $dN_i(u)$ is distributed as Bernoulli with probability $\lambda_0(u)du \exp\{\gamma^T G(u)\alpha_i + \eta^T Z_i\}$, and $dN_i(u)$ and $G(u)\hat{\alpha}_i(u)$ are conditionally independent. Hence the conditional density

$$\begin{aligned} &\text{pr}\{dN_i(u) = a, G(u)\hat{\alpha}_i(u) = h | Y_i(u) = 1, \alpha_i, Z_i, t_i(u)\} \\ &\quad = \text{pr}\{dN_i(u) = a | Y_i(u) = 1, \alpha_i, Z_i, t_i(u)\}\text{pr}\{G(u)\hat{\alpha}_i(u) = h | Y_i(u) = 1, \alpha_i, Z_i, t_i(u)\} \end{aligned}$$

$$= [\lambda_0(u)du\exp\{\gamma^T G(u)\alpha_i + \eta^T Z_i\}]^a [1 - \lambda_0(u)du\exp\{\gamma^T G(u)\alpha_i + \eta^T Z_i\}]^{1-a}$$
$$\times \frac{1}{(2\pi)^{s/2}|\Sigma_i(u,\omega)|^{1/2}} \exp\left[-\frac{\{h - G(u)\alpha_i\}^T \Sigma_i^{-1}(u,\omega)\{h - G(u)\alpha_i\}}{2}\right].$$

Thus, the conditional likelihood of $\{dN_i(u), G(u)\hat{\alpha}_i(u)\}$ given $\{Y_i(u) = 1, \alpha_i, Z_i, t_i(u)\}$ up to order du, is equal to

$$\exp\left\{S_i^T(u,\gamma,\omega)\Sigma_i^{-1}(u,\omega)G(u)\alpha_i\right\} \frac{\{\lambda_0(u)du\exp(\eta^T Z_i)\}^{dN_i(u)}}{(2\pi)^{s/2}|\Sigma_i(u,\omega)|^{1/2}}$$
$$\times \exp\left\{-\frac{\hat{\alpha}_i^T(u)G^T(u)\Sigma_i^{-1}(u,\omega)G(u)\hat{\alpha}_i(u) + \alpha_i^T G^T(u)\Sigma_i^{-1}(u,\omega)G(u)\alpha_i}{2}\right\}, \quad (2.1)$$

where

$$S_i(u,\gamma,\omega) = dN_i(u)\Sigma_i(u,\omega)\gamma + G(u)\hat{\alpha}_i(u).$$

This representation implies that, conditional on $Y_i(u) = 1$, $S_i(u,\gamma,\omega)$ is a complete sufficient statistic for α_i, which suggests that conditioning on $S_i(u,\gamma,\omega)$ would remove the dependence on α_i. Hence, the conditional hazard

$$\lambda_i\{u|S_i(u,\gamma,\omega)\} = \lim_{du\to 0} du^{-1} Pr\{dN_i(u) = 1|S_i(u,\gamma,\omega), Z_i, t_i(u), Y_i(u)\},$$

which, by some algebra in Section 2.4.1, is equal to

$$\lambda_0(u)\exp\left\{\gamma^T S_i(u,\gamma,\omega) - \gamma^T\Sigma_i(u,\omega)\gamma/2 + \eta^T Z_i\right\} Y_i(u). \quad (2.2)$$

We now outline the reasoning underlying the conditional score estimator, which follows by analogy to that for estimators for the proportional hazards model with no measurement error. The conditional intensity of $dN(u) = \sum_{i=1}^n dN_i(u)$ given $\{S_i(u,\gamma,\omega), Z_i, t_i(u), Y_i(u), i = 1,\ldots,n\}$ is

$$\lambda_0(u)E_0(u,\gamma,\eta,\omega),$$

where

$$\begin{aligned} E_0(u,\gamma,\eta,\omega) &= \sum_{i=1}^{n} E_{0i}(u,\gamma,\eta,\omega), \\ E_{0i}(u,\gamma,\eta,\omega) &= \exp\{\gamma^T S_i(u,\gamma,\omega) - \gamma^T \Sigma_i(u,\omega)\gamma/2 + \eta^T Z_i\} Y_i(u) \end{aligned}$$

from (2.2). This suggests that an estimator for $\lambda_0(u)du$ is

$$\hat{\lambda}_0(u)du = dN(u)E_0(u,\gamma,\eta,\omega).$$

By analogy to the usual score equations derived from the partial likelihood in a proportional hazards model, the estimating equations are

$$\sum_{i=1}^{n} \int \{S_i^T(u,\gamma,\omega), Z_i^T\}^T \{dN_i(u) - E_{0i}(u,\gamma,\eta,\omega)\lambda_0(u)du\} = 0,$$

which, upon substituting $\hat{\lambda}_0(u)du$ for $\lambda_0(u)du$, may be written as

$$\sum_{i=1}^{n} \int \{S_i^T(u,\gamma,\omega), Z_i^T\}^T \left\{ dN_i(u) - \frac{E_{0i}(u,\gamma,\eta,\omega)}{E_0(u,\gamma,\eta,\omega)} dN(u) \right\} = 0. \tag{2.3}$$

Interchanging the sums in (2.3), the estimating equations may be expressed as

$$\sum_{i=1}^{n} \int \left[\{S_i^T(u,\gamma,\omega), Z_i^T\}^T - \frac{E_1(u,\gamma,\eta,\omega)}{E_0(u,\gamma,\eta,\omega)} \right] dN_i(u) = 0, \tag{2.4}$$

where

$$\begin{aligned} E_1(u,\gamma,\eta,\omega) &= \sum_{i=1}^{n} E_{1i}(u,\gamma,\eta,\omega), \\ E_{1i}(u,\gamma,\eta,\omega) &= \{S_i^T(u,\gamma,\omega), Z_i^T\}^T E_{0i}(u,\gamma,\eta,\omega). \end{aligned}$$

When none of the M covariates is error-prone, (2.4) reduces to the usual score equations for the maximum partial likelihood estimator of Cox (1975).

The above developments rely on $G(u,\alpha_i) = G(u)\alpha_i$, so that the predictor in the hazard depends linearly on α_i and $G(u)\hat{\alpha}_i(u)$ is exactly normally distributed.

For $G(u, \alpha_i)$ that may be nonlinear in α_i, we propose invoking a linear approximation. By the delta method, $G\{u, \hat{\alpha}_i(u)\}$ is asymptotically normally distributed with mean $G(u, \alpha_i)$ and variance $\Sigma_i(u, \alpha_i, \omega) = \dot{G}(u, \alpha_i)\Gamma_i(u, \omega)\dot{G}^T(u, \alpha_i)$ $(s \times s)$, where $\dot{G}(u, \alpha_i) = \partial G(u, \alpha_i)/\partial \alpha_i$. This suggests substituting $\dot{G}\{u, \hat{\alpha}_i(u)\}$ for $G(u)$ and $\Sigma_i\{u, \hat{\alpha}_i(u), \omega\}$ for $\Sigma_i(u, \omega)$ in (2.4) to obtain estimating equations for general $G(u, \alpha_i)$. We evaluate the performance of this approximation in Section 3.2.

2.3 Large Sample Properties

Generally ω is unknown; however, under our assumptions, it may be estimated based on least squares fits to all the data on each covariate for each subject when possible (i.e., $m_{ik} > q_k$). It is shown in Section 2.4.2 that an unbiased estimator for ω is $\hat{\omega}$, with element $\sigma_{kk'}$ estimated by

$$\hat{\sigma}_{kk'} = \frac{\sum_{i=1}^{n} I(m_{ik} > q_k, m_{ik'} > q_{k'}, m_{ikk'} > 0) R_{ik}^T A_{ikk'}^* R_{ik'}}{\sum_{i=1}^{n} I(m_{ik} > q_k, m_{ik'} > q_{k'}, m_{ikk'} > 0) \mathrm{tr}\{P_{ik} A_{ikk'}^* P_{ik'} A_{ikk'}^{*T}\}}, \tag{2.5}$$

where R_{ik}, $A_{ikk'}^*$, P_{ik}, and $m_{ikk'}$ are defined in Section 2.4.2.

Let $S_i^*(u, \gamma, \omega) = \{S_i^T(u, \gamma, \omega), Z_i^T\}^T$ and $\bar{S}^*(u, \gamma, \eta, \omega) = E_1(u, \gamma, \eta, \omega)/E_0(u, \gamma, \eta, \omega)$. The combined centered estimating equations for $\tau = (\gamma^T, \eta^T, \omega^T)^T$ are

$$\sum_{i=1}^{n} \phi_i(\gamma, \eta, \omega) = 0, \tag{2.6}$$

where

$$\phi_i(\gamma, \eta, \omega) = \begin{bmatrix} \int \{S_i^*(u, \gamma, \omega) - \bar{S}^*(u, \gamma, \eta, \omega)\}\{dN_i(u) - \dfrac{E_{0i}(u, \gamma, \eta, \omega)}{E_0(u, \gamma, \eta, \omega)} dN(u)\} \\ \psi_i(\omega) \end{bmatrix},$$

and $\psi_i(\omega)$ is a vector with the elements

$$I(m_{ik} > q_k, m_{ik'} > q_{k'}, m_{ikk'} > 0)[R_{ik}^T A_{ikk'}^* R_{ik'} - \sigma_{kk'} \mathrm{tr}\{P_{ik} A_{ikk'}^* P_{ik'} A_{ikk'}^{*T}\}], \quad k \geq k'.$$

It is shown in Section 2.4.3 that the estimator $\hat{\tau} = (\hat{\gamma}^T, \hat{\eta}^T, \hat{\omega}^T)^T$ solving these equations is consistent and asymptotically normal. An estimator for the variance of $\hat{\tau}$ is the sandwich matrix

$$n^{-1} A^{-1} B (A^{-1})^T,$$

where $A = n^{-1} \sum_{i=1}^n \partial / \partial \tau^T \{\phi_i(\hat{\gamma}, \hat{\eta}, \hat{\omega})\}$, and $B = n^{-1} \sum_{i=1}^n \phi_i(\hat{\gamma}, \hat{\eta}, \hat{\omega}) \phi_i^T(\hat{\gamma}, \hat{\eta}, \hat{\omega})$.

2.4 Technical Details

In this section, we present the details underlying arguments in earlier sections.

2.4.1 Derivation of (2.2)

After some simple algebra,

$$\begin{aligned} &\hat{\alpha}_i^T(u) G^T(u) \Sigma_i^{-1}(u, \omega) G(u) \hat{\alpha}_i(u) \\ &= S_i^T(u, \gamma, \omega) \Sigma_i^{-1}(u, \omega) S_i(u, \gamma, \omega) - 2 dN_i(u) \gamma^T S_i(u, \gamma, \omega) + dN_i(u) \gamma^T \Sigma_i(u, \omega) \gamma \end{aligned} \quad (2.7)$$

Substituting (2.7) in (2.1), (2.1) can be expressed as

$$\begin{aligned} &\{\lambda_0(u) \exp(\eta^T Z_i) du\}^{dN_i(u)} H\{u, \alpha_i, S_i(u, \gamma, \omega)\} \\ &\quad \times \exp\left[-\frac{1}{2} \{S_i^T(u, \gamma, \omega) \Sigma_i^{-1}(u, \omega) S_i(u, \gamma, \omega) - 2 dN_i(u) \gamma^T S_i(u, \gamma, \omega) \right. \\ &\quad \left. + dN_i(u) \gamma^T \Sigma_i(u, \omega) \gamma \} \right] \end{aligned} \quad (2.8)$$

where $H\{u, \alpha_i, S_i(u, \gamma, \omega)\}$ is given by

$$\frac{1}{(2\pi)^{s/2}|\Sigma_i(u,\omega)|^{1/2}} \exp\left\{ S_i^T(u,\gamma,\omega)\Sigma_i^{-1}(u,\omega)G(u)\alpha_i - \frac{\alpha_i^T G^T(u)\Sigma_i^{-1}(u,\omega)G(u)\alpha_i}{2} \right\}.$$

The conditional probability

$$\text{pr}\left\{dN_i(u) = 1 | S_i(u,\gamma,\omega) = v, Y_i(u) = 1, Z_i, t_i(u)\right\} = P_1/(P_0 + P_1),$$

where for $a = 0, 1$,

$$\begin{aligned} P_a &= \text{pr}\left\{dN_i(u) = a, S_i(u,\gamma,\omega) = v | Y_i(u) = 1, Z_i, t_i(u)\right\} \\ &= \int \text{pr}\left\{dN_i(u) = a, S_i(u,\gamma,\omega) = v | Y_i(u) = 1, \alpha_i, Z_i, t_i(u)\right\} \\ &\quad \times p\left\{\alpha_i | Y_i(u) = 1, Z_i, t_i(u)\right\} d\alpha_i \end{aligned} \tag{2.9}$$

Substituting (2.8) into (2.9),

$$\begin{aligned} P_a &= \{\lambda_0(u)\exp(\eta^T Z_i)du\}^a \\ &\quad \times \exp[-\{v^T\Sigma_i^{-1}(u,\omega)v - 2a\gamma^T S_i(u,\gamma,\omega) + a\gamma^T\Sigma_i(u,\omega)\gamma\}/2] \\ &\quad \times \int H(u,\alpha_i,v) p\left\{\alpha_i | Y_i(u) = 1, Z_i, t_i(u)\right\} d\alpha_i. \end{aligned}$$

Therefore,

$$\begin{aligned} &\text{pr}\left\{dN_i(u) = 1 | S_i(u,\gamma,\omega) = v, Y_i(u) = 1, Z_i, t_i(u)\right\} \\ &= \frac{\lambda_0(u)\exp(\eta^T Z_i)du \, \exp\left\{\gamma^T v - \gamma^T\Sigma_i(u,\omega)\gamma/2\right\}}{1 + \lambda_0(u)\exp(\eta^T Z_i)du \, \exp\left\{\gamma^T v - \gamma^T\Sigma_i(u,\omega)\gamma/2\right\}} \\ &= \lambda_0(u)du\exp\{\gamma^T v - \gamma^T\Sigma_i(u,\omega)\gamma/2 + \eta^T Z_i\} + o_p(du), \end{aligned}$$

which implies (2.2).

2.4.2 Derivation of (2.5)

For subject i, let $W_{ik} = \{W_{ik}(t_{ik1}), \ldots, W_{ik}(t_{ikm_{ik}})\}^T$ be the full set of observations on the kth covariate. For each k,

$$W_{ik} = F_{ik}\alpha_{ik} + e_{ik},$$

where $F_{ik} = [\, f_k(t_{ik1}), \ldots, f_k(t_{ikm_{ik}})\,]^T$, $(m_{ik} \times q_k)$. Assume that $\text{rank}(F_{ik}) = q_k$ if $m_{ik} > q_k$, as would ordinarily be the case with longitudinal modeling. The residuals from a least squares fit of this model to all the data on covariate k for subject i may be written as

$$R_{ik} = P_{ik}W_{ik} = P_{ik}e_{ik},$$

where $P_{ik} = I_{m_{ik}} - F_{ik}(F_{ik}^T F_{ik})F_{ik}^T$, and $I_{m_{ik}}$ is a m_{ik}-dimensional identity matrix. Suppose covariates k and k' are observed in common at $m_{ikk'} > 0$ time points. Let A_{ik} be the $(m_{ikk'} \times m_{ik})$ matrix of zeroes and ones that identifies the residuals for covariate k at the common time points, so that $A_{ik}R_{ik}$ is the $(m_{ikk'} \times 1)$ vector of least squares residuals at the common time points, and define $A_{ik'}$ similarly. Then the sum of the product of residuals for covariates k and k' over the common times is

$$R_{ik}^T A_{ik}^T A_{ik'} R_{ik'} = e_{ik}^T P_{ik} A_{ikk'}^* P_{ik'} e_{ik'}$$

, where $A_{ikk'}^* = A_{ik}^T A_{ik'}$ $(m_{ik} \times m_{ik'})$, which has expectation equal to

$$\text{tr}\{P_{ik}A_{ikk'}^* P_{ik'} E(e_{ik'}e_{ik}^T)\} = \sigma_{kk'}\text{tr}\{P_{ik}A_{ikk'}^* P_{ik'} A_{ikk'}^{*T}\};$$

note that $A_{ikk'}^* = I_{ikk'}(\infty)$. The estimator in (2.5) follows.

2.4.3 Heuristic proof of asymptotic properties of the conditional score estimator

We assume first ω is known. By the Weak Law of Large Numbers, $E_1(u,\gamma,\eta,\omega)/n$ and $E_0(u,\gamma,\eta,\omega)/n$, and hence $\bar{S}^*(u,\gamma,\eta,\omega)$, converges in probability. Denote the probabilistic limit of $\bar{S}^*(u,\gamma,\eta,\omega)$ by $\mu(u,\gamma,\eta,\omega)$. Because

$$\sum_{i=1}^{n}\int \bar{S}^*(u,\gamma,\eta,\omega)\{dN_i(u) - E_{0i}(u,\gamma,\eta,\omega)\lambda_0(u)du\} = 0,$$

(2.3) can be reexpressed as

$$\begin{aligned}
&\sum_{i=1}^{n}\int \{S_i^*(u,\gamma,\omega) - \bar{S}^*(u,\gamma,\eta,\omega)\}\{dN_i(u) - E_{0i}(u,\gamma,\eta,\omega)\lambda_0(u)du\} \\
= &\sum_{i=1}^{n}\int \{S_i^*(u,\gamma,\omega) - \mu(u,\gamma,\eta,\omega)\}\{dN_i(u) - E_{0i}(u,\gamma,\eta,\omega)\lambda_0(u)du\} \quad (2.10)\\
+&\sum_{i=1}^{n}\int \{\mu(u,\gamma,\eta,\omega) - \bar{S}^*(u,\gamma,\eta,\omega)\}\{dN_i(u) - E_{0i}(u,\gamma,\eta,\omega)\lambda_0(u)du\}. \quad (2.11)
\end{aligned}$$

Equation (2.10) evaluated at the true values (γ_0,η_0) is a sum of independent and identically distributed mean zero random vectors. This may be shown by taking expectation inside the integral for the ith summand and conditioning on $\{S_i^*(u,\gamma,\omega)$, Z_i,$t_i(u)$,$Y_i(u)\}$, which yields

$$\begin{aligned}
\int E\Big(&\{S_i^*(u,\gamma,\omega) - \mu(u,\gamma,\eta,\omega)\} \\
&\times[E\{dN_i(u)|S_i^*(u,\gamma,\omega),Z_i,t_i(u),Y_i(u)\} - E_{0i}(u,\gamma,\eta,\omega)\lambda_0(u)du]\Big) \\
&= 0,
\end{aligned}$$

as $E\{dN_i(u)|S_i^*(u,\gamma,\omega),Z_i,t_i(u),Y_i(u)\} = E_{0i}(u,\gamma,\eta,\omega)\lambda_0(u)du$ is the conditional intensity in (2.2). Also, n^{-1} times (2.11) converges in probability to zero uniformly

in a neighborhood $\mathcal{N}(\gamma_0, \eta_0)$ of (γ_0, η_0), which follows from the inequality

$$\sup_{\mathcal{N}(\gamma_0,\eta_0)} \left| \int \{\mu(u,\gamma,\eta,\omega) - \bar{S}^*(u,\gamma,\eta,\omega)\} n^{-1} \sum_{i=1}^{n} \{dN_i(u) - E_{0i}(u,\gamma,\eta,\omega)\lambda_0(u)du\} \right|$$

$$\leq \sup_{\mathcal{N}(\gamma_0,\eta_0)} \left\{ \sup_u |\mu(u,\gamma,\eta,\omega) - \bar{S}^*(u,\gamma,\eta,\omega)| \right\} \tag{2.12}$$

$$\times \left[n^{-1} \sum_{i=1}^{n} \int dN_i(u) + n^{-1} \sum_{i=1}^{n} \sup_{\mathcal{N}(\gamma_0,\eta_0)} \left\{ \int E_{0i}(u,\gamma,\eta,\omega)\lambda_0(u)du \right\} \right]. \tag{2.13}$$

Now (2.12) converges in probability to zero from uniform convergence of $n^{-1}E_0(u,\gamma,\eta,\omega)$ and $n^{-1}E_1(u,\gamma,\eta,\omega)$ to their probabilistic limits, which can be shown by a modification of the Glivenko-Cantelli lemma, and (2.13) is bounded in probability by the Weak Law of Large Numbers. Combining these arguments, the behavior of the estimators solving (2.4) will be dictated by (2.10). Because (2.10) is an unbiased estimating equation, under regularity conditions, a consistent solution to it exists, indicating the existence of a consistent solution $(\hat{\gamma}, \hat{\eta})$ to (2.4).

We now argue that $(\hat{\gamma}, \hat{\eta})$ is asymptotically normal. Evaluated at (γ_0, η_0), $n^{-1/2}$ times (2.10) is asymptotically normal by the central limit theorem, and $n^{-1/2}$ times (2.11) may be written as

$$\int n^{1/2}\{\mu(u,\gamma_0,\eta_0,\omega) - \bar{S}^*(u,\gamma_0,\eta_0,\omega)\} n^{-1} \sum_{i=1}^{n} \{dN_i(u) - E_{0i}(u,\gamma_0,\eta_0,\omega)\lambda_0(u)du\}. \tag{2.14}$$

By arguments similar to those in Breslow & Crowley (1974, Theorem 4) or Tsiatis (1981, Theorem 5.1), we expect that, under regularity conditions, (2.14) converges in probability to zero. Therefore, $n^{-1/2}$ times (2.4) is asymptotically normal. The asymptotic normality of $(\hat{\gamma}, \hat{\eta})$ follows.

When ω is unknown, it is straightforward to show that

$$
\begin{aligned}
n^{1/2}(\hat{\sigma}_{kk'} - \sigma_{kk'}) = \\
& n^{-1/2} \sum_{i=1}^{n} [E\{I(m_{ik} > q_k, m_{ik'} > q_{k'}, m_{ikk'} > 0)\mathrm{tr}(P_{ik} A^*_{ikk'} P_{ik'} A^{*T}_{ikk'})\}]^{-1} \\
& \times I(m_{ik} > q_k, m_{ik'} > q_{k'}, m_{ikk'} > 0)\{R^T_{ik} A^*_{ikk'} R_{ik'} - \mathrm{tr}(P_{ik} A^*_{ikk'} P_{ik'} A^{*T}_{ikk'})\sigma_{kk'}\} \\
& + o_p(1).
\end{aligned}
$$

Thus, because (2.4) is asymptotically equivalent to (2.10), standard M-estimator arguments apply, from which it may be concluded that there exists a consistent and asymptotically normal solution $(\hat{\gamma}^T, \hat{\eta}^T, \hat{\omega}^T)^T$ to (2.6).

Chapter 3

Application and Performance of the Conditional Score Approach

In this chapter, we illustrate the utility and finite-sample performance of the conditional score approach. Section 3.1 presents an analysis of the ACTG 175 data, showing the utility of this method in the evaluation of several potential surrogate markers. Section 3.2 demonstrates the performance of the method via simulation.

3.1 Application to AIDS Clinical Trials Group 175

To demonstrate the utility of the methods for investigating associations among multiple time-dependent covariates and clinical endpoint and for elucidating the joint role of multiple covariates as potential surrogate markers, we apply the methods to the ACTG 175 data. As discussed by Xu and Zeger (2001b), there is scientific rationale for considering several time-dependent covariates simultaneously. The association between covariates and endpoint may reflect underlying mechanisms whose elucidation

may enhance understanding of the disease and allow better prediction. Moreover, it may be fruitful to consider several covariates in evaluating surrogacy. Because of time and cost issues, interest has focused on identifying surrogate markers that could be substituted for the clinical endpoint in evaluation of treatment efficacy. According to Prentice (1989), (i) a surrogate marker should be prognostic for clinical outcome and (ii) the risk of progression given the marker should be independent of treatment. As Xu and Zeger (2001b) point out, associations with several covariates may reflect multiple biological pathways of treatment action. Even in the case where there is a single, predominant pathway, it may be characterized more effectively with more than one covariate.

These considerations suggest it would be advantageous to be able to entertain different models for the relationship between several time-dependent covariates and prognosis. An appealing feature of the conditional score approach is computational ease and speed, even for a large data set such as that for ACTG 175. The fit of each model in the analyses below took only a few minutes, demonstrating the feasibility of screening numerous, complex models in practice.

For the ACTG 175 data, we focus on features of the time trajectories of CD4 and CD8 (count per cubic millimeter) in combination and their association with the clinical endpoint (progression to AIDS or death) and their potential combined surrogacy. There were 308 events, with on average, 8.2 CD4 and 8.1 CD8 measurements per subject. CD4 was available with CD8 in all but 2 cases. The original analysis (Hammer *et al.*, 1996) found zidovudine alone to be inferior to the other three therapies; thus, for simplicity, we consider two treatment groups, zidovudine alone and the combina-

tion of the other three therapies, and let $Z_i = I($ treatment $\neq$ zidovudine). Figure 3.1 presents CD4 and CD8 profiles for 10 randomly selected subjects and shows an apparent initial increase in both measures, with a peak at week 12, followed by a decline. Because only 9 events occurred before week 12, for simplicity, we consider the data including and after week 12.

To achieve approximate within-subject normality and constant variance of CD4 and CD8 measurements, it is customary to transform these measures. We investigated several transformations, including square-, cube-, fourth-root, and logarithmic. Here, we report on results using base-10 logarithmic transformation for CD4 and CD8; use of base-10 logarithm is standard in the medical literature. The results for other transformations are qualitatively similar to those reported below.

Here, $M = 2$. From Figure 3.1, both log CD4, with inherent trajectory $X_{i1}(u)$ and log CD8, with inherent trajectory $X_{i2}(u)$, seem to follow approximate straight-line relationships $X_{ik}(u) = \alpha_{ik0} + \alpha_{ik1}u$, $k = 1, 2$, after week 12. To assess if such simple linear mixed effects models suffice to represent inherent log CD4 and log CD8, we investigated formally whether a quadratic relationship as in (1.7) provides a better characterization using a conditional F-test, which is in the spirit of leaving the distribution of the α_{ik}, $k = 1, 2$, unspecified. That is, for each k, letting W_{ik} denote the vector of log-transformed measurements for subject i, we regarded the α_{ik}, $i = 1, \ldots, n$, as fixed parameters in an overall model for W_{ik}, $i = 1, \ldots, n$, and fitted the linear ("reduced") and quadratic ("full") versions of the model by ordinary least squares using the data for subjects with $m_{ik} > 3$ observations. We then constructed the usual F-statistic based on the error sums of squares for the "full" and "reduced"

models. As discussed in the Appendix, under our assumption on the conditional distribution of e_i given $(T_i, C_i, \alpha_i, t_i, m_i)$, the estimators for the variance σ_{kk} for each k will be unbiased. Thus, because the usual F-statistic is a function of this estimator for each k, the resulting test should have the usual properties. In particular, letting N_k and n_k denote the total number of observations and total number of subjects with $m_{ik} > 3$, the test statistic under the null hypothesis ("reduced" model) and normal errors has an F distribution with $(n_k, N_k - 3n_k)$ degrees of freedom. Because N_k and n_k are large, the appropriate critical value is virtually 1. For log CD4 ($k = 1$) and log CD8 ($k = 2$), the statistics are 2.576 and 1.854, respectively, suggesting that the quadratic model (1.7) provides a better characterization for each, leading us to adopt this model for both log CD4 and log CD8.

Figure 3.2 shows Q-Q plots of studentized residuals for each quadratic model fit. Under our assumptions on the conditional distribution of e_i, these residuals should have the usual linear model properties for each k, so should reflect deviation from normality in the conventional manner. The plots suggest that the error distribution may be short-tailed relative to the normal but symmetric. Although the derivation of Section 2.2 requires the errors e_{ikj} to be normally distributed, evidence in Section 3.2 shows that the conditional score method is not affected by short-tailed error distributions. The sample correlation coefficient for studentized residuals of log CD4 and log CD8 is 0.58, indicating that it would be inappropriate to assume that the covariance $\sigma_{12} = 0$.

To assess the association of log CD4 and log CD8 with prognosis and their role as surrogate markers, we carried out several analyses.

1. We fitted the model with treatment only, $\lambda_i(u) = \lambda_0(u)\exp\{\eta Z_i\}$, by usual partial likelihood methods (Model 1).

2. We investigated various combinations of features of the inherent trajectories of log CD4 and log CD8 to determine which combination(s) are most strongly associated with the endpoint. In particular, we considered a series of models of the general form $\lambda_i(u) = \lambda_0(u)\exp[\gamma^T\{G(u)\alpha_i^T\}]$, where $G(u)\alpha_i^T$ was chosen to represent different linear combinations of current value and intercept of log CD4 and log CD8, and fitted them using the conditional score method.

   Because the conditional score approach is not likelihood-based, we used the Wald statistic for the null hypothesis $\gamma = 0$ for each model, which has approximately a $\chi^2_{\dim(\gamma)}$ distribution, to reflect strength of association. The estimates and Wald statistics for several models are shown in Table 3.1 (Models 2.1–2.8). The table excludes models that resulted in unstable estimation, gauged by a very large ratio of largest to smallest eigenvalues of the estimated covariance matrix for $\hat{\gamma}$. With the exception of Model 2.8, this occurred only for models with $\dim(\gamma) > 2$. Among models with $\dim(\gamma) > 2$, none showed appreciable improvement in the Wald statistic over those in the table. In all models except 2.8, the sign of the coefficients for current value and intercept of log CD4 and log CD8 is consistent with scientific understanding; as both measures reflect status of the body"s immune response, the higher the better, decrease of the hazard with increased CD4/CD8 is biologically plausible.

   Among all the models, Model 2.5 with $\lambda_i(u) = \lambda_0(u)\exp\{\gamma_1 X_{i1}(u) + \gamma_2 X_{i2}(u)\}$ involving the combination of current values of log CD4 and log CD8 leads to

the largest Wald statistic. From Table 3.1, Model 2.1 with log CD4 alone yields the second-largest Wald statistic; adding intercept of log CD8 to this model (Model 2.6) results in virtually no improvement, while adding current value as in Model 2.5 yields a considerably larger statistic. Models 2.3 and 2.4 indicate that features of log CD8 may be associated with prognosis on their own.

All of the models considered embody the assumption of proportional hazards. We assessed the adequacy of this assumption using the following approach, which we illustrate using Model 2.5. Following a suggestion of Fleming and Harrington (1991, p. 173), we re-fitted the model separately for the data in two time intervals, corresponding roughly to the events occurring before and after the median event time. According to these authors, under proportional hazards, we would expect the estimates for γ to be comparable in each interval. The results in the first part of Table 3.2 indicate that the proportional hazards assumption may be reasonable, but there is some suggestion of possible deviation from it. To investigate further, we followed Fleming and Harrington (1991, p. 173) and considered two additional models, Model 2.5*, which includes log CD4, log CD8, and the interaction of time and log CD4, and Model 2.5**, including the terms in Model 2.5 plus an additional interaction of time and log CD8. Evidence that the coefficient of an interaction term is non-zero may reflect a deviation from the proportional hazards assumption; thus, unlike the diagnostic above, this approach allows formal significance testing. As noted in Section 1.4.1, fitting these models may be implemented straightforwardly in our framework, and the interaction parameters may be estimated and tested us-

ing the conditional score method with no additional difficulty, as in (1.8). The second part of Table 3.2 shows the results. The fit of Model 2.5** is entirely similar to that of Model 2.5, with no evidence of an interaction of log CD8 with time (p-value = 0.908), indicating no deviation from proportional hazards. The fit of Model 2.5* suggests a decreasing hazard relationship to log CD4 over time (p-value for the interaction term = 0.016). Although the interaction effect of log CD4 and time is significant at level 0.05, the practical implication is modest. From the fit of Model 2.5*, predicted log hazard ratios for log CD4 at the 10th, 50th, and 90th percentiles of the observed distribution of failure times are -2.620, -1.972, and -1.514, respectively, compared to the constant estimate of -1.838 from Model 2.5. Thus, whether the interaction term is included or not, the relationship with log CD4 is qualitatively similar. Results for other models in Table 3.2 resemble those here.

Different data analysts may make different subjective judgments over whether ease of interpretation of a proportional hazards outweighs the importance of the possible, practically modest departure from this assumption in gauging the association of the covariates with prognosis. In the sequel, we focus for illustration on proportional hazards models but also indicate results for models including interaction terms.

3. We fitted several of the models showing evidence of strong associations in step 2 augmented by an additive effect of treatment in the hazard. E.g., for Model 2.5, we considered $\lambda_i(u) = \lambda_0(u) \exp\{\gamma_1 X_{i1}(u) + \gamma_2 X_{i2}(u) + \eta Z_i\}$. (These models are denoted as Model 3.X below.)

Models for step 2 were used to deduce the most productive combination for describing the association between covariates and prognosis. The results for Models 2.5 suggest that, although current value of log CD4 is strongly prognostic on its own, consideration of current value of log CD8 offers additional explanatory value.

Models from step 2 also were used to verify Prentice"s condition (i), which would be indicated by $\gamma = (\gamma_1, \gamma_2)^T \neq 0$. Model 1 and those in step 3 were used to investigate condition (ii); $\eta \neq 0$ in Model 1 but $\eta = 0$ in models in step 3 would suggest that the treatment effect is mediated through the relevant combination of features of log CD4 and log CD8. The estimate of the parameters of the error covariance matrix, $\omega^T = (\sigma_{11}, \sigma_{12}, \sigma_{22})$, is (0.00928, 0.00558, 0.01009). For comparison, models from step 2 and 3 were also fitted using the naive regression method.

To illustrate, results for Models 1, 2.5, and 3.5 and involving current values of log CD4 and log CD8 in combination, are shown in Table 3.3; p-values are for the corresponding Wald statistic for the null hypothesis that each entry is equal to zero. Treatment alone (Model 1) yields a p-value of 0.003, indicating a strong effect, as deduced by Hammer *et al.* (1996). For Model 2.5, the results from the conditional score and naive regression methods differ considerably; the naive regression analysis suggests no association with current log CD8, while the conditional score shows strong association with both covariates. The Model 3.5 conditional score results show that the treatment effect is not significant after adjustment for log CD4 and log CD8, which supports Prentice"s condition (ii) that the treatment effect is mediated through a combination of log CD4 and log CD8. These results are in marked contrast to the conclusion that would be reached from the naive regression fit, which shows a

significant treatment effect even after adjustment. Table 3.3 also includes results for Models 2.5* and 3.5*, which yield the same qualitative conclusions.

The conditional score results suggest that some combination of current values of log CD4 and log CD8 is a surrogate for the treatment considered in ACTG 175. We carried out similar analyses (not shown) for Models 2.1 (current log CD4 alone) and 2.3 (current log CD8 alone) and the augmented versions of these models including a treatment effect (Models 3.1 and 3.3, respectively). In Model 3.1, the estimate of the treatment effect is insignificant, indicating that current value of log CD4 on its own represents a potential surrogate marker. As with Model 3.5, the naive regression analysis of this model yields a highly significant treatment effect. In contrast, the conditional score estimate of treatment effect in Model 3.3 is highly significant (p-value=0.006), indicating that current log CD8 on its own may not be a plausible surrogate.

Given these results, we investigated whether there is added benefit to considering both measures as a surrogate relative to current log CD4 only. Following a procedure similar to that of Xu and Zeger (2001b, p. 83), we inspected point estimates and 95% Wald confidence intervals for the treatment effect in models 3.1 and 3.5, -0.074 (-0.514,0.366) and -0.194 (-0.570,0.189), respectively. The intervals overlap considerably, and the estimated treatment effect is greater in magnitude for Model 3.5, suggesting that adding current log CD8 offers little benefit. Analyses using models incorporating interaction terms yield similar qualitative conclusions.

Summarizing, the results of competing model fits imply that while the combination of current log CD4 and log CD8 is highly associated with prognosis, the strength of

the combination of current log CD4 and log CD8 as a surrogate is derived primarily through log CD4. Use of the naive approach to carry out these analyses would potentially result in erroneous inferences on both association and surrogacy. Although in this example there was no benefit to considering more than one covariate as a surrogate, the availability of accessible methods for fitting multiple-covariate models allowed this possibility to be considered. Had log CD4 and log CD8 considered separately not shown evidence of surrogacy, consideration of their joint association would permit evaluation of their potential surrogacy in combination.

3.2 Simulation Studies

We conducted several simulation experiments, based roughly on the situation of the ACTG 175 data. For each of n subjects, the treatment indicator Z_i was generated according to Bernoulli distribution with probability 0.5. For simplicity, we took log CD4 and log CD8 to follow a linear mixed effects model, i.e. $X_{i1}(u) = \alpha_{i10} + \alpha_{i11}u$ for log CD4 and $X_{i2}(u) = \alpha_{i20} + \alpha_{i21}u$ for log CD8, with $t_{ijk} = 0, 12, 24, 36, \ldots, 168$ weeks. The true hazard relationship was $\lambda_i(u) = \lambda_0(u) \exp\{\gamma_1 X_{i1}(u) + \gamma_2 X_{i2}(u) + \eta Z_i\}$, with $\gamma_1 = -2.0$, $\gamma_2 = -1.0$, and $\eta = 0$. The baseline hazard $\lambda_0(u) = 2$ for $u > 12$ and 0 otherwise. To mimic the censoring of the ACTG 175 data, censoring was generated from an exponential distribution with mean 250, leading to 86.3% and 87.1% censoring, respectively, for the normal and bimodal random effects distributions described below. To represent missed visits that are often seen in practice, after week 12, (CD4,CD8) pairs were missing with probability 0.1 at each time point for all subjects.

We considered several scenarios representing different magnitudes of error and distributions for the random effects as follows. True values for the variances of the errors e_{i1j} and e_{i2j} and covariance matrices of α_{i1} and α_{i2} were specified from separate linear mixed effects model fits to the ACTG 175 data for log CD4 and log CD8. We added correlation between errors at the same time using the value observed for the correlation among least squares residuals, and chose arbitrary but nonnegligible values for cross-correlations among elements of α_{i1} and α_{i2}. The errors $e_{ij} = (e_{i1j}, e_{i2j})^T$ were initially taken to be independent and identically normally distributed with covariance matrix D such that vech(D)=(0.01089,0.00628,0.01114)^T. The random effects α_i were initially taken to be jointly normal with mean $(2.5915, -0.00315, 2.9329, -0.00048)^T$ for $Z_i = 0$ and mean $(2.5915, -0.00198, 2.9329, -0.00038)^T$ for $Z_i = 1$ and covariance matrix

$$\begin{bmatrix} 0.02408 & -0.00008 & 0.01036 & -0.00006 \\ -0.00008 & 0.000014 & -0.00021 & 6.113 \times 10^{-6} \\ 0.01036 & -0.00021 & 0.03540 & -0.00011 \\ -0.00006 & 6.113 \times 10^{-6} & -0.00011 & 3.451 \times 10^{-6} \end{bmatrix} \tag{3.1}$$

We carried out simulations for the scenario assuming D and (3.1) for $n = 250$ and 500; results for larger sample sizes are similar to those for $n = 500$. For each n, we also considered the cases of a four-fold increase in measurement error, i.e. taking the covariance matrix of e_{ij} to be four times D, replacing the normal specification for α_i by a bimodal mixture of normals described below, and both of these together. For the bimodal specification, α_{i1} and α_{i2} were taken to be independently distributed, each following a bimodal mixture of normals as described in Davidian and Gallant (1993) with mixing proportion $p = 0.5$ and sep=4 for α_{i1} and $p = 0.3$ and sep=3 for α_{i2} with same means and covariance matrices as in the normal case above.

For each scenario, 1000 Monte Carlo data sets were simulated. For each data set, we fitted the model above four ways: (i) using the "ideal" approach in which the true values of $X_{i1}(u)$ and $X_{i2}(u)$ at each failure time were used (fitting using the usual partial likelihood method); (ii) using the conditional score estimator, (iii) using naive regression; (iv) using LVCF. For (i), (iii), and (iv), we estimated standard errors as for the usual time-dependent Cox model and used the approach in Section 2.3 for the conditional score. For all methods, 95% Wald confidence intervals for γ_1, γ_2, and η were constructed.

The results are given in Tables 3.4 and 3.5. In all cases and for both sample sizes, the conditional score estimator shows negligible bias close to that of the unachievable "ideal" estimator. For $n = 500$, coverage probabilities are close to or achieve the nominal level. For $n = 250$ and 4-fold error, the coverage is a little below the nominal level. This is not surprising; with such large censoring rates, $n = 250$ corresponds to a case with roughly 33 events. That the conditional score estimator yields reasonable results under these difficult conditions is encouraging. For all sample sizes, the conditional score estimator provides unbiased estimation regardless of the distribution of α_i. In contrast, the naive regression and LVCF approaches can yield biased estimates and coverage probabilities well below the nominal level. For 1-fold error, the results for conditional score and naive regression are similar to the ideal, but LVCF performance is degraded. For 4-fold error, differences among the methods are more dramatic.

To assess the performance of conditional score method when the error distribution is short tailed compared to the normal, as in the ACTG 175 analysis, we conducted

simulations taking the error distribution as the normal truncated at 3 standard deviations. The results for $n = 500$ and 4-fold error are shown in Table 3.6. The conditional score performs well, while the naive methods are subject to bias. For $G(u, \alpha_i)$ nonlinear in α_i, we also carried out simulations to investigate the performance of the linear approximation in Section 2.2, using the same scenarios as above except that the true hazard was taken to be $\lambda_i(u) = \lambda_0(u) \exp\{\gamma_1 X_{i1}(u) + \gamma_2 \alpha_{i20}^2 + \eta Z_i\}$, with $\gamma_1 = -2.0$, $\gamma_2 = -0.2$ and $\eta = 0$. Results for the "ideal", conditional score and naive regression methods with $n = 500$ and 4-fold error are in Table 3.7, and show that the linear approximation yields good performance. The estimates of the parameter ω containing the error variances σ_{11} and σ_{22} and the covariance σ_{12} for all simulations are given in Table 3.8 and demonstrate that the proposed estimator for this parameter given in Section 2.3 is unbiased.

Overall, the simulation evidence suggests that the conditional score estimator yields reliable inferences under a broad range of conditions.

3.3 Discussion

We have proposed an estimator for a joint model for survival and longitudinal data for which the hazard relationship may depend on multiple time-dependent covariates measured at potentially different time points whose observation is subject to error. The method offers a feasible approach to a problem that entails considerable computational complexity otherwise. The estimator is easy to compute and does not require an assumption on the distribution of underlying random effects nor the need

to model or identify this distribution. Application of the method and simulation evidence show that it is accessible for routine use and leads to reliable inference.

We implemented the conditional score method in the C++ language Newton-Raphson to solve the conditional score estimating equations; code is available upon request.

Table 3.1: Results for several models fitted to the ACTG 175 data. Entries are the estimate and its estimated standard error in parentheses. The fit of Model 2.8 is suspect, as the Wald statistic decreases relative to that for Model 2.2 and the coefficient of intercept of log CD8 is positive. This model yielded a large eigenvalue ratio, suggesting instability.

	current log CD4	intercept of log CD4	current log CD8	intercept of log CD8	Wald statistic
Model 2.1	−2.033 (0.146)	—	—	—	193.577
Model 2.2	—	−1.568 (0.405)	—	—	22.095
Model 2.3	—	—	−1.722 (0.593)	—	8.440
Model 2.4	—	—	—	−0.994 (0.342)	8.442
Model 2.5	−1.838 (0.147)	—	−0.982 (0.215)	—	312.370
Model 2.6	−2.034 (0.150)	—	—	0.037 (0.668)	193.598
Model 2.7	—	−1.842 (0.324)	−1.532 (0.321)	—	43.490
Model 2.8	—	−1.819 (0.452)	—	0.209 (0.055)	18.529*

Table 3.2: Diagnostic checks of the proportional hazards assumption in Model 2.5.

(a) Fits of Model 2.5 using data from the first and second time intervals

	$\hat{\gamma}_1$	SE($\hat{\gamma}_1$)	$\hat{\gamma}_2$	SE($\hat{\gamma}_2$)
Using first half events	−2.072	0.233	−0.897	0.318
Using second half events	−1.533	0.179	−1.362	0.275

(b) Fits of Models 2.5* and 2.5**

	log CD4	log CD8	log CD4 ×time	log CD8 ×time	Wald statistic
Model 2.5*	−3.030 (0.552)	−1.073 (0.250)	0.012 (0.005)	—	327.178
Model 2.5**	−1.844 (0.169)	−1.032 (0.435)	—	0.001 (0.006)	312.133

Table 3.3: Fits of Models 1, 2.5 and 2.5*, and 3.5 and 3.5* for the ACTG 175 data.

Terms Included	Est.	SE	p-value	Est.	SE	p-value
			(a) Model 1			
treatment	−0.375	0.128	0.003	—	—	—
			(b) Models 2.5 and 2.5*			
			Conditional Score			
current log CD4	−1.838	0.147	0.000	−3.030	0.552	0.000
current log CD8	−0.982	0.215	0.000	−1.073	0.250	0.000
log CD4 × time	—	—	—	0.012	0.005	0.016
			Naive Regression			
current log CD4	−0.116	0.013	0.000	−0.801	0.107	0.000
current log CD8	−0.008	0.008	0.308	0.006	0.009	0.490
log CD4 × time	—	—	—	0.006	0.001	0.000
			(c) Models 3.5 and 3.5*			
			Conditional Score			
current log CD4	−1.831	0.150	0.000	−3.003	0.551	0.000
current log CD8	−0.999	0.212	0.000	−1.077	0.247	0.000
log CD4 × time	—	—	—	0.012	0.005	0.019
treatment	−0.190	0.194	0.326	−0.066	0.207	0.749
			Naive Regression			
current log CD4	−0.119	0.014	0.000	−0.816	0.110	0.000
current log CD8	−0.009	0.008	0.274	0.006	0.009	0.518
log CD4 × time	—	—	—	0.006	0.001	0.000
treatment	−0.408	0.138	0.003	−0.416	0.139	0.003

Table 3.4: Simulation results for sample size $n = 250$. "1-fold" and "4-fold" refer to error according to D and $4D$, as described in the text. I, "ideal" method; CS, Conditional Score; N, naive regression; LV, last value carried forward; SD, Monte Carlo standard deviation; SE, average of estimated standard errors; CP, coverage probability of 95% Wald confidence interval.

Error			1 fold			4 fold	
Method	I	CS	N	LV	CS	N	LV
			(a) Normal α_i				
$\hat{\gamma}_1$	-2.047	-2.055	-2.029	-2.060	-2.080	-1.971	-1.748
SD($\hat{\gamma}_1$)	0.588	0.615	0.592	0.573	0.712	0.599	0.494
SE($\hat{\gamma}_1$)	0.565	0.581	0.560	0.558	0.674	0.543	0.496
CP($\hat{\gamma}_1$)	0.954	0.944	0.946	0.954	0.942	0.932	0.920
$\hat{\gamma}_2$	-1.056	-1.065	-0.972	-0.667	-1.136	-0.753	-0.193
SD($\hat{\gamma}_2$)	0.958	1.081	0.994	0.831	1.717	1.035	0.653
SE($\hat{\gamma}_2$)	0.940	1.025	0.917	0.848	1.335	0.854	0.673
CP($\hat{\gamma}_2$)	0.944	0.934	0.919	0.941	0.928	0.876	0.782
$\hat{\eta}$	0.004	0.005	0.002	-0.012	0.004	-0.004	-0.041
SD($\hat{\eta}$)	0.377	0.381	0.377	0.371	0.397	0.376	0.366
SE($\hat{\eta}$)	0.362	0.363	0.362	0.361	0.378	0.362	0.358
CP($\hat{\eta}$)	0.949	0.947	0.953	0.958	0.953	0.953	0.953
			(b) Bimodal α_i				
$\hat{\gamma}_1$	-2.070	-2.074	-2.021	-1.956	-2.108	-1.904	-1.580
SD($\hat{\gamma}_1$)	0.496	0.524	0.498	0.483	0.630	0.507	0.435
SE($\hat{\gamma}_1$)	0.492	0.509	0.487	0.491	0.597	0.472	0.441
CP($\hat{\gamma}_1$)	0.962	0.945	0.954	0.961	0.935	0.929	0.833
$\hat{\gamma}_2$	-1.016	-1.029	-0.910	-0.604	-1.089	-0.663	-0.077
SD($\hat{\gamma}_2$)	0.852	0.958	0.884	0.766	1.271	0.918	0.617
SE($\hat{\gamma}_2$)	0.824	0.910	0.801	0.748	1.170	0.744	0.598
CP($\hat{\gamma}_2$)	0.951	0.940	0.926	0.910	0.921	0.869	0.652
$\hat{\eta}$	-0.008	-0.008	-0.013	-0.037	-0.004	-0.026	-0.076
SD($\hat{\eta}$)	0.387	0.390	0.386	0.381	0.403	0.385	0.376
SE($\hat{\eta}$)	0.373	0.374	0.373	0.371	0.387	0.372	0.370
CP($\hat{\eta}$)	0.955	0.952	0.957	0.953	0.954	0.957	0.953

Table 3.5: Simulation results for sample size $n = 500$. "1-fold" and "4-fold" refer to error according to D and $4D$, as described in the text. I, "ideal" method; CS, Conditional Score; N, naive regression; LV, last value carried forward; SD, Monte Carlo standard deviation; SE, average of estimated standard errors; CP, coverage probability of 95% Wald confidence interval.

Error			1 fold			4 fold	
Method	I	CS	N	LV	CS	N	LV
			(a) Normal α_i				
$\hat{\gamma}_1$	-2.052	-2.054	-2.034	-2.061	-2.065	-1.978	-1.756
SD($\hat{\gamma}_1$)	0.406	0.429	0.413	0.389	0.492	0.422	0.339
SE($\hat{\gamma}_1$)	0.388	0.398	0.384	0.384	0.452	0.373	0.343
CP($\hat{\gamma}_1$)	0.945	0.936	0.934	0.944	0.944	0.921	0.891
$\hat{\gamma}_2$	-0.978	-0.991	-0.900	-0.630	-1.042	-0.696	-0.181
SD($\hat{\gamma}_2$)	0.684	0.757	0.711	0.611	0.994	0.748	0.483
SE($\hat{\gamma}_2$)	0.653	0.717	0.638	0.592	0.929	0.597	0.472
CP($\hat{\gamma}_2$)	0.939	0.938	0.923	0.895	0.941	0.850	0.596
$\hat{\eta}$	0.005	0.006	0.003	-0.012	0.008	-0.003	-0.042
SD($\hat{\eta}$)	0.256	0.258	0.256	0.253	0.264	0.255	0.250
SE($\hat{\eta}$)	0.252	0.252	0.252	0.251	0.259	0.251	0.250
CP($\hat{\eta}$)	0.946	0.947	0.948	0.946	0.945	0.951	0.949
			(b) Bimodal α_i				
$\hat{\gamma}_1$	-2.029	-2.032	-1.987	-1.929	-2.058	-1.878	-1.567
SD($\hat{\gamma}_1$)	0.348	0.365	0.348	0.338	0.432	0.349	0.302
SE($\hat{\gamma}_1$)	0.336	0.347	0.333	0.338	0.408	0.323	0.306
CP($\hat{\gamma}_1$)	0.950	0.942	0.937	0.938	0.939	0.906	0.684
$\hat{\gamma}_2$	-1.050	-1.068	-0.964	-0.651	-1.142	-0.738	-0.119
SD($\hat{\gamma}_2$)	0.592	0.660	0.607	0.519	0.876	0.622	0.410
SE($\hat{\gamma}_2$)	0.570	0.625	0.555	0.520	0.825	0.518	0.417
CP($\hat{\gamma}_2$)	0.944	0.944	0.930	0.900	0.930	0.871	0.426
$\hat{\eta}$	0.004	0.003	-0.001	-0.024	0.004	-0.013	-0.061
SD($\hat{\eta}$)	0.261	0.263	0.261	0.258	0.270	0.260	0.255
SE($\hat{\eta}$)	0.258	0.259	0.258	0.257	0.265	0.258	0.257
CP($\hat{\eta}$)	0.950	0.947	0.951	0.951	0.947	0.954	0.945

Table 3.6: Simulation results with $n = 500$ and "4-fold" error according to $4D$ as described in the text: Short-tailed (truncated normal) error. I, "ideal" method; CS, Conditional Score; N, naive regression; LV, last value carried forward; SD, Monte Carlo standard deviation; SE, average of estimated standard errors; CP, coverage probability of 95% Wald confidence interval.

α_i			Normal				Bimodal	
Method	I	CS	N	LV	I	CS	N	LV
$\hat{\gamma}_1$	−2.039	−2.057	−1.971	−1.762	−2.039	−2.073	−1.890	−1.578
SD($\hat{\gamma}_1$)	0.408	0.490	0.416	0.336	0.357	0.452	0.362	0.302
SE($\hat{\gamma}_1$)	0.388	0.450	0.374	0.345	0.337	0.410	0.324	0.306
CP($\hat{\gamma}_1$)	0.946	0.937	0.920	0.883	0.934	0.936	0.902	0.706
$\hat{\gamma}_2$	−0.976	−1.028	−0.698	−0.192	−1.008	−1.080	−0.695	−0.106
SD($\hat{\gamma}_2$)	0.687	0.943	0.722	0.472	0.568	0.839	0.601	0.405
SE($\hat{\gamma}_2$)	0.651	0.914	0.597	0.474	0.566	0.815	0.516	0.419
CP($\hat{\gamma}_2$)	0.934	0.940	0.857	0.599	0.948	0.950	0.848	0.411
$\hat{\eta}$	0.013	0.017	0.008	−0.029	0.009	0.014	−0.008	−0.058
SD($\hat{\eta}$)	0.248	0.258	0.248	0.242	0.249	0.261	0.250	0.244
SE($\hat{\eta}$)	0.251	0.258	0.251	0.250	0.257	0.265	0.257	0.256
CP($\hat{\eta}$)	0.956	0.955	0.959	0.958	0.959	0.958	0.955	0.955

Table 3.7: Simulation results with $n = 500$ and "4-fold" error according to $4D$ as described in the text: Nonlinear $G(u, \alpha_i)$. I, "ideal" method; CS, Conditional Score; N, naive regression; LV, last value carried forward; SD, Monte Carlo standard deviation; SE, average of estimated standard errors; CP, coverage probability of 95% Wald confidence interval.

α_i		Normal			Bimodal	
Method	I	CS	N	I	CS	N
$\hat{\gamma}_1$	−2.020	−2.033	−1.816	−2.023	−2.050	−1.850
SD($\hat{\gamma}_1$)	0.240	0.315	0.250	0.242	0.311	0.242
SE($\hat{\gamma}_1$)	0.243	0.303	0.230	0.243	0.305	0.230
CP($\hat{\gamma}_1$)	0.951	0.941	0.849	0.946	0.949	0.871
$\hat{\gamma}_2$	−0.203	−0.211	−0.130	−0.206	−0.213	−0.134
SD($\hat{\gamma}_2$)	0.076	0.103	0.065	0.078	0.111	0.069
SE($\hat{\gamma}_2$)	0.076	0.103	0.063	0.077	0.106	0.063
CP($\hat{\gamma}_2$)	0.950	0.954	0.791	0.946	0.953	0.780
$\hat{\eta}$	−0.002	−0.001	−0.018	0.006	0.010	−0.006
SD($\hat{\eta}$)	0.165	0.169	0.163	0.166	0.170	0.165
SE($\hat{\eta}$)	0.165	0.169	0.164	0.164	0.169	0.164
CP($\hat{\eta}$)	0.956	0.962	0.964	0.946	0.943	0.948

Table 3.8: Simulation results for the estimation of $\omega^T = (\sigma_{11}, \sigma_{12}, \sigma_{22})$. For 1-fold normal error, the true value of $\omega^T = (\sigma_{11}, \sigma_{12}, \sigma_{22}) = (10.89, 6.28, 11.14) \times 10^{-3}$. For 4-fold normal error, the true value of $\omega^T = (43.56,\ 25.12,\ 44.56) \times 10^{-3}$. For 4-fold normal error truncated at 3 standard deviations, the true value of $\omega^T = (42.13,\ 23.89,\ 43.09) \times 10^{-3}$. Mean$(\hat{\omega}^T \times 10^3)$ and SD$(\hat{\omega}^T \times 10^3)$ are the Monte Carlo mean and standard deviation of $\hat{\omega}^T \times 10^3$, respectively.

n	α_i	Error	Mean$(\hat{\omega}^T \times 10^3)$	SD$(\hat{\omega}^T \times 10^3)$
		(a) Linear $G(u, \alpha_i)$		
250	normal	1-fold	(10.877, 6.268, 11.121)	(0.012, 0.010, 0.012)
250	bimodal	1-fold	(10.897, 6.285, 11.145)	(0.011, 0.009, 0.011)
250	normal	4-fold	(43.509, 25.071, 44.484)	(0.047, 0.039, 0.047)
250	bimodal	4-fold	(43.587, 25.141, 44.581)	(0.044, 0.037, 0.046)
500	normal	1-fold	(10.891, 6.273, 11.135)	(0.008, 0.007, 0.008)
500	bimodal	1-fold	(10.894, 6.280, 11.146)	(0.008, 0.006, 0.008)
500	normal	4-fold	(43.565, 25.093, 44.541)	(0.033, 0.027, 0.033)
500	bimodal	4-fold	(43.575, 25.120, 44.584)	(0.032, 0.026, 0.033)
500	normal	truncated 4-fold	(42.186, 23.865, 43.035)	(0.031, 0.026, 0.031)
500	bimodal	truncated 4-fold	(42.135, 23.863, 43.060)	(0.030, 0.024, 0.030)
		(b) Nonlinear $G(u, \alpha_i)$		
500	normal	4-fold	(43.608, 25.161, 44.597)	(0.034, 0.028, 0.036)
500	bimodal	4-fold	(43.631, 25.132, 44.583)	(0.033, 0.028, 0.035)

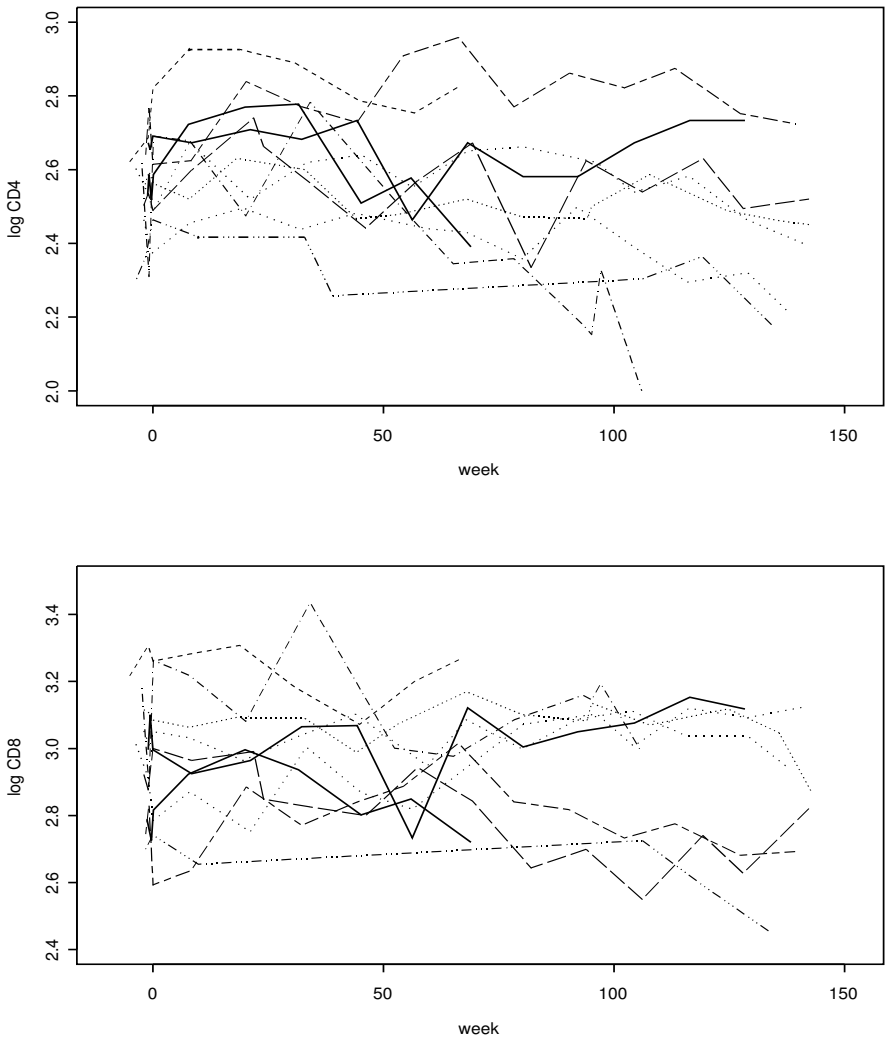


Figure 3.1: Trajectories of log CD4 and log CD8 for 10 randomly selected subjects. The same type line is used for each subject in each panel.

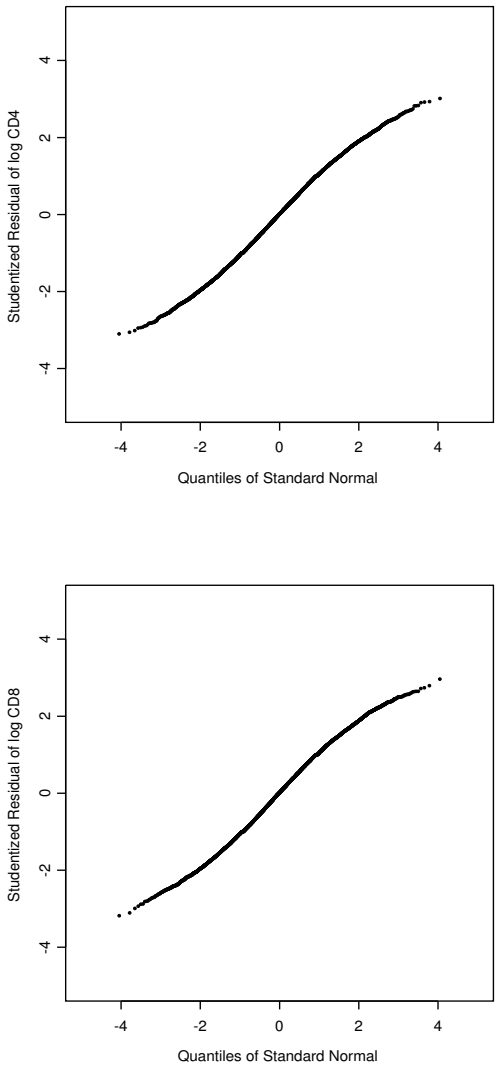


Figure 3.2: Q-Q plots for studentized residuals for log CD4 and log CD8.

Chapter 4

Semiparametric Likelihood Approach

4.1 Introduction

As we have stated in Section 1.4, two issues may be of interest for joint models: characterizing the longitudinal process and characterizing the time-to-event. A drawback of the conditional score approach is that it might be less efficient compared to the likelihood methods, since it makes inefficient use of the information in the longitudinal data. Also, the conditional score approach places no restrictions on the nature of the random effects. Random effects associated with continuous covariates such as CD4 are likely to have a "smooth" distribution, so the conditional score considers too wide a class. In addition, no insight into the nature of the true distribution is possible from the conditional score method, and inference on the longitudinal data process is not available.

In this chapter, we propose a likelihood approach to inference in joint models requiring only that the random effects have distribution in a plausible class with

"smooth" densities. The approach is "semiparametric" in that the density is thus not assumed of a specific parametric form and yields an estimator for the density that may yield insight into features of the population.

The approach accommodates inference on both the longitudinal and time-to-event processes. In Section 4.2, we define the particular case of the joint model on which we focus and derive a likelihood-based EM algorithm for implementation in Section 4.3. In Section 5.1, we exploit the availability of this method to investigate the consequences of violation of the normality assumption and to evaluate performance via simulation. We apply the methods to the ACTG 175 data in Section 5.2.

4.2 Model and Density Approximation

For simplicity, we consider a scalar time-dependent covariate and a single time-independent covariate (e.g. CD4 and indicator for treatment group in ACTG data). Longitudinal observations W_{ij} at times t_{ij}, $j = 1, \ldots, m_i$, are assumed to satisfy

$$W_{ij} = X_i(t_{ij}) + e_{ij}, \tag{4.1}$$

where the $e_{ij} \sim \mathcal{N}(0, \sigma^2)$ are independent deviations reflecting intra-individual variation including measurement error, and the t_{ij} are assumed sufficiently separated that within-subject serial correlation is negligible. Here, $X_i(u)$ is the "inherent" value of the longitudinal response at time u, which depends on a vector of individual-specific random effects α_i. For definiteness, we take $X_i(u) = \alpha_{i0} + \alpha_{i1}u$, so $\alpha_i = (\alpha_{i0}, \alpha_{i1})^T$, but this may be extended to more complicated relationships; see Section 5.3. Let Z_i be a vector of time-independent covariates for subject i. With these definitions,

the observed data on subject i are $(V_i, \Delta_i, W_i, t_i, Z_i)$, independent across i, where $W_i = (W_{i1}, \ldots, W_{im_i})^T$, and $t_i = (t_{i1}, \ldots, t_{im_i})^T$.

The random effects α_i are usually assumed normally distributed, possibly differently according to values of Z_i. We relax this to require only that the α_i given Z_i have density belonging to a class of "smooth" densities studied by Gallant and Nychka (1987). As discussed by Zhang and Davidian (2001), densities in this class are sufficiently differentiable to rule out behavior such as kinks, jumps, or oscillation, and may be skewed, multi-modal, and fat- or thin-tailed relative to the normal density, which is also a member of the class. Gallant and Nychka (1987) provided a theoretical basis for approximating densities in this class by a truncated series expansion, which also suggests a density estimator, and referred to the approximation and estimator as "seminonparametric" (SNP).

For the ACTG 175 data, Z_i is a scalar, dichotomous treatment indicator. Figure 1.1(b) and the analogous plot for combination treatment suggest that variation in $\log_{10}$ CD4 intercepts and slopes is similar for $Z_i = 0, 1$. Thus, for definiteness, we discuss the SNP representation for α_i assuming the densities for scalar $Z_i = 0, 1$ belong to the same location family and let

$$\alpha_i = \mu_0(1 - Z_i) + \mu_1 Z_i + R\xi_i, \tag{4.2}$$

where $\mu = (\mu_0^T, \mu_1^T)^T$; R is lower triangular with $\text{vech}(R) = \theta = (R_{00}, R_{01}, R_{11})^T$ not depending on Z_i; and ξ_i is independent of Z_i with "smooth" density approximated by

$$h_K(z) = P_K^2(z)\varphi(z) = \left(\sum_{0 \le i_1 + i_2 \le K} a_{i_1 i_2} z_1^{i_1} z_2^{i_2} \right)^2 \varphi(z), \quad z = (z_1, z_2)^T, \tag{4.3}$$

for fixed $K \geq 0$. Here $\varphi(z)$ is the bivariate standard normal density, and $P_K(z)$ is a polynomial of all powers and cross products of (z_1, z_2) up to order K with coefficients $a_{i_1 i_2}$, $i_1, i_2 = 0, 1, \ldots, K$ and $0 \leq i_1 + i_2 \leq K$. E.g., for $K = 2$, $P_2(z) = a_{00} + a_{10}z_1 + a_{01}z_2 + a_{20}z_1^2 + a_{11}z_1z_2 + a_{02}z_2^2$. For general K, the coefficients may be collected into the $d = (K+1)(K+2)/2$-dimensional vector $a = (a_{00}, \ldots, a_{0K}, a_{10}, \ldots, a_{1(K-1)}, \ldots, a_{K0})^T$.

To ensure that $h_K(z)$ is a density, the $a_{i_1 i_2}$ must be such that $\int h_K(z)dz = 1$. Zhang and Davidian (2001) show that this may be accomplished by requiring $E\{P_K^2(U)\} = 1$, where $U = (U_1, U_2)^T$ is bivariate standard normal, which is equivalent to requiring $a^T A a = 1$ for $(d \times d)$ positive definite matrix A with (i, j) element $E(U_1^{i_1+j_1})E(U_1^{i_2+j_2})$, where the ith and jth elements of a have subscripts (i_1, i_2) and (j_1, j_2). These moments may be found using recursive formulas as in Johnson and Kotz (1994). Writing $A = BB^T$ and $c = B^T a$, $a^T A a = c^T c = 1$. Because $-a$ and hence $-c$ yield the same density, consider c on the half-unit d-dimensional sphere. By a polar coordinate transformation, c satisfying $c^T c = 1$ may be written as $c = (c_1, \ldots, c_d)^T$ with

$$\begin{aligned}
c_1 &= \sin(\phi_1), \\
c_2 &= \cos(\phi_1)\sin(\phi_2), \\
&\vdots \\
c_{d-1} &= \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{d-2})\sin(\phi_{d-1}), \\
c_d &= \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{d-2})\cos(\phi_{d-1}),
\end{aligned}$$

where $-\pi/2 < \phi_r \leq \pi/2$, $r = 1, \ldots, d-1$. Zhang and Davidian (2001) thus suggest

parameterizing $h_K(z)$ in terms of $\phi = (\phi_1, \ldots, \phi_{d-1})^T$, so imposing $\int h_K(z)dz = 1$ "automatically."

As shown by Zhang and Davidian (2001), (4.3) is capable of approximating a variety of shapes; thus, the density representation for $\alpha_i|Z_i$ in (4.2) is also a flexible form, which for fixed K depends on the parameters $(\mu^T, \theta^T, \phi^T)^T$. It is straightforward to observe that if $K = 0$, ξ_i is standard normal, so that, conditional on Z_i, α_i is $\mathcal{N}\{\mu_0(1 - Z_i) + \mu_1 Z_i, RR^T\}$, the usual assumption. For $K > 0$, $E(\alpha_i|Z_i) = \mu_0(1 - Z_i) + \mu_1 Z_i + RE(\xi_i)$, and $\text{var}(\alpha_i) = R\text{var}(\xi_i)R^T$.

To link the longitudinal response to the time-to-event, we assume for each i that

$$\begin{aligned} \lambda_i(u) &= \lim_{du \to 0} du^{-1} \text{pr}\{u \leq T_i < u + du | T_i \geq u, \alpha_i, Z_i, C_i, e_i, t_i\} \\ &= \lambda_0(u) \exp\{\gamma X_i(u) + \eta Z_i\}, \end{aligned} \tag{4.4}$$

where $\lambda_0(u)$ is an unspecified baseline hazard function, and $e_i = (e_{i1}, \ldots, e_{im_i})^T$; (4.4) incorporates the assumption that censoring, timing of measurements, and covariate errors are noninformative. In addition, we assume C_i is independent of e_i and t_i given Z_i. Thus, (V_i, Δ_i) and W_i are independent given α_i and Z_i.

When interest lies in the longitudinal data process, inference on functions of $(\mu^T, \theta^T, \phi^T)^T$ is the objective. Inference on (γ, η) is of primary interest when the relationship between prognosis and longitudinal response is the focus.

4.3 Estimation Procedure

4.3.1 Likelihood Function and Existence of Estimators

The observed data are (V, Δ, W, t, Z), where $V = (V_1, \ldots, V_n)^T$, $\Delta = (\Delta_1, \ldots, \Delta_n)^T$, $W = (W_1^T, \ldots, W_n^T)^T$, $t = (t_1^T, \ldots, t_n^T)^T$, and $Z = (Z_1, \ldots, Z_n)^T$. Letting $\Omega = (\gamma, \eta, \sigma^2, \mu^T, \theta^T, \phi^T, \lambda_0)^T$ be the vector of all parameters for fixed K, similar to Wulfsohn and Tsiatis (1997), the observed-data likelihood is

$$\begin{aligned} L(\Omega; V, \Delta, W, t, Z) &= \prod_{i=1}^{n} \int \left\{ \prod_{j=1}^{m_i} p(W_{ij}|\alpha_i, \sigma^2, t_{ij}) \right\} p(V_i, \Delta_i|\alpha_i, Z_i, \gamma, \eta, \lambda_0) \\ &\quad \times p(\alpha_i|Z_i, \mu, \theta, \phi) d\alpha_i \\ &= \prod_{i=1}^{n} \int L_{ci}(\Omega; V_i, \Delta_i, W_i, t_i, Z_i, \alpha_i) d\alpha_i, \end{aligned} \tag{4.5}$$

where, noting $\mu_{Z_i} = \mu_0(1 - Z_i) + \mu_1 Z_i$,

$$\begin{aligned} p(W_{ij}|\alpha_i, \sigma^2, t_{ij}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(W_{ij} - \alpha_{i0} - \alpha_{i1} t_{ij})^2}{2\sigma^2} \right\}, \\ p(V_i, \Delta_i|\alpha_i, Z_i, \gamma, \eta, \lambda_0) &= [\lambda_0(V_i) \exp\{\gamma(\alpha_{i0} + \alpha_{i1} V_i) + \eta Z_i\}]^{\Delta_i} \\ &\quad \times \exp\left[-\int_0^{V_i} \lambda_0(u) \exp\{\gamma(\alpha_{i0} + \alpha_{i1} u) + \eta Z_i\} du \right], \\ p(\alpha_i|Z_i, \mu, \theta, \phi) &= h_K\{R^{-1}(\alpha_i - \mu_{Z_i})\}|R|^{-1} \\ &= P_K^2\{R^{-1}(\alpha_i - \mu_{Z_i})\} \varphi\{R^{-1}(\alpha_i - \mu_{Z_i})\}|R|^{-1}. \end{aligned}$$

Here, the Z_i and potential observation times are ancillary; this is reflected in the form of (4.5).

We now establish that an estimator for Ω maximizing (4.5) for fixed K exists and is well-defined. Let $\mathcal{U} = \{u_{1,\ldots,} u_L\}$ denote the set of death times, and assume that $L > 0$ and that each treatment group has at least one subject. Similar to Johansen

(1983), we have the following results, whose proofs are sketched in Sections 4.4.1 and 4.4.2.

Theorem 1. $L(\Omega; V, \Delta, W, t, Z)$ is maximized when $\lambda_0(u) = 0$ for $u \notin \mathcal{U}$.

Thus the parameters we need to consider are in the set $\mathcal{A} = \{\Omega : \gamma \in \mathbb{R}, \eta \in \mathbb{R}, \sigma^2 > 0, \mu_0 \in \mathbb{R}^2, \mu_1 \in \mathbb{R}^2, R_{11} \neq 0, R_{22} \neq 0, R_{21} \in \mathbb{R}, -\pi/2 < \phi_r \leq \pi/2 \text{ for } r = 1, \ldots, d-1, \lambda_0(u) = 0 \text{ for } u \notin \mathcal{U}, \text{ and } \lambda_0(u_l) \geq 0 \text{ for } l = 0, \ldots, L\}$.

Theorem 2. $\sup_\Omega L(\Omega; V, \Delta, W, t, Z) < \infty$, and there exists $\widehat{\Omega} \in \mathcal{A}$ such that

$$L(\widehat{\Omega}; V, \Delta, W, t, Z) = \sup_{\Omega} L(\Omega; V, \Delta, W, t, Z).$$

4.3.2 EM Algorithm

In principle, (4.5) may be maximized in Ω for fixed K given a means of carrying out the integration. However, such high-dimensional direct optimization is prone to instability. We thus consider use of the EM algorithm to find $\widehat{\Omega}$, treating the α_i as "missing data." From this perspective, the "complete data" are $(V, \Delta, W, t, Z, \alpha)$, where $\alpha = (\alpha_1^T, \ldots, \alpha_n^T)^T$, and the "complete data likelihood" is, from (4.5),

$$L_c(\Omega; V, \Delta, W, t, Z, \alpha) = \prod_{i=1}^{n} L_{ci}(\Omega; V_i, \Delta_i, W_i, t_i, Z_i, \alpha_i). \tag{4.6}$$

The estimator $\widehat{\Omega}$ is obtained by iterating between the E-step and the M-step. In the E-step, the expectation of the logarithm of (4.6) is computed, conditional on the observed data and current parameter estimate, and, in the M-step, the parameter

estimate is updated by maximizing this expectation. Let $\widehat{\Omega}^{(k)}$ be the estimate at the kth iteration, and let $E_{ik}\{g(\alpha_i)\} = E_{\widehat{\Omega}^{(k)}}\{g(\alpha_i)|V_i, \Delta_i, W_i, t_i, Z_i\}$ for any function $g(\cdot)$, where this notation implies evaluation at $\widehat{\Omega}^{(k)}$. Then the estimates of $\sigma^2$and λ_0 at the $(k+1)$th iteration may be shown to be $(\widehat{\sigma}^2)^{(k+1)} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} E_{ik}\{(W_{ij} - \alpha_{i0} - \alpha_{i1}t_{ij})^2\}/\sum_{i=1}^{n} m_i$, and

$$\widehat{\lambda}_0^{(k+1)}(u_l) = \frac{\sum_{i=1}^{n}\Delta_i I(V_i = u_l)}{\sum_{i=1}^{n} E_{ik}\left[\exp\{\gamma^{(k)}(\alpha_{i0} + \alpha_{i1}u_l) + \eta Z_i\}\right] Y_i(u_l)},$$

where $Y_i(u) = I(V_i \geq u)$ is the "at risk" process. The $(k+1)$th iterates $(\widehat{\gamma}^{(k+1)}, \widehat{\eta}^{(k+1)})^T$ are not available in closed form; however, following Wulfsohn and Tsiatis (1997), may be approximated by one-step Newton-Raphson estimators. Letting $\beta = (\gamma, \eta)^T$, the update is $\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} + I^{-1}_{\hat{\beta}^{(k)}} S_{\hat{\beta}^{(k)}}$, where the score vector for β at iteration k has sth element

$$S_{\hat{\beta}_s^{(k)}} = \sum_{i=1}^{n}\Delta_i\left\{E_{ik}(G_{sii}) - \frac{\sum_{j=1}^{n} E_{jk}(G_{sij}H_{kij})Y_j(V_i)}{\sum_{j=1}^{n} E_{jk}(H_{kij})Y_j(V_i)}\right\},$$

$H_{kij} = \exp\left\{\widehat{\gamma}^{(k)}(\alpha_{j0} + \alpha_{j1}V_i) + \eta^{(k)}Z_j\right\}$, $G_{sij} = \alpha_{j0} + \alpha_{j1}V_i$ for $s = 1$ and $G_{sij} = Z_j$ for $s = 2$, and $I_{\hat{\beta}^{(k)}}$ is the information for β at the kth iteration, which is equal to the negative derivative of $S_{\hat{\beta}^{(k)}}$, with (s, r) element

$$\begin{aligned} I_{\hat{\beta}^{(k)},s,r} \;=\; & \sum_{i=1}^{n}\Delta_i\left[\frac{\sum_{j=1}^{n} E_{jk}(G_{sij}G_{rij}H_{kij})Y_j(V_i)}{\sum_{j=1}^{n} E_{jk}(H_{kij})Y_j(V_i)}\right. \\ & \left. - \frac{\left\{\sum_{j=1}^{n} E_{jk}(G_{sij}H_{kij})Y_j(V_i)\right\}\left\{\sum_{j=1}^{n} E_{jk}(G_{rij}H_{kij})Y_j(V_i)\right\}}{\left\{\sum_{j=1}^{n} E_{jk}(H_{kij})Y_j(V_i)\right\}^2}\right], \end{aligned}$$

for $s, r = 1, 2$. Because the parameters $(\mu^T, \theta^T, \phi^T)^T$ that characterize the random effects densities are separated from the others in (4.6), their $(k+1)$th iterates may be found using any standard optimization software to maximize

$$\sum_{i=1}^{n} E_{ik}\left\{\log p(\alpha_i|Z_i, \mu_{Z_i}, \theta, \phi)\right\} =$$

$$\sum_{i=1}^{n} E_{ik}\left(\log\left[P^2\{R^{-1}(\alpha_i-\mu_{Z_i})\}\varphi\{R^{-1}(\alpha_i-\mu_{Z_i})\}|R|^{-1}\right]\right). \tag{4.7}$$

The expressions for the iterative updates for all parameters involve intractable integrals of the form

$$E_{ik}\{g(\alpha_i)\} = \int g(\alpha_i)p(\alpha_i|V_i,\Delta_i,W_i,t_i,Z_i,\widehat{\Omega}^{(k)})d\alpha_i,$$

where $p(\alpha_i|V_i,\Delta_i,W_i,t_i,Z_i,\Omega)$ is the density of α_i given the observed data, that must be computed numerically. In Sections 5.1 and 5.2, we use Gauss-Hermite quadrature; see Section 4.4.3.

When $K = 0$ and there are no covariates Z_i, the algorithm reduces to that of Wulfsohn and Tsiatis (1997) for assumed normally-distributed random effects. Implementation of more flexible models for various choices of K thus involves only the additional requirement of maximizing (4.7), which involves just moderate increase in computational burden.

As a by-product of the approach, we can easily estimate the baseline cumulative hazard $\Lambda_0(u) = \int_0^u \lambda_0(t)dt$ at the final iteration as $\widehat{\Lambda}_0(u) = \sum_{u_l \le u} \widehat{\lambda}_0(u_l)$.

To obtain starting values for Ω for the EM algorithm, we have had success using the following approach. We use the conditional score estimates of γ, η, λ_0, and σ^2 for these elements of Ω, as the conditional score is fast and straightforward. To obtain starting values for the SNP parameters $(\mu^T,\theta^T,\phi^T)^T$, we compute the "maximum likelihood estimate" of $(\mu^T,\theta^T,\phi^T)^T$ treating the individual ordinary least squares estimates of the α_i as "data" and maximizing the density corresponding to (4.2) for the particular choice of K under consideration. This may be accomplished using

standard optimization software, for which we take as starting values randomly chosen values for ϕ, with μ and R computed by moment methods for the given ϕ. As suggested by Davidian and Gallant (1993), it is essential when fitting SNP densities to try a series of starting values in order to ensure that the maximum has been found. In Sections 5.1 and 5.2, we used 100 initial starting values to search for the "maximum likelihood estimate" to use as the starting value for the EM algorithm.

There are several possible approaches to deriving approximate standard error estimates. When interest focuses on explicit model parameters, such as γ and η, following Wulfsohn and Tsiatis (1997), for parameter δ, say, with final estimate $\widehat{\delta}$, $\text{var}(\widehat{\delta})$ may be estimated by the negative inverse of the derivative (slope) of the profile score at $\widehat{\delta}$. Let $\Omega_{-\delta}$ be the set of parameters in Ω except δ. The profile score $S_\delta\{\widehat{\Omega}_{-\delta}(\delta)\}$ is the derivative of the profile log likelihood with δ fixed and $\Omega_{-\delta}$ estimated by the restricted likelihood estimator $\widehat{\Omega}_{-\delta}$ computed by a separate EM algorithm identical to the above with δ held fixed. $S_\delta\{\widehat{\Omega}_{-\delta}(\delta)\}$ is of the same form as $S_{\hat{\delta}}$ with $\widehat{\delta}$ replaced by the given δ and $\Omega_{-\delta}$ replaced by $\widehat{\Omega}_{-\delta}$. The derivative of the profile score may be approximated by the difference method.

Alternatively, a nonparametric bootstrap in which individuals are resampled at random may be used, although this may be computationally prohibitive for large data sets. Standard errors may also be based on the inverse of the observed Hessian; i.e., the second derivative matrix of (4.5) evaluated at $\widehat{\Omega}$. This may be calculated via numerical approximation, where the integral in (4.5) is carried out by Gauss-Hermite quadrature. The dimension of this matrix can be immense when the number of distinct times-to-event, and hence number of parameters associated with $\lambda_0(u)$, is

large. We have found that treating the $\widehat{\lambda}_0(u_l)$, $l = 1, \ldots, L$, as fixed in this calculation yields standard error estimates comparable to those obtained from the full matrix; this may reflect an approximate "orthogonality" of the baseline hazard and other parameters. Standard errors for $E(\alpha_i|Z_i)$, for example, which are nonlinear functions of $(\mu^T, \theta^T, \phi^T)^T$, maybe obtained directly via the bootstrap or follow from application of the delta method using the appropriate submatrix of the inverse Hessian.

4.3.3 Choosing K

The value of K controls the degree of flexibility of the random effects density representation; $K > 0$ allows departures from normality. The value of K acts similarly to a bandwidth in density estimation, and hence does not represent, for example, the number of "subpopulations" in the underlying population. The foregoing developments take K to be fixed, so one possibility for practical use is to base inference on a fixed $K > 0$ to accommodate possible nonnormality. Alternatively, following Davidian and Gallant (1993), K may be chosen objectively based on various information criteria that have the general form

$$\{-\ell(\widehat{\Omega}; V, \Delta, W, t, Z) + C(N)p_{net}\}/N, \tag{4.8}$$

where $\ell(\Omega; V, \Delta, W, t, Z)$ is the logarithm of (4.5), $N = \sum_{i=1}^{n} m_i$ is the total number of longitudinal observations, p_{net} is the number of free parameters in Ω, and $C(N)p_{net}$ is the penalty factor to compensate for small $-\ell(\widehat{\Omega}; V, \Delta, W, t, Z)$ achieved by overparameterization, with $C(N) = 1$ for the Akaike criterion (AIC), $C(N) = \log\log N$ for the Hannan-Quinn criterion (HQ), and $C(N) = \log N/2$ for the Schwarz criterion (BIC). For a given criterion, (4.8) is computed for different K, and K minimizing

(4.8) is preferred. AIC tends to select larger K, BIC tends to select smaller K with HQ intermediate. In simpler settings than that here, Eastwood and Gallant (1991) show that confidence intervals for model parameters based on such adaptive rules for choosing K achieve nominal coverage. We demonstrate the performance of these criteria in our model in the next chapter.

4.4 Technical details

In this section, we present details underlying the developments in previous sections.

4.4.1 Proof of Theorem 1

Note first that

$$\begin{aligned}
& \exp\left[-\int_0^{V_i} \lambda_0(u)\exp\{\gamma(\alpha_{i0}+\alpha_{i1}u)+\eta Z_i\}du\right] \\
= \;& \exp\left[-\int_0^{V_i} \exp\{\gamma(\alpha_{i0}+\alpha_{i1}u)+\eta Z_i\}d\Lambda_0(u)\right] \\
= \;& \exp\left[-\sum_{u_l\le V_i}\int_{\{u_l\}} \exp\{\gamma(\alpha_{i0}+\alpha_{i1}u)+\eta Z_i\}d\Lambda_0(u)\right. \\
& \left.-\int_{[0,V_i]-\cup_{u_l\le V_i}\{u_l\}} \exp\{\gamma(\alpha_{i0}+\alpha_{i1}u)+\eta Z_i\}d\Lambda_0(u)\right].
\end{aligned}$$

It is straightforward to see that this is

$$\le \exp\left[-\sum_{u_l\le V_i}\int_{\{u_l\}} \exp\{\gamma(\alpha_{i0}+\alpha_{i1}u)+\eta Z_i\}d\Lambda_0(u)\right],$$

with the equality when $d\Lambda_0(u)=0$ for $u\in[0,V_i]-\cup_{u_l\le V_i}\{u_l\}$.

4.4.2 Proof of Theorem 2

we first show $\sup_\Omega(L(\Omega; V, \Delta, W, t, Z)) < \infty$. For $i = 1, \ldots, n$, clearly

$$\prod_{j=1}^{m_i} p(W_{ij}|\alpha_i, \sigma^2, t_{ij}) \leq \prod_{j=1}^{m_i} p(W_{ij}|\tilde{\alpha}_i, \tilde{\sigma}_i^2, t_{ij}),$$

where $\tilde{\alpha}_i$ is the ordinary least squares estimator of α_i, and $\tilde{\sigma}_i^2 = \sum_{j=1}^{m_i}(W_{ij} - \tilde{\alpha}_{i0} - \tilde{\alpha}_{i1}t_{ij})^2/m_i$. $p(V_i, \Delta_i|\alpha_i, Z_i, \gamma, \eta, \lambda_0) \leq 1$ can be shown by noting that if $\Delta_i = 0$, $p(V_i, \Delta_i|\alpha_i, Z_i, \gamma, \eta, \lambda_0) = \exp\left[-\int_0^{V_i} \lambda_0(u)\exp\{\gamma(\alpha_{i0} + \alpha_{i1}u) + \eta Z_i\}du\right] \leq 1$, as the integrand ≥ 0; otherwise $\Delta_i = 1$, $p(V_i, \Delta_i|\alpha_i, Z_i, \gamma, \eta, \lambda_0) \leq \lambda_0(V_i)\exp\{\gamma(\alpha_{i0}+\alpha_{i1}V_i)+ \eta Z_i\}]\exp[-\lambda_0(V_i)\exp\{\gamma(\alpha_{i0} + \alpha_{i1}V_i) + \eta Z_i\}] \leq 1$, which follows from $xe^{-x} \leq 1$ for any $x \in \mathbb{R}$. These results imply that $L_{c_i}(\Omega; V_i, \Delta_i, W_i, t_i, Z_i, \alpha_i)$ is bounded from above by $\left\{\prod_{j=1}^{m_i} p(W_{ij}|\tilde{\alpha}_i, \tilde{\sigma}_i^2, t_{ij})\right\} p(\alpha_i|Z_i, \mu_{Z_i}, \theta, \phi)$, so $L(\Omega; V, \Delta, W, t, Z) \leq \prod_{i=1}^n \left\{\prod_{j=1}^{m_i} p(W_{ij}|\tilde{\alpha}_i, \tilde{\sigma}_i^2)\right\} \times \int p(\alpha_i|Z_i, \mu_{Z_i}, \theta, \phi)d\alpha_i = \prod_{i=1}^n \left\{\prod_{j=1}^{m_i} p(W_{ij}|\tilde{\alpha}_i, \tilde{\sigma}_i^2)\right\}$, so $\sup_\Omega L(\Omega; V, \Delta, W, t, Z) < \infty$.

We now show that the supremum is attained in $\mathcal{A}$. Because $L_{c_i}(\Omega; V_i, \Delta_i, W_i, t_i, Z_i, \alpha_i)$ is continuous in Ω, so is $L(\Omega; V, \Delta, W, t, Z)$ by the Dominated Convergence Theorem, and $L(\Omega; V, \Delta, t, W, Z)$ for $\phi_r = -\pi/2$ is the same for $\phi_r = \pi/2$, $r = 1, 2, \ldots, d-1$. Let $\Omega_\star$ be the set of parameters in Ω except ϕ, and

$$\begin{aligned}\mathcal{A}_\star \;=\; &\{\Omega_\star : \gamma \in \mathbb{R}, \eta \in \mathbb{R}, \sigma^2 > 0, \mu_0 \in \mathbb{R}^2, \mu_1 \in \mathbb{R}^2, R_{00} \neq 0, R_{11} \neq 0, R_{01} \in \mathbb{R},\\ &\lambda_0(u) = 0 \text{ for } u \notin \{u_{1,\ldots,}u_L\}, \text{ and } \lambda_0(u_l) \geq 0 \text{ for } l = 0, \ldots, L\}.\end{aligned}$$

The result follows if we can show $L(\Omega; V, \Delta, W, t, Z) \to 0$ when $\Omega_\star$ goes to points in $\mathcal{D} = \partial\mathcal{A}_\star \cap \mathcal{A}_\star{}^c$, the boundary points in $\mathcal{A}_\star$ not belonging to $\mathcal{A}_\star$. By the Dominated Convergence Theorem, we need only show there exists $L_{c_i}(\Omega; V_i, \Delta_i, t_i, W_i, Z_i, \alpha_i) \to 0$

when $\Omega_\star$ goes to points in $\mathcal{D}$, which follows from $\prod_{j=1}^{m_i} p(W_{ij}|\alpha_i, \sigma^2, t_{ij}) \to 0$ as $\sigma^2 \to 0$ or $+\infty$; $\prod_{i=1}^{n} p(V_i, \Delta_i|\alpha_i, Z_i, \gamma, \eta, \lambda_0) \to 0$ as $\gamma \to \pm\infty$ for $\alpha_i \notin \mathcal{B}$, $\to 0$ as $\eta \to \pm\infty$, $= 0$ as $\lambda_0(u_l) = 0$ for $l = 0, \ldots, L$, $\alpha_i \notin \mathcal{B}$; and $p(\alpha_i|Z_i, \mu_{Z_i}, \theta, \phi) \to 0$ as $\mu_s \to \pm\infty$, or $R_{km} \to \pm\infty$, or $R_{kk} \to 0$, $k = 0, 1$, $m = 0, 1$, for $Z_i = s, s = 0, 1$; using $xe^{-x} \to 0$ as $x \to 0$ or $+\infty$, and where $\mathcal{B} = \{\alpha_i : \alpha_{i0} + \alpha_{i1}u_l = 0 \text{ for some } u_l, l = 0, \ldots, L\}$, which has Lebesgue measure 0.

4.4.3 Computation of $E_{ik}\{g(\alpha_i)\}$

Note that

$$p(\alpha_i|V_i, \Delta_i, W_i, t_i, Z_i, \widehat{\Omega}^{(k)}) \propto L_{ci}(\widehat{\Omega}^{(k)}; V_i, \Delta_i, W_i, t_i, Z_i, \alpha_i).$$

By some straightforward algebra, for $Z_i = s$,

$$\begin{aligned} L_{ci}(\Omega; V_i, \Delta_i, W_i, t_i, Z_i, \alpha_i) \quad \propto \quad & \exp[\Delta_i\{\gamma(\alpha_{i0} + \alpha_{i1}V_i) + \eta Z_i\}] \\ & \times \exp\left[-\sum_{l=1}^{L} \lambda_0(u_l) \exp\{\gamma(\alpha_{i0} + \alpha_{i1}u_l) + \eta Z_i\} Y_i(u_l)\right] \\ & \times P_K^2\{R^{-1}(\alpha_i - \mu_s)\}|R|^{-1}\varphi\{D_i(\alpha_i - c_{si})\}. \end{aligned}$$

where D_i satisfies

$$D_i D_i^T = B_i = \begin{bmatrix} B_{i00} & B_{i01} \\ B_{i01} & B_{i11} \end{bmatrix},$$

$$B_{i00} = (R_{01}^2 + R_{00}^2)/(R_{00}^2 R_{11}^2) + m_i/\sigma^2,$$

$$B_{i01} = -(R_{00}R_{01})/(R_{00}^2 R_{11}^2) + \sum_{j=1}^{m_i} t_{ij}/\sigma^2,$$

$$B_{i11} = R_{00}^2/(R_{00}^2 R_{11}^2) + \sum_{j=1}^{m_i} t_{ij}^2/\sigma^2,$$

and

$$c_{si} = B_i^{-1} \begin{bmatrix} \sum_{j=1}^{m_i} W_{ij}/\sigma^2 + \{(R_{01}^2 + R_{11}^2)\mu_{s0} - R_{00}R_{01}\mu_{s1}\}/(R_{00}^2 R_{11}^2) \\ \sum_{j=1}^{m_i} W_{ij}t_{ij}/\sigma^2 + \{R_{00}^2\mu_{s1} - R_{00}R_{01}\mu_{s0}\}/(R_{00}^2 R_{11}^2) \end{bmatrix}.$$

Let

$$Q(\alpha_i; V_i, \Delta_i, W_i, t_i, Z_i, \Omega) = \exp[\Delta_i\{\gamma(\alpha_{i0} + \alpha_{i1}V_i) + \eta Z_i\}]$$
$$\times \exp\left[-\sum_{l=1}^{L} \lambda_0(u_l) \exp\{\gamma(\alpha_{i0} + \alpha_{i1}u_l) + \eta Z_i\} Y_i(u_l)\right] P_K^2\{R^{-1}(\alpha_i - \mu_s)\}|R|^{-1}.$$

Then it follows that

$$\begin{aligned} E_{ik}\{g(\alpha_i)\} &= \frac{\int g(\alpha_i) Q(\alpha_i; V_i, \Delta_i, t_i, W_i, Z_i, \widehat{\Omega}^{(k)}) \varphi\{D_i^{(k)}(\alpha_i - c_{si}^{(k)})\} d\alpha_i}{\int Q(\alpha_i; V_i, \Delta_i, t_i, W_i, Z_i, \widehat{\Omega}^{(k)}) \varphi\{D_i^{(k)}(\alpha_i - c_{si}^{(k)})\} d\alpha_i} \\ &= \frac{\int g\{\alpha_i(\beta_i)\} Q\{\alpha_i(\beta_i); V_i, \Delta_i, t_i, W_i, Z_i, \widehat{\Omega}^{(k)}\} \exp(-\beta_i^2) d\beta_i}{\int Q\{\alpha_i(\beta_i); V_i, \Delta_i, t_i, W_i, Z_i, \widehat{\Omega}^{(k)}\} \exp(-\beta_i^2) d\beta_i} \end{aligned}$$

where $\beta_i = (\sqrt{2})^{-1}(D_i^{(k)})^{-1}(\alpha_i - c_{si}^{(k)})$ and $\alpha_i(\beta_i)$ means α_i expressed in terms of β_i; and $D_i^{(k)}$ and $c_{si}^{(k)}$ are D_i and c_{si}, respectively, with Ω replaced by $\widehat{\Omega}^{(k)}$.

Now $E_{ik}\{g(\alpha_i)\}$ can be approximated by M-point Gauss-Hermite quadrature (e.g. Abramowitz and Stegun, 1970, p. 924). If x_s, $s = 1, \ldots, M$ are the abscissas with corresponding weights w_s, the approximation to the integral over β_i is

$$\frac{\sum_{s=1}^{M}\sum_{t=1}^{M} g\{\alpha_i(x_s, x_t)\} Q\{\alpha_i(x_s, x_t); V_i, \Delta_i, t_i, W_i, Z_i, \widehat{\Omega}^{(k)}\} w_s w_t}{\sum_{s=1}^{M}\sum_{t=1}^{M} Q\{\alpha_i(x_s, x_t); V_i, \Delta_i, t_i, W_i, Z_i, \widehat{\Omega}^{(k)}\} w_s w_t},$$

where $\alpha_i(x_s, x_t)$ indicates that α_i in this expression is a function of x_s and x_t.

Chapter 5

Application and Performance of the Semiparametric Likelihood Approach

In Section 5.1, we demonstrate the performance of the semiparametric likelihood approach via simulation. In Section 5.2, we illustrate the use of the method for the ACTG 175 data. We compare our approach to the conditional score method.

5.1 Simulation Studies

The availability of a likelihood method that requires only a mild, plausible assumption on the random effects both provides a benchmark against which robustness to departures from normality may be gauged and a flexible framework for addressing questions of interest.

To evaluate robustness and performance of the approach, we conducted simulations under several scenarios similar to that in Tsiatis and Davidian (2001), where 200 Monte Carlo data sets were generated for each scenario. In all cases, for each of

$n = 200$ subjects, longitudinal observations W_{ij} of $X_i(u) = \alpha_{0i} + \alpha_{1i}u$ were generated at times $t_{ij} = 0, 2, 4, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80$ weeks, with a 10% missingness rate at any time after week 16, according to the linear mixed effects model (4.1) with independent, normally-distributed errors e_{ij} with $\sigma^2 = 0.60$. We report here on three scenarios; qualitative results for others were similar. In the first two situations, there was no time-independent covariate Z_i, so $E(\alpha_i) = (4.173, -0.0103)^T$ for all i, and random effects followed either (i) a bivariate normal or (ii) bivariate bimodal mixture of normals as described in Davidian and Gallant (1993) with mixing proportion $p = 0.5$ and sep=4. For both, the covariance matrix of the random effects D satisfied $\text{vech}(D) = (4.96, -0.0456, 0.012) = \{\text{var}(\alpha_{i0}), \text{cov}(\alpha_{i0}, \alpha_{i1}), \text{var}(\alpha_{i1})\}^T$. Figure 5.1(a) shows the true density for the bimodal case (ii). The hazard relationship was taken to be $\lambda_i(u) = \lambda_0(u)\exp\{\gamma X_i(u)\}$ with $\gamma = -1.0$ and $\lambda_0 = 1$ for $u \geq 16$ and 0 otherwise. The censoring distribution was exponential with mean 110 weeks, leading to censoring rates 52.9% and 52.4%, respectively, for both cases. The third scenario, (iii), was the same as (ii) except that a treatment variable Z_i with $P(Z_i = 1) = 0.5$ was generated, the hazard followed (4.4) with $\eta = 0.0$, and the random effects were from the same location family as the bimodal mixture of normals with $E(\alpha_i|Z_i = 0) = (4.173, -0.0103)^T$ and $E(\alpha_i|Z_i = 1) = (4.173, 0.0412)^T$.

To address the situation where inference on the hazard is of interest, γ [and η in (iii)] were estimated using the unattainable "ideal" estimator for which the true value of $X_i(u)$ is known at each failure time for all subjects and the hazard model fit via ordinary partial likelihood; the conditional score estimator of Tsiatis and Davidian (2001); and the semiparametric likelihood (SNP) estimator with $K = 0, 1, 2$, where

two-point Gauss-Hermite quadrature was used to calculate $E_{ik}[g(\alpha_i)]$. Preliminary studies revealed that increasing the number of abscissæ beyond two did not improve performance appreciably. In all cases, 95% Wald confidence intervals for γ (and η) were constructed as the estimated value plus or minus 1.96 times an estimated standard error. For the ideal method, the usual standard error formula for partial likelihood was used; for the conditional score, standard errors were calculated as described by Tsiatis and Davidian (2001); and the profile score approach in Section 4.3.2 was used for the likelihood estimators. We also estimated the baseline cumulative hazard function $\Lambda_0(u)$, where the Breslow estimator (Breslow, 1972) was used for the "ideal" method. As suggested by Tsiatis and Davidian (2001), the corresponding conditional score estimator is $\int_0^u [\sum_{i=1}^n \exp\{\widehat{\gamma}_{CS} S_i(u, \widehat{\gamma}, \widehat{\sigma}^2_{CS}) - \widehat{\gamma}^2_{CS}\Sigma_i(u, \widehat{\sigma}^2_{CS})/2 + \widehat{\eta}_{CS} Z_i\} Y_i(u)]^{-1} \sum_{i=0}^n dN_i(u)$, where $N_i(u) = I(V_i \leq u, \Delta_i = 1)$ is the counting process; $\widehat{\gamma}_{CS}$, $\widehat{\eta}_{CS}$, and $\widehat{\sigma}_{CS}$ are the conditional score estimators proposed by Tsiatis and Davidian (2001); $S_i(u, \widehat{\gamma}_{CS}, \widehat{\sigma}^2_{CS}) = \widehat{X}_i(u) + dN_i(u)\Sigma_i(u, \widehat{\sigma}^2_{CS})\widehat{\gamma}_{CS}$; and $\widehat{X}_i(u)$ is the ordinary least squares estimator for $X_i(u)$ based on the observations prior to and including u with variance $\Sigma_i(u, \sigma^2)$.

Inference on the longitudinal data process was assessed via the performance of the SNP estimators for $E(\alpha_i|Z_i)$, $\text{var}(\alpha_i)$, and the densities for α_i. Standard errors for $E(\alpha_i|Z_i)$ were obtained using the delta method and inverse Hessian treating the $\widehat{\lambda}_0(u_l)$ as fixed, as described in Section 4.3.2, with 95% Wald confidence intervals constructed as above.

Results for estimation of γ and η are given in Table 5.1. All estimators show small or negligible bias close to that of the unachievable "ideal" estimator. For the

ideal and SNP estimators, standard errors track the Monte Carlo standard deviations regardless of the distribution of the random effects, and coverage probabilities are close to the nominal level. The conditional score estimator, which takes the random effects distribution to be completely unrestricted, is notably inefficient relative to the SNP estimators, demonstrating the potential for improved precision offered by the likelihood-based approach with restriction to the class of "smooth" densities when the true density is smooth. Table 5.1 also summarizes the good performance of the information criteria, which correctly select $K = 0$ close to 100% of the time under the normal scenario (i) and fail to select $K = 0$ for any of the data sets under the bimodal scenarios (ii) and (iii), with the exception of the conservative BIC procedure under (iii). In practice, examination of the criteria would be augmented by inspection of plots of the estimated densities, allowing the analyst to incorporate subject-matter knowledge and the visual pattern in the selection.

Estimated cumulative hazards for the bimodal case (ii) are shown in Figure 5.2; the pattern for other scenarios is similar. For the SNP approach, the Monte Carlo average is shown for estimates preferred by HQ; the average of estimates preferred by AIC and BIC and over fixed $K = 0, 1, 2$ are virtually the same and identical to the true baseline cumulative hazard and the ideal estimate, while the conditional score yields an overestimate.

Table 5.2 summarizes performance for estimation of the first two moments of the random effects distribution for the longitudinal model. Estimates are close to the true values, and estimated standard errors for the mean track the empirical sampling variation. Coverage probabilities are close to the nominal level in all cases.

Figure 5.1 shows estimated densities chosen by the information criteria for the bimodal case (ii). Panel (b) shows the Monte Carlo average of full joint density estimates using HQ, and (c) and (d) present the averages of marginal densities chosen by each criterion, while panels (e) and (f) show the 200 estimates preferred by HQ; corresponding plots for panels (b), (e), and (f) for AIC and BIC are virtually identical. In almost all cases, the SNP estimators capture accurately the shape of the underlying random effects density. In a few data sets, a spurious "extra mode" is evident in the selected density estimate. In practice, as with any density estimation procedure, visual inspection of the estimate in such circumstances would likely lead the analyst to take a pragmatic approach and choose K to avoid overinterpretation of such behavior.

A striking feature of Tables 5.1 and 5.2 is that using the SNP estimator with $K = 0$, so taking the random effects to be normal, does not involve bias or loss of efficiency in estimating γ, η, $E(\alpha_i|Z_i)$, or $\text{var}(\alpha_i)$ when the true distribution is nonnormal [cases (ii) and (iii)] relative to the estimators where K is chosen adaptively. Likewise, in case (i), choosing K via information criteria when the true distribution is normal entails no deleterious effects. A prominent exception is estimation of the intercept component of $E(\alpha_i|Z_i)$ in the bimodal case (iii), where there is a subject-level covariate Z_i; here, the estimators for $E(\alpha_{i0}|Z_i = z)$ for $z = 0, 1$ using $K = 0$ are roughly 60-65% efficient relative to those with K chosen on the basis of the information criteria, acknowledging nonnormality of the random effects, so that inference on, for example, $E(\alpha_{i0} + \alpha_{i1}u|Z_i)$ will be inefficient. This coincides with previously-seen behavior in ordinary mixed effects models with nonnormal random effects and no informative censoring. In particular, as discussed by Zhang and Davidian (2001) and Tao et al.

(1999), estimation of mean parameters in longitudinal data models corresponding to subject-level covariates seems to be subject to inefficiency if normality is assumed but the true distribution is nonnormal, whereas estimation of those involving within-subject effects such as time appears to be robust to such misspecification. These phenomena persisted over other scenarios not reported here and suggest remarkable robustness to the assumption of normal random effects for most model parameters, contrary to intuition. We discuss this issue further in Section 5.3. However, as we demonstrate in the next section, the ability to estimate the random effects density may provide insights that may aid subject-matter understanding and come at no sacrifice of precision.

5.2 Application to ACTG 175

We apply the approach to the ACTG 175 data. Based on the discussion in Section 1.4.2, we consider the post-twelve-week data and assume $X_i(u) = \alpha_{i0} + \alpha_{i1}u$ represents "inherent" $\log_{10}$ CD4 count for subject i at time u and model (4.1), and we take the hazard to be as in (4.4) with $Z_i = I(\text{ treatment} \neq \text{zidovudine})$. We fitted the full model using the SNP likelihood-based approach with $K = 0, 1, 2, 3, 4$. For inference on the hazard relationship, we also estimated γ and η using the conditional score approach. Estimated standard errors for each parameter were obtained using the same methods as in the simulations.

Results are shown in Table 5.3; estimates for $K = 1$ are almost identical to $K = 0$ with a modest increase in loglikelihood and are omitted for brevity. The SNP estimates of γ and η are similar across K; the conditional score estimate of γ is a bit

smaller in magnitude. Estimated standard errors for the SNP estimates of γ and η are much smaller than those for the conditional score estimator, reflecting "efficiencies" relative to the conditional score between 4.00 and 5.20, and likelihood estimates γ and η and standard errors are similar across K, echoing the simulation results. The results suggest that there is a strong relationship between hazard and $\log_{10}$ CD4, with hazard decreasing with increasing $\log_{10}$ CD4; moreover, once this is taken into account, there appears to be no association between prognosis and treatment. A partial likelihood fit of the usual proportional hazards model to these data with treatment indicator Z_i as the only covariate (i.e., $\gamma = 0$) yields a strong treatment effect ($\widehat{\eta} = -0.375$ with standard error 0.128). Hence, a possible interpretation is that $\log_{10}$ CD4 may have value as a potential surrogate marker (e.g., Tsiatis et al., 1995).

Estimation of $E(\alpha_i|Z_i)$ and the elements of $\text{var}(\alpha_i)$ from the SNP fits are similar across K and suggest that the longitudinal data process is similar in both treatment groups. Estimates of the within-subject error variance σ^2 are virtually identical and equal to 0.11.

Table 5.3 presents the information criteria for each K. AIC and HQ prefer $K = 4$ and, BIC prefers $K = 3$, suggesting that there is evidence that the true random effects distribution deviates from bivariate normality, as suggested by Figure 1.1(b). Figure 5.3(a) depicts the estimated density of the random effects for $K = 4$ in the zidovudine only group, which indicates the presence of a obvious second mode; an extreme, modest third mode is likely an artifact of tail estimation. Figure 5.3(b) and (c) show the corresponding estimated marginal densities superimposed on the histograms of least squares estimates from Figure 1.1(b), and reinforce the existence

of a second mode for the slope. A possible interpretation is that there may be a subpopulation of individuals whose $\log_{10}$ CD4 decrease at a faster rate. It may be that an important covariate has failed to be taken into account that helps to identify these "sicker" patients.

This may be investigated in part by adapting the ad hoc approach of Davidian and Gallant (1992) to this setting. One may fit the basic model presented here and obtain "empirical Bayes" estimates of the individual α_i as

$$\widehat{\alpha}_i = \text{argmax}_{\alpha_i} p(\alpha_i | V_i, \Delta_i, W_i, t_i, Z_i, \widehat{\Omega}).$$

Plots of each component versus baseline covariates, using box plots for categorical covariates, may be inspected for relationships that may suggest refinements of model (4.2) to include additional covariates. For ACTG 175, a number of such covariates, including indicators of homosexual activity, hemophilia, intravenous drug use, prior antiretroviral therapy, and continuous measures such as weight and age were collected on each participant. Figure 5.4 shows representative plots for four covariates, which suggest no such relationships; plots for all other covariates are similar. Covariates that may be associated with the bimodality were apparently unmeasured in the study.

Figure 5.3(d) presents the estimated baseline cumulative hazards for each method. For the SNP likelihood approach, estimates for $K = 0, 2, 4$ are shown; the estimates for $K = 1$ and $K = 3$ are virtually the same. The conditional score estimate seems ill-behaved for the first several weeks, which may be attributable to the way this estimator uses early information; specially-defined risk sets during this period may be non-monotone decreasing. The simulation results suggest that the likelihood-based

estimates may provide a more reliable representation than the others.

5.3 Discussion

We have proposed a semiparametric likelihood approach for a joint model for survival and longitudinal data in which parametric assumptions on the distribution of random effects may be relaxed to that of a "smooth" density. Estimation may be based on an EM algorithm; C code implementing the approach is available from the first author. An important feature of the procedure is that it makes possible the study of robustness to parametric assumptions on the random effects in joint models. Estimation of the densities, a by-product of the approach, offers the potential for guiding subject-matter insight on population.

For smaller data sets, such as those in the simulations, the fit of the model for fixed K takes only a few minutes, including the initial search for starting values. For the ACTG 175, with over 2400 subjects, fits were on the order of a few hours.

The model as presented here may be extended to more complicated situations. For example, it is in principle possible to consider models for individual inherent longitudinal trajectories involving higher dimensional α_i; trajectories that are nonlinear in individual random effects may be entertained. Similarly, the model may be extended to incorporate vector-valued covariates Z_i with both continuous and discrete components. Here, the model for α_i given in (4.2) could be extended to allow the mean and the covariance matrix (through R) to depend on components of Z_i in more general ways. In all these cases, increased model complexity involves both the need

to make assumptions on model components and a likely increase in computational burden.

A major finding, facilitated by the availability of a likelihood procedure that relaxes the normality assumption, is that, as noted in Sections 5.1 and 5.2, inference seems remarkably and surprisingly robust to the assumption on the random effects distributions. In particular, estimation of hazard parameters and certain longitudinal data model parameters under the assumption of normality results in estimators that are apparently approximately unbiased and as efficient as those that accommodate nonnormality even when the true distribution is nonnormal, and estimated cumulative hazards are virtually identical. A similar phenomenon has been observed in other settings; e.g., nonlinear mixed models (Lai and Shih, unpublished technical report 2000-42, Stanford University) and general structural measurement error models (L.A. Stefanski, personal communication). Faced with this result, we speculated that perhaps it is possible that the likelihood-based approach using normality yields consistent estimators even when normality is a misspecification under certain "nice" conditions. As a test of this premise, we carried out additional simulations in which the true random effects distribution was taken to be discrete with mass at only a few points. We found that the apparent robustness of the likelihood-based estimator assuming normality can break down.

We believe that a theoretical basis for this robustness property may exist. Further research is needed.

Table 5.1: Simulation results, hazard parameters. I represents the "ideal" method; CS is conditional score; and SNP is the likelihood method, where $K = 0$ denotes normal random effects, AIC, HQ, BIC are K preferred by AIC, HQ, and BIC, respectively. SD is Monte Carlo standard deviation, SE is average of estimated standard errors, CP is coverage probability of 95% Wald confidence interval, and RE is square of SD for CS divided by that for the indicated estimator. $\%K$ shows percentage of simulated data sets with $K = (0, 1, 2)$ preferred by each information criterion.

			SNP			
	I	CS	$K = 0$	AIC	HQ	BIC
			(i) Normal α_i			
γ	−1.011	−1.041	−0.998	−0.998	−0.998	−0.998
SD	0.081	0.150	0.097	0.097	0.096	0.097
SE	0.081	0.123	0.103	0.103	0.103	0.103
CP	0.955	0.920	0.960	0.960	0.960	0.960
RE	—	1.000	2.422	2.406	2.424	2.420
$\%K$	—	—	—	(91.5,5.0,3.5)	(99.0,1.0,0.0)	(100.0,0.0,0.0)
			(ii) Bimodal α_i			
γ	−1.017	−1.044	−1.000	−0.997	−0.997	−0.996
SD	0.079	0.157	0.105	0.106	0.105	0.105
SE	0.083	0.128	0.106	0.105	0.105	0.104
CP	0.970	0.925	0.950	0.955	0.955	0.955
RE	—	1.000	2.233	2.204	2.229	2.249
$\%K$	—	—	—	(0.0,70.5,29.5)	(0.0,93.5,6.5)	(0.0,100.0,0.0)
			(iii) Bimodal α_i with treatment			
γ	−1.021	−1.064	−1.005	−1.002	−1.003	−1.003
SD	0.086	0.169	0.098	0.099	0.099	0.098
SE	0.088	0.136	0.112	0.110	0.110	0.111
CP	0.955	0.930	0.975	0.970	0.970	0.970
RE	—	1.000	2.977	2.956	2.955	2.984
η	0.005	0.025	0.010	0.013	0.013	0.012
SD	0.234	0.348	0.267	0.265	0.265	0.266
SE	0.236	0.297	0.285	0.284	0.284	0.285
CP	0.950	0.915	0.980	0.975	0.970	0.970
RE	—	1.000	1.695	1.730	1.731	1.715
$\%K$	—	—	—	(0.0,79.5,20.5)	(0.0,81.5,18.5)	(16.0,81.5,2.5)

Table 5.2: Simulation results, means and covariance matrix of random effects distribution and intra-subject variance using the SNP method. For means, entries are defined as in Table 5.1. The Monte Carlo means and SD of the second (slope) entries of $E(\alpha_i)$ and $E(\alpha_i|Z_i)$, along with SE, have been multiplied by 10^2. For entries of $\text{var}(\alpha_i)$ and σ^2, entries are average of Monte Carlo estimates (Monte Carlo standard deviation).

	$K=0$	AIC	HQ	BIC
		(i) Normal α_i		
$E(\alpha_i)^T$	$(4.155, -0.906)$	$(4.155, -0.901)$	$(4.155, -0.905)$	$(4.155, -0.906)$
SD	$(0.153, 0.891)$	$(0.153, 0.888)$	$(0.153, 0.891)$	$(0.153, 0.891)$
SE	$(0.159, 0.860)$	$(0.205, 0.969)$	$(0.169, 0.888)$	$(0.159, 0.860)$
CP	$(0.940, 0.950)$	$(0.940, 0.955)$	$(0.940, 0.950)$	$(0.940, 0.950)$
$\text{var}(\alpha_{i0})$	$4.902\,(0.532)$	$4.901\ (0.531)$	$4.902\ (0.532)$	$4.902\ (0.532)$
$\text{cov}(\alpha_{i0}, \alpha_{i1})$	$-0.045\,(0.020)$	$-0.045\ (0.020)$	$-0.045\ (0.020)$	$-0.045\ (0.020)$
$\text{var}(\alpha_{i1})$	$0.012\,(0.002)$	$0.012\,(0.002)$	$0.012\,(0.002)$	$0.012\,(0.002)$
σ^2	$0.596\,(0.026)$	$0.595\,(0.026)$	$0.596\,(0.026)$	$0.596\,(0.026)$
		(ii) Bimodal α_i		
$E(\alpha_i)^T$	$(4.172, -0.972)$	$(4.166, -0.918)$	$(4.166, -0.924)$	$(4.166, -0.924)$
SD	$(0.163, 0.869)$	$(0.163, 0.871)$	$(0.163, 0.871)$	$(0.163, 0.871)$
SE	$(0.160, 0.859)$	$(0.161, 0.856)$	$(0.160, 0.853)$	$(0.160, 0.853)$
CP	$(0.940, 0.945)$	$(0.940, 0.945)$	$(0.940, 0.945)$	$(0.940, 0.945)$
$\text{var}(\alpha_{i0})$	$4.934\,(0.347)$	$4.888\,(0.352)$	$4.896\,(0.351)$	$4.897\,(0.351)$
$\text{cov}(\alpha_{i0}, \alpha_{i1})$	$-0.048\,(0.021)$	$-0.045\,(0.021)$	$-0.045\,(0.021)$	$-0.045\,(0.021)$
$\text{var}(\alpha_{i1})$	$0.012\,(0.002)$	$0.011\,(0.002)$	$0.012\,(0.002)$	$0.012\,(0.002)$
σ^2	$0.600\,(0.027)$	$0.583\,(0.027)$	$0.585\,(0.026)$	$0.585\,(0.026)$
		(iii) Bimodal α_i with trt		
$E(\alpha_i\|Z_i=0)^T$	$(4.160, -0.992)$	$(4.158, -0.960)$	$(4.157, -0.960)$	$(4.158, -0.955)$
SD	$(0.231, 1.214)$	$(0.186, 1.224)$	$(0.186, 1.222)$	$(0.198, 1.219)$
SE	$(0.226, 1.203)$	$(0.181, 1.188)$	$(0.181, 1.188)$	$(0.188, 1.188)$
CP	$(0.955, 0.940)$	$(0.965, 0.935)$	$(0.965, 0.935)$	$(0.955, 0.935)$
$E(\alpha_i\|Z_i=1)^T$	$(4.190, 4.015)$	$(4.181, 4.086)$	$(4.182, 4.084)$	$(4.183, 4.059)$
SD	$(0.225, 1.208)$	$(0.174, 1.212)$	$(0.175, 1.215)$	$(0.189, 1.211)$
SE	$(0.226, 1.191)$	$(0.181, 1.178)$	$(0.181, 1.178)$	$(0.187, 1.177)$
CP	$(0.950, 0.955)$	$(0.960, 0.955)$	$(0.960, 0.955)$	$(0.955, 0.955)$
$\text{var}(\alpha_{i0})$	$4.900\,(0.339)$	$4.878\,(0.342)$	$4.878\,(0.343)$	$4.887\,(0.341)$
$\text{cov}(\alpha_{i0}, \alpha_{i1})$	$-0.043\,(0.020)$	$-0.041\,(0.020)$	$-0.041\,(0.020)$	$-0.041\,(0.020)$
$\text{var}(\alpha_{i1})$	$0.012\,(0.001)$	$0.011\,(0.001)$	$0.011\,(0.001)$	$0.011\,(0.001)$
σ^2	$0.598\,(0.024)$	$0.583\,(0.022)$	$0.583\,(0.022)$	$0.587\,(0.024)$

Table 5.3: Fit of the model in (4.1) and (4.4) to the ACTG 175 data using CS (conditional score) and SNP with $K = 0, 2, 3, 4$. Values under each estimate in parentheses are estimated standard errors. The second entry of estimates of $E(\alpha_i|Z_i = z)^T$ and its standard error are multiplied by 10^2 in each case. Estimates of $\text{cov}(\alpha_{i0}, \alpha_{i1})$ and $\text{var}(\alpha_{i1})$ are multiplied by 10^3.

	CS	$K = 0$	$K = 2$	$K = 3$	$K = 4$
γ	-2.214	-2.487	-2.487	-2.490	-2.498
	(0.207)	(0.091)	(0.092)	(0.092)	(0.092)
η	0.145	0.003	0.002	0.001	-0.007
	(0.264)	(0.132)	(0.132)	(0.132)	(0.131)
$E(\alpha_i\|Z_i = 0)^T$	—	$(2.523, -0.295)$	$(2.525, -0.223)$	$(2.525, -0.238)$	$(2.527, -0.234)$
	—	$(0.007, 0.017)$	$(0.007, 0.017)$	$(0.007, 0.014)$	$(0.008, 0.019)$
$E(\alpha_i\|Z_i = 1)^T$	—	$(2.579, -0.207)$	$(2.572, -0.195)$	$(2.575, -0.214)$	$(2.578, -0.229)$
	—	$(0.004, 0.010)$	$(0.005, 0.014)$	$(0.004, 0.012)$	$(0.006, 0.017)$
$\text{var}(\alpha_{i0})$	—	0.024	0.024	0.023	0.023
$\text{cov}(\alpha_{i0}, \alpha_{i1})$	—	0.093	0.094	0.094	0.126
$\text{var}(\alpha_{i1})$	—	0.014	0.013	0.013	0.012
loglike	—	8558.465	9018.782	9310.945	9347.163
AIC	—	-0.412	-0.435	-0.449	-0.450
HQ	—	-0.397	-0.418	-0.432	-0.433
BIC	—	-0.364	-0.385	-0.397	-0.396

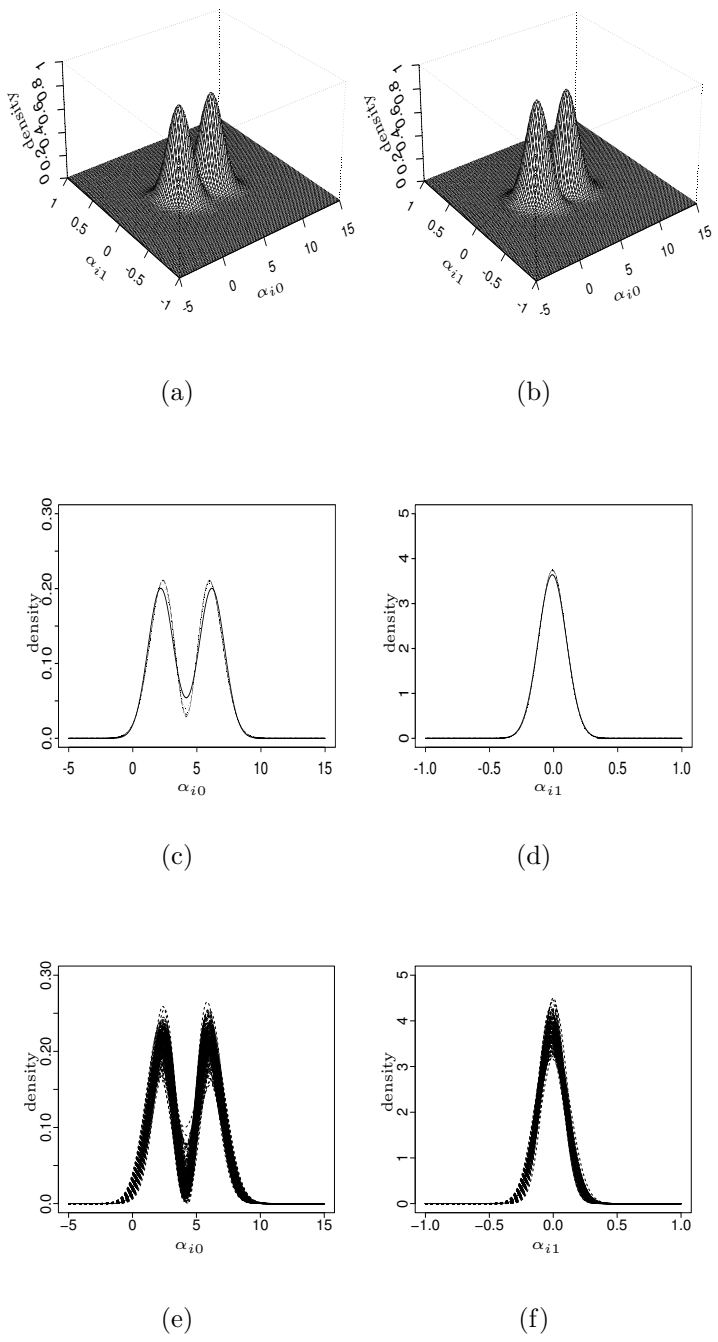


Figure 5.1: Simulation results for the bimodal case. **a.** True density of α_i. **b.** Monte Carlo average of estimated densities of α_i preferred by HQ. **c and d.** True marginal density (solid line) and Monte Carlo average of estimated marginal densities preferred by AIC (dotted line), HQ (dash-dotted line) and BIC (long dashed line) for α_{i0} and α_{i1}. **e and f.** Estimated marginal densities for α_{i0} and α_{i1} preferred by HQ for all 200 Monte Carlo data sets.

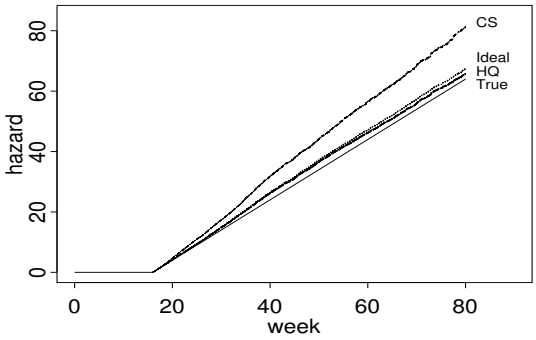
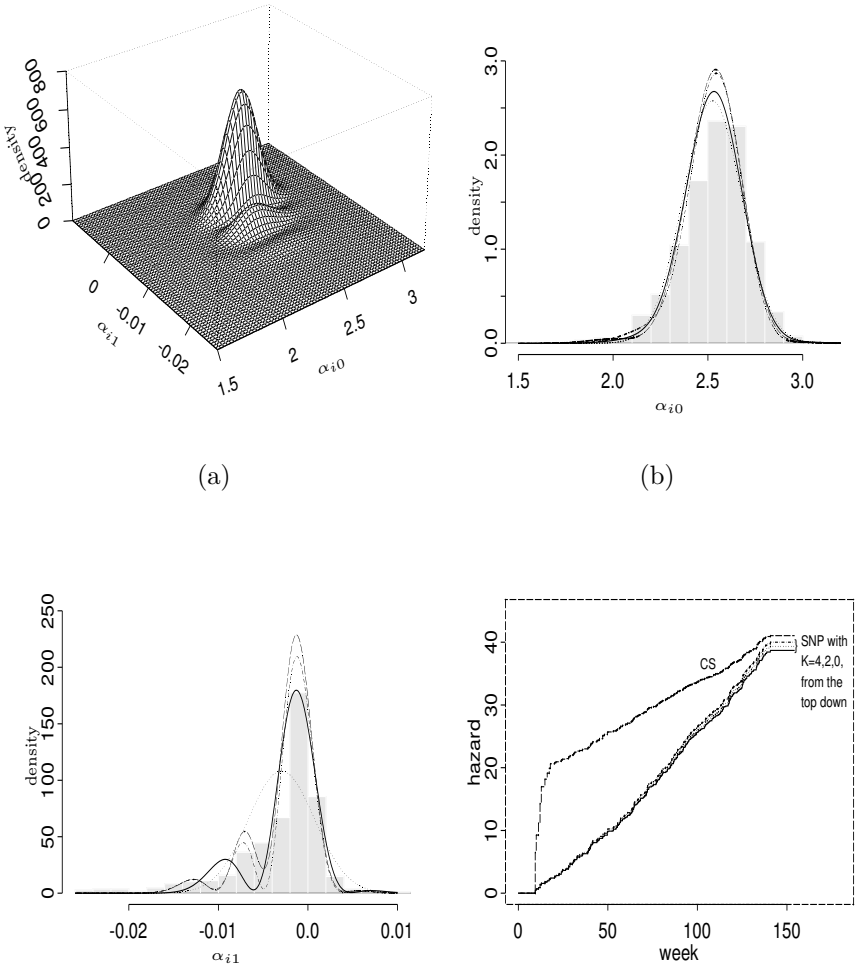


Figure 5.2: Monte Carlo average of estimated baseline cumulative hazard functions for the 200 data sets in the bimodal case. CS, conditional score; HQ, SNP likelihood estimates preferred by HQ.



(a) (b)

(c) (d)

Figure 5.3: Fit of Model (4.1),(4.4) to the ACTG data using the SNP estimator for $Z_i = 0$. **a.** Estimated density of α_i for $K = 4$. **b and c.** Estimated marginal densities for $K = 0$ (dotted line), $K = 2$ (solid line), $K = 3$ (dash-dotted line), $K = 4$ (long dashed line) with the histogram of ordinary least square estimates of the intercept and slope superimposed, respectively. **d.** Estimated baseline cumulative hazard function.

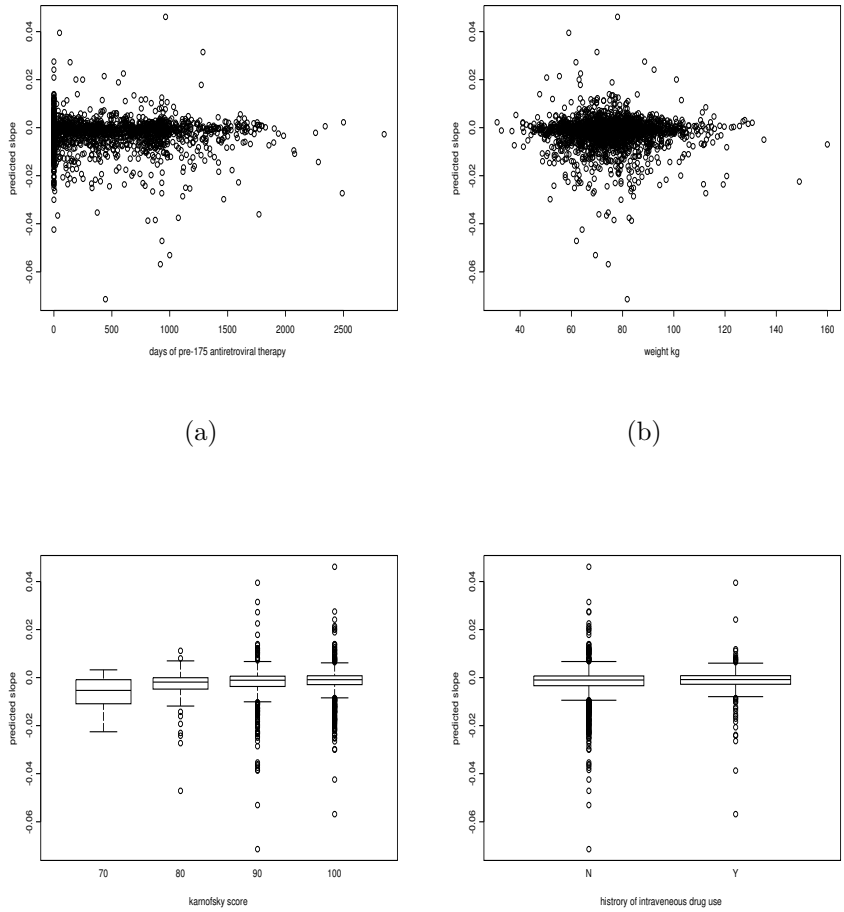


Figure 5.4: Plots of $\widehat{\alpha}_{i1}$ versus various covariates for the ACTG 175 data: **a.** Days of antiretroviral therapy prior to enrollment. **b.** Weight. **c.** Karnofsky score. **d.** Intravenous drug use.

Bibliography

Abramowitz, M. and I. Stegun (1970). Handbook of mathematical functions.

Breslow, N. E. and J. J. Crowley (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics 2*, 437–453.

Bycott, P. and J. Taylor (1998). A comparison of smoothing techniques for cd4 data measured with error in a time-dependent cox proportional hazards model. *Statistics in Medicine 17*, 2061–2077.

Carroll, R., D. Ruppert, and L. Stefanski (1995). Measurement error in nonlinear models.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B 34*, 187–200.

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

Dafni, U. G. and A. A. Tsiatis (1998). Evaluating surrogate markers of clinical outcome measured with error. *Biometrics 54*, 1445–1462.

Davidian, M. and A. R. Gallant (1992). Smooth nonparametric maximum likelihood for population pharmacokinetics, with application to quinidine. *Journal*

of Pharmacokinetics and Biopharmaceutics 20, 529–556.

Davidian, M. and A. R. Gallant (1993). The nonlinear mixed effects model with a smooth ranom effects density. *Biometrika 80*, 475–488.

Davidian, M. and D. M. Giltinan (1995). *Nonlinear Models for Repeated Measurement Data*. London: Chapman and Hall.

DeGruttola, V. and X. M. Tu (1994). Modeling progression of cd-4 lymphocyte count and its relationship to survival time. *Biometrics 50*, 1003–1014.

Eastwood, B. and A. Gallant (1991). Adaptive truncation rules for seminonparametric estimators that achieve asymptotic normality. *Econometric Theory 7*, 307–340.

Faucett, C. J. and D. C. Thomas (1996). Simultaneously modeling censored survival data and repeatedly measured covariates: A gibbs sampling approach. *Statistics in Medicine 15*, 1663–1685.

Gallant, A. R. and D. W. Nychka (1987). Seminonparametric maximum likelihood estimation. *Econometrica 55*, 363–390.

Hammer, S. M., D. A. Katzenstein, M. D. Hughes, H. Gundaker, R. T. Schooley, R. H. Haubrich, W. K. Henry, M. M. Lederman, J. P. Phair, M. Niu, M. S. Hirsch, and T. C. Merigan, for the AIDS Clinical Trials Group Study 175 Study Team (1996). A trial comparing nucleoside monotherpy with combination therapy in hiv-infected adults with cd4 cedll counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine 335*, 1081–1089.

Heagerty, P. and B. Kurland (2001). Misspecified maximum likelihood estimates

and generalised linear mixed models. *Biometrika 88*, 973–985.

Henderson, R., P. Diggle, and A. Dobson (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics 4*, 465–480.

Hogan, J. and N. Laird (1997). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine 16*, 259–272.

Hu, P., A. A. Tsiatis, and M. Davidian (1998). Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics 54*, 1407–1419.

Johansen, S. (1983). An extension of Cox's regression model. *International Statistical Review 51*, 165–174.

Johnson, N. and S. Kotz (1994). Continuous univariate distributions.

Pawitan, Y. and S. Self (1993). Modeling disease marker preocesses in aids. *Journal of the American Statistical Association 83*, 719–726.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimates in a failure time regression model. *Biometrika 69*, 331–342.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operation criteria. *Statistics in Medicine 8*, 431–440.

Raboud, J., N. Reid, R. Coates, and V. Farewell (1993). Estimating risks of progressing to aids when covariates are measured with error. *Journal of the Royal Statistical Society, Series A 156*, 396–406.

Schluchter, M. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine 11*, 1861–1870.

Stefanski, L. A. and R. J. Carroll (1987). Conditional scores and optimal scores in generalized linear measurement error models. *Biometrics 74*, 703–716.

Tao, H., M. Palta, B. Yandell, and M. Newton (1999). An estimation method for the semiparametric mixed effects model. *Biometrics 55*, 102–110.

Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *Annals of Statistics 9*, 93–108.

Tsiatis, A. A. and M. Davidian (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika 88*, 447–458.

Tsiatis, A. A., V. DeGrutolla, and M. S. Wulfsohn (1995). Modeling the relationship of survival to longitudinal data measured with error. *Journal of the American Statistical Association 90*, 27–37.

Verbeke, G. and E. Lesaffre (1997). The effect of misspecifying the random effects distribution in linear mixed effects models for longitudinal data. *Computational Statistics and Data Analysis 23*, 541–556.

Wu, M. and R. Carroll (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics 44*, 175–188.

Wulfsohn, M. S. and A. A. Tsiatis (1997). A joint model for survival and longitudinal data measured with error. *Biometrics 53*, 330–339.

Xu, J. and S. Zeger (2001a). The evaluation of multiple surrogate endpoints. *Biometrics*.

Xu, J. and S. Zeger (2001b). Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics*.

Zhang, D. and M. Davidian (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics 57*, 795–802.