A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data

Xiao Song*, Marie Davidian and Anastasios A. Tsiatis
Department of Statistics, Box 8203, North Carolina State University, Raleigh NC 27695-8203, USA
*email: xsong@stat.ncsu.edu

SUMMARY. Joint models for a time-to-event (e.g., survival) and a longitudinal response have generated considerable recent interest. The longitudinal data are assumed to follow a mixed effects model, and a proportional hazards model depending on the longitudinal random effects and other covariates is assumed for the survival endpoint. Interest may focus on inference on the longitudinal data process, which is informatively censored, or on the hazard relationship. Several methods for fitting such models have been proposed, most requiring a parametric distributional assumption (normality) on the random effects. A natural concern is sensitivity to violation of this assumption; moreover, a restrictive distributional assumption may obscure key features in the data. We investigate these issues through our proposal of a likelihood-based approach that requires only the assumption that the random effects have a “smooth” density. Implementation via the EM algorithm is described, and performance and the benefits for uncovering noteworthy features are illustrated by application to data from an HIV clinical trial and by simulation.

KEY WORDS: Informative censoring; Mixed model; Proportional hazards; SNP density; Survival.

1. Introduction

In clinical research, it is common for both longitudinal measurements on a continuous response and an observation on a possibly censored time-to-event (e.g., survival) to be recorded for each participant, along with additional covariate information. An example is given by
AIDS Clinical Trials Group (ACTG) Protocol 175, a randomized trial to compare zidovudine alone, zidovudine plus didanosine, zidovudine plus zalcitabine, or didanosine alone, in HIV-infected subjects on the basis of time to progression to AIDS or death (Hammer et al., 1996). A total of 2467 participants were recruited between December 1991 and October 1992 and followed until November 1994, leading to 308 events. The primary analysis found zidovudine alone to be inferior to the other three therapies; thus, further investigations focused on two treatment groups, zidovudine alone and the combination of the other three. CD4 counts were also collected on each subject per protocol about every 12 weeks after randomization, with an average of 8.2 CD4 measurements per subject. It is well-known that biological markers such as CD4 are subject to considerable intra-subject variation and measurement error.

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 informative censoring of longitudinal profiles (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.

A standard approach to these problems is to postulate a joint model for the longitudinal and time-to-event processes in which both depend on common, underlying random effects. A popular model assumes that the longitudinal data follow a mixed effects model, where a subject’s “inherent trajectory” is characterized by individual-specific random effects, and normal deviations characterize intra-subject variation and measurement error. The hazard for survival is assumed to depend on the random effects and covariates via a proportional hazards relationship. Conditional on time-independent covariates such as treatment group,
depending on the focus (longitudinal or survival process), various approaches to implementation and inference in this and related models have been proposed, for example, based on using individual estimates as “building blocks” (e.g., Pawitan and Self, 1993; Tsiatis, DeGruttola, and Wulfsohn, 1995) or based on a joint likelihood (e.g., Schluchter, 1992; De-Gruttola and Tu, 1994; Wulfsohn and Tsiatis, 1997; Faucett and Thomas, 1996; Henderson, Diggle, and Dobson, 2000; Xu and Zeger, 2001). These procedures are generally predicated on the assumption of normality or other parametric model for the unobserved random effects.

Figure 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(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 not; the plot for the combined treatment group is virtually identical. This informal evidence indicates that the normality assumption for log_{10} CD4 random effects may be inappropriate, raising concern that inference using the foregoing methods may be compromised.

These observations suggest both the need for investigation of sensitivity of inference to violations of the normality assumption and for procedures that do not require a parametric random effects assumption. When interest focuses on the hazard, Tsiatis and Davidian (2001) proposed a conditional score approach to estimation of the hazard parameters in which the random effects are treated as nuisance parameters and are “conditioned away” via a sufficiency argument. This leads to estimating equations that are easily solved and yield consistent estimators regardless of the random effects distribution. However, the approach
makes inefficient use of the longitudinal data and, as it is unlikely that random effects associated with a continuous response like CD4 have discrete distribution, accommodates implausible distributional models. Moreover, because the random effects are removed from consideration, no insight into the nature of the true distribution is possible. 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.

In this paper, 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 provide insight into features of the population. In Section 2, we define the joint model and derive the likelihood-based EM algorithm in Section 3. In Section 4, we exploit the availability of this method to investigate the consequences of violation of the normality assumption, and we evaluate performance via simulation. We apply the methods to the ACTG 175 data in Section 6.

2. Model and Density Approximation

Let $T_i$ and $C_i$ denote time-to-event and censoring time, respectively, for subject $i = 1, \ldots, n$. The observed survival data on $i$ are $V_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$. Longitudinal observations $W_{ij}$ at times $t_{ij}$, $j = 1, \ldots, m_i$, are assumed to satisfy

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

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 \( \alpha_i \). For simplicity, 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 6. 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 \( \alpha_i \) are usually assumed normally distributed, possibly differently according to values of \( Z_i \). We relax this to require only that the \( \alpha_i \) 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(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 \( \alpha_i \) assuming the densities for scalar \( Z_i = 0, 1 \) belong to the same location family and let

\[
\alpha_i = \mu_0 (1 - Z_i) + \mu_1 Z_i + R \xi_i, \tag{2}
\]

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 \( \xi_i \) is independent of \( Z_i \) with “smooth” density approximated by

\[
h_K(z) = P_K^2(z) \varphi(z) = \left( \sum_{0 \leq i_1 + i_2 \leq K} a_{i_1,i_2} z_1^{i_1} z_2^{i_2} \right)^2 \varphi(z), \quad z = (z_1, z_2)^T, \tag{3}
\]

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 \(K = 2\), \(P_2(z) = a_{00} + a_{10}z_1 + a_{01}z_2 + a_{20}z_1^2 + a_{11}z_1z_2 + a_{20}z_2^2\). 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 \(a_{i_1i_2}\) must be such that \(\int h_K(z)dz = 1\). Zhang and Davidian (2001) show that this may be accomplished by requiring \(E\{P^2_K(U)\} = 1\), 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 \(i\)th and \(j\)th 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 \(A = B^T B\) 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 \(c_1 = \sin(\phi_1), c_2 = \cos(\phi_1)\sin(\phi_2), \ldots, c_{d-1} = \cos(\phi_1)\cos(\phi_2)\ldots \cos(\phi_{d-2})\sin(\phi_{d-1}), c_d = \cos(\phi_1)\cos(\phi_2)\ldots \cos(\phi_{d-2})\cos(\phi_{d-1})\), 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), (3) is capable of approximating a variety of shapes; thus, the density representation for \(\alpha_i|Z_i\) in (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 \(K = 0\), \(\xi_i\) is standard normal, so that, conditional on \(Z_i\), \(\alpha_i\) is \(N\{\mu_0(1 - Z_i) + \mu_1 Z_i, RR^T\}\), 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

\[
\lambda_i(u) = \lim_{du \to 0} du^{-1} \Pr\{u \leq T_i < u + du | T_i \geq u, \alpha_i, Z_i, C_i, e_i, t_i\}
= \lambda_0(u) \exp\{\gamma X_i(u) + \eta Z_i\}, \tag{4}
\]

where \(\lambda_0(u)\) is an unspecified baseline hazard function, and \(e_i = (e_{i1}, \ldots, e_{imi})^T\); (4) incor-
porates the assumption that censoring, covariate errors, and timing of measurements are noninformative; the latter is often reasonable in trials like ACTG 175 where frequent longitudinal data collection is dictated by protocol. We assume \( C_i \) is independent of \( T_i, \alpha_i, \epsilon_i \), and \( t_i \) given \( Z_i \). Thus, \((V_i, \Delta_i)\) and \( W_i \) are independent given \( \alpha_i \) 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 \((\gamma, \eta)\) is of primary interest when the relationship between prognosis and longitudinal response is the focus.

3. Estimation Procedure

3.1 Likelihood Function and Existence of Estimators

The observed data are \((V, \Delta, W, t, Z)\), 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

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

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

\[
p(W_{ij}|\alpha_i, \sigma^2, t_{ij}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(W_{ij} - \alpha_{i0} - \alpha_{i1} t_{ij})^2}{2\sigma^2} \right\},
\]

\[
p(V_i, \Delta_i|\alpha_i, Z_i, \gamma, \eta, \lambda_0) = [\lambda_0(V_i) \exp\{\gamma(\alpha_{i0} + \alpha_{i1} V_i) + \eta Z_i\}]^{\lambda_i} \\
\times \exp \left[ -\int_{V_i}^{V_i} \lambda_0(u) \exp\{\gamma(\alpha_{i0} + \alpha_{i1} u) + \eta Z_i\} du \right],
\]

\[
p(\alpha_i|Z_i, \mu, \theta, \phi) = h_K \{R^{-1}(\alpha_i - \mu_{Z_i})\}|R|^{-1}
\]

\[
= P_K \{R^{-1}(\alpha_i - \mu_{Z_i})\} \varphi\{R^{-1}(\alpha_i - \mu_{Z_i})\}|R|^{-1}.
\]
Under our assumptions, (5) may be regarded as a partial likelihood (e.g., Andersen et al., 1993); we refer to (5) as a “likelihood” for brevity.

We now establish that an estimator for \( \Omega \) maximizing (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; proofs are in the first author’s unpublished Ph.D. thesis.

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

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

### 3.2 EM Algorithm

In principle, (5) may be maximized in \( \Omega \) 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 \( \hat{\Omega} \), treating the \( \alpha_i \) 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 (5),

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

The estimator \( \hat{\Omega} \) is obtained by iterating between the E-step and the M-step. In the E-step, the expectation of the logarithm of (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 \( \hat{\Omega}^{(k)} \) be the estimate at the \( k \)th iteration, and let
$E_{ik}\{g(\alpha_i)\} = E_{i\lambda(k)}\{g(\alpha_i)|V_i, \Delta_i, W_i, t_i, Z_i\}$ for any function $g(\cdot)$, where this notation implies evaluation at $\hat{\Omega}(k)$. Then the estimates of $\sigma^2$ and $\lambda_0$ at the $(k + 1)$th iteration may be shown to be $(\hat{\sigma}^2)^{(k+1)} = \sum_{i=1}^{n} \frac{\sum_{j=1}^{m_i} E_{ik}\{(W_{ij} - \alpha_{i0} - \alpha_{i1}t_{ij})^2\}}{\sum_{i=1}^{m_i} m_i}$, and

$$\widehat{\lambda_0}^{(k+1)}(u_t) = \frac{\sum_{i=1}^{n} \Delta_i I(V_i = u_t)}{\sum_{i=1}^{n} E_{ik}\{\exp\{\gamma(k)(\alpha_{i0} + \alpha_{i1}u_t) + \eta Z_i\}\} Y_i(u_t)},$$

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), these may be approximated by one-step Newton-Raphson estimators. Letting $\beta = (\gamma, \eta)^T$, the update is

$$\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} + I_{\beta(k)}^{-1} S_{\beta(k)},$$

where the score vector for $\beta$ at iteration $k$ has $s$th element

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

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

$$I_{\beta(s,k), r} = \sum_{i=1}^{n} \Delta_i \left[ \frac{\sum_{j=1}^{n} E_{jk}(G_{sjj}G_{rij}H_{kij})Y_j(V_i)}{\sum_{j=1}^{n} E_{jk}(H_{kij})Y_j(V_i)} \right]$$

$$- \left\{ \frac{\sum_{j=1}^{n} E_{jk}(G_{sjj}H_{kij})Y_j(V_i)}{\sum_{j=1}^{n} E_{jk}(H_{kij})Y_j(V_i)} \right\} \left\{ \frac{\sum_{j=1}^{n} E_{jk}(G_{rij}H_{kij})Y_j(V_i)}{\sum_{j=1}^{n} E_{jk}(H_{kij})Y_j(V_i)} \right\}^2,$$

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 (6), their $(k + 1)$th iterates may be found using any standard optimization software to maximize

$$\sum_{i=1}^{n} E_{ik}\{\log p(\alpha_i|Z_i, \mu, \phi)\} = \sum_{i=1}^{n} E_{ik}\{\log \left[ P^2\{R^{-1}(\alpha_i - \mu Z_i)\}\{R^{-1}(\alpha_i - \mu Z_i)\}|R|^{-1} \right] \}.$$

(7)

The expressions for the iterative updates for all parameters involve intractable integrals of the form $E_{ik}\{g(\alpha_i)\} = \int g(\alpha_i)p(\alpha_i|V_i, \Delta_i, W_i, t_i, Z_i, \hat{\Omega}(k))d\alpha_i$, where $p(\alpha_i|V_i, \Delta_i, W_i, t_i, Z_i, \Omega)$
is the density of $\alpha_i$ given the observed data, that must be computed numerically. In Sections 4 and 5, we use Gauss-Hermite quadrature; see the Appendix.

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 (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 $\tilde{\Lambda}_0(u) = \sum_{u_i \leq u} \tilde{\lambda}_0(u_i)$.

To obtain starting values for $\Omega$ for the EM algorithm, we have had success using the following approach. We use the conditional score estimates of $\gamma$, $\eta$, $\lambda_0$, and $\sigma^2$ for these elements of $\Omega$, 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 $\alpha_i$ as “data” and maximizing the density corresponding to (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 $\phi$, with $\mu$ and $R$ computed by moment methods for the given $\phi$. 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 4 and 5, 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 $\gamma$ and $\eta$, following Wulfsohn and Tsiatis (1997), for parameter $\delta$, say, with final estimate $\tilde{\delta}$, var($\tilde{\delta}$) may be estimated by the negative inverse of the derivative (slope) of the profile score at $\tilde{\delta}$. Let $\Omega_{-\delta}$ be the set of parameters in $\Omega$ except $\delta$. The profile score $S_\delta(\tilde{\Omega}_{-\delta}(\delta))$ is the derivative of the profile
log likelihood with $\delta$ fixed and $\Omega_{-\delta}$ estimated by the restricted likelihood estimator $\widehat{\Omega}_{-\delta}$ computed by a separate EM algorithm identical to the above with $\delta$ held fixed. \( S_{\tilde{\delta}} \{ \widehat{\Omega}_{-\delta} (\delta) \} \) is of the same form as \( S_{\tilde{\delta}} \) with $\tilde{\delta}$ replaced by the given $\delta$ 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 (5) evaluated at $\hat{\Omega}$. This may be calculated via numerical approximation, where the integral in (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 $\tilde{\lambda}_0(u_l)$, $l = 1, \ldots, L$, 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.

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

$$
\{-\ell(\hat{\Omega}; V, \Delta, W, t, Z) + C(N)p_{net}\}/N, \tag{8}
$$

where $\ell(\Omega; V, \Delta, W, t, Z)$ is the logarithm of (5), $N = \sum_{i=1}^{n} m_i$ is the total number of longitudinal observations, $p_{net}$ is the number of free parameters in $\Omega$, and $C(N)p_{net}$ is the penalty factor to compensate for small $-\ell(\hat{\Omega}; V, \Delta, W, t, Z)$ 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, (8) is computed for different $K$, and $K$ minimizing (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 section.

4. 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 $X_i(u) = \alpha_0 + \alpha_1 u$ 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 (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
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 has distinct elements $\{\operatorname{var}(\alpha_{i0}), \operatorname{cov}(\alpha_{i0}, \alpha_{i1}), \operatorname{var}(\alpha_{i1})\} = (4.96, -0.0456, 0.012)$. Figure 2(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) 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, $\gamma$ [and $\eta$ 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 semi-parametric likelihood (SNP) estimator with $K = 0, 1, 2$, where two-point Gauss-Hermite quadrature was used to calculate $E_{i_k}\{g(\alpha_i)\}$. Preliminary studies revealed that increasing the number of abscissae beyond two did not improve performance appreciably. In all cases, 95% Wald confidence intervals for $\gamma$ (and $\eta$) 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 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”
methods. As suggested by Tsiatis and Davidian (2001), the corresponding conditional score estimator is 
\[ \int_{0}^{u} \left[ \sum_{i=1}^{n} \exp \left\{ \gamma_{CS} S_i(u, \tilde{\gamma}, \tilde{\sigma}_{CS}^2) - \tilde{\gamma}_{CS}^2 \Sigma_i(u, \tilde{\sigma}_{CS}^2)/2 + \tilde{\eta}_{CS} Z_i \right\} Y_i(u) - 1 \right] \sum_{i=0}^{n} dN_i(u), \]
where \( N_i(u) = I(V_i \leq u, \Delta_i = 1); Y_i(u) = I(V_i \geq u, t_{i2} \leq u); \gamma_{CS}, \tilde{\gamma}_{CS}, \) and \( \tilde{\sigma}_{CS} \) are the conditional score estimators proposed by Tsiatis and Davidian (2001); \( S_i(u, \tilde{\gamma}_{CS}, \tilde{\sigma}_{CS}^2) = \tilde{X}_i(u) + dN_i(u) \Sigma_i(u, \tilde{\sigma}_{CS}^2) \gamma_{CS} \); and \( \tilde{X}_i(u) \) 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 \( \alpha_i \). Standard errors for \( E(\alpha_i|Z_i) \) were obtained using the delta method and inverse Hessian treating the \( \lambda_0(u_i) \) as fixed, as described in Section 3.2, with 95% Wald confidence intervals constructed as above.

Results for estimation of \( \gamma \) and \( \eta \) are given in Table 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 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 3; 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 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 2 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 1 and 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 $\gamma, \eta, E(\alpha_i|Z_i)$, 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 inefficient 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 6. 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. Application to ACTG 175

We apply the approach to the ACTG 175 data. Based on the discussion in Section 1, 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 (1), and we take the hazard for AIDS or death to be as in (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 $\gamma$ and $\eta$ 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 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 $\gamma$ and $\eta$ are similar across $K$; the conditional score estimate of $\gamma$ is a bit smaller in magnitude.
Estimated standard errors for the SNP estimates of $\gamma$ and $\eta$ 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 $\gamma$ and $\eta$ and standard errors are similar across $K$, echoing the simulation results. The results suggest that there is a strong relationship between hazard and $\log_{10} \text{CD4}$, with hazard decreasing with increasing $\log_{10} \text{CD4}$; 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 ($\hat{\eta} = -0.375$ with standard error 0.128). Hence, a possible interpretation is that $\log_{10} \text{CD4}$ 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 $\sigma^2$ are virtually identical and equal to 0.11.

Table 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(b). Figure 4(a) depicts the estimated density of the random effects for $K = 4$ in the zidovudine only group, which indicates the presence of an obvious second mode; an extreme, modest third mode is likely an artifact of tail estimation. Figure 4(b) and (c) show the corresponding estimated marginal densities superimposed on the histograms of least squares estimates from Figure 1(b). A possible interpretation is that there may be a subpopulation of individuals whose $\log_{10} \text{CD4}$ 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 longitudinal model presented here with the se-
lected $K$ and obtain “empirical Bayes” estimates of the $\alpha_i$ as $\hat{\alpha}_i = \arg\max_{\alpha_i} p(\alpha_i | V_i, \Delta_i, W_i, t_i, Z_i, \hat{\Omega})$. Plots of each component versus baseline covariates, using box plots for categorical covariates, may be inspected for relationships that may suggest refinements of model (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 shows representative plots for four covariates based on the $K = 4$ fit, 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 4(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.

6. 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 study populations.
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 possible to consider models for individual inherent longitudinal trajectories involving higher dimensional $\alpha_i$, and trajectories that are nonlinear in individual random effects may be entertained; see Zhang and Davidian (2001) for the general SNP representation for $\dim(\alpha_i) > 2$. Similarly, the model may be extended to incorporate vector-valued covariates $Z_i$ with both continuous and discrete components. Here, the model for $\alpha_i$ given in (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. The assumption that the observation times $t_i$ are noninformative could also be relaxed; under this condition, $L_{ei}$ in the observed-data likelihood (5) would require modification and additional modeling assumptions on the relationship between $t_i$, random effects, and other components. In all these cases, increased model complexity would involve 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 4 and 5, 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.

ACKNOWLEDGEMENTS

This work was supported by NIH grants R01-AI31789, R01-CA85848, and R01-CA51962. The authors are grateful to Michael Hughes, Heather Gorski, and the AIDS Clinical Trials Group for providing the data from ACTG 175.

REFERENCES


APPENDIX

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

Note that $p(\alpha_i | V_i, \Delta_i, W_i, t_i, Z_i, \widetilde{\Omega}^{(k)}) \propto L_{ci}(\widetilde{\Omega}^{(k)}; V_i, \Delta_i, W_i, t_i, Z_i, \alpha_i)$. By some straightforward algebra, for $Z_i = s$,

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

where $D_i$ satisfies

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

$B_{i00} = (R^2_{01} + R^2_{00})/(R^2_{00} R^2_{11}) + m_i/\sigma^2$, $B_{i01} = -(R_{00} R_{01})/(R^2_{00} R^2_{11}) + \sum_{j=1}^{m_i} t_{ij}/\sigma^2$, $B_{i11} = R^2_{00}/(R^2_{00} R^2_{11}) + \sum_{j=1}^{m_i} t^2_{ij}/\sigma^2$, and

$$c_{si} = B_i \left[ \frac{\sum_{j=1}^{m_i} W_{ij} \sigma^2 + \{(R^2_{01} + R^2_{00}) \mu_{i0} - R_{00} R_{01} \mu_{i1}\} / (R^2_{00} R^2_{11})}{\sum_{j=1}^{m_i} W_{ij} t_{ij} / \sigma^2 + \{R^2_{00} \mu_{i1} - R_{00} R_{01} \mu_{i0}\} / (R^2_{00} R^2_{11})} \right].$$

Let

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

Then it follows that

$$E_{ik}\{g(\alpha_i)\} = \frac{\int g(\alpha_i)Q(\alpha_i; V_i, \Delta_i, t_i, W_i, Z_i, \widetilde{\Omega}^{(k)}) \varphi\{(D^k_{i})^{-1}(\alpha_i - c_{ki})\} \, d\alpha_i}{\int Q(\alpha_i; V_i, \Delta_i, t_i, W_i, Z_i, \Omega^{(k)}) \varphi\{(D^k_{i})^{-1}(\alpha_i - c_{ki})\} \, d\alpha_i} = \frac{\int g(\alpha_i(\beta_i)) Q\{\alpha_i(\beta_i); V_i, \Delta_i, t_i, W_i, Z_i, \widetilde{\Omega}^{(k)}\} \exp(-\beta_i^2) \, d\beta_i}{\int Q\{\alpha_i(\beta_i); V_i, \Delta_i, t_i, W_i, Z_i, \widetilde{\Omega}^{(k)}\} \exp(-\beta_i^2) \, d\beta_i}$$

where $\beta_i = (\sqrt{2})^{-1}(D^k_{i})^{-1}(\alpha_i - c_{ki})$ and $\alpha_i(\beta_i)$ means $\alpha_i$ expressed in terms of $\beta_i$; and $D^k_{i}$ and $c_{ki}$ are $D_i$ and $c_{si}$, respectively, with $\Omega$ replaced by $\widetilde{\Omega}^{(k)}$.
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 $\beta_i$ is
\[
\frac{\sum_{s=1}^{M} \sum_{t=1}^{M} g(\alpha_i(x_s, x_t)) Q(\alpha_i(x_s, x_t); V_i, \Delta_i, t_i, W_i, Z_i, \tilde{\Omega}^{(k)}) w_s w_t}{\sum_{s=1}^{M} \sum_{t=1}^{M} Q(\alpha_i(x_s, x_t); V_i, \Delta_i, t_i, W_i, Z_i, \tilde{\Omega}^{(k)}) w_s w_t},
\]
where $\alpha_i(x_s, x_t)$ indicates that $\alpha_i$ in this expression is a function of $x_s$ and $x_t$. 
Table 1

Simulation results, hazard parameters $\gamma$ and $\eta$ of formula (4). 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.

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

25
Table 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 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 var($\alpha_i$) and $\sigma^2$, entries are average of Monte Carlo estimates (Monte Carlo standard deviation).

<table>
<thead>
<tr>
<th></th>
<th>$K = 0$</th>
<th>AIC</th>
<th>HQ</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Normal $\alpha_i$</td>
<td>$E(\alpha_i)^T$</td>
<td>(4.155, -0.906)</td>
<td>(4.155, -0.906)</td>
<td>(4.155, -0.905)</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>(0.153, 0.891)</td>
<td>(0.153, 0.891)</td>
<td>(0.153, 0.891)</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>(0.159, 0.860)</td>
<td>(0.205, 0.969)</td>
<td>(0.169, 0.888)</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>(0.940, 0.950)</td>
<td>(0.940, 0.955)</td>
<td>(0.940, 0.950)</td>
</tr>
<tr>
<td></td>
<td>var($\alpha_{i0}$)</td>
<td>4.902 (0.532)</td>
<td>4.901 (0.531)</td>
<td>4.902 (0.532)</td>
</tr>
<tr>
<td></td>
<td>$\text{cov}(\alpha_{i0}, \alpha_{i1})$</td>
<td>-0.045 (0.020)</td>
<td>-0.045 (0.020)</td>
<td>-0.045 (0.020)</td>
</tr>
<tr>
<td></td>
<td>var($\alpha_{i1}$)</td>
<td>0.012 (0.002)</td>
<td>0.012 (0.002)</td>
<td>0.012 (0.002)</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2$</td>
<td>0.596 (0.026)</td>
<td>0.595 (0.026)</td>
<td>0.596 (0.026)</td>
</tr>
<tr>
<td>(ii) Bimodal $\alpha_i$</td>
<td>$E(\alpha_i)^T$</td>
<td>(4.172, -0.972)</td>
<td>(4.166, -0.918)</td>
<td>(4.166, -0.924)</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>(0.163, 0.869)</td>
<td>(0.163, 0.871)</td>
<td>(0.163, 0.871)</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>(0.160, 0.859)</td>
<td>(0.161, 0.856)</td>
<td>(0.160, 0.853)</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>(0.940, 0.945)</td>
<td>(0.940, 0.945)</td>
<td>(0.940, 0.945)</td>
</tr>
<tr>
<td></td>
<td>var($\alpha_{i0}$)</td>
<td>4.934 (0.347)</td>
<td>4.888 (0.352)</td>
<td>4.896 (0.351)</td>
</tr>
<tr>
<td></td>
<td>$\text{cov}(\alpha_{i0}, \alpha_{i1})$</td>
<td>-0.048 (0.021)</td>
<td>-0.045 (0.021)</td>
<td>-0.045 (0.021)</td>
</tr>
<tr>
<td></td>
<td>var($\alpha_{i1}$)</td>
<td>0.012 (0.002)</td>
<td>0.011 (0.002)</td>
<td>0.012 (0.002)</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2$</td>
<td>0.600 (0.027)</td>
<td>0.583 (0.027)</td>
<td>0.585 (0.026)</td>
</tr>
<tr>
<td>(iii) Bimodal $\alpha_i$ with trt</td>
<td>$E(\alpha_i</td>
<td>Z_i = 0)^T$</td>
<td>(4.160, -0.992)</td>
<td>(4.158, -0.960)</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>(0.231, 1.214)</td>
<td>(0.186, 1.224)</td>
<td>(0.186, 1.222)</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>(0.226, 1.203)</td>
<td>(0.181, 1.188)</td>
<td>(0.181, 1.188)</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>(0.955, 0.940)</td>
<td>(0.965, 0.935)</td>
<td>(0.965, 0.935)</td>
</tr>
<tr>
<td></td>
<td>$E(\alpha_i</td>
<td>Z_i = 1)^T$</td>
<td>(4.190, 4.015)</td>
<td>(4.181, 4.086)</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>(0.225, 1.208)</td>
<td>(0.174, 1.212)</td>
<td>(0.175, 1.215)</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>(0.226, 1.191)</td>
<td>(0.181, 1.178)</td>
<td>(0.181, 1.178)</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>(0.950, 0.955)</td>
<td>(0.960, 0.955)</td>
<td>(0.960, 0.955)</td>
</tr>
<tr>
<td></td>
<td>var($\alpha_{i0}$)</td>
<td>4.900 (0.339)</td>
<td>4.878 (0.342)</td>
<td>4.878 (0.343)</td>
</tr>
<tr>
<td></td>
<td>$\text{cov}(\alpha_{i0}, \alpha_{i1})$</td>
<td>-0.043 (0.020)</td>
<td>-0.041 (0.020)</td>
<td>-0.041 (0.020)</td>
</tr>
<tr>
<td></td>
<td>var($\alpha_{i1}$)</td>
<td>0.012 (0.001)</td>
<td>0.011 (0.001)</td>
<td>0.011 (0.001)</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2$</td>
<td>0.598 (0.024)</td>
<td>0.583 (0.022)</td>
<td>0.583 (0.022)</td>
</tr>
</tbody>
</table>
Table 3
Fit of the model in (1) and (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 $10^3$

<table>
<thead>
<tr>
<th></th>
<th>$K = 0$</th>
<th>$K = 2$</th>
<th>$K = 3$</th>
<th>$K = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>-2.214</td>
<td>-2.487</td>
<td>-2.490</td>
<td>-2.498</td>
</tr>
<tr>
<td></td>
<td>(0.207)</td>
<td>(0.091)</td>
<td>(0.092)</td>
<td>(0.092)</td>
</tr>
<tr>
<td>$\eta$</td>
<td>0.145</td>
<td>0.003</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>(0.264)</td>
<td>(0.132)</td>
<td>(0.132)</td>
<td>(0.131)</td>
</tr>
<tr>
<td>$E(\alpha_i</td>
<td>Z_i = 0)^T$</td>
<td>(2.523, -0.295)</td>
<td>(2.525, -0.233)</td>
<td>(2.525, -0.238)</td>
</tr>
<tr>
<td></td>
<td>(0.007, 0.017)</td>
<td>(0.007, 0.017)</td>
<td>(0.007, 0.014)</td>
<td>(0.008, 0.019)</td>
</tr>
<tr>
<td>$E(\alpha_i</td>
<td>Z_i = 1)^T$</td>
<td>(2.579, -0.207)</td>
<td>(2.572, -0.195)</td>
<td>(2.575, -0.214)</td>
</tr>
<tr>
<td></td>
<td>(0.004, 0.010)</td>
<td>(0.005, 0.014)</td>
<td>(0.004, 0.012)</td>
<td>(0.006, 0.017)</td>
</tr>
<tr>
<td>$\text{var}(\alpha_{i0})$</td>
<td>—</td>
<td>0.024</td>
<td>0.024</td>
<td>0.023</td>
</tr>
<tr>
<td>$\text{cov}(\alpha_{i0}, \alpha_{i1})$</td>
<td>—</td>
<td>0.093</td>
<td>0.094</td>
<td>0.094</td>
</tr>
<tr>
<td>$\text{var}(\alpha_{i1})$</td>
<td>—</td>
<td>0.014</td>
<td>0.013</td>
<td>0.013</td>
</tr>
<tr>
<td>loglike</td>
<td>—</td>
<td>8558.465</td>
<td>9018.782</td>
<td>9310.945</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>9347.163</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>—</td>
<td>-0.412</td>
<td>-0.435</td>
<td>-0.449</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>-0.450</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HQ</td>
<td>—</td>
<td>-0.397</td>
<td>-0.418</td>
<td>-0.432</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>-0.433</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>—</td>
<td>-0.364</td>
<td>-0.385</td>
<td>-0.397</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>-0.396</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1. a. Trajectories of $\log_{10} \text{CD4}$ for 10 randomly selected subjects. b. Histograms of subject-specific intercept and slope estimates from simple least square fits.
Figure 2. Simulation results for the bimodal case. a. True density of $\alpha_i$. b. Monte Carlo average of estimated densities of $\alpha_i$ 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 $\alpha_{i0}$ and $\alpha_{i1}$. e and f. Estimated marginal densities for $\alpha_{i0}$ and $\alpha_{i1}$ preferred by HQ for all 200 Monte Carlo data sets.
Figure 3. Monte Carlo average of estimated baseline cumulative hazard functions for the 200 data sets in the bimodal case. HQ is the average of likelihood estimates preferred by HQ.
Figure 4. Fit to the ACTG data using the SNP estimator for $Z_i = 0$. a. Estimated density of $\alpha_i$ 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 squares estimates of the intercept and slope superimposed, respectively. d. Estimated baseline cumulative hazard function.
Figure 5. Plots of $\tilde{\alpha}_{43}$ from the fit with $K = 4$ 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.